لجنة الميكانيك
تقدم لكم..

[المكتبة التخصصية]

لجنة
الميكانيك
Polytechnic

Mech.MuslimEngineer.Net

FB.com/Groups/Mid.Group

0789434018 MechFet

FUNDAMENTALS OF RENEWABLE ENERGY PROCESSES

2nd Edition

Aldo V. Da Rosa

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper. ∞

Library of Congress Cataloging-in-Publication Data
Application Submitted.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: 978-0-12-374639-9

For information on all Academic Press publications
visit our Web site at *www.books.elsevier.com*

Printed in the United States of America
09 10 11 12 13 14 15 16 5 4 3 2 1

Foreword to the Second Edition

The public's widespread desire to become informed about energy has been, in part, satisfied by excellent media coverage and by a plethora of good books on the subject. Most of these books are, quite naturally, journalistically slanted and treat technology superficially. Granted that of the various components of the problem—technology, economics, politics—technology represents only a small fraction of the total, but it is the one fraction that must be tackled first.

Those who need to understand the limitations of technical solutions require a good scientific grasp of what is being proposed. This book tries to explain how each energy process discussed actually works. A reasonable degree of mathematics is used to unify and clarify the explanations. By discussing fundamentals more than the state of art, it is hoped to delay the obsolescence of this writing, especially in this moment of very fast evolution of ideas. Those who wanting to labor in this field may find this book useful in preparing themselves to comprehend more specialized articles on whatever energy process that especially interests them.

In spite of its fundamentalist approach, this book will eventually become dated, not because fundamentals change but because different fundamentals will be invoked. This second edition discusses several scientific areas that only recently have been recruited to resolve energy problems.

After more than two centuries of intense development, even very mature technologies such as heat engines (Chapter 2) can still find new and improved forms. This is the case of the free-piston Stirling engine whose high efficiency and very long mantenance-free life has made it now a favorite for generating electricity in remote, unmanned locations, such as in spacecraft and in planetary exploration. This second edition expands the seven pages of the first edition dedicated to Stirling engines, and these ultramodern free-piston devices are included.

Thermoelectrics (Chapter 5) has also progressed in recent years with a better understanding of artificially created nano materials and superlattices that, in a way, get around the limitations of the Wiedemann–Franz–Lorenz law, allowing the synthesis of materials that have large electric conductivity but small heat conductivity.

Fuel cells have matured substantially. Those described in the first edition, though adequately light and efficient, were short-lived and expensive. Catalysis problems were responsible for these shortcomings. The second edition has a much expanded discussion of chemical kinetics and describes very recent work (late 2008) that completely avoids precious metals as catalysts, while substantially outperforming these metals.

Hydrogen production, a fairly old technique, is now beginning to lean on photolytic processes that were of only marginal interest when the first edition was prepared.

It is perhaps in biomass that the most dramatic evolution has occurred. Public enthusiasm for ethanol and biodiesel has propelled biomass from a minor energy source into one that can contribute markedly to the fueling of our vehicles. Biomass will be firmly entrenched in such a role if the economical hydrolysis of cellulose can be achieved. The second edition delves deeper into the mysteries of the required biochemistry.

Utility-size photovoltaic plants expanded in the last few years at a sustained rhythm of over 40% per year. They now face a moment of decision: to continue with efficient but expensive silicon devices or to adopt cheap, though much less efficient, plastic cells. It may all hinge on finding a way to improve the life span of plastic cells. The second edition discusses the chemistry and technology of these polymer cells.

Finally, wind energy has established itself as a major player in energy production. Wind farms are expanding at the same 40% per year rate as photovoltaics, but having started from a much higher base are now beginning to make significant contributions to the energy mix. When the first edition was prepared, wind energy played a minor role, and it was not entirely clear which type of turbine (horizontal or vertical axis) would win out. It is now clear that the horizontal axis (propeller-type) is the dominant solution. The second edition treats the fundamentals of these machines (Betz limit, Rankine–Froude law, wake rotation, etc.), subjects that were omitted in the first edition.

This book is based on class notes created in the teaching of Fundamentals of Energy Processes at Stanford since 1976. As both the cost of energy and our dependence on foreign suppliers have risen, so has the interest in these lectures, reflecting the mood of the American people.

Aldo Vieira da Rosa
<*darosa@ee.stanford.edu*>
Palo Alto, CA
August 2008

Foreword to the First Edition

This book examines the fundamentals of some nontraditional energy processes. Little effort is made to describe the "state of the art" of the technologies involved because, owing to the rapidity with which these technologies change, such description would soon become obsolete. Nevertheless, the underlying principles are immutable and are essential for the comprehension of future developments. An attempt is made to present clear physical explanations of the pertinent principles.

The text will not prepare the student for detailed design of any specific device or system. However, it is hoped that it will provide the basic information to permit the understanding of more specialized writings.

The topics were not selected by their practicability or by their future promise. Some topics are discussed solely because they represent good exercises in the application of physical principles, notwithstanding the obvious difficulties in their implementation.

Whenever necessary, rigor is sacrificed in favor of clarity. Although it is assumed that the reader has an adequate background in physics, chemistry, and mathematics (typical of a senior science or an engineering student), derivations tend to start from first principles to permit the identification of basic mechanisms.

Energy problems are only partially technical problems—to a large extent economics and politics dominate the picture. In a limited fashion, these considerations are included in the discussions presented here.

The organization of the book is arbitrary and certainly not all-encompassing. Processes that can be considered "traditional" are generally ignored. On the other hand, the list of "nontraditional" processes considered is necessarily limited.

Acknowledgements

I wrote this book. Without Aili, I could not.

My thanks to Dr. Edward Beardsworth who, incessantly scanning the literature, alerted me to many new developments.

My gratitude also to the hundreds of students who, since 1976, have read my notes and corrected many typos and errors.

Chapter 1
Generalities

1.1 Units and Constants

Although many different units are employed in energy work, whenever possible we shall adopt the Système International (SI). This means **joules** and **watts**. If we are talking about large energies, we'll speak of MJ, GJ, TJ, PJ, and EJ—that is, 10^6, 10^9, 10^{12}, 10^{15}, and 10^{18} joules, respectively, (See Table 1.1).

One might wish for greater consistency in the choice of names and symbols of the different prefixes adopted by the SI. The symbols for *submultiplier* prefixes are all in lowercase letters, and it would make sense if the *multipliers* were all in uppercase letters, which they are not. All symbols are single letters, except the one for "deca" which has two letters ("da"). Perhaps that explains why deciliters are popular and decaliters are extremely rare. Unlike the rest of the multipliers, "deca," "hecta," and "kilo" start with lowercase letters. The names of the prefixes are derived mostly from Greek or Latin with some severe corruptions, but there are also Danish words and one "Spanish" word—"pico"—which is not listed in most Spanish dictionaries. Some prefixes allude to the power of 1000 of the multiplier—"exa" (meaning six), for instance, refers to 1000^6: others allude to the multiplier itself—"kilo" (meaning one thousand) indicates the multiplier directly.

We cannot entirely resist tradition. Most of the time we will express pressures in **pascals**, but we will occasionally use **atmospheres** because most of the existing data are based on the latter. Sometimes **electronvolts** are more convenient than joules. Also, expressing energy in **barrels of oil** or **kWh** may better convey the idea of cost. On the whole, however, we shall avoid quads, BTUs, calories, and other non-SI units. The reason for this choice is threefold: SI units are easier to use, they have been adopted by most countries, and they are frequently better defined.

Consider, for instance, the calorie, a unit preferred by chemists. Does one mean the international steam table calorie (4.18674 J)? or the mean calorie (4.19002 J)? or the thermochemical calorie (4.18400 J)? or the calorie measured at 15 C (4.18580 J)? or at 20 C (4.18190 J)?

Americans like to use the BTU, but, again, there are numerous BTUs: steam table, mean, thermochemical, at 39 F, at 60 F. The ratio of the BTU to the calorie of the same species is about 251.956, with some variations in the sixth significant figure. Remember that 1 BTU is roughly equal to 1 kJ, whereas 1 quad equals roughly 1 EJ. The conversion factors between

Table 1.1 SI Prefixes and Symbols

Multiplier	Symbol	Prefix	Etymology
10^{24}	Y	yotta	corrupted Italian *otto* = eight, 1000^8
10^{21}	Z	zetta	corrupted Italian *sette* = seven, 1000^7
10^{18}	E	exa	corrupted Greek *hexa* = six, 1000^6
10^{15}	P	peta	corrupted Greek *penta* = five, 1000^5
10^{12}	T	tera	from Greek *teras* = monster
10^{9}	G	giga	from Greek *gigas* = giant
10^{6}	M	mega	from Greek *megas* = great
10^{3}	k	kilo	from Greek *khilioi* = thousand
10^{2}	h	hecto	from Greek *hekton* = ten
10^{1}	da	deca	from Greek *deka* = ten
10^{-1}	d	deci	from Latin *decimus* = tenth
10^{-2}	c	centi	from Latin *centum* = hundred
10^{-3}	m	milli	from Latin *mille* = thousand
10^{-6}	μ	micro	from Greek *mikros* = small
10^{-9}	n	nano	from Latin *nanus* = dwarf
10^{-12}	p	pico	from Spanish *pico* = little bit
10^{-15}	f	femto	from Danish *femten* = fifteen
10^{-18}	a	atto	from Danish *atten* = eighteen
10^{-21}	z	zepto	adapted Latin *septem* = seven, 1000^{-7}
10^{-24}	y	yocto	adapted Latin *octo* = eight, 1000^{-8}

Table 1.2 Fundamental Constants

Quantity	Symbol	Value	Units
Avogadro's number	N_0	6.0221367×10^{26}	per kmole
Boltzmann's constant	k	1.380658×10^{-23}	J K^{-1}
Charge of the electron	q	$1.60217733 \times 10^{-19}$	C
Gas constant	R	8314.510	J $\text{kmole}^{-1}\text{K}^{-1}$
Gravitational constant	G	6.67259×10^{-11}	$\text{m}^3\text{s}^{-2}\text{kg}^{-1}$
Planck's constant	h	$6.6260755 \times 10^{-34}$	J s
Permeability of free space	μ_0	$4\pi \times 10^{-7}$	H/m
Permittivity of free space	ϵ_0	$8.854187817 \times 10^{-12}$	F/m
Speed of light	c	2.99792458×10^{8}	m s^{-1}
Stefan–Boltzmann constant	σ	5.67051×10^{-8}	W $\text{K}^{-4}\text{m}^{-2}$

the different energy and power units are listed in Table 1.3. Some of the fundamental constants used in this book are listed in Table 1.2.

1.2 Energy and Utility

In northern California, in a region where forests are abundant, one cord of wood sold in 2008 for about $150. Although one cord is a stack of 4 by 4 by

Table 1.3 Conversion Coefficients

To convert from	to	multiply by
Energy		
Barrel of oil	GJ	≈ 6
British thermal unit (Int. Steam Table)	joule	1055.04
British thermal unit (mean)	joule	1055.87
British thermal unit (thermochemical)	joule	1054.35
British thermal unit (39 F)	joule	1059.67
British thermal unit (60 F)	joule	1054.68
Calorie (International Steam Table)	joule	4.18674
Calorie (mean)	joule	4.19002
Calorie (thermochemical)	joule	4.1840
Calorie (15 C)	joule	4.1858
Calorie (20 C)	joule	4.1819
Cubic foot (methane, STP)	MJ	≈ 1
Electron volt	joule	1.60206×10^{-19}
ERG	joule	1.0×10^{-7}
Foot LBF	joule	1.3558
Foot poundal	joule	4.2140×10^{-2}
kWh	joule	3.6×10^{6}
Quad	BTU	1.0×10^{15}
Ton of TNT	joule	4.2×10^{9}
Power		
Foot LBF/second	watt	1.3558
Foot LBF/minute	watt	2.2597×10^{-2}
Foot LBF/hour	watt	3.7662×10^{-4}
Horsepower (550 foot LBF/sec)	watt	745.70
Horsepower (electric)	watt	746
Horsepower (metric)	watt	735
Other		
Atmosphere	pascal	1.0133×10^{5}
Dalton	kg	1.660531×10^{-27}

LBF stands for pounds (force).

8 ft (128 cubic feet), the actual volume of wood is only 90 cubic feet—the rest is empty space between the logs. Thus, one cord contains 2.5 m^3 of wood, or about 2200 kg. The heat of combustion of wood varies between 14 and 19 MJ/kg. If one assumes a mean of 16 MJ per kilogram of wood burned, one cord delivers 35 GJ. Therefore, the cost of energy from wood was $4.3/GJ in northern California.

Still in 2008, the price of gasoline was about $3 per gallon ($1.2 per kg) although if fluctuated wildly. Since the heat of combustion of gasoline is 49 MJ/kg, gasoline energy used to cost \$24/GJ, over five times the cost from burning wood.

In California, the domestic consumer of electricity paid \$0.12 per kWh, or \$33/GJ.

From these statistics, it is clear that when we buy energy, we are willing to pay a premium for energy that is in a more convenient form—that is, for energy that has a higher **utility**. Utility is, of course, relative. To stoke a fireplace in a living room, wood has higher utility than gasoline and, to drive a car, gasoline has higher utility than electricity, at least for the time being. For small vehicles, liquid fuels have higher utility than gaseous ones. For fixed installations, the opposite is true.

The relative cost of energy is not determined by utility alone. One barrel contains 159 liters, or 127-kg of oil. With a heat of combustion of 47 MJ/kg, this corresponds to 6 GJ of energy. In mid-1990, at a price of \$12/barrel or \$2/GJ, oil cost less than wood (then at \$3.2/GJ) notwithstanding oil being, in general, more useful. However, oil prices are highly unstable depending on global political circumstances. The 2008 price of oil (that peaked well above \$100/barrel, or \$17/GJ) is now, as one might expect, substantially higher than that of wood and is one of the driving forces toward the greening of energy sources. Perhaps more importantly, there is the dangerous dependence of developed nations on oil from countries whose interests clashes with those of the West.

Government regulations tend to depress prices below their free market value. During the Carter era, natural gas was sold in interstate commerce at the regulated price of \$1.75 per 1000 cubic feet. This amount of gas yields 1 GJ when burned. Thus, natural gas was cheaper than oil or wood.

1.3 Conservation of Energy

Energy can be utilized but not consumed.[†] It is a law of nature that energy is conserved. We degrade or randomize energy, just as we randomize mineral resources when we process ores into metal and then discard the product as we do, for example, with used aluminum cans. All energy we use goes into heat and is eventually radiated out into space.

The consumable is not energy; it is the fact that energy has not yet been randomized. The degree of randomization of energy is measured by the entropy of the energy. This is discussed in some detail in Chapter 2.

[†] It is convenient to distinguish *consumption* from *utilization*. Consumption implies destruction—when oil is consumed, it disappears, being transformed mainly into carbon dioxide and water, yielding heat. On the other hand, energy is never consumed; it is utilized but entirely conserved (only the entropy is increased).

Source	Process	Power	Radiation
Solar radiation 173,000 TW	Direct reflection	52,000 TW (30%)	Short-wave radiation
	Direct conversion to heat	78,000 TW (45%)	Long-wave radiation
	Evaporation of water	39,000 TW (22%)	
	Wind & waves	3,600 TW (2%)	
	Photosynthesis	40 TW (0.02%)	
Tides		3 TW	
Geothermal	Volcanos & hot springs	0.3 TW	
	Rock conduction	32 TW	

Figure 1.1 Planetary energy balance.

1.4 Planetary Energy Balance

The relative stability of Earth's temperature suggests a near balance between planetary input and output of energy. The input is almost entirely solar radiation, which amounts to 173,000 TW (173,000 × 10^{12} W).

Besides solar energy, there is a contribution from tides (3 TW) and from heat sources inside the planet, mostly radioactivity (32 TW).

Some 52,000 TW (30% of the incoming radiation) is reflected back to the interplanetary space: it is the **albedo** of Earth. All the remaining energy is degraded to heat and reemitted as long-wave infrared radiation. Figure 1.1 shows the different processes that take place in the planetary energy balance mechanism.

The recurrence of ice ages shows that the equilibrium between incoming and outgoing energy is oscillatory. It is feared that the observed secular increase in atmospheric CO_2 might lead to a general heating of the planet, resulting in a partial melting of the Antarctic glaciers and consequent flooding of sea-level cities. The growth in CO_2 concentration is the result of the combustion of vast amounts of *fossil*[†] fuels and the destruction of forests in which carbon had been locked.

1.5 The Energy Utilization Rate

The energy utilization rate throughout the ages can only be estimated in a rough manner. In early times, humans were totally nontechnological, not even using fire. They used energy only as food, probably at a rate somewhat below the modern average of 2000 kilocalories per day, equivalent to 100 W. Later, with the discovery of fire and an improved diet involving cooked foods, the energy utilization rate may have risen to some 300 W/capita.

[†] Fuels derived from recent biomass, such as ethanol from sugarcane, do not increase the amount of carbon dioxide in the atmosphere; such fuels only recycle this gas.

In the primitive agricultural Mesopotamia, around 4000 B.C., energy derived from animals was used for several purposes, especially for transportation and for pumping water in irrigation projects. Solar energy was employed for drying cereals and building materials such as bricks. Per capita energy utilization may have been as high as 800 W.

Harnessing wind, water, and fire dates from early times. Sailboats have been used since at least 3000 B.C. and windmills were described by Hero of Alexandria around A.D. 100. By A.D. 300, windmills were used in Persia and later spread to China and Europe. Hero's toy steam engines were apparently built and operated. Vitruvius, the Roman architect whose book, first published in Hero's time, is still on sale today, discusses waterwheels used to pump water and grind cereals. In spite of available technology, ancients limited themselves to human or animal power. Lionel Casson (1981), a professor of ancient history at New York University, argues that this was due to cultural rather than economic constraints. Only at the beginning of the Middle Ages did the use of other energy sources become "fashionable." The second millennium exploded with windmills and waterwheels.

The widespread adoption of advanced agriculture, the use of fireplaces to heat homes, the burning of ceramics and bricks, and the use of wind and water led to an estimated energy utilization rate in Europe of 2000 watts per capita in A.D. 1200. Since the popular acceptance of such activities, energy utilization has increased rapidly. Figure 1.2 illustrates (a wild estimate) the number of kilowatts utilized per capita as a function of the date. If we believe these data, we can conclude that the annual rate of increase of the per capita energy utilization rate behaved as indicated in Figure 1.3. Although the precision of these results is doubtful, it is probable that the general trend is correct: for most of our history, the growth of the energy utilization rate was steady and quite modest. With the start of the industrial revolution at the beginning of the nineteenth century, this growth accelerated dramatically and has now reached a worrisome level.

The increase of the worldwide per capita energy utilization rate was driven by the low cost of oil before 1973 when it was substantially lower than now.[†] Perez Alfonso, the Venezuelan minister of oil in 1946, was among those who recognized that this would lead to future difficulties. He was instrumental in creating the Organization of the Petroleum Exporting Countries (OPEC) in 1954, not as a cartel to squeeze out higher profits but to "reduce the predatory oil consumption to guarantee humanity enough time to develop an economy based on renewable energy sources." Alfonso also foresaw the ecological benefits stemming from a more rational use of oil.

[†] In 1973, before the OPEC crisis, petroleum was sold at between $2 and $3 per barrel. The price increased abruptly, traumatizing the economy. In 2000 dollars, the pre-1973 petroleum cost about $10/bbl (owing to a 3.8-fold currency devaluation), a price that prevailed again in 1999. However, in 2006, the cost had risen to over $70/bbl. In 2008, the price of an oil barrel peaked at more than $140.

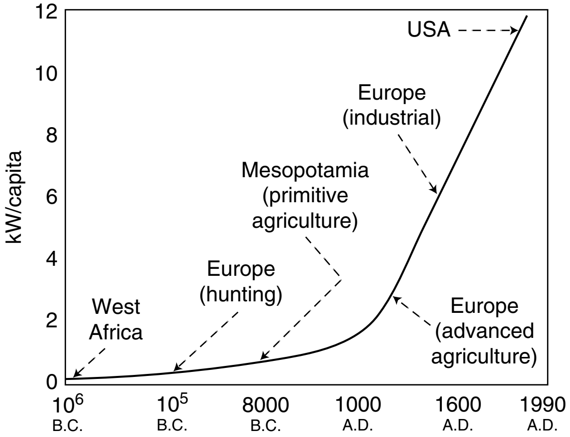


Figure 1.2 A very rough plot of the historical increase in the per capita energy utilization rate.

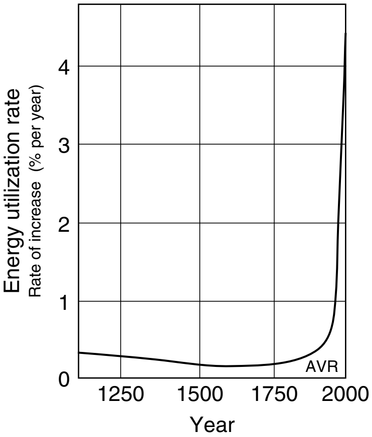


Figure 1.3 The annual rate of increase of per capita energy utilization was small up to the nineteenth century.

OPEC drove the oil prices high enough to profoundly alter the world economy, causing the overall energy utilization rate to slow its increase. Owing to the time delay between the price increase and the subsequent response from the system, several years elapsed before a new equilibrium was established. We witnessed a major overshooting of the oil-producing capacity and a softening of prices up to the 1991 Iraqi crisis.

The recent effort of less developed countries (LDCs) to catch up with developed ones has been an important factor in the increase in energy demand. Figure 1.4 shows the uneven distribution of energy utilization rate throughout the world; 72% of the world population uses less than 2 kW/ capita, whereas 6% of the population uses more than 7 kW/ capita.

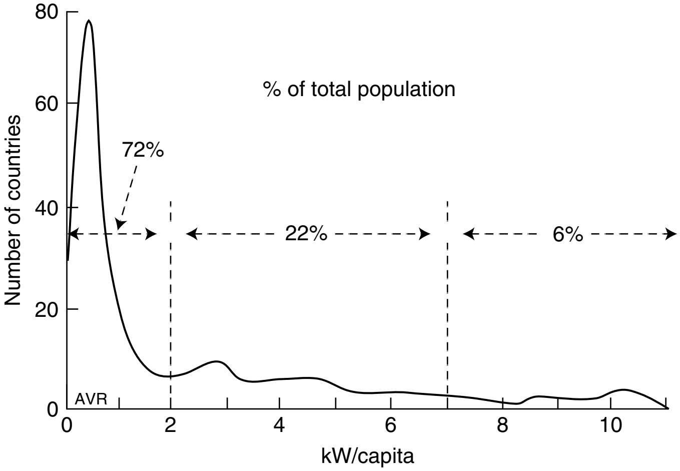


Figure 1.4 Most countries use little energy per capita, while a few developed ones use a lot.

There is a reasonable correlation between the total energy utilization rate and the annual gross national product. About 2.2 W are used per dollar of yearly GNP. To generate each dollar, 69 MJ are needed. These figures, based on 1980 dollars, vary with time, owing to the devaluation of the currency and to changing economic circumstances. In fact, it has been demonstrated that the number of megajoules per dollar decreases, during an energy crisis, while the opposite trend occurs during a financial crisis.

Further industrialization of developed countries may not necessarily translate into an increase in the per capita energy utilization rate—the trend toward higher efficiency in energy use may have a compensating effect. However, in the United States, the present decline in energy utilization[†] is due mainly to a change in the nature of industrial production. Energy-intensive primary industries (such as steel production) are phasing out owing to foreign competition, while sophisticated secondary industries (such as electronics and genetic engineering) are growing.

Technological innovation has led to more efficient energy use. Examples include better insulation in houses and better mileage in cars. Alternate energy sources have somewhat alleviated the demand for fossil fuels. Bioethanol is replacing some gasoline. It is possible that the development of fusion reactors will, one day, bring back the times of abundant energy.

Introduction of a more efficient device does not immediately result in energy economy because it takes considerable time for a new device to be widely accepted. The reaction time of the economy tends to be long. Consider the privately owned fleet of cars. A sudden rise in gasoline price

[†] American industry used less energy in 1982 than in 1973.

has little effect on travel, but it increases the demand for fuel efficiency. However, car owners don't rush to buy new vehicles while their old ones are still usable. Thus, the overall fuel consumption will only drop many years later, after a significant fraction of the fleet has been updated.

Large investments in obsolete technologies substantially delay the introduction of more efficient systems. A feeling for the time constants involved can be obtained from study of the market penetration function, discussed in Section 1.7.

1.6 The Population Explosion

In the previous section we discussed the *per capita* energy utilization rate. Clearly, the total rate of energy utilization is proportional to the planetary population, which has been growing at an accelerated rate:[†]

The most serious problem that confronts humankind is the rapid growth in population. The planet has a little more than 6 billion inhabitants, and the growth rate these last few decades has been around 1.4% per year. Almost all projections predict a population of about 7 billion by the year 2010 even if, right now, everyone were to agree on a limit of two children per family. Under present-day actuarial conditions, the population would eventually stabilize at around 11 billion by the year 2050. Population growth alone could account for a 1.4% annual increase in energy demand. In fact, the recent growth rate of energy use exceeded the population growth rate. The worldwide rate of energy use was 9 TW in 1980 and 15.2 TW in 2008, a yearly growth of 1.9%. The Energy Information Administration (EIA) has used this constant 1.9% per year growth rate to estimate an energy usage rate of slightly over 22 TW in 2030. Clearly, supplying this much energy will not be an easy task.

The constant population increase has its Malthusian side. About 10% of the world's land area is used to raise crops—it is **arable land**. (See "Farming and Agricultural Technology: Agricultural Economics: Land, Output, and Yields," Britannica Online.) Roughly 15 million km^2 or 1.5×10^9 hectares are dedicated to agriculture. Up to the beginning of the twentieth century, on average, each hectare was able to support 5 people (*Smil*), thus limiting the population to 7.4 billion people. More arable land can be found, but probably not enough to sustain 11 billion people. What limits agricultural productivity is nitrogen, one kilogram of which is (roughly) needed to produce one kilogram of protein. Although it is the major constituent of air, it is, in its elemental form, unavailable to

[†] On October 12 1999, a 3.2-kg baby was born in Bosnia. Kofi Annan, general secretary of the United Nations, was on hand and displayed the new Bosnian citizen to the TV cameras because, somewhat arbitrarily, the baby was designated as the 6 billionth inhabitant of this planet.

plants and must either be fixed by appropriate microorganisms or be added as fertilizer.

Nitrogen fertilizers are produced almost exclusively from ammonia. When used in adequate amounts, they can increase productivity by nearly an order of magnitude. The present-day and the future excess population of the planet can exist only if sufficient ammonia is produced. Although there is no dearth of raw materials (it is made from air and water), its intensive use has a serious adverse environmental effect as discussed by Smil (1997).

1.7 The Market Penetration Function

The enormous body of literature accumulated throughout the centuries makes it impossible for even the most assiduous readers to have studied the writings of all scientists and philosophers of the past. Hence, modern writers have built up a large roster of "often cited, rarely read" authors whose ideas are frequently mentioned even when only nebulously understood. This is, for instance, the case of Thomas Robert Malthus. We all have an idea that he had something to say about population growth. In 1846, Pierre François Verhulst put this population growth idea in the plausible mathematical form known now as the **Verhulst equation**. This equation is an excellent starting point to understand the problem of technological substitution, that is, the question of how a more advanced technology will replace a more cumbersome older one.

Adapting the Verhulst equation to this problem, we have

$$\frac{1}{f}\frac{df}{dt} = a(1-f), \tag{1.1}$$

where f is the fraction of the market supplied by the new technology (hence constrained to $0 \leq f \leq 1$), t is time, and a is a constant. In words, the Verhulst equation states that *the fractional rate of change in f (represented by $\frac{1}{f}\frac{df}{dt}$) must be proportional to that fraction of the market, $(1-f)$, not yet taken over by the new technology.* This makes intuitive sense. The Verhulst equation is a nonlinear differential equation whose solution is

$$\ln\left(\frac{f}{1-f}\right) = at + b, \tag{1.2}$$

b being an integration constant.

One can solve for f:

$$f = \frac{\exp(at+b)}{1+\exp(at+b)}, \tag{1.3}$$

and plot f as a function of t.

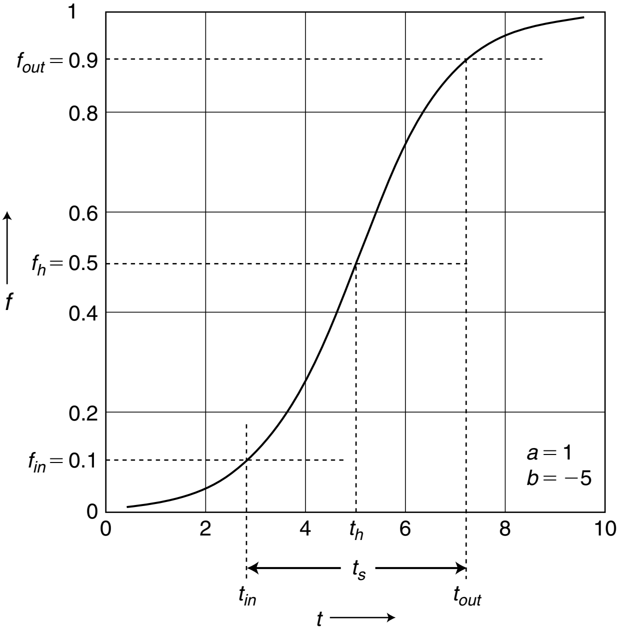


Figure 1.5 The Verhulst function.

Figure 1.5 shows the result for arbitrarily chosen $a = 1$, $b = -5$.

Equation 1.3, illustrated in Figure 1.5, is an example of the **logistics function**. One defines a **takeover time interval**, t_s, as the length of time for the function to go from f_{in}, when its value is 0.1 (at time t_{in}) to, f_{out}, when its value is 0.9 (at time t_{out}).

In 1970, two General Electric scientists, J. C. Fisher and R. H. Pry, applied the Verhulst equation to the problem of new technology market penetration. Each case considered involved only two competing technologies—an old one having a fraction, f_1, of the market and a modern one having a fraction, f_2, of the market. They used the Verhulst equation in the form of Equation 1.2 ($\ln \frac{f}{1-f}$ as a function of t), not in the form of Equation 1.3, because Equation 1.2 yields a straight-line plot and is thus much easier to handle analytically and to extrapolate.

Figure 1.6 shows the rates of penetration of the oxygen steel process into the market previously dominated by open-hearth and Bessemer technologies. It can be seen that the penetration rate was fastest in Japan (a takeover time interval of 8.3 years). In Germany and the United States, the takeover time was 11.8 years, while in the former Soviet Union it was 14.0 years. The results show surprisingly small dispersion—the observed points fall all very nearly on the straight-line regression.

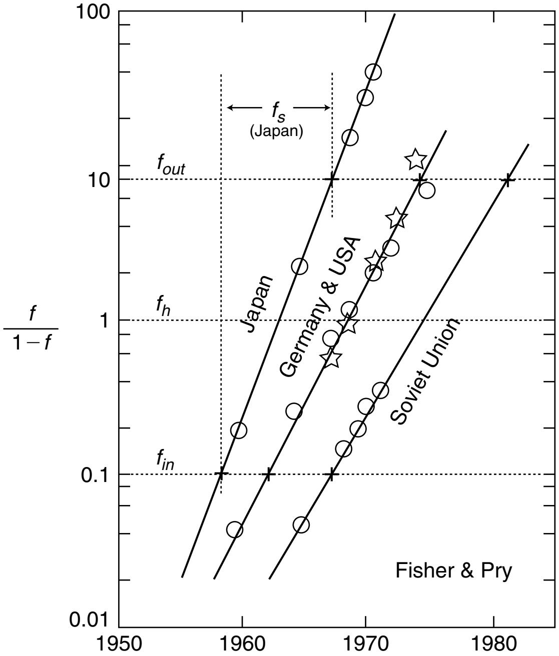


Figure 1.6 The penetration of the oxygen–steel technology in the steel production market.

As an example, in the case of Germany and the United States, the regression is

$$\ln\left(\frac{f}{1-f}\right) = 0.368t - 725, \tag{1.4}$$

where t is the time in years, expressed as 19xx.

The graph shows that the takeover times were different in different countries, and, as Fisher and Pry (1970) point out, the technique appears to work even in a centralized command economy. In addition, it turns out, as expected, that even in the same country, substitutions of different types of products have different penetration rates. Some substitutions occur rapidly, whereas others take a long time. Fisher and Pry list a number of substitutions with takeover times ranging from just over 8 years for the case of detergents replacing natural soaps in Japan to 58 years for the case of synthetic rubber replacing natural rubber in the United States.

The plots of Figure 1.6 depend on two parameters, a and b. Parameter a is related to the takeover time and determines the slope of the graph. If instead of using the actual time, t, as the ordinate, one uses a dimensionless normalized variable, τ, then all plots will have the same slope and can be integrated into a single graph. See Figure 1.7 showing how 17 different substitutions fit a single straight line with surprising accuracy.

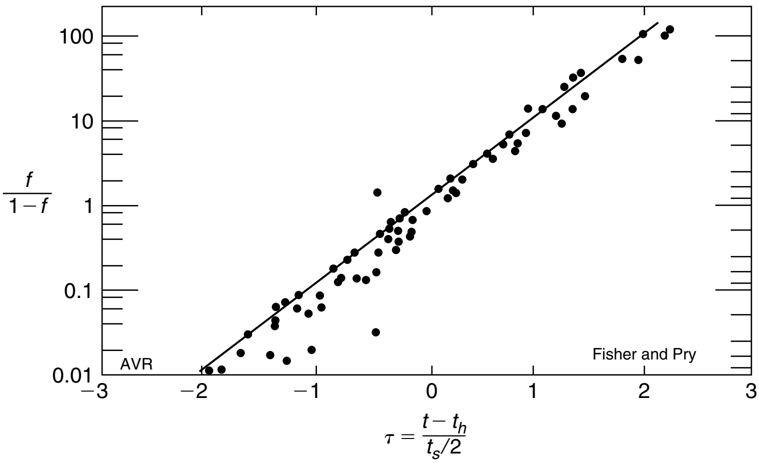


Figure 1.7 Fisher–Pry plot for 17 different substitutions.

The normalized time variable, $\tau \equiv \frac{t-t_h}{t_s/2}$, is introduced as follows. For $t = t_h, f = 0.5$ and

$$\ln \frac{f}{1-f} = at_h + b = 0 \qquad \therefore \qquad b = -at_h. \tag{1.5}$$

For $t = t_{in}, f = 0.1$ and

$$\ln \frac{f}{1-f} = at_{in} + b = -2.2. \tag{1.6}$$

Owing to the symmetry of the logistics curve, the takeover time, t_s, is twice the interval, $t_h - t_{in}$, hence, subtracting Equation 1.6 from Equation, 1.5,

$$2.2 = a(t_h - t_{in}) = at_s/2 \qquad \therefore \qquad a = \frac{2.2}{t_s/2}. \tag{1.7}$$

Thus, the market penetration formula can be written as

$$\ln \frac{f}{1-f} = 2.2\frac{(t-t_h)}{t_s/2} \equiv 2.2\tau. \tag{1.8}$$

A remarkable and useful property of the market penetration function is its insensitivity to many factors that profoundly affect the overall market. Thus, although variations in political or geopolitical circumstances can substantially affect market volume, they frequently have a minimal effect on the fractional market share, probably because they influence simultaneously all competing technologies.

In the case of a simple substitution of an old technology by a newer one, it is clear that if the newer is progressively increasing its market share, then the older must be progressively abandoning the market. If $\ln(\frac{f_2}{1-f_2}) = at+b$

then $\ln(\frac{f_1}{1-f_1}) = -at - b$, because the total market is $\sum f = 1$. Since the behavior of the technology being replaced is, in this case, simply a mirror image of the advancing technology, graphs are shown only for the latter. This is not true for the case of multivariate competition when several different technologies compete to fill market needs.

Fisher and Pry's results are entirely empirical. Cesare Marchetti (1976), working at the International Institute for Applied System Analysis (IIASA) in Austria, still using an empirical approach, extended the Fisher–Pry idea in two meaningful ways. One was to make it possible to consider cases in which the market for a given product was supplied by more than two technologies. To this effect, Marchetti introduced the rule of first in, first out. The other was the application of the Fisher–Pry procedure to the case of various *energy* sources. This allows the prognostication of the market share of individual energy-supplying technologies. In his 1976 paper, Marchetti presents the available data on the market share of different fuels used in the United States during the 1850 to 1975 period. From these data, he obtained the different coefficients of Equation 1.2 for each fuel. Wood was already abandoning the market with a characteristic **abandonment time** of 60 years being replaced by coal (takeover time of 66 years). This substitution was driven by the much greater usefulness of coal in driving the locomotives of the expanding railroad system. Increasing use of oil and natural gas, beginning at the turn of the twentieth century, caused the turnaround of the coal share (but not of the total coal use). Coal started abandoning the market with a characteristic time of 99 years. Initially, oil's takeover time was 52 years, but then rose to 135 years and showed signs of turning around in the early 1970s. All this is illustrated in Figure 1.8.

As an exercise in prognostication, Marchetti, using the trend lines of Figure 1.8 *derived only from data before 1935*, calculated the behavior of the oil market share, employing the formula,

$$f_{oil} = 1 - (f_{coal} + f_{gas}). \tag{1.9}$$

The results are displayed in Figure 1.9. They are very accurate, a fact that led Marchetti to comment: "we were able to predict the fractional market share of oil in the U.S.A. up to 1970 with a precision of better than one percent." Alas, "It's tough to make predictions, especially about the future."[†] If we extend Marchetti's graph to 2008, we find that the nice regularity of the behavior of the coal and gas shares, on which the prognostication is based, breaks down badly in modern times (see Figure 1.10). To understand why, we must refer to the work of Václav Peterka (1977), which, dropping the empiricism of previous authors, tried to put the analysis of market penetration on a firmer theoretical basis. Peterka carefully defines the conditions under which the empirical models hold.

[†] Attributed to numerous sources, from Mark Twain to Niels Bohr to Yogi Berra.

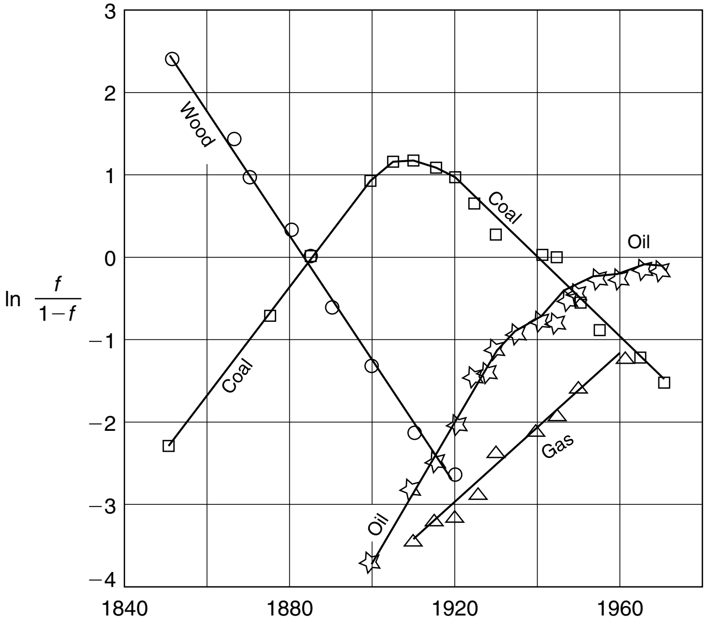


Figure 1.8 Market share of different fuels in the United States reported by Marchetti (1976). Plots of ln(f/1-f) vs. time exhibit surprising regularity.

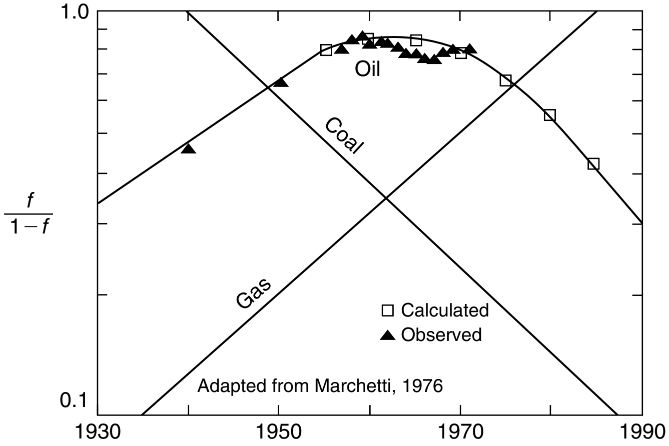


Figure 1.9 Prognostication of the behavior of the oil market share up to 1990 using only observed data from before 1935. The results are surprisingly accurate and predict the peak that actually occurred in the 1960s. This is totally unrelated to the OPEC crisis in the early 1970s.

Peterka argues that any scientific forecasting must be based on certain a priori assumptions. In the case being discussed here, a fundamental assumption is that there be no *external* infusion of capital once the

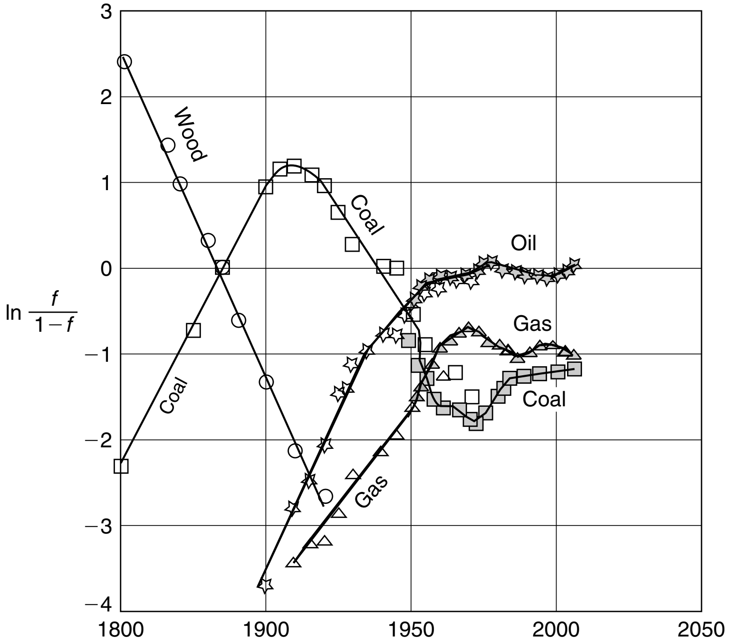


Figure 1.10 The regular and predictable behavior of the market penetration function for coal and gas used by Marchetti breaks down if modern data are added to his graph of Figure 1.8.

technology has established itself. This is self evident—if during the penetration period, a substantial increase in capital becomes available, this will alter the rate of penetration, even though it may not increase the profitability of the enterprise. It would be of great value if it were possible to estimate how much it would cost to accelerate the penetration by a given amount. Unfortunately, this is not yet possible. The assumption above implies that when a technology starts to penetrate the market, it must already be well developed and its degree of maturity determines the eventual penetration rate. Thus, "the magnitude of the original external investment actually determines the initial conditions for the model" (Peterka, 1977).

The market penetration rules discussed in this subsection provide a powerful tool for planning, but must be used with caution and with close attention to possible violations of implicit assumptions.

1.8 Planetary Energy Resources

In Section 1.5, we pointed out that the rate of per capita energy utilization rose rapidly in the last century. This, combined with the fast increase in population mentioned in Section 1.6, leads one to the inescapable conclusion that we are facing a serious challenge if we hope to maintain

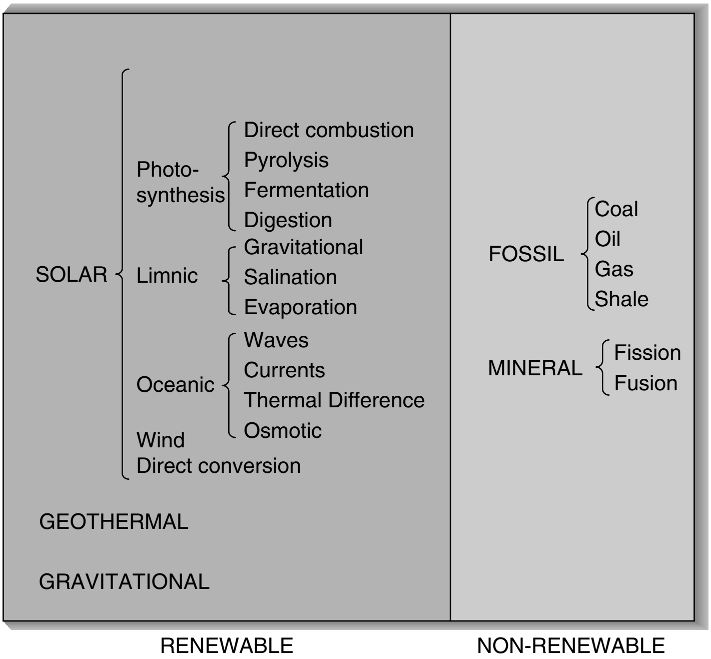


Figure 1.10 The energy resources of Earth.

these trends in the foreseeable future. To investigate what can be done to resolve this difficulty, we must first inquire what energy resources are available (Section 1.8) and next (Section 1.9) how we are using the resources at present.

Figure 1.10 shows the planetary energy resources. These can be renewable or nonrenewable.

Geothermal energy has been used for a very long time in Iceland and more recently in Italy, New Zealand, and the United States. In many places, it is possible to take advantage of the stability of the ground temperature a few meters below the surface. The ground can thus be used as a source of heat in the winter and of cold in the summer.

Gravitational energy—that is, energy from tides (see Chapter 16)—has been used in France. Tides can only be harnessed in certain specific localities of which there is a limited number in the world. Gravitational energy is also important in all hydroelectric plants.

Of the renewable resources, solar energy is by far the most abundant. A small part of it has been absorbed by plants and, over the eons, has been stored as coal, oil, and gas.

Estimates of reserves, fossil or nuclear, are extremely uncertain and are sure to be greatly underestimated because of incomplete prospecting. Table 1.4 gives us only a very rough idea of our fossil fuel reserves, and Table 1.5 shows an even more uncertain estimate of reserves of fissile

Table 1.4 Known Fossil Fuel Reserves

Methane clathrate	>100,000 EJ (1998)
Coal	39,000 EJ (2002)
Oil	18,900 EJ (2002)
Gas	15,700 EJ (2002)
Liquefied gas	2,300 EJ (2002)
Shale	16,000 EJ (?)

Table 1.5 Known Reserves of Fissionable Materials[†]

^{235}U	2,600 EJ
^{238}U	320,000 EJ
^{232}Th	11,000 EJ

[†] *Does not include the USSR and China.*

materials. The estimates of nuclear fuels do not include the reserves of the former Soviet Union and China. Values given in the tables are very far from precise. They probably represent a *lower* limit, because people who estimate these numbers tend to be conservative, as testified by the secular *increase* in proved reserves: proved reserves of dry natural gas, 2200 EJ in 1976, rose to 6200 EJ in January 2007, notwithstanding the substantial consumption of gas in the intervening years. A similar situation exists with respect of proved oil reserves: 7280 EJ in 2002 and 7900 EJ in 2007. For oil and gas, the table lists the sum of proved reserves, reserve growth, and undiscovered reserves.

Proved reserves are fuels that have been discovered but not yet produced. Proved reserves for oil and gas are reported periodically in the *Oil and Gas Journal.*

Reserve growth represents the increase in the reserves of existing fields owing to further development of these fields and to the introduction of better technology for their extraction.

Undiscovered reserves represent the best possible guess of the magnitude of plausible new discoveries.

Reserve growth and undiscovered reserves are estimated by the U.S. Geological Survey (USGS). For example, in 2002 the *Oil and Gas Journal* reported proved reserves of oil of 7280 EJ, and the USGS estimated a growth of 4380 EJ and undiscovered oil reserves amounting to 5630 EJ, adding up to the total of 18,900 EJ listed in the table. For coal, the table shows only proved reserves. The total reserves for this fuel are thus substantially larger than listed.

One number that is particularly uncertain is that referring to hydrated methane. William P. Dillon, a geologist of the USGS, testified in the

U.S. House of Representatives in 1998 that "the amount of methane contained in the world's gas hydrate accumulations is enormous, but estimates of the amounts are speculative and range over three orders-of-magnitude from about 100,000 to 270,000,000 trillion cubic feet [100,000 to 270,000,000 EJ] of gas." We, being ultraconservative, listed the lower figure.

Methane Clathrate

Clathra is the Latin word for bar or cage.

Atoms in a number of molecules group themselves in such a fashion that a cavity (or cage) is left in the center. The most famous of these arrangements is the "buckyball," a molecule consisting of 60 carbon atoms arranged as a hollow sphere capable of engulfing a number of substances. Buckyballs, discovered in the early 1980s, are not alone among "hollow" molecules. Under appropriate circumstances, water will freeze, forming a cage consisting sometimes of 20 water molecules, but more commonly of 46 water molecules. The configuration is unstable (it decays into a common ice crystal) unless certain gases become trapped in the central cage of the large molecule. Gases commonly trapped are methane, ethane, propane, isobutane, n-butane, nitrogen, carbon dioxide, and hydrogen sulfide.

The ice crystal consisting of 46 water molecules is able to trap up to 8 "guest" gas molecules (a water-to-gas ratio of 5.75:1). In natural deposits, methane is by far the most abundant and the one of greatest interest to the energy field. Usually, up to 96% of the cages are fully occupied. These solid hydrates are called **clathrates**.

The density of the clathrate is about 900 kg/m^3. This means that the methane is highly compressed. See Problem 1.28. Notwithstanding its low density, water ice clathrate does not float up from the bottom of the ocean because it is trapped beneath the ocean sediment.

Clathrates form at high pressure and low temperature under sea and are stable at sufficient depth. The methane is the result of anaerobic digestion of organic matter that continuously rains down on the ocean floor. See Chapter 13.

There is no mature technology for the recovery of methane from clathrates. Proposed processes all involve destabilizing the clathrate and include:

1. Raising the temperature of the deposits.
2. Depressurizing the deposits.
3. Injecting methanol or other clathrate inhibitors.

The third process may be environmentally undesirable.

(*Continues*)

(*Continued*)

There are dangers associated with methane clathrate extraction. The most obvious ones are the triggering of seafloor landslides and the accidental release of large volumes of methane into the Earth's atmosphere where it has a powerful greenhouse effect. Some scientists attribute the extinction that marked the end of the Permian era (300 to 250 megayears ago) to an enormous bubbling up of methane. The Permian-Triassic extinction (P-Tr extinction) was the worst catastrophe to hit the biosphere of Earth—96% of all ocean species disappeared, together with 70% of land species.

1.9 Energy Utilization

Most of the energy currently used in the world comes from nonrenewable sources as shown in Figures 1.11 and 1.12, which display energy sources in 2001 for the whole world and for the United States, in 2008, respectively. The great similarity between these two charts should not come as a surprise in as much as the United States uses such a large fraction of the total world consumption.

What may be unexpected is that most of the renewable resources (geothermal, biomass, solar, and wind) make such a small contribution to the overall energy picture. Figure 1.13 shows that as late as 2008 only 8.3% of the energy used to generate electricity in the United States came from renewable sources. Of these, 83% came from hydroelectrics. Thus, only 2% of the total came from the remaining renewables.

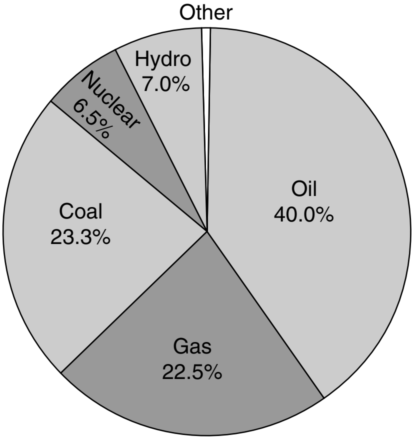


Figure 1.11 Energy sources in the world.

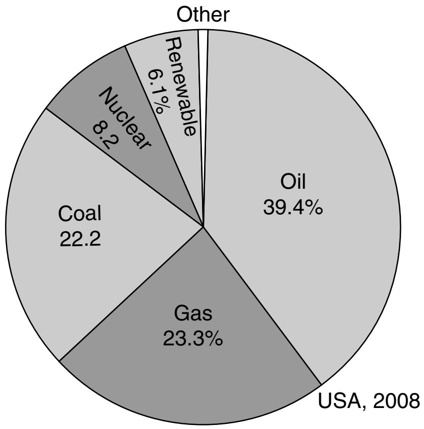
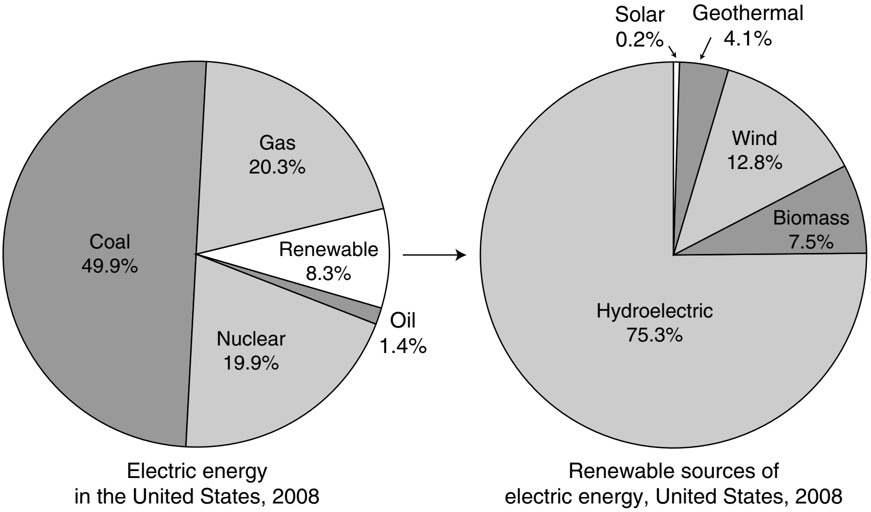


Figure 1.12 Energy sources in the United States.



Data from EIA

Figure 1.13 Sources of electric energy in the United States.

Disappointingly so far, the contribution of solar and wind energy has been very small. But since about 1964, this has begun to change significantly.[†]

For all sources of energy, the cost of the plant is proportional to the installed capacity, while the revenue is proportional to the energy generated. The **plant utilization factor** is the ratio of the energy produced

[†] For fairly up-to-date statistics on the production of renewable energy, consult <http://www.earth-policy.org/Indicators/index.htm>.

to that which would be produced if the plant operated uninterruptedly at full capacity (Table 1.6). Observe the extremely high utilization factor of nuclear plants. Wind generators operate with a rather small plant factor (≈ 30%) as a result of the great variability of wind velocity. Although specific data for solar plants are not available, they also suffer from a low utilization factor owing to the day/night cycle and the vagaries of meteorological conditions.

It is of interest to know which are the main users of energy in the United States. (See Figure 1.14) American residences account for nearly 20% of all energy used. Most of it is employed for ambient heating, an area in which considerable economy can be realized, especially through better home design and use of geothermal sources.

Waste heat from electric power plants can be used for residential and commercial water and space heating and constitutes a form of **district heating**. The other form uses dedicated centrally located boilers, not a cogeneration scheme. Some decades ago, when steam plants had an average efficiency of 30%, a whopping 70% of the fuel energy was either thrown away or was piped as hot water to consumers. The latter option increased the overall system efficiency to more than 50%. Currently, steam plants

Table 1.6 Electric Energy Use, United States 2005

Source	Used (EJ)	Capacity (GW)	Utilization factor
Thermal	11.01	840	41.5%
Nuclear	2.81	106	84.4%
Hydro	0.97	97	31.8%
Other renewable	0.34		

These data are from the EIA.

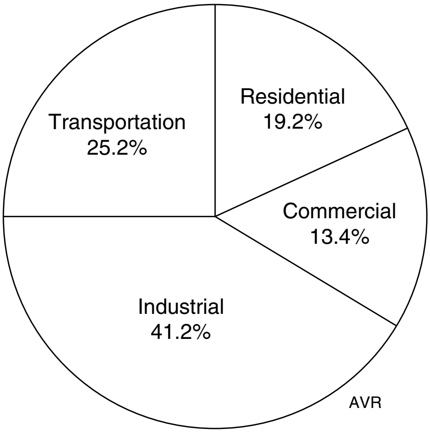


Figure 1.14 The different users of energy in the United States.

Table 1.7 Relative Merit of Different Light Sources

Type	**Efficiency**	**Lifetime**	**Merit**
100 W incandescent	2.5%	1,000	2,500
Compact fluorescent	7%	10,000	70,000
white LED[†]	22%	50,000	1,000,000

[†] *The actual efficiency of commercially available white LEDs is substantially less than the listed 22%; it is about 13%. However, laboratory prototypes have demonstrated the efficiency shown in the table.*

have, by themselves, efficiencies that exceed 50%,[†] and district heating can boost the overall efficiency to some 90%. This comes at some considerable initial cost, so the economic benefits are only realizable in the long term. District heating requires the location of power plants in densely populated areas; consequently, it is nonadvisable in the case of nuclear plants and large fossil-fueled installations. However, fuel cell plants (see Chapter 9), being noiseless and pollution free, can be placed in a downtown area.

Although the largest district heating system in the world is the one operated by Con Edison Steam Operations, active since 1882, a subsidiary of Consolidated Edison of New York, the technology is, in relative terms, much more popular in Europe.

It is probably in the transportation sector (25% of the total energy use) that modern technology can have the most crucial impact. We are still enamored of heavy, overpowered cars that realize less than half the fuel efficiency possible with *current* technology. Thus, a cultural change would be desirable. The transition to more rational personal transportation has started with the introduction of hybrid cars, soon to be followed by plug-in hybrids and perhaps by both electric and fuel cell cars, both of which promise to increase automobile efficiency while reducing pollution.

Another area in which efficiency can be improved by nearly *one order of magnitude* is illumination, an important use in both residential and commercial energy use. About 10% of the energy used by residences go toward illumination. We are in a period of transition from the extremely inefficient incandescent bulbs to the compact fluorescent lamp to the super-efficient white light-emitting diodes (LEDs).

In Table 1.7, "Merit" is defined as the product of the efficiency (expressed in percent) by the lifetime in hours. According to this arbitrary criterion, LEDs have the potential of being 4000 "better" than incandescent bulbs. Of course, a number of technical problems as well as the high cost of the LEDs must be addressed before they become the standard source of light.

[†] The following is a quote from General Electric: "GE's H System—one of the world's most advanced combined cycle system and the first capable of breaking the 60 percent efficiency barrier—integrates the gas turbine, steam turbine and heat recovery steam generator into a seamless system."

American industry's relative use of energy may decrease even in the face of an expansion of this sector because of the progressive shift of emphasis from an energy-intensive industry, such as iron and steel, to more sophisticated activities that have a low-energy demand per dollar produced.

1.10 The Ecology Question

We have shown that there is an almost unavoidable trend toward increasing energy utilization. We have also pointed out that at present the energy used is at least 85% of fossil origin. Finally, we have observed that the fossil fuel reserves seem ample to satisfy our needs for a good fraction of the next millennium. So, what is the problem?

Most of the easily accessible sources of oil and gas have already been tapped. What is left is getting progressively more expensive to extract. Thus, one part of the problem is economical. Another is political—most of the fuel used by developed nations is imported (using the large American reserves is unpopular, and politicians hesitate to approve such exploration). This creates an undesirable vulnerability. The major problem, however, is ecological. Fossil fuels are still the most inexpensive and most convenient of all energy resources, but their use pollutes the environment, and we are quickly approaching a situation in which we can no longer dismiss the problem or postpone the solution.

By far, the most undesirable gas emitted is carbon dioxide whose progressively increasing concentration in the atmosphere (from 270 ppm in the late 1800s to some 365 ppm at present) constitutes a worrisome problem. It is sad to hear influential people (among them, some scientists) dismiss this problem as inconsequential, especially in view of the growing signs of a possible runaway ecological catastrophe. For instance, in the last few decades, the thickness of the north polar ice has decreased by 40% and in the first year of the current millennium, a summertime hole appeared in the polar ice. Since increased concentrations of CO_2 can lead to global warming, some people have proposed increasing the emission of SO_2 to stabilize the temperature because of the cooling effect of this gas. Even ignoring the vegetation-killing acid rain that would result, this proposal is equivalent to balancing a listing boat by piling stones on the other side.

Public indifference to the CO_2 problem may partially be due to the focus on planetary temperature rise. Although the growth in CO_2 concentration is very easily demonstrated, the conclusion that the temperature will rise, though plausible, is not easy to prove. There are mechanisms by which an increase of greenhouse gases would actually result in a *cooling* of Earth. For instance, increasing greenhouse gases would result in enhanced evaporation of the tropical oceans. The resulting moisture, after migrating toward the poles, would fall as snow, thereby augmenting the albedo of the planet and thus reducing the amount of heat absorbed from the sun.

The Kyoto Treaty aims at curbing excessive carbon dioxide emissions. It is worth noting that China, the world's major CO_2 emitter,[†] is exempt from the treaty's restrictions, as are numerous other countries, such as India and Brazil. The United States and Australia have never ratified the treaty and are, therefore, also exempt. There are many contributors to CO_2 pollution, but by far the largest single culprit is the coal-fired power plant, which emits 30% of the amount of carbon dioxide dumped into the atmosphere.[††] The problem will not go away unless the coal-to-electricicity situation is corrected.

Some scientists and engineers who are less concerned with political correctness are investigating techniques to reduce (or at least, to stabilize) the concentration of atmospheric carbon dioxide. This can, in principle, be accomplished by reducing emissions or by disposing of carbon dioxide in such a way as to avoid its release into the air. Emissions can be reduced by diminishing overall energy consumption (an utopian solution), by employing alternative energy sources, by increasing the efficiency of energy use, and by switching to fuels that yield more energy per unit amount of carbon emitted. It is known that 1 kmole of methane, CH_4, when burned yielding liquid water and carbon dioxide, releases 889.6 MJ and emits 1 kilomole of carbon: it generates heat at a rate of 889.6 MJ per kilomole of carbon. n-heptane, C_7H_{16}, which can represent gasoline, releases 4820 MJ of heat per kilomole burned and emits 7 kilomoles of CO_2—a rate of 688.6 MJ per kilomole of carbon. Clearly, the larger the number of carbon atoms in the hydrocarbon molecule, the lower the ratio of the heat of combustion to the amount of carbon dioxide emitted because the ratio of hydrogen to carbon decreases. This is one reason for preferring methane to oil and oil to coal.

Renewable forms of energy are attractive but, at least for the present, they are too expensive to seriously compete with fossil fuels. Hence, methods for reducing carbon dioxide emission are under intense investigation. All these methods have two stages: carbon dioxide capture and carbon dioxide disposal or sequestration. The capture stage is described, superficially, in Chapter 10. In lieu of sequestration, the captured and purified gas gan be sold to, for instance, the carbonated drink industry. But this can only take care of a minute fraction of the total CO_2 involved.

In order to select a technique, for carbon dioxide disposal, it is important to inquire where nature stores the existing carbon. Table 1.8 shows the estimated amount of carbon stored in different places.

Methods to dispose of CO_2 could include the following.

[†] In mid-2007, China surpassed the United States as the major carbon dioxide emitter.

[††] It is somewhat surprising that a 1 GW coal-fired plant can emit 100 times more radioactive isotopes (because of the radioactive traces in common coal) than a nuclear plant of the same power.

Table 1.8 Stored Carbon on Earth

Oceans	45×10^{15} kg
Fossil fuels	10×10^{15} kg
Organic matter	2.4×10^{15} kg
Atmosphere	0.825×10^{15} kg

1.10.1 Biological

For photosynthesis to remove carbon dioxide from the air, the biomass produced must be preserved; it cannot be burned or allowed to rot. There is a limited capacity for this method of CO_2 disposal. The current biological uptake rate of carbon is only 0.002×10^{15} kg year.

1.10.2 Mineral

CO_2 is removed naturally from the air by forming carbonates (mainly of magnesium and calcium). The gas is removed by reacting with abundant silicates, a process too slow to cope with human-made emissions.

Ziock et al. (2000) propose the use of magnesium silicates to sequester carbon dioxide at the point where fossil fuels are burned. Enormous deposits of magnesium oxide-rich silicates exist in the form of olivines and serpentines.

For serpentine, the net reaction involved is

$$\mathrm{Mg_3Si_2O_5(OH)_4 + 3CO_2 \rightarrow 3MgCO_3 + 2SiO_2 + 2H_2O}$$

Notice that the end products are materials that already exist naturally in great abundance. Substantial additional research is needed to improve the proposed disposal system and to make it economical.

1.10.3 Subterranean

CO_2 can be sequestered underground as the oil industry has been doing (for secondary oil recovery) for more than 50 years. The volume of the exhaust gases of a combustion engine is too large to be economically stored away. It is necessary to separate out the carbon dioxide, a task that is not easy to accomplish. One solution is proposed by Clean Energy Systems, Inc. of Sacramento, California. The suggested equipment extracts oxygen from air (a well-developed process) and mixes this gas with the fuel. Combustion produces steam and CO_2 at high temperature and pressure and drives several turbines at progressively lower temperatures. The water in the final exhaust is condensed and recycled leaving the carbon dioxide to be pumped, at 200 atmospheres, into an injection well. At present, no turbines exist capable of operating at the high temperature (over 3000 C) of the combustion products. See Anderson et al. (1998).

1.10.4 Undersea

The Norwegian government imposes a stiff carbon dioxide emission tax that has made it economical to install disposal systems that pump the gas deep into the ocean. It appears that liquid carbon dioxide can be injected into the seas at great depth and that it will stay there for a long time. Again, more work is required to determine the feasibility of the scheme.

Carbon in the Atmosphere

How much carbon is there in the atmosphere?

The surface area of Earth is 510×10^{12} m^2, while the scale height of the atmosphere is around 8800 m (see the section on Boltzmann's law in Chapter 2). Consequently, the volume of air (all of it compressed to 1 atmosphere pressure) is $510 \times 10^{12} \times 8800 = 4.5 \times 10^{18}$ m^3.

Present-day atmospheric CO_2 concentration is 13.5×10^{-6} kmol/m^3. The atmosphere contains $13.5 \times 10^{-6} \times 4.5 \times 10^{18} = 61 \times 10^{12}$ kmol of CO_2 and, therefore, 61×10^{12} kmol of carbon. Since the atomic mass of carbon is 12 daltons, the mass of carbon in the atmosphere is 0.73×10^{15} kg. Compare with the 0.825×10^{15} kg of the table.

A simpler way to achieve about the same result is to consider that the atmospheric pressure at sea level is 1 kg/cm^2 or 10^4 kg/m^2. Consequently, the total mass of the atmosphere is $510 \times 10^{12} \times 10^4 = 510 \times 10^{16}$. Of this, 360×10^{-6} is carbon dioxide and 12/44 is carbon. The carbon content of the atmosphere is $510 \times 10^{16} \times 365 \times 10^{-6} \times 12/44 = 0.51 \times 10^{15}$ kg, a result comparable with the previous one.

1.11 Nuclear Energy

Chemical fuels, such as oil or methane, release energy when the atoms in their molecules are rearranged into lower energy configurations. The energies involved are those of molecular binding and are of the order of tens of MJ/kg. When the *components* of an atom are arranged into lower energy configurations, then the energy released is orders of magnitude larger (hundreds of TJ/kg) because of the much larger intra-atomic binding energies.

The internal structure of atoms can be changed in different ways:

1. An atomic nucleus can be bombarded with a neutron, absorbing it. A different atom emerges.
2. An atom can spontaneously change by emitting either electrons (beta-rays) or helium nuclei (alpha-rays). Such radioactive decay releases energy, which can be harvested as, for instance, it is done in **Radioisotope Thermal Generators** (RTGs). (See Chapter 5).

3. Atoms with large atomic numbers can be made to break up into smaller atoms with the release of energy. This is called **nuclear fission** and requires that the atomic number, Z, be larger than 26.
4. Atoms with low atomic numbers can be assembled into a heavier one, releasing energy. This is called **nuclear fusion** and requires that the final product have an atomic number smaller than 26.[†]

Currently, only two techniques are used to produce energy from nuclear sources: the RTG mentioned above and nuclear fission reactors.[††] But, nuclear energy has developed a bad reputation, especially after the Chernobyl accident in 1986. Nevertheless, it is a source of substantial amounts of energy in many countries. According to the Energy Information Administration, EIA, since 1998, the number of nuclear plants in the United States has remained unaltered at 104. Nevertheless, there has been a 2% per year secular increase in the generation of nuclear electricity owing mostly to an improvement of the plant utilization factor from 78.2% in 1998 to over 94% in 2007. It appears that after 2008, a number of new reactors may be purchased.In 2007, the United States led the world in installed capacity—104 GW—followed by France (63 GW) and Japan (47.6 GW). The utilization factor of nuclear plants that year was excellent. In the United States, it was over 94%, in France, 77.5%, and in Japan, 68.9%.

Of the total electricity generated, nuclear plants in the United States (2008) contributed a relatively modest 19.9%, while in France, heavily reliant on this form of energy, the contribution was 76.1%. In Japan, it was 34.6%. In 2000, Germany decided to phase out its 19 nuclear power plants. Each one was assigned a 32-year life after which they would be deactivated. Many plants have already operated more than half of their allotted lifetime.

The cost of nuclear electricity is high, about double that from fossil fuel. In the United States (1996), it was 7 cents/kWh, whereas that of a state of the art natural gas plant was 3 cents/kWh (Sweet, William #1). Advanced reactor designs may bring these costs down considerably while ensuring greater safety (Sweet, William #2). This promised reduced cost combined with the ecological advantage of no greenhouse gas emission—a growing concern—may lead to a renewed popularity for nuclear generators.

The major objection to fission-type reactors is not so much the danger of the operation of the power plants (the Chernobyl accident was perfectly avoidable), but rather the problem of disposing of large amounts of long-lived radioactive by-products. If the need for such disposal can be avoided, then there is good reason to reconsider fission generators as an important contributor to the energy supply system, especially if they are not restricted to the use of the rare ^{236}U fuel the way present-day reactors are.

[†] All are transmutations, the age-old dream of medieval alchemists.

[††] Cannons preceded by centuries the invention of heat engines. Nuclear bombs were used before nuclear reactors—fusion has for decades been used in thermonuclear bombs, but its use in reactors still seems far into the future.

Specifications of new-generation nuclear fission reactors might include (not necessarily in order of priority), the following items:

1. Safety of operation (including resistance to terrorist attacks)
2. Affordability
3. Reliability
4. Absence of weaponizable subproducts
5. Absence of long-lived waste products
6. Ability to transmute long-lived radioactive waste products from old reactors into short-lived radioactive products

The U.S. Department of Energy is funding research (2004) in technologies that might realize most of these specifications. One of these is the **heavy-metal fast breeder reactor** technology. It appears that this type of reactor may be able not only to produce waste with relatively short half-lives (100 years contrasted with 100,000 years of the current waste), but in addition may be able to use current-type waste as fuel. Furthermore, because heavy-metal reactors operate at high temperatures (yet at low pressures), the thermolytic production of hydrogen (see Chapter 10) for use in fuel cell-driven automobiles looms as a good possibility. For further reading on this topic see Loewen (2004).

The waste disposal problem is absent in fusion devices. Unfortunately, it has been impossible to demonstrate a working prototype of a fusion machine, even after several decades of concerted research.

To do even a superficial analysis of the technical aspects of nuclear reactions, we need to know the masses of the atoms involved (see Table 1.9). Most of the mass values are from Richard B. Firestone. Those marked with a ♣ are from Audi and Wapstra (1993), and the one marked with a • is from a different source. It can be seen that the precision of the numbers is very large. This is necessary because, in calculating the energy released in a nuclear reaction, one uses the small difference between large numbers, which is, of course, extremely sensitive to uncertainties in the latter.

The listed values for the masses of the nucleons (the proton and the alpha in the table) are nearly the values of the masses of the corresponding atoms minus the mass the electron(s). On the other hand, there is a large difference between the the mass of a nucleon and the sum of the masses of the component protons and neutrons. Indeed, for the case of the alpha, the sum of the two protons and the two neutrons (4.03188278 daltons) exceeds the mass of the alpha (4.001506175 daltons) by 0.030376606 daltons—about 28 MeV of mass. This is, of course, the large **nuclear binding energy** necessary to overcome the great electrostatic repulsion between the protons.

1.11.1 Fission

There are at least four fissile elements of practical importance: ^{233}U, ^{235}U, ^{239}Pu, and ^{241}Pu. Of these, only ^{235}U is found in nature in usable

Table 1.9 Masses of Some Particles Important to Nuclear Energy

Particle	Symbol	Mass	Mass
		(daltons[†])	(kg)
electron ♣	e	0.00054579903	$9.1093897 \times 10^{-31}$
muon ·	μ	0.1134381	1.883566×10^{-28}
proton ♣	p	1.007276467	1.672648×10^{-27}
neutron ♣	n	1.008664909	$1.6749286 \times 10^{-27}$
$^1_1$H		1.007825032	$1.673533967 \times 10^{-27}$
$^2_1$D		2.014101778	$3.344496942 \times 10^{-27}$
$^3_1$T		3.016049278	$5.008271031 \times 10^{-27}$
$^3_2$He		3.016029319	$5.008237888 \times 10^{-27}$
$^4_2$He		4.002603254	$6.646483555 \times 10^{-27}$
alpha ♣	α	4.001506175	$6.644661810 \times 10^{-27}$
$^5_3$Li		5.01254	$8.323524107 \times 10^{-27}$
$^6_3$Li		6.015122794	$9.988353127 \times 10^{-27}$
$^7_3$Li		7.01600455	$1.165035751 \times 10^{-26}$
$^{10}_5$B		10.012937	$1.662688428 \times 10^{-26}$
$^{11}_5$B		11.009305	1.82814×10^{-26}

[†] *The dalton is not yet the official name for the atomic mass unit.*

Table 1.10 Uranium Isotopes

Isotope	Abundance	Lifetime
	(%)	(years)
^{238}U	99.283	4.5×10^9
^{235}U	0.711	7.1×10^8
^{234}U	0.005	2.5×10^5

quantities; ^{233}U, ^{239}Pu, and ^{241}Pu must be created by transmutation of "fertile" materials, respectively, ^{232}Th, ^{238}U and ^{240}Pu. The ^{240}Pu element must itself be created artificially from ^{239}Pu.

Uranium isotopes cover the range from 227 to 240 daltons, but natural uranium contains only a small percentage of the fissile material:

It is estimated that the Western world has reserves of uranium oxide (U_3O_8) amounting to some 6×10^9 kg, but only 34×10^6 kg are fissile, corresponding to an available energy of 2600 EJ. Compare this with the 40,000 EJ of available coal energy.

Nuclear fission reaction (with a corresponding release of energy) occurs when a fissile material interacts with neutrons. Consider ^{235}U:

$$^{235}_{92}\mathrm{U} + {}^1_0\mathrm{n} \rightarrow {}^{236}_{92}\mathrm{U}. \tag{1.10}$$

The resulting ^{236}U decays with the emission of alpha particles (lifetime 7.5 seconds). More importantly, the uranium also suffers spontaneous fission; that is, under the proper circumstances, ${}^{235}_{92}\mathrm{U}$ absorbs a neutron, and the resulting atom splits into smaller nuclei simultaneously releasing, on average, 2.5 neutrons and about 3×10^{-11} joules of energy:

$$ {}^{235}_{92}\mathrm{U} + {}^{1}_{0}\mathrm{n} \rightarrow 2.5\ {}^{1}_{0}\mathrm{n} + \text{fission products} + 3 \times 10^{-11}\ \mathrm{J}. \tag{1.11} $$

Per kilogram of ${}^{235}_{92}\mathrm{U}$, the energy released is

$$ \frac{3 \times 10^{-11} \frac{\mathrm{J}}{\mathrm{atom}} \times 6 \times 10^{26} \frac{\mathrm{atoms}}{\mathrm{kmol}}}{235 \frac{\mathrm{kg}}{\mathrm{kmol}}} = 77\ \mathrm{TJ/kg}. $$

However, the situation is somewhat more complicated than suggested by the equation above because more energy and additional neutrons are produced by the radioactive decay of the fission products. These additional neutrons are called **delayed neutrons**. Compare this with chemical reactions that involve energies of the order of a few tens of MJ/kg.

When Otto Hahn, demonstrated uranium fission in 1939, it became immediately obvious that a sustained "chain" reaction might be achievable—all that was needed was to cause one of the emitted neutrons to split a new uranium atom. Using natural uranium, this proves difficult because of the small percentage of the fissile ${}^{235}\mathrm{U}$. The emitted neutrons have a much greater probability of being absorbed by the abundant ${}^{238}\mathrm{U}$—the reaction simple dies out. The solution is to "enrich" the uranium by increasing the percentage of ${}^{235}\mathrm{U}$. This is a complicated and expensive process because one cannot use chemistry to separate the two isotopes since they are chemically identical. Any separation method must take advantage of the minute mass difference of the two isotopes. If the enrichment is carried out far enough, you can build a nuclear bomb. Reactors in the United States use uranium typically enriched to 3.7%; this is insufficient to sustain a chain reaction. An additional trick must be used.

Neutrons resulting from a ${}^{235}\mathrm{U}$ fission are high-energy particles (some 1 MeV), and their absorption cross section is about the same for both uranium isotopes. However, slow thermal neutrons (say at 0.05 eV) happen to be absorbed much more readily by the fissile uranium than by the more stable isotope. Thus, some of the emitted neutrons have to be slowed down by making them move through a low atomic mass substance called a **moderator**. Graphite or water will do. If water is used as a coolant and heat-extraction medium, then it contributes to the moderation process.[†]

Fast neutrons may be absorbed by impurities in the fuel or in the moderator. Of course, ${}^{238}_{92}\mathrm{U}$ is a major "impurity" in the fuel; it absorbs

[†] No enrichment is needed if the moderator is heavy water (D_2O) as used in the CANDU reactor. This is the **CAN**adian **D**euterium **U**ranium, pressurized heavy water reactor that uses natural (unenriched) uranium and heavy water as both moderator and coolant.

some of the fast neutrons. To reduce neutron losses, it is necessary to place the fuel into a number of long rods embedded in a mass of moderator. This configuration allows most of the fast neutrons to escape the fuel region and reach the moderator where they are slowed and may eventually reenter one of the fuel rods. They now interact with the ^{235}U perpetuating the reaction.

It is essential that exactly one of the released neutrons is, on average, used to trigger a new fission. If more than one, the reaction will grow exponentially; if less, it will die out. Control systems are used to adjust this number to precisely one.

When all is said and done, the only useful output of a fission reactor is heat, which has to be removed by a coolant and transferred to a turbine. Most American reactors use liquid water for this purpose. This limits the temperature to about 300 C, leading to low thermal efficiency. Even then, pressurization is required to keep the water in the liquid phase (hence the label **pressurized water reactor**). Remember that the vapor pressure of water at 300 C is 85 atmospheres. Any rupture can cause loss of coolant and can lead to a meltdown. Reactors of the class operating in the United States have a number of disadvantages that may be absent in more modern designs:

1. Scarcity of fuel because only the rare isotope, $^{235}_{92}$U, is burned
2. Production of dangerously radioactive “ashes”
3. Safety concerns
4. Production of weaponizable materials such as plutonium

The scarcity of fuel problems can be circumvented by using one of the other fissile materials such as plutonium, $^{239}_{94}$Pu, and uranium, $^{233}_{92}$U. These elements are not found in nature but can be obtained by the transmutation that occurs in any type of reactor.

Take ^{238}U:

$$^{238}_{92}\mathrm{U} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{239}_{92}\mathrm{U} \rightarrow {}^{0}_{-1}\mathrm{e} + {}^{239}_{93}\mathrm{Np} \rightarrow {}^{0}_{-1}\mathrm{e} + {}^{239}_{94}\mathrm{Pu} \tag{1.12}$$

or take ^{232}Th:

$$^{232}_{90}\mathrm{Th} + {}^{1}_{0}\,\mathrm{n} \rightarrow {}^{233}_{90}\mathrm{Th} \rightarrow {}^{0}_{-1}\mathrm{e} + {}^{233}_{91}\mathrm{Pa} \rightarrow {}^{0}_{-1}\mathrm{e} + {}^{233}_{92}\mathrm{U}. \tag{1.13}$$

By using plutonium, all uranium can be made to yield energy: 320,000 EJ become available. Even larger amounts of energy could be derived from thorium. One type of reactor that can greatly improve the efficiency of fuel use (by some two orders of magnitude) is the **heavy-metal fast breeder reactor**. Its initial fuel load must contain enough enriched uranium or plutonium not to need moderators; it must operate with fast neutrons so as to transmute ^{235}U into plutonium (Equation 1.12) at a rate larger than that at which fissile fuel is used—it breeds more fuel than it uses (as long as the abundant supply of fertile uranium lasts). Subsequent fuel loads (during the 60-year life expectancy of the machine) can contain waste fuel, natural uranium, or even **depleted** uranium. Because fast neutrons are needed and moderators must actually be avoided the coolant must have a large atomic mass; otherwise, it would itself act as a moderator.[†] In addition, the coolant must be liquid (have a low melting point) and must have a high boiling point so that high temperatures can be achieved at low pressures. It is also desirable that it be relatively inert chemically. This is one of the disadvantages of using sodium as a coolant, since it reacts explosively when it comes in contact with water. One material that fulfills these requirements is a lead–bismuth alloy, Pb-Bi that, notwithstanding its elevated boiling point of 1679 C, melts at slightly over 100 C.

Eutectics

From my online dictionary:

Relating to or denoting a mixture of substances (in fixed proportions) that melts and solidifies at a single temperature that is lower than the melting points of the separate constituents or of any other mixture of them.

Wood's metal, 50% bismuth, 25% lead, 12.5% tin, and 12.5% cadmium, melts at 73 C. It is used to, among other things, fashion teaspoons that melt when dipped into a hot cup of brew, startling the unweary.

Lead melts at 327.5 C, and bismuth at 271.5 C. The alloy proposed for the heavy-metal reactor melts at $\approx$ 100 C.

Pb-Bi reactors operate at 800 C ensuring good efficiency. In addition, they may contribute to the solution of the difficult nuclear waste problem.

In current nuclear plants, ^{235}U is consumed until the amount left in the fuel rods becomes insufficient to sustain the chain reaction (typically below 1%. Remember that American reactors use enriched uranium in which the percentage of the fissile variety is, say, some 3.7%.) It is then necessary to replace the fuel rods, the spent ones being immersed in a boric acid pool where they cool down for a number of months until the short-life radioactive materials have decayed sufficiently. Then they are classified as waste, in spite of consisting mostly of ^{238}U and some plutonium and other transuranic elements (**actinides**, such as neptunium, amercium, and curium). Discarded fuel rods also contain some medium lifetime fission products dominated by ^{90}Sr and ^{137}Cs. The short-lived fission products

[†] A fast neutron will lose more energy when colliding with a light nucleus, which recoils a lot, than with a heavy one.

have, by now, mostly died out. So the waste consists mostly of fuel that can be used by the heavy-metal reactors and do not have to be stored away for thousands of years as dictated by the current U.S. policy.

Heavy-metal fast breeders probably can be designed for passive safety, so that their safety does not depend on control systems or human decisions. Loss of coolant must interrupt the chain reaction automatically.

1.11.2 Fusion

Fusion reactors may overcome many of the objections to fission. The reaction that is, by far, the easiest to ignite is[†]

$$ {}^{3}_{1}\mathrm{T} + {}^{2}_{1}\mathrm{D} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}. \tag{1.14}$$

To estimate the reaction energy released, one calculates the amount of mass lost. The mass of ${}^{3}_{1}\mathrm{T}$ ion is $5.008271 \times 10^{-27} - 9.10939 * 10^{-31} = 5.007360 \times 10^{-27}$ kg as given by the table at the beginning of this subsection. Notice that we subtracted the mass of the electron from the mass of the tritium atom. The mass of the deuterium ion is $3.344497 \times 10^{-27} - 9.10939 \times 10^{-31} = 3.344497 \times 10^{-27}$ kg, so that the mass of the left side of the equation is 8.350946×10^{-27} kg. On the right-hand side of the equation, the sum of the masses of the alpha particle (the helium ion) and the neutron is 8.319590×10^{-27} kg, a deficit of 3.135569×10^{-29} kg. When multiplied by c^2, this yields an energy of 2.818×10^{-12} joules per deuterium/tritium pair. The reaction yields 337 TJ per kg of tritium/deuterium alloy or 562 TJ per kg of tritium.

The energy released by the reaction is carried by both the alphas and the neutrons. The conversion of the neutron energy to usable forms has an efficiency of only some 40% because the particles are uncharged and heat management and mechanical heat engines are involved. On the other hand, the alphas can be directly converted to electricity at a much higher efficiency (≈ 90%). See Rostoker, Monkhorst, and Binderbauer (1997); Moir and Barr (1973); Momota et al. (1992); Yoshikawa, Noma, and Yamamoto (1992); and Bloch and Jeffries (1950). In addition, the heavy neutron flux creates serious radioactivity and material destruction problems. Consequently, it is important to know how the released energy is divided between the alphas and the neutrons. This can be done by assuming that the momenta are equally divided between the two types of particles:

$$ m_\alpha v_\alpha = m_n v_n, \tag{1.15}$$

[†] The larger the atomic number, Z, the greater the difficulty of causing a reaction owing to the large electrostatic repulsion between nuclei.

and combining this with with the energy equation,

$$\frac{1}{2}m_\alpha v_\alpha^2 + \frac{1}{2}m_n v_n^2 = W = W_\alpha + W_n. \tag{1.16}$$

Here, m_α is the mass of the alpha, m_n is the mass of the neutron, v_α is the velocity of the alpha, v_n is the velocity of the neutron, and W is the energy released by one pair of reacting atoms. Solving these simultaneous equations leads to

$$W_\alpha = \frac{W}{\frac{m_\alpha}{m_n} + 1} \tag{1.17}$$

and

$$W_n = \frac{W}{\frac{m_n}{m_\alpha} + 1}. \tag{1.18}$$

For the reaction under consideration, it is found that neutrons carry about 14 MeV, while the more massive alphas carry only some 3.5 MeV.

The T + D reaction is popular because of its high reactivity, which should facilitate ignition, and because the atomic number of the fuel is $Z = 1$, thus minimizing radiation losses. This is because radiation is a function of Z^2. However, it has drawbacks:

1. One neutron is emitted for each 2.8×10^{-12} J generated, whereas, in fission, the rate is one neutron per 10^{-11} J. Thus, the neutron bombardment is serious: it radioactivates substances and weakens structures by causing dislocations in the crystal lattice and by generating hydrogen bubbles inside materials.
2. As pointed out before, most of the energy is in the neutron stream reducing the recovery efficiency.
3. Although deuterium is not radioactive, tritium is radioactive with a lifetime of 12 years. It has the tendency to stick around by replacing normal hydrogen in water molecules.
4. There is no natural source of tritium; it must be obtained from lithium:

$${}^6_3\mathrm{Li} + {}^1_0\mathrm{n} \rightarrow {}^3_1\mathrm{T} + {}^4_2\mathrm{He} + \cdots 7.7 \times 10^{-13} \text{ joules}. \tag{1.19}$$

Thus, each lithium atom yields $2.8 \times 10^{-12} + 7.7 \times 10^{-13} = 3.57 \times 10^{-12}$ J. One kg of lithium yields 350 TJ.

The world reserves of lithium are not known accurately. Conservative estimates are of 10^{10} kg. However, most of this is ^{7}Li. The desired isotope, ^{6}Li, has a relative abundance of 7.4%. Consequently, one can count on only 740×10^6 kg of this material, or 260,000 EJ of energy.

In order of ease of ignition, the next two reactions are

$$ {}_1^2\mathrm{D} + {}_1^2\mathrm{D} \rightarrow {}_2^3\mathrm{He} + {}_0^1\mathrm{n} + \cdots 5.23 \times 10^{-13} \text{ joules,} \tag{1.20}$$

$$ {}_1^2\mathrm{D} + {}_1^2\mathrm{D} \rightarrow {}_1^3\mathrm{T} + {}_1^1\mathrm{H} + \cdots 6.45 \times 10^{-13} \text{ joules.} \tag{1.21}$$

These reactions have equal probability of occurring.

The tritium produced will react with the deuterium according to Reaction 1.14. The average energy of the $\mathrm{D}+\mathrm{D}$ reaction is

$$\frac{(5.23 + 6.45 + 28.0) \times 10^{-13}}{5} = 7.94 \times 10^{-13} \text{ J per D atom.} \tag{1.22}$$

The $\mathrm{D}+\mathrm{D}$ reaction is still dirty (neutronwise) and still involves a radioactive gas (tritium). However, it does not use a fuel of limited abundance, such as lithium. It uses only deuterium, which is available in almost unlimited amounts. In common water, there is one $\mathrm{D_2O}$ molecule for every 6700 $\mathrm{H_2O}$ molecules. One can estimate roughly how much deuterium is available:

The oceans cover about 2/3 of the Earth's surface, which is 5.1×10^{14} m^2. Assuming an average depth of 3000 m, the ocean has a volume of 10^{18} m^3 and a mass of 10^{21} kg. Of this, 1/9 is the mass of hydrogen, and 2/6700 of the hydrogen is the mass of deuterium, amounting to some 3.3×10^{16} kg, or about 10^{31} J—an amount of energy that, for practical purposes, can be considered unlimited.

Next, in order of ignition difficulty is the ${}^2\mathrm{D} + {}^3\mathrm{He}$ reaction that burns cleanly: no radioactive substances are involved, and no neutrons are generated. Also clean is the ${}^3\mathrm{He} + {}^3\mathrm{He}$ reaction.

The catch in these reactions is that there is no natural ^{3}He on Earth; it must be made from the (dirty) fusion of Li and H. However, it is estimated that over a billion tons of the material exists on the moon. One day this may justify a mining operation on our satellite.

The ^{3}H on the moon comes from the solar wind that has, for billions of years, deposited it there. The ^{3}H on Earth is trapped by the atmosphere and is eventually evaporated away.

An interesting reaction involves ^{11}B, the common isotope of boron:

$$ {}_5^{11}\mathrm{B} + {}_1^1\mathrm{H} \rightarrow {}_6^{12}\mathrm{C}^* \rightarrow {}_2^4\mathrm{He} + {}_4^8\mathrm{Be}. \tag{1.23}$$

${}_6^{12}\mathrm{C}^*$ is nuclearly excited carbon, which spontaneously decays into an alpha and ${}_4^8$Be, a very unstable atom with a lifetime of 2×10^{-16} seconds. Fortunately, it is an alpha emitter:

$$ {}_4^8\mathrm{Be} \rightarrow 2{}_2^4\mathrm{He}. \tag{1.24}$$

The overall reaction is

$$ {}_5^{11}\mathrm{B} + {}_1^1\mathrm{H} \rightarrow 3{}_2^4\mathrm{He}, \tag{1.25}$$

Table 1.11 Neutron Yields

Reaction	**% of energy carried by neutrons**
D + T	65–75
D + T	65–75
D + D	20–45
B + H	<0.1

or, using a different notation,

$$^{1}_{1}\mathrm{H} + ^{11}_{5}\mathrm{B} = 3\alpha. \tag{1.26}$$

This **triple alpha** reaction may be able to sustain itself in a **colliding beam fusion reactor** (see Rostoker, Binderbauer, and Monkhorst, 1997) but this has not yet been demonstrated. If it does work, we would have a clean fusion reactor using abundantly available fuel and capable of operating in units of moderate size, in contrast with the T + D reaction in a Tokamak, which must be 10 GW or more if it can be made to work at all.

It should be noted that ^{10}B will also yield a triple alpha reaction when combined with a deuteron:

$$^{10}_{5}\mathrm{B} + ^{2}_{1}\mathrm{D} \rightarrow 3^{4}_{2}\mathrm{He}. \tag{1.27}$$

Both isotopes of boron are abundant, stable, and nonradioactive. Natural boron consists essentially of 20% ^{10}B and 80% ^{11}B.

The triple alpha reaction may also be an important player in the cold fusion process, if such a process can be made to work. (See the next subsection.)

Table 1.11 lists the percentage of the energy of a reaction that is carried away by neutrons.

Although fusion reactors have not yet been demonstrated,[†] there is a possibility that they will become the main source of energy some 50 years from now. If so, they may provide the bulk of the energy needed by humanity, and the energy crunch will be over.

1.11.3 Cold Fusion

At the beginning of the millennium, when this subsection was being rewritten, the cold fusion question remained unresolved. So far, no one has been able to reproduce the claims of Fleishmann and Pons (1989) but, on the

[†] Fusion research dates back to at least 1938 when Eastman N.Jacobs and Arthur Kantrowitz built the first magnetic confinement fusion reactor at NACA's (now NASA) Langley Memorial Aeronautical Laboratory.

other hand, no one has been able to disprove the existence of cold fusion. As a matter of fact, cold fusion can and has been demonstrated. Let us review what we know for sure of this topic.

As indicated in Subsection 1.11.2, a deuteron will react spontaneously with a deuteron in one of these two reactions:

$$ {}^{2}_{1}\mathrm{D} + {}^{2}_{1}\mathrm{D} \rightarrow {}^{3}_{1}\mathrm{T} + {}^{1}_{1}\mathrm{H} \tag{1.28}$$

$$ {}^{2}_{1}\mathrm{D} + {}^{2}_{1}\mathrm{D} \rightarrow {}^{3}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n} \tag{1.29}$$

These reactions have about the same probability of occurrence and produce a substantial amount of energy. The problem is that the probability of occurrence (under normal conditions) is extremely small, being of the order of one fusion per galaxy per century according to a good-humored scientist.

It is easy to understand the reluctance of the ${}^{2}\mathrm{D}$ atoms to join: they carry positive charges and therefore repel one another. This problem can be overcome by imparting sufficient kinetic energy to the atoms, by, for instance, heating them to extreme temperatures as in thermonuclear fusion.

Luis W.Alvarez (late professor of the University of California at Berkeley and Nobel Prize winner) suggested a neat trick that increases by 85 orders of magnitude the reaction cross section (read probability). Replacing the orbital electron of the deuterium by a muon, which is 207 times heavier, collapses the orbital by a large factor.

Muon-mediated fusion can be observed in the laboratory as Steven Earl Jones (1989; Brigham Young University) demonstrated. The catch is that it takes more energy to create the muon than what one gets from the fusion.

Muon-catalyzed Fusion

Muonic Atoms

The radius of a normal deuteron is 37,000 fm. A muon can be regarded as a heavy electron: it has a charge of −1, a mass of 207 electrons, and a lifetime of 2.2 microseconds.

In the ground state of an atom, the angular momentum of the orbiting particle must be equal to Planck's constant divided by 2π:

$$mr^2\omega = \frac{h}{2\pi} \equiv C_1, \tag{1.30}$$

$$\omega^2 = \frac{C_1^2}{m^2r^4}. \tag{1.31}$$

(*Continues*)

(*Continued*)

The centrifugal force must equal the centripetal one, which is inversely proportional to r (the radius of the orbit):

$$mr\omega^2 = \frac{C_2}{r^2}. \tag{1.32}$$

Combining Equations 1.31 and 1.32, and solving for r,

$$r = \frac{C_3}{m}. \tag{1.33}$$

In the above equations, C_1, C_2, and C_3, are various constants. Thus, *since the muon is 207 times more massive than the electron, the radius of the muonic deuteron must be 207 smaller than that of the normal deuteron; that is, it must be (approximately) 180 fm.*

Deuterium-Muon-Tritium Fusion

A positive **heavy hydrogen ion** (D-muon-tritium ion or D-μ-T$^+$) can be synthesized. The presence of the muon effectively shields the deuterium from the tritium, causing the distance between these atoms to be some 207 times smaller than in the D-electron-tritium case. This is close enough to allow tunneling, though the electrostatic barrier and the two atoms fuse, releasing the energy discussed when the D-T reaction (Equation 1.14) was discussed. Thus, the output of the system is 2.81×10^{-12} J (17.6 MeV) per muonic molecule. As we saw, 14 MeV are carried by the neutron, which can yield an additional 4.81 MeV when it is exothermally absorbed by lithium (Equation 1.19).

The muon survives the fusion and may live to catalyze more reactions. However, two mechanisms require the continuous replenishment of muons: its short lifetime and alpha-sticking, the fact that the muons tend to stick to the alphas from the fusion, being carried away. The energy balance pits the useful output against the cost of replenishing the muons. Muons can be produced by accelerating deuterons (800 MeV) and causing them to impact deuterium gas.

Thus, cold fusion certainly does occur. More than that, cold fusion occurs (almost certainly) even when not mediated by muons. S. E. Jones (1989) described an experiment that appears to prove just that. He used an electrolytic cell with a platinum anode and a palladium (sometimes, titanium) cathode, and D_2O (heavy water) as electrolyte. Since water is a poor conductor, salts were added. Here is Jones's extraordinary recipe:

> "The electrolyte is a mixture of about 160 g of deuterium oxide (D_2O) plus various metal salts in about 0.2 g amounts each:

$FeSO_4$, $NiCl_2$, $PdCl_2$, $CaCO_3$, $LiSO_4$, $NaSO_4$, $CaH_4(PO_4)_2$, $TiOSO_4$, and a very small amount of AuCN."

A chemist might be horrified by the cocktail above—it would be hard to tell what is going on.[†]

When a current was forced through the cell, a small flux of neutrons with a characteristic energy of 2.5 Mev was observed. Jones, a physicist, did a good job of neutron detection. Since 2.5 Mev is the energy of the neutrons in Reaction 1.29, this experiment tends to show that, indeed, fusion is going on.

Jones observed that some 8 hours after the start of operation, the neutron "signal" turned off by itself. This effect was attributed to the poisoning of the palladium electrode by deposition of metals from the solution. In fact, etching the electrode revived the cell.

The reaction rate observed by Jones was small, perhaps 10^{-20} fusions per deuterium pair per second. This could be explained if the deuterium molecules were somehow squeezed from 74,000 fm to half this distance by their residence in the palladium lattice.[††] Jones dubs this **piezonuclear fusion**.

Stanley Pons and Martin Fleishmann (1989) ran similar experiments but, being chemists not physicists, adopted a simpler electrolyte: an LiOH solution in D_2O (heavy water). They also failed to make careful neutron measurements. What they reported is that, after prolonged precooking, some cells suddenly developed a great deal of heat, billions of times more than in the Jones experiment. Unfortunately, these results were never reproduced by other experimenters, and this casts severe doubts on their validity. Here is where I will don my devil's advocate mantle and, just for the fun of it, will defend the P&F results.

In a lecture delivered at the Utah University on March 31, 1989, Stanley Pons relates the most spectacular of his results. "A cube of palladium with a volume of 1 cm^3 was used as cathode of an electrolyzer with lithium hydroxide dissolved in D_2O as an electrolyte. A current of 250 mA/cm^2 was applied for several weeks/months [*sic*] with nothing remarkable happening. A Geiger counter detected no radiation. The current was cut to 125 mA/cm^2 late one day, and next morning the cube of palladium and the electrolysis cell were gone. A nearby Geiger counter was also ruined.[†††]

There was a long delay (several days, at least) before heat evolved. Since the Jones cell poisons itself in 8 hours, this cell will never reach the primed state and no heat can be observed.

[†] Jones was trying to create a chemical environment somewhat like the one in the soil because he was trying to show that some of the internal heat of our planet is generated by deuterium fusion.

[††] A possible cause of the squeezing would be the increase of the electron mass to a few times its free mass.

[†††] As related by Patrick Nolan, 1989 (paraphrased).

Why such a delay? Hydride hydrogen storage systems (see Chapter 11) are well known and are commercially available. One popular system uses a TiFe alloy to absorb H_2. Many other metals and alloys will do the same. Palladium, in particular, is a notorious H_2 absorber. It is not used commercially owing to its high price.

When TiFe powder (after duly activated) is exposed to hydrogen, it will form a (reversible) hydride, TiFeH. If the amount of hydrogen is small, there will be a mixture of TiFe and TiFeH in the powder. This mixture, called β-phase, has the empirical formula TiFeH_x, where x becomes 1 when all the material has been hydrided.

After full hydridization, addition of more hydrogen will cause the formation of a di-hydride, TiFeH_2 (γ-phase). Clearly, the hydrogen is more densely packed in the (di-hydride) γ-phase than in the β-phase. It is, therefore, plausible that the fusion will proceed faster once the γ-phase is reached. How long does it take to reach this γ-phase?

In the described experiment, Pons used a current density of 250 mA/cm^{-2}—a total current of 1.5 A when added over the six sides of the cubic cathode. This corresponds to a production of 9.4×10^{18} deuterons/second. Each cubic centimeter of palladium contains 68×10^{21} atoms. Thus, it takes 7200 seconds, or 2 hours, for the palladium, in this particular experiment, to start becoming di-hydrided. This assumes that all the deuterons produced are absorbed by the palladium, and, thus, the time calculated is a rough lower limit.

Could the heat have resulted from a chemical reaction? The highest enthalpy of formation of any palladium salt seems to be 706 MJ/kmole, for palladium hydroxide. Atomic mass of palladium is 106 daltons, and density is 12 g cm^{-3}. This means that one gets 80 kJ cm^{-3} chemically. Pons and Fleishmann have (they say) gotten 5 MJ cm^{-3}, two orders of magnitude more than chemistry allows.

Conclusion:

1. The heat produced cannot be due to classical fusion reaction (insufficient neutrons, tritium, and γ-rays).
2. The heat produced cannot be due to chemical reaction.
3. Then, simplistically, the heat was not produced.

There is at least one more possible reaction that occurs very rarely:

$$ {}^2_1\text{D} + {}^2_1\text{D} \rightarrow {}^4_2\text{He}. \tag{1.34}$$

As written above, this reaction cannot take place because two particles are converted into a single particle, and it is impossible to conserve simultaneously energy and momentum under such conditions. For the reaction to proceed, it is necessary to shed energy, and, in classical physics, this is done by emitting a 16-MeV γ-ray. Pons did not report γ-rays. There is still an outside possibility that the energy can be shed by some other mechanism such as a phonon, although physicists tell me that this is nonsense. Observe that Reaction 1.34 produces one order of magnitude more energy per fusion than do Reactions 1.28 and 1.29.

So far, we have attempted to explain the hypothetical cold fusion as the result of deuteron-deuteron reaction. It has been difficult to account for the absence of the expected large fluxes of neutrons or gamma rays. It is even more difficult to imagine such a reaction proceeding when common water is used in place of heavy water. Nevertheless, some experimentalists make exactly such a claim.

There have been suggestions that cold fusion actually involves nuclear reactions other than those considered so far. Let us recapitulate what has been said about cold fusion.

1. The results, if any, are not easily reproduced.
2. No substantial neutron flux has been detected. This seems to eliminate the deuteron-deuteron reactions of Equations 1.28 and 1.29.
3. No substantial gamma ray flux has been detected. This eliminates the classical form of the deuteron–deuteron reaction of Equation 1.34.
4. Reactions are reported to be highly dependent on the exact nature of the palladium electrode.
5. Reactions have been reported with an H_2O instead a D_2O electrolyte.

The following cold fusion mechanism fitting the above observations has been proposed.

Boron is a common impurity in palladium. Natural boron exists in the form of two isotopes, with the relative abundance of 20% for ^{10}B and 80% for ^{11}B. Thus, under some special circumstances, the two triple-alpha reactions of Equations 1.25 and 1.27 might occur. They emit neither neutrons nor gamma rays and can occur with either normal water or heavy water.

The boron impurity may be interstitial, or it may collect in grain boundaries. The reaction may only occur if the boron is in one or the other of these distributions. It may also only occur when the amount of impurity falls within some narrow range. Thus, a palladium rod may become "exhausted" after some time of operation if the boron concentration falls below some given limiting concentration.

Perhaps the worst indictment of the P&F experiment is its irreproducibility. No one has claimed to have seen the large heat production reported from Utah. Pons himself states that his experiment will only work occasionally; he claims that there is *live* palladium and *dead* palladium. This could be interesting. Hydrogen absorbed in metals is known to accumulate in imperfections in the crystal lattice. It is possible that such defects promote the high concentrations of deuterium necessary to trigger the reaction.

I still have an old issue of the CRC handbook that lists the thermoelectric power of silicon as both +170 mV/K and −230 mV/K. How can it be both positive and negative? Notice that the determination of the sign of the Seebeck effect is trivial; this cannot be the result of an experimental error. In both cases "chemically pure" silicon was used. So, how come? We have a good and classical example of irreproducibility. That was back in the 1930s. Now any EE junior knows that one sample must have been p-silicon, while the other, n-silicon. Both could be "chemically pure"—to change the Seebeck sign, all it takes is an impurity concentration of 1 part in 10 million. Is there an equally subtle property in the palladium that will allow fusion in some cases?

In April 1992, Akito Takahashi of Osaka University revealed that his cold fusion cell produced an average excess heat of 100 W over periods of months. The electric power fed to the cell was only 2.5 W. The main difference between the Takahashi cell and that of other experimenters is the use of palladium sheets (instead of rods) and of varying current to cause the cell to operate mostly under transient conditions. The excess heat measured is far too large to be attributed to errors in calorimetry. Disturbing to theoreticians is the absence of detectable neutrons. See D. H. Freedman's (1992) report.

In spite of being saddled with the stigma of "pseudo-science," cold fusion does not seem to go away. The September 2004 issue of *IEEE Spectrum* published a report titled "Cold Fusion Back from the Dead," in which recent work on cold fusion by reputable laboratories is mentioned. It quotes the U.S. Navy as revealing that the Space and Naval Systems Center (San Diego) was working on this subject.[†] It also mentioned the Tenth International Conference on Cold Fusion that took place in Cambridge, Massachnets, in August 2003. It appears that by 2004, "a number of groups around the world have reproduced the original Pons-Fleishmann excess heat effect." Mike McKubre of SRI International maintains that the effect requires that the palladium electrode be 100% packed with deuterium (one deuterium-to-one-palladium atom). This coincides with our wild guess at the beginning of this subsection.

At the moment, cold fusion research has gone partially underground, at least as far as the media are concerned. Yet, the consensus is that it merits further study. This is also the opinion of independent scientists such as Paul Chu and Edward Teller who have been brought in as observers.

Osaka University, mentioned a couple of paragraphs above, seems to be keeping the cold fusion flame alive. In two papers (February and March 2008) published in the *Journal of the High Temperature Society of Japan*, physics professor Yoshiaki Arata claims the reproducible production of

[†] It is reported that Stanislaw Szpak (of the SNSC) has taken infrared pictures of mini-explosions on the surface of the palladium, when cold fusion appears to be taking place.

excess heat and helium-4 from samples of zirconium oxide/palladium nano powder charged with deuterium gas. In May 2008, Professor Arata made a public demonstration of this phenomenon, which could be interpreted as a confirmation of cold fusion.

It may be that cold fusion will one day prove practical. That is almost too good to be true and, for the classical fusion researchers, almost too bad to be true.

1.12 Financing

Finding new sources of energy is not difficult. What is difficult is finding new sources of *economically attractive* energy. It is, therefore, important to estimate the cost of the energy produced by different methods. One of the main ingredients of the cost formula is the cost of financing, examined below.

Engineers can estimate roughly how the investment cost will affect the cost of the product by using a simple rule of thumb:

"The yearly cost of the investment can be taken as 20%[†] of the overall amount invested."

Thus, if a 1 million dollar power plant is to be built, one must include in the cost of the generated energy a sum of $200,000 per year.

To allow a comparison of the costs of energy produced by different alternative sources, the Department of Energy has recommended a standard method of calculating the cost of the capital investment.

We will here derive an expression for the cost of a direct reduction loan.

Assume that the payment of the loan is to be made in N equal installments. We will consider a $1.00 loan. Let x be the interest rate of one payment period (say, one month) and let p be the value of the monthly payment. At the end of the first month, the amount owed is

$$1 + x - p. \tag{1.35}$$

at the end of the second month, it is

$$(1 + x - p)(1 + x) - p = (1 + x)^2 - p(1 + 1 + x). \tag{1.36}$$

At the end of the third month, it is

$$\begin{aligned}&[(1 + x)^2 - p(1 + 1 + x)](1 + x) - p \\ &\quad = (1 + x)^3 - p[1 + (1 + x) + (1 + x)^2].\end{aligned} \tag{1.37}$$

[†] This percentage is, of course, a function of the current interest rate. In the low interest rate regimen of the early years of this millennium, the percentage is lower than 20%.

At the end of N months, the amount owed is zero because the loan has been repaid. Thus,

$$(1+x)^N - p[1+(1+x)+(1+x)^2+\cdots+(1+x)^{N-1}] = 0, \tag{1.38}$$

hence

$$p = \frac{1}{(1+x)^{1}+(1+x)^{-2}+\cdots+(1+x)^{-N}} = \left[\sum_{\gamma=1}^{N} z^{\gamma}\right]^{-1}, \tag{1.39}$$

where

$$z \equiv (1+x)^{-1}. \tag{1.40}$$

But

$$\sum_{\gamma=1}^{N} z^{\gamma} = \frac{1-z^{N+1}}{1-z} - 1; \tag{1.41}$$

hence

$$p = \frac{1-z}{z-z^{N+1}} = \frac{x}{1-(1+x)^{-N}}. \tag{1.42}$$

Formula 1.42 yields the magnitude of the monthly payment as a function of the interest rate (per month) and the number of payments.

As an example, consider a small entrepreneur who owns a Diesel-electric generating plant in which he has invested \$1000 per kW. The utilization factor is 50%—that is, 4380 kWh of electricity are produced yearly for each kW of installed capacity. Taxes and insurance amount to \$50 year^{-1} kW^{-1}. Fuel, maintenance, and personnel costs are \$436 kW^{-1} year^{-1}. In order to build the plant, the entrepreneur borrowed money at 12% per year and is committed to monthly payments for 10 years. What is the cost of the generated electricity?

The monthly rate of interest is

$$(1+x)^{12} = 1.12 \qquad \therefore \qquad x = 0.009489. \tag{1.43}$$

The number of payments is

$$N = 10 \text{ years} \times 12 \text{ months/year} = 120. \tag{1.44}$$

The monthly payment is

$$p = \frac{0.009489}{1-(1+0.009489)^{-120}} = \$0.013995 \text{ month}^{-1}. \tag{1.45}$$

The yearly payment is

$$P = 12p = \$0.167937 \text{ year}^{-1}. \tag{1.46}$$

If there were no interest, the yearly payment would be \$0.1. Thus, the yearly cost of interest is \$0.067937.

All the above is on a loan of \$1.00. Since the plant cost \$1000 kW^{-1}, the cost of the investment is \$167.94 kW^{-1} year^{-1}. But on a per kW basis, there is an additional expense of \$50 for taxes and insurance, raising the yearly total to \$217.94. Thus, in this example, the yearly investment cost is 21.79% of the total amount.

A total of 4380 kWh per kW installed are generated (and sold) per year. The fixed cost per kWh is therefore

$$\frac{217.94}{4380} = 0.0497 \text{ \$ kWh}^{-1}, \tag{1.47}$$

whereas the fuel, maintenance, and personnel cost is

$$\frac{436}{4380} = 0.0995 \text{ \$ kWh}^{-1}. \tag{1.48}$$

Total cost is 0.1492 \$ kWh^{-1}. This is commonly expressed as 149.2 mils/kWh, an awkward unit. It is better to use 149.2 \$/MWh or, to stick to the conventional SI units of measure, \$41.4 GJ^{-1}.

When the loan is paid after 10 years, does the entrepreneur own the plant? Maybe. The Diesel-generator may have only a 10-year life, and a new one may have to be acquired.

References

Anderson, R, H. Brandt, H. Mueggenburg, J. Taylor, and F. Viteri, A power plant concept which minimizes the cost of carbon dioxide sequestration and eliminates the emission of atmospheric pollutants, Clean Energy Systems, Inc., 1812 Silica Avenue, Sacramento, CA 95815, **1998**.

Audi, G., and A. H. Wapstra, The 1993 atomic mass evaluation, *Nuclear Physics A,* p. 565, **1993**.

Bloch, F., and C. D. Jeffries, *Phys. Rev.* **77**, p. 305, **1950**.

Casson, Lionel, Godliness & work, *Science 81* (2), p. 36, **1981**.

Energy Information Administration, International Energy Outlook 2007. <http://www.eia.doe.gov/oiaf/ieo/world.html>, **2007**.

Firestone, Richard B., <http://ie.lbl.gov/toi2003/MassSearch.asp>

Fisher, J. C., and R. H. Pry, A simple substitution model of technological change, *Report 70-C-215*, General Electric, R. & D. Center, June **1970**.

Fleishmann, M., and S. Pons, Electrochemically induced nuclear fusion of deuterium, *J. Electroanal. Chem., 261*, pp. 301–308, **1989**.

Freedman, D. H., A Japanese claim generates new heat, *News and Comments, Science, 256*, April 24, **1992**.

Hafele, W., and W. Sassin, Resources and endowments: An outline on future energy systems, IIASA, NATO Science Comm. Conf., Brussels, April **1978**.

Jones, S. E., et al., Observation of cold nuclear fusion in condensed matter, *Nature 378*, pp. 737–740, April **1989**.

Loewen, Eric P., Heavy-metal nuclear power, *American Scientist*, November–December **2004**.

Marchetti, C., Primary energy substitution models, *Int. Inst. Appl. Syst. An.(IIASA),* internal paper WP-75-88, June **1975**.

Marchetti, C., On strategies and fate, in W. Häfele et al., Second status report on the IIASA project on energy systems, 1975, RR-76-1, *Int. Inst, Appl. Sys. An. (IIASA)*, Laxenburg, Austria, **1976**.

Moir R. W., and W. L. Barr,"Venetian-Blind Direct Energy Converter for Fusion Reactors", *Nuclear Fusion*, **13**, p. 35–45, **1973**.

Momota, H., A. Ishida, Y. Kohzaki, G. H. Miley, S. Ohi, M. Ohnishi, K. Yoskikawa, K. Sato, L. C. Steinhauer, Y. Tomita and M. Tuszewski, Fusion Technol. **21**, p. 2307, **1992.**

Nolan, Patrick, *e-mail circular*, March 31, **1989**.

Peterka, V., Macrodynamics of technological change: Market penetration by new technologies, *Int. Inst. Appl. Syst. An. (IIASA)*, RR-77-22, November **1977**.

Rafelski, J., et al., Theoretical limits on cold fusion in condensed matter, *AZPH-TH/89-19*, March 27, **1989**.

Rostoker, N., H. Monkhorst, and M. Binderbauer, *Office of Naval Research Reports*, February, May, and August **1997** (available upon request).

Rostoker, Norman, Michl W. Binderbauer, and Hendrik J. Monkhorst, *Colliding Beam Fusion Reactor, Science 278*, 1419, November 21, **1997**.

Smil, Vaclav, Global population and the nitrogen cycle, *Scientific American*, p. 76, July **1997**.

Sweet, William #1, A nuclear reconnaissance, *IEEE Spectrum* 23, November **1997**.

Sweet, William #2, Advanced reactor development rebounding, *IEEE Spectrum* 23, November **1997**.

Yoshikawa, K., T. Noma, and Y. Yamamoto, *Fusion Technol.* **19**, 870, **1991**.

Ziock, Hans-J., Darryl P. Butt, Klaus S. Lackner, and Christopher H. Wendt, *Reaction Engineering for Pollution Prevention,* Elsevier Science, **2000**.

Abundant statistical information on energy:http://www.eia.doe.gov/

For more detailed information on some topics in this chapter, read: Sørensen, Bent, *Renewable energy*, Academic Press, **2003**.

PROBLEMS

1.1 Assume that from 1985 on the only significant sources of fuel are:

1. coal (direct combustion),
2. oil,
3. synthetic liquid fuel (from coal), and
4. natural gas.

Sources a, b, and c are assumed to follow the market penetration rule:

$$\ln \frac{f}{1-f} = at + b$$

where f is the fraction of the market supplied by the fuel in question and t is the year (expressed as 1988, for instance, not as simply 88). The coefficients are:

	a	b
for coal:	-0.0475,	92.14;
for oil:	-0.0436,	86.22.

The above coefficients are derived from historical data up to 1975.

The objective of this exercise is to predict what impact the (defunct) federal coal liquefaction program would have had on the fuel utilization pattern.

According to the **first in, first out** rule, the "free" variable, that is, the one that does not follow the market penetration rule, is the natural gas consumption fraction, f_{ng}. The questions are:

– In what year will f_{ng} peak?
– What is the maximum value of f_{ng}?

Assume that f_{syn} (the fraction of the market supplied by synthetic fuel) is 0.01 in 1990 and 0.0625 in 2000. Please comment.

1.2 The annual growth rate of energy utilization in the world was 3.5% per year in the period between 1950 and 1973. How long would it take to consume all available resources if the consumption growth rate of 3.5% per year is maintained?

Assume that the global energy resources at the moment are sufficient to sustain, at the current utilization rate,

a. 1000 years
b. 10,000 years

1.3 A car moves on a flat horizontal road with a steady velocity of 80 km/h. It consumes gasoline at a rate of 0.1 liter per km. Friction of the tires on the road and bearing losses are proportional to the velocity and, at 80 km/h, introduce a drag of 222 N. Aerodynamic drag is proportional to the square of the velocity with a coefficient of proportionality of 0.99 when the force is measured in N and the velocity in m/s.

What is the efficiency of fuel utilization? Assuming that the efficiency is constant, what is the "kilometrage" (i.e., the number of kilometers per liter of fuel) if the car is driven at 50 km/h?

The density of gasoline is 800 kg per cubic meter, and its heat of combustion is 49 MJ per kg.

1.4 Venus is too hot in part because it is at only 0.7 AU from the sun. Consider moving it to about 0.95 AU. One AU is the distance between the Earth and sun and is equal to 150 million km.

To accomplish this feat, you have access to a rocket system that converts mass into energy with 100% efficiency. Assume that all the energy of the rocket goes into pushing Venus. What fraction of the mass of the planet would be used up in the project? Remember that you are changing both kinetic and potential energy of the planet.

1.5 Consider the following arrangement:

A bay with a narrow inlet is dammed up so as to separate it from the sea, forming a lake. Solar energy evaporates the water, causing the level inside the bay to be h meters lower than that of the sea.

A pipeline admits sea water in just the right amount to compensate for the evaporation, thus keeping h constant (on the average). The inflow water drives a turbine coupled to an electric generator. Turbine plus generator have an efficiency of 95%.

Assume that there is heat loss neither by conduction nor by radiation. The albedo of the lake is 20% (20% of the incident radiation is reflected, the rest is absorbed). The heat of vaporization of water (at STP) is 40.6 MJ per kilomole. Average solar radiation is 250 W/square meter.

If the area of the lake is 100 km^2, what is the mean electric power generated? What is the efficiency? Express these results in terms of h.

Is there a limit to the efficiency? Explain.

1.6 The thermonuclear (fusion) reaction,

$$^{11}_{5}\mathrm{B} + {}^{1}_{1}\mathrm{H} \rightarrow 3\,^{4}_{2}\mathrm{He},$$

is attractive because it produces essentially no radiation and uses only common isotopes.

How much energy does 1 kg of boron produce? Use the data of Problem 1.11.

1.7 The efficiency of the photosynthesis process is said to be below 1% (assume 1%). Assume also that, in terms of energy, 10% of the biomass produced is usable as food. Considering a population of 6 billion people, what percentage of the **land** area of this planet must be planted to feed these people.

1.8 Each fission of ^{235}U yields, on average, 165 MeV and 2.5 neutrons. What is the mass of the fission products?

1.9 There are good reasons to believe that in early times, the Earth's atmosphere contained no free oxygen.

Assume that all the oxygen in the Earth's atmosphere is of photosynthetic origin and that all oxygen produced by photosynthesis is in the atmosphere. How much fossil carbon must there be in the ground (provided no methane has evaporated)? Compare with the amount contained in the estimated reserves of fossil fuels. Discuss the results.

1.10 What is the total mass of carbon in the atmosphere?

CO_2 concentration is currently 330 ppm but is growing rapidly!

If all the fossil fuel in the estimated reserves (see Section 1.8) is burned, what will be the concentration of CO_2

1.11 Here are some pertinent data:

Particle	Mass (daltons)	Particle	Mass (daltons)
electron	0.00054579903	alpha	4.001506175
muon	0.1134381	${}^5_3$Li	5.01254
proton	1.007276467	${}^6_3$Li	6.015122794
neutron	1.008664909	${}^7_3$Li	7.01600455
${}^1_1$H	1.007825032	${}^{10}_5$B	10.012937
${}^2_1$D	2.014101778	${}^{11}_5$B	11.009305
${}^3_1$T	3.016049278		
${}^3_2$He	3.016029319		
${}^4_2$He	4.002603254		
	Constants		
	c	2.998×10^8 m/s	
	h	6.625×10^{-34} joule-sec	

To convert daltons to kg, divide by $6.02213670 \times 10^{26}$.

Deuterium is a very abundant fusion fuel. It exists in immense quantities in Earth's oceans. It is also relatively easy to ignite. It can undergo three different reactions with itself:

$$ {}^2_1\mathrm{D} + {}^2_1\mathrm{D} \rightarrow {}^3_1\mathrm{T} + {}^1_1\mathrm{H} \qquad (1.1)$$

$$ {}^2_1\mathrm{D} + {}^2_1\mathrm{D} \rightarrow {}^3_2\mathrm{He} + {}_0\mathrm{n}^1 \qquad (1.2)$$

$$ {}^2_1\mathrm{D} + {}^2_1\mathrm{D} \rightarrow {}^4_2\mathrm{He} + h\nu \qquad (1.3)$$

For each reaction, calculate the energy released and, assuming equipartition of momenta of the reaction products, the energy of each product.

What is the energy of the photon released in Reaction 3?

1.12 Random access memories (RAMs) using the Zing Effect were first introduced in 1988 but only became popular in 1990 when they accounted for 6.3% of total RAM sales. In 1994 they represented \$712 million of a total of \$4.75 billion. Sales of all types of RAMs reached \$6 billion in 1997.

A company considering the expansion of Z-RAM production needs to have an estimate of the overall (all manufacturers) sales volume of this type of memory in the year 2000. Assume that the growth rate of the overall dollar volume of RAM sales between 1900 and 2000 is constant (same *percentage* increase every year)

1.13 A 1500-kg Porsche 912 was driven on a level highway on a windless day. After it attained a speed of 128.7 km/h, it was put in neutral and allowed to coast until it slowed down to almost standstill. The coasting speed was recorded every 10 seconds and resulted in the following table.

From the given data, derive an expression relating the decelerating force to the velocity.

Calculate how much horsepower the motor has to deliver to the wheel to keep the car at a constant 80 mph.

Coasting time (s)	**Speed (km/h)**	**Coasting time (s)**	**Speed (km/h)**
0	128.7	100	30.6
10	110.8	110	25.9
20	96.2	120	20.4
30	84.0	130	16.2
40	73.0	140	12.2
50	64.2	150	9.2
60	56.4	160	5.1
70	48.0	170	2.0
80	41.8	180	0
90	35.8		

1.14 The California Air Resources Board (CARB) mandated, for 1995, an upper limit of 200 g/km for the emission of CO_2 from a minivan.

This could be achieved by bubbling the exhaust through a $Ca(OH)_2$ bath or through a similar CO_2 sequestering substance. However, this solution does not seem economical. Assume that all the produced CO_2 is released into the atmosphere.

What is the minimum mileage (miles/gallon) that a minivan had to have by 1995. Assume gasoline is pentane (C_5H_{12}) which has a density of 626 kg m^{-3}. A gallon is 3.75 liters and a mile is 1609 meters. The atomic mass of H is 1, of C is 12, and of O is 16.

1.15 A geological survey revealed that the rocks in a region of northern California reach a temperature of 600 C at a certain depth. To exploit this geothermal source, a shaft was drilled to the necessary depth, and a spherical cave with 10-m diameter was excavated. Water at 30 C is injected into the cave where it reaches the temperature of 200 C (still in liquid form, owing to the pressure) before being withdrawn to run a steam turbine.

Assume that the flow of water keeps the cave walls at a uniform 200 C. Furthermore, assume that, at 100 m from the cave wall, the rocks are at their 600 C temperature. Knowing that the heat conductivity, λ, of the rocks is 2 W m^{-1}K^{-1}, what is the flow rate of the water?

The heat capacity of water is 4.2 MJ m^{-3}K^{-1}, and the heat power flux (W m^{-2}) is equal to the product of the heat conductivity times the temperature gradient.

1.16 The following data are generally known to most people:

a. The solar constant, C (the solar power density), at Earth's orbit is 1360 W m^{-2}.
b. The astronomical unit (AU, the average sun–Earth distance) is about 150 million km.
c. The angular diameter of the moon is 0.5°.

Assume that the sun radiates as a black body. From these data, estimate the sun's temperature.

1.17 Using the results from Problem 1.16, compare the sun's volumetric power density (the number of watts generated per m^3) with that of a typical *Homo sapiens.*

1.18 Pollutant emission is becoming progressively the limiting consideration in the use of automobiles. When assessing the amount of pollution, it is important take into account not only the emissions from the vehicle but also those resulting from the fuel production processes. Gasoline is a particularly worrisome example. Hydrocarbon emission at the refinery is some 4.5 times larger than that from the car itself. Fuel cell cars (see Chapter 9) when fueled by pure hydrogen are strictly a zero emission vehicle. However, one must inquire how much pollution results from the production of the hydrogen. This depends on what production method is used (see Chapter 10). The cheapest hydrogen comes from reforming fossil fuels, and that generates a fair amount of pollution. A clean way of producing hydrogen is through the electrolysis of water; but, then, one must check how much pollution was created by the generation of the electricity. Again, this depends on how the electricity was obtained: if from a fossil fuel steam plant, the pollution is substantial; if from hydroelectric plants, the pollution is zero.

The technical means to build and operate a true zero emission vehicle are on hand. This could be done immediately but would, at the present stage of the technology result in unacceptably high costs.

Let us forget the economics and sketch out roughly one possible ZEV combination. Consider a fuel cell car using pure hydrogen (stored, for instance, in the form of a hydride—Chapter 11). The hydrogen is produced by the electrolysis of water, and the energy required for this is obtained from solar cells (Chapter 14). Absolutely no pollution is produced. The system is to be dimensioned so that each individual household is independent. In other words, the solar cells are to be installed on the roof of each home.

Assume that the car is to be driven an average of 1000 miles per month and that its gasoline-driven equivalent can drive 30 miles/gallon. The fuel cell version, being much more efficient, will drive three times farther using the same energy as the gasoline car.

How many kilograms of hydrogen have to be produced per day?

How large an area must the solar cell collector have?

You must make reasonable assumptions about the solar cell efficiency, the efficiency of the electrolyzer, and the amount of insolation (Chapter 12).

1.19 From a fictitious newspaper story:

> A solar power plant in the Mojave Desert uses 1000 photovoltaic panels, each "40 meters square." During the summer, when days are invariably clear, the monthly sale of electricity amounts to $22,000. The average price charged is 3 cents per kWh. The plant is able to sell all the electricity produced.

There is an unfortunate ambiguity in the story: "40 meters square" can be interpreted as a square with 40 meters to its side or as an area of 40 m^2.

From the data in the story, you must decide which is the correct area.

1.20 Sport physiologists have a simple rule of thumb: Any healthy person uses about 1 kilocalorie per kilometer per kilogram of body weight when running.

It is interesting to note that this is true independently of how well trained the runner is. A trained athlete will cover 1 km in much less time than an occasional runner but will use about the same amount of energy. Of course, the trained athlete uses much more power.

The overall efficiency of the human body in transforming food intake into mechanical energy is a (surprisingly high) 25%!

A good athlete can run 1 (statute) mile in something like 4 minutes and run the Marathon (42.8 km) in a little over 2 hours.

1. Calculate the power developed in these races. Repeat for a poor performer who runs a mile in 8 minutes and the Marathon in 5 hours. Assume a body weight of 70 kg.
2. Evaporation of sweat is the dominant heat removal mechanism in a human body. Is this also true for a dog? for a horse?
3. Assuming that all the sweat evaporates (i.e., none of it drips off the body), how much water is lost by the runners in one hour. The latent heat of vaporization of water is 44.1 MJ/kmole.

1.21 One major ecological concern is the emission of hothouse gases, the main one being CO_2.

A number of measures can be taken to alleviate the situation. For instance, the use of biomass-derived fuels does not increase the carbon dioxide content of the atmosphere.

Fossil fuels, on the other hand are a major culprit. Suppose you have the option of using natural gas or coal to fire a steam turbine to generate electricity. Natural gas is essentially methane, CH_4, while coal can be taken (for the purposes of this problem only) as eicosane, $C_{20}H_{42}$. The higher heat of combustion of methane is 55.6 MJ/kg, and that of eicosane is 47.2 MJ/kg.

For equal amounts of generated heat, which of the two fuels is preferable from the CO_2 emission point of view? What is the ratio of the two emission rates?

1.22 A planet has a density of 2500 kg/m^3 and a radius of 4000 km. Its "air" consists of 30% ammonia, 50% carbon dioxide, and 20% nitrogen.

Note that the density, δ_{earth}, of Earth is 5519 kg/m^3.

What is the acceleration of gravity on the surface of the planet?

1.23 At 100 million km from a star, the light power density is 2 kW/m^2. How much is the total insolation on the planet of Problem 1.22 if it is 200 million km from the star. The total insolation on Earth is 173,000 TW.

1.24 $^3_2$He can be used as fuel in "dream" fusion reactions—that is, in reactions that involve neither radioactive materials nor neutrons. Two possible reactions are

$$^2_1\mathrm{D} + {}^3_2\mathrm{He} \rightarrow {}^1_1\mathrm{H} + {}^4_2\mathrm{He} \tag{1.1}$$

and

$$^3_2\mathrm{He} + {}^3_2\mathrm{He} \rightarrow {}^1_1\mathrm{H} + {}^1_1\mathrm{H} + {}^4_2\mathrm{He} \tag{1.2}$$

1. For each of the above reactions, calculate the energy (in kWh) released by 1 kg of $^3_2$He.

   On Earth, $^3_2$He represents 0.00013% of the naturally occurring helium. The U.S. helium production amounts, at present, to 12,000 tons per year.

2. If all this helium were processed to separate the helium-three, what would be the yearly production of this fuel?

   There are reasons to believe that the moon has a substantial amount of $_2He^3$. Let us do a preliminary analysis of the economics of setting up a mining operation on our satellite.

   One advantage of using "dream" reactions is that only charged particles (protons and alphas) are produced. The energy associated with charged particles can be more efficiently transformed into electricity than when the energy is carried by neutrons, which must first produce heat that is then upgraded to mechanical and electric energy by inefficient heat engines. Thus, it is not necessarily optimistic to assign a 30% efficiency for the conversion of fusion energy into electricity.

3. How many kWh of electricity does 1 kg of ${}^3_2$He produce? Use the most economical of the two reactions mentioned.

   Assume that the plant factor is 70% (the reactor delivers, on average, 70% of the energy it would deliver if running constantly at full power). Assume further that the cost of the fusion reactor is \$2000/kW and that the cost of borrowing money is 10% per year. Finally, the cost of running the whole operation is \$30 $\text{kW}^{-1}\text{year}^{-1}$.

4. How much would the electricity cost (per kWh) if the fuel were free?

5. How much can we afford to pay for 1 kg of ${}^3_2$He and still break even when electricity is sold at 5 cents per kWh?

1.25 Between 1955 and 1995, the ocean temperature (Atlantic, Pacific, and Indian) increased by 0.06 C.

Estimate how much energy was added to the water.

What percentage of the solar energy incident on Earth during these 40 years was actually retained by the ocean?

1.26 It seems possible that climate changes will cause the polar ice caps to melt. The amount of ice in Antarctica is so large that if it were to melt, it would submerge all ports such as New York and Los Angeles.

Estimate by how much the sea level would rise if only the north pole ice is melted, leaving Greenland and Antarctica untouched.

1.27 Refueling a modern ICV with 50 liters of gasoline may take, say, 5 minutes. A certain amount of energy was transferred from the pump to the car in a given time. What is the power represented by this transfer? Assume that the overall efficiency of a gasoline car is 15% and that of an electric car is 60%. How much power is necessary to charge the batteries of the electric car in 5 minutes (as in the ICV case)? Assume that the final drive train energy is the same in both the ICV and the EV. Is it practical to *recharge* a car as fast as *refueling* one?

1.28 Some of the more attractive fuels happen to be gases. This is particularly true of hydrogen. Thus, storage of gases (Chapter 11) becomes an important topic in energy engineering. Lawrence Livermore Labs, for instance, has proposed glass micro-balloons, originally developed for housing minute amounts of tritium-deuterium alloy for laser fusion experiments. When heated, the glass becomes porous and hydrogen under pressure can fill the balloons. Cooled, the gas is trapped.

Clathrate is one of nature's ways of storing methane, even though no one is proposing it as a practical method for transporting the gas.

Methane clathrate frequently consists of cages of 46 H_2O trapping 8 CH_4 molecules.

1. What is the gravimetric concentration, GC, of methane in the clathrate? Gravimetric concentration is the ratio of the mass of the stored gas to the total mass of gas plus container.

   Consider a hermetic container with 1 m^3 internal volume filled completely with the clathrate described, which has a density of 900 kg/m^3. Assume that by raising the temperature to 298 K, the material will melt and methane will evolve. Assume also (although this is not true) that methane is insoluble in water.

2. What is the pressure of the methane in the container?

1.29 A Radioisotope Thermal Generator (RTG) is to deliver 500 W of dc power to a load at 30 V. The generator efficiency (the ratio of the dc power out to the heat power in) is 12.6%. The thermoelectric generator takes heat in at 1200 K and rejects it at 450 K. The heat source is plutonium-241. This radioactive isotope has a half-life of 13.2 years and decays emitting α and β^- particles. These particles have an aggregate energy of 5.165 MeV.

Only 85% of the power generated by the plutonium finds its way to the thermoelectric generator. The rest is lost.

How many kilograms of plutonium are required? Note that radioactive substances decay at a rate proportional to the amount of undecayed substance and to a constant decay rate, λ:

$$\frac{dN}{dt} = -\lambda N.$$

1.30 In the United States we burn (very roughly) an average of 150 GW of coal, 40 GW of oil, and 70 GW of natural gas.

Assume that

Coal is (say) $C_{20}H_{44}$ and that it yields 40 MJ per kg.
Oil is (say) $C_{10}H_{22}$ and yields 45 MJ per kg.
Natural gas is CH_4 and yields 55 MJ per kg.

How many kg of carbon are released daily by the combustion of coal alone? (Clearly, after you have handled coal, the other two fuels can be handled the same way. But, for the sake of time, don't do it.)

1.31 The photovoltaic plant at Serpa (southeast Portugal) has a peak output of 11 MW. Since the collectors track the sun, the power output is steady (at 11 MW) from 09:00 to 15:00, independently of the date. Assume the power output is zero outside this time interval. (Note that this is not actually true: the output ramps up in the morning before 09:00 and down in the evening, after 15:00.)

Serpa is notorious for its sunny climate. The average insolation for a surface facing the sun over the 09:00 to 15:00 period (and then averaged over the year) is 900 W/m^2.

The lifetime of the system is 25 years. Yearly operating cost is US$1 million. Total capital investment is US$78 million. Cost of capital is 10% per year for the 25 years of the plant's life. The average efficiency of the collectors and distribution system is 15%. Of the generated electricity, 90% is delivered to the customers.

a. If the solar plant did not exist, the electric power would have to be generated by an oil-fired steam turbine with 30% efficiency. Oil has a heat of combustion of, say, 40 MJ/kg, and, for the purpose of this problem can be considered pure $C_{12}H_{26}$. How many tons of CO_2 would such a plant emit per year? Assume sunny days throughout the year.

b. What is the cost of the electricity generated by the photovoltaic plant ($/kWhr)? This includes the 10% of the electricity consumed in house, that is, not delivered to the customers. Compare with typical costs for nuclear and gas-fired turbines. What is the per kilowatt capital cost and the utilization factor of this photoelectric system? Compare with a thermal plant that costs around $1000/kW.

c. What is the capital cost of the square meter of solar collector? Include all the capital cost, not only that of the collectors.

Chapter 2
A Minimum of Thermodynamics and of the Kinetic Theory of Gases

2.1 The Motion of Molecules

A gas is a collection of particles (molecules) that, to a first approximation, interact with one another solely through elastic collisions—in other words, through collisions that conserve both energy and momentum. If molecules were dimensionless, point-like objects, their thermal energy would be only that of linear motion in three dimensions—they would have only *three degrees of freedom*. In reality, molecules are more complicated. Even a simple monatomic one, such as helium, may be able to spin (because it has a finite dimension) and may, therefore, have more than three degrees of freedom. Multiatomic molecules can also vibrate, and this confers on them additional degrees of freedom.

At a given moment, some molecules have large kinetic energy, while others have little. However, over a sufficiently long period of time, each has the same *average kinetic energy*, $<W_{mol}>$. This intuitive result is called the **principle of equipartition of energy**. What is not so immediately obvious is that the principle applies even to a collection of molecules with different masses: the more massive ones will have smaller average velocities than the lighter ones, but their average energy will be the same. According to this principle, the energy associated with any degree of freedom is the same. The instantaneous velocities have a **Maxwellian** distribution, as discussed in Section 2.18.

2.2 Temperature

It is useful to distinguish the two components of the average molecular energy: $<W_{mol,\ linear}>$ and $<W_{mol,\ spin\ \&\ vibr.}>$,

$$<W_{mol}>=<W_{mol,\ linear}> + <W_{mol,\ spin\ \&\ vibr.}> \tag{2.1}$$

The pressure a gas exerts on an obstacle is the result of molecules colliding with the obstacle. Clearly, only the *linear* motion of the molecules can contribute to pressure; spin and vibration cannot.

Temperature is a measure of $<W_{mol,\ linear}>$. It is defined by

$$T = \frac{2}{3k} <W_{mol,\ linear}> \tag{2.2}$$

The factor $\frac{1}{3}$ in the proportionality constant, $\frac{2}{3k}$, results from the three degrees of freedom. k is **Boltzmann's constant** and has, of course, the dimensions of energy per temperature (in the SI, $k = 1.38 \times 10^{-23}$ joules/kelvin).

In terms of temperature, the average energy of linear molecular motion in a three-dimensional gas is

$$<W_{mol,\ linear}>= 3\frac{k}{2}T \tag{2.3}$$

and per degree of freedom,

$$<W_{mol,\ linear,\ per\ deg.\ of\ freed.}>= \frac{k}{2}T. \tag{2.4}$$

Since each degree of freedom (associated with linear motion, spin, or with vibration) has the same energy, the average total molecular energy is

$$<W_{mol}>= \frac{\nu}{2}kT, \tag{2.5}$$

where ν is the number of degrees of freedom.

2.3 The Perfect-Gas Law

It is obvious that a simple relationship must exist between pressure and temperature. Consider motion in a single dimension normal to a surface. Upon impact, the molecule deposits a momentum of $2mv$ on the wall (the factor 2 accounts for the impinging velocity being v and the reflected velocity being another v). The flux of molecules (i.e., the number of molecules moving through a unit area in unit time) is $\frac{1}{2}nv$, where n is the **concentration** of molecules (i.e., the number of molecules per unit volume). Here the $\frac{1}{2}$ accounts for half the molecules moving in one direction, while the other half moves in the opposite direction because we are assuming that there is no net flux—that is, no bulk gas motion. The rate of change of momentum per unit area per unit time (i.e., the pressure exerted by the gas) is thus,

$$p = nmv^2. \tag{2.6}$$

Since the kinetic energy of the molecules moving in the direction being considered is $\frac{1}{2}mv^2$ and since this energy is $\frac{1}{2}kT$,

$$p = nkT. \tag{2.7}$$

The pressure is proportional to both the gas concentration and to its temperature. This is the **perfect-gas law** applicable to ideal gases, in which particles neither attract nor repel one another. Practical gases may not follow this law exactly owing to weak Van der Waals's forces between molecules and to the finite size of molecules (which causes the volume available to the gas to be smaller than the volume of the vessel containing it). The higher the concentration of the gas, the greater the error when using the perfect gas law. But in many situations the error tends to be small—the volume of air at 300 K and 100 kPa is overestimated by only 0.07%.

It proves convenient to count particles in terms of *kilomoles* (just as it might be useful to count loaves of bread in terms of *dozens*). While a dozen is equal to 12, a kilomole is 6.022×10^{26}. This latter quantity is called **Avogadro's number,** N_0.[†]

Observe that $n = \mu N_0/V$ (μ is the number of kilomoles, and V is the volume of the container).

$$p = \mu \frac{N_0 kT}{V} = \mu \frac{RT}{V}. \tag{2.8}$$

This is another form of the perfect-gas law.

$$R \equiv kN_0 = 1.38 \times 10^{-23} \times 6.022 \times 10^{26} = 8314 \quad \text{J K}^{-1}\text{kmole}^{-1}. \tag{2.9}$$

R is the **gas constant**.

2.4 Internal Energy

The total **internal energy**, U, of a gas is the sum of the energy of all molecules:

$$U \equiv \sum_i W_{mol_i} = \mu N_0 <W_{mol}>= \mu N_0 \frac{\nu}{2} kT = \mu \frac{\nu}{2} RT \text{ J}. \tag{2.10}$$

Thus, the internal energy, U, of a quantity, μ, of gas depends only on the temperature, T, and on the number, ν, of degrees of freedom of its constituent molecules.

2.5 Specific Heat at Constant Volume

When, in a fixed amount of gas, the temperature is changed, the internal energy also changes. When the volume is kept constant (as is the case of a mass of gas confined inside a rigid container), the rate of change of its

[†] If Avogadro's number is taken as the number of molecules per mole (instead of kilomoles as one does when using the SI), then its value is 6.022×10^{23}.

internal energy with a change of temperature (per kilomole of gas) is called the **specific heat at constant volume**, c_v:

$$c_v = \frac{1}{\mu}\frac{dU}{dT} = \frac{\nu}{2}R \quad \text{J kmole}^{-1}\text{K}^{-1}. \tag{2.11}$$

Heat energy added to a gas is equally divided among the various degrees of freedom—hence the larger the number of degrees of freedom the more energy is necessary to increase the energy of linear motion—that is, to increase the temperature. For this reason, the specific heats of a gas are proportional to ν.

We can see from Equation 2.11, that when a quantity, μ, of gas changes its temperature, then its internal energy changes by

$$\Delta U = \mu \int_{T_0}^{T} c_v dT. \tag{2.12}$$

Notice that we have left c_v inside the integral to cover the possibility that it may be temperature dependent, although this is not obvious from our derivations so far.

2.6 The First Law of Thermodynamics

Introduce an amount, ΔQ, of heat energy into an otherwise adiabatic gas-filled cylinder equipped with a frictionless piston. **Adiabatic** means that no heat is exchanged between the gas in the cylinder and the environment. If the piston is allowed to move, it can do external work, ΔW, by lifting a weight. If held immobile, no work will be done. In general, $\Delta W \neq \Delta Q$. In fact, since energy cannot be created from nothing, $\Delta W \leq \Delta Q$. What happens to the excess energy, $\Delta Q - \Delta W$?

The principle of conservation of energy requires that the internal energy of the system increase by just the correct amount. A “bookkeeping” equation is written:

$$\Delta U = \Delta Q - \Delta W, \tag{2.13}$$

where ΔU is the increase in internal energy. In differential form, the change in internal energy can be related to the incremental heat added and to the incremental work done by the system,

$$dU = dQ - dW. \tag{2.14}$$

Equations 2.13 and 2.14 are the mathematical statement of the **first law of thermodynamics**. It is a statement of conservation of energy. In all cases of interest here, the internal energy is the energy associated with the random motion—the **thermal** energy. A more complicated system may increase its internal energy through such additional mechanisms as atomic or molecular excitations, ionization, and others.

When heat is added at constant volume, there is no external work ($dW = 0$) and, consequently, $dU = dQ$. Since the specific heat at constant volume is dU/dT (per kilomole),

$$\mu c_v = \frac{dU}{dT} = \frac{dQ}{dT}. \tag{2.15}$$

2.7 The Pressure-Volume Work

In the previous subsection, we mentioned that work can be extracted from a closed cylinder-with-piston system. How much work is generated?

The force on the piston is pA, where A is the area of the piston face. If the piston moves a distance, dx, it does an amount of work:

$$dW = pAdx. \tag{2.16}$$

The volume of the cylinder is changed by

$$dV = Adx. \tag{2.17}$$

Thus,

$$dW = p\,dV \tag{2.18}$$

and

$$W = \int p\,dV. \tag{2.19}$$

2.8 Specific Heat at Constant Pressure

c_v is the amount of heat that has to be delivered to 1 kilomole of gas to increase its temperature by 1 K, provided that the *volume* is kept unaltered. In a system like the one depicted in Figure 2.1, this corresponds to immobilizing the piston. On the other hand, if the *pressure* is kept constant, then, in order to increase the temperature by the same 1 K, more energy is needed. The extra energy is required because in addition to increasing the internal energy, heat must also do work lifting the piston. This work (per unit temperature rise) is pdV/dT, or

$$p\frac{dV}{dT} = p\frac{d}{dT}\ \frac{RT}{p} = R \tag{2.20}$$

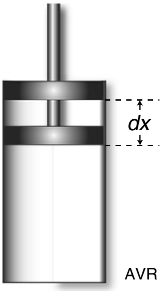


Figure 2.1 Cylinder with frictionless piston.

because p is constant. It follows that

$$c_p = c_v + R = \frac{\nu}{2}R + R = \left(1 + \frac{\nu}{2}\right) R \quad \text{J K}^{-1}\text{kmole}^{-1}. \tag{2.21}$$

The ratio of the two specific heats is

$$\gamma \equiv \frac{c_p}{c_v} = \frac{R(1+\nu/2)}{R\nu/2} = 1 + \frac{2}{\nu}. \tag{2.22}$$

2.9 Adiabatic Processes

In the closed system we have considered so far, we described the interplay between the internal energy, U, the work, W, and the heat, Q. The simplest possible system is one in which the cylinder is so well insulated that heat can neither enter nor leave. In such an **adiabatic** system, $\Delta Q = 0$. As the piston moves down, the work done on it is entirely transformed into an increase in internal energy: $\Delta U = W$. The compression can be accomplished in a gradual manner so that at any given instant the pressure exerted by the piston is only infinitesimally larger than that of the gas—the compression is a succession of quasi-equilibrium states, and the pressure is always uniform throughout the gas. Such is the case, for instance, when the piston is pressed down by the connecting rod of a mechanical heat engine, even though the action may appear to be very rapid. It is also possible to compress a gas abruptly as when an immobilized piston loaded with a heavy weight is suddenly released. In this case, the pressure of the gas immediately under the piston will rise rapidly, but there is no time to transmit this change to the rest of the gas. A nonequilibrium situation is created. The former case—gradual compression—is by far the most common and most important. Nevertheless, we will first consider the abrupt compression because gradual compression can be treated as an infinite succession of infinitely small abrupt steps.

2.9.1 Abrupt Compression

We will start with a qualitative description of what happens, and then we will consider an example.

Assume that a cylinder-and-piston system is in equilibrium. The piston, with a face area, A, is at a height, h_0, above the bottom of the cylinder enclosing a volume, $V_0 = h_0 A$. The force on the piston is F_0,[†] so that the pressure of the gas is $p_0 = F_0/A$. Next, the piston is clamped into place so that it cannot move and an additional mass is added to it. This increases the force to a value, F_1. If the piston is released, it will exert a pressure, $p_1 = F_1/A$, on the gas, but, at the initial instant, the latter will still be at the substantially lower pressure, p_0. The piston will descend explosively to a height, h, and, after a while, will settle at a new height, h_1, when the gas pressure has risen to p_1. An amount of work, $W_{0\to1} = F_1(h_0 - h_1)$ has been done on the gas, and, owing to the adiabatic conditions, this work is entirely translated into an increase, $\Delta U = \mu c_v(T_1 - T_0)$, in internal energy. The compression caused a reduction in volume and an increase in pressure and temperature of the gas.

If, next, the force on the piston is returned to its original value, $F_2 = F_0$, the piston will shoot up, and it is found that it will settle at a height, $h_2 > h_0$. The temperature will fall from the value, T_1, after the compression, to a new value, $T_2 > T_0$. The system does not return to its original state, and the reason is obvious: The compression was caused by a force, F_1, but the expansion was against a smaller force, F_0. Thus, an amount of energy, $W_{0\to1} - W_{1\to2}$ was left over. This particular cycle extracted some energy from the environment. Hence, by definition, it is an irreversible process.

Example

Consider the adiabatic cylinder-and-piston system shown in Figure 2.1. In our example, it contains $\mu = 40.09 \times 10^{-6}$ kilomoles of a gas whose $\gamma = 1.4$, independently of temperature. Its temperature is $T_0 = 300$ K. The cross-sectional area of the cylinder is $A = 0.001$ m^2.

The piston slides with no friction, exerting a force, $F_0 = 1000$ N. Consequently, the piston causes a pressure, $p_0 = \frac{F_0}{A} = \frac{1000}{0.001} = 10^6$ Pa.

The volume of the gas is

$$V_0 = \frac{\mu R T_0}{p} = \frac{40.09 \times 10^{-6} \times 8314 \times 300}{10^6} = 0.0001 \ \text{m}^3. \qquad (2.23)$$

The piston hovers at $h_0 = V_0/A = 0.0001/0.001 = 0.1$ m above the bottom of the cylinder.

(*Continues*)

[†] The force on the piston is the sum of the force exerted by the atmosphere plus the force owing to the weight of the piston.

(*Continued*)

At equilibrium, the pressure of the gas is equal to the pressure the piston exerts. The specific heat at constant volume is

$$c_v = \frac{R}{\gamma - 1} = 20{,}785 \quad \text{J K}^{-1}\text{kmole}^{-1}. \tag{2.24}$$

The internal energy of the gas is

$$U_0 = \mu c_v T_0 = 40.09 \times 10^{-6} \times 20,785 \times 300 = 250 \ \text{ J}, \tag{2.25}$$

and the $p_0 V_0^\gamma$ product is

$$p_0 V_0^\gamma = 10^6 \times 0.0001^{1.4} = 2.51. \tag{2.26}$$

The fixed characteristics of the system are

Area, $A = 0.001$ m^2.
Gas amount, $\mu = 40.09 \times 10^{-6}$ kmoles.
Gamma, $\gamma = 1.4$.
Specific heat at constant volume, c_v =20,785 JK^{-1}kmole^{-1}.

and the initial data are

Force, $F_0 = 1000$ N.
Volume, $V_0 = 0.0001$ m^3.
Pressure, $p_0 = 10^6$ Pa.
Temperature, $T_0 = 300$ K.
Internal energy, $U_0 = 250$ J.
Height, $h_0 = 0.1$ m.
$p_0 V_0^\gamma = 2.51$.

Data on all the phases of this exercise are displayed in Table 2.1 at the end of the next subsection.

What happens if the force exerted by the piston is abruptly increased so that F_1 is now 10,000 N? The piston will go down stopping at a height, h_1. (Actually, the piston will initially overshoot its mark and then oscillate up and down until the internal losses of the gas dampen out these oscillations.) When at equilibrium, the pressure of the gas is

$$p_1 = \frac{F_1}{A} = \frac{10,000}{0.001} = 10^7 \text{ Pa}. \tag{2.27}$$

In moving from h_0 to h_1, the piston did an amount of work, $W_{0\to 1}$,

$$W_{1\to 2} = F_1(h_0 - h_1). \tag{2.28}$$

(*Continues*)

(*Continued*)

Since the cylinder is adiabatic, the internal energy of the gas must increase by an amount,

$$\Delta U_0 = \mu c_v (T_1 - T_0) = F_1(h_0 - h_1), \tag{2.29}$$

where T_1 is the temperature of the gas after compression. It is

$$T_1 = \frac{p_1 V_1}{\mu R} = \frac{p_1 A h_1}{\mu R} = \frac{F_1}{\mu R} h_1. \tag{2.30}$$

Introducing Equation 2.30 into Equation 2.29,

$$F_1(h_0 - h_1) = \frac{c_v}{R} F_1 h_1 - \mu c_v T_0. \tag{2.31}$$

Solving for h_1,

$$h_1 = \frac{\gamma - 1}{\gamma}\left(h_0 + \mu c_v \frac{T_0}{F_1}\right). \tag{2.32}$$

Using the values of the example, $h_1 = 0.0357$ m.
The volume of the gas is now

$$V_1 = A h_1 = 0.001 \times 0.0357 = 35.7 \times 10^{-6} \ \ \mathrm{m}^3. \tag{2.33}$$

The gas temperature is

$$T_1 = \frac{p_1 V_1}{\mu R} = \frac{10^7 \times 35.7 \times 10^{-6}}{40.09 \times 10^{-6} \times 8314} = 1071.4 \ \ \mathrm{K}. \tag{2.34}$$

and the pV^γ product is

$$p_1 V_1^\gamma = 10^7 \times (35.7 \times 10^{-6})^{1.4} = 5.94. \tag{2.35}$$

Collecting these data, we obtain the following values after the abrupt compression:

$F_1 = 10,000$ N.
Volume, $V_1 = 35.7 \times 10^{-6}$ m^3.
Pressure, $p_1 = 10^7$ Pa.
Temperature, $T_1 = 1071.4$ K.
Height, $h_1 = 0.0357$ m.
$p_1 V_1^\gamma = 5.94$.

The amount of energy the piston delivered to the gas is

$$W_{0\rightarrow 1} = F_1(h_0 - h_1) = 10,000(0.1 - 0.0357) = 643 \quad \mathrm{J}. \tag{2.36}$$

(*Continues*)

(*Continued*)

In this example, the sudden application of 10,000 N (an increase of 9000 N) resulted in a strongly nonequilibrium situation. At the moment this additional force was applied, the piston exerted a pressure of 10^7 Pa, while the opposing pressure of the gas was only 10^6 Pa. The piston descended explosively, seeking a new equilibrium. We will attempt to reverse the situation, starting with the values above, also listed in the second column ("abrupt compression") of Table 2.1. We suddenly remove 9000 N (leaving the 1000 N we had originally). Calculations entirely parallel to the one we just did would lead to the new final values, as follows:

$F_2 = 1000$ N.
Volume, $V_2 = 268 \times 10^{-6}$ m^3.
Pressure, $p_2 = 10^6$ Pa.
Temperature $T_2 = 795.6$ K.
Height $h_2 = 0.265$ m.
$p_2 V_2^\gamma = 9.98$.

The force on top of the piston is back to its original value of 1000 N, but the state of the gas is very far from that at the beginning of the experiment. This, as we pointed out, is to be expected. We compressed the gas with a 10,000 N force and then lifted the piston against a much smaller 1000 N force. Although the final height is larger than it was initially, some energy is left over. Indeed, the internal energy of the gas is now $U_2 = \mu c_v T_2 = 663$ J, an increase of 413 J over the initial value of $U_0 = 250$ J. This is, of course, the difference between the mechanical input energy, $W_{0\to1}$ and the mechanical output energy, $W_{1\to2}$.

2.9.2 Gradual Compression

Can a gas be compressed adiabatically in such a way that when expanded it returns exactly to the same state? In other words, can an adiabatic compression be reversible? The answer is yes, provided the force is applied gradually. The compression (or expansion) must proceed in a number of steps, each of which maintains the gas in quasi-equilibrium. This can be demonstrated numerically in a simple way by using a spreadsheet such as Excel.

Starting with the conditions we had at the beginning of the experiment, increment the force by a small amount, ΔF, so that $F_i = F_{i-1} + \Delta F$. Calculate h_i from

$$h_i = \frac{\gamma - 1}{\gamma}\left(h_{i-1} + \mu c_v \frac{T_{i-1}}{F_i}\right). \tag{2.37}$$

Table 2.1 Variables in Different Phases of the Compression Experiment

Phase	**Subscript**	**Vol.** **liters**	**Press.** **MPa**	**Temp.** **K**	**Height** **cm**	pV^γ	**Force** **N**
Initial	"0"	100	1	300	10	2.51	1000
Abrupt compression	"a"	35.7	10	1071	3.57	5.94	10,000
Abrupt expansion	"$a\ _{rever.}$"	265	1	796	26.5	9.98	1000
Gradual compression	"$_{reversible}$"	19.3	10	579	1.93	2.51	10,000

Calculate T_i from

$$T_i = \frac{F_1 h_i}{\mu R}. \tag{2.38}$$

Iterate until $F_i = F_{final}$. Here, $F_{final}(10,000$ N, in this example) is the final value of the force.

For sufficiently small ΔF, it is found that $h_{final} = 0.0193$ m and $T_{final} = 579.2$ K. It is also found that if we decompress in the same manner, we return to the original values of h, V, and T. The process is reversible. In addition, it turns out that the final value of pV^γ is the same as the initial one. Indeed, pV^γ is the same in all steps of the calculation. This is not a coincidence. Later on in this chapter, we will demonstrate that in a reversible adiabatic process, pV^γ is constant. In Chapter 4 we will demonstrate that this so-called **polytropic law** applies to all **isentropic processes**. Use of the polytropic law allows the calculation of reversible adiabatic processes in a simple way, not requiring the iteration technique mentioned above.

The results of the reversible compression are as follows:

Volume, $V_{reversible} = 19.3 \times 10^{-6}$ m^3.
Pressure, $p_{reversible} = 10^7$ Pa.
Temperature $T_{reversible} = 579.2$ K.
Height $h_{reversible} = 0.0193$ m.
$(pV^\gamma)_{reversible} = 2.51$.
$F_{reversible} = 10,000$ N.

2.9.3 p-V Diagrams

It is easier to understand the process by plotting the pressure versus volume behavior of the gas as illustrated in Figure 2.2.

Since reversible compression is the result of a large number of consecutive equilibrium steps, we can calculate the pressure and volume after each step. This is indicated by the smooth (exponential-looking) curve in the figure. Notice that the shaded area under the curve represents the amount of work done during the compression.

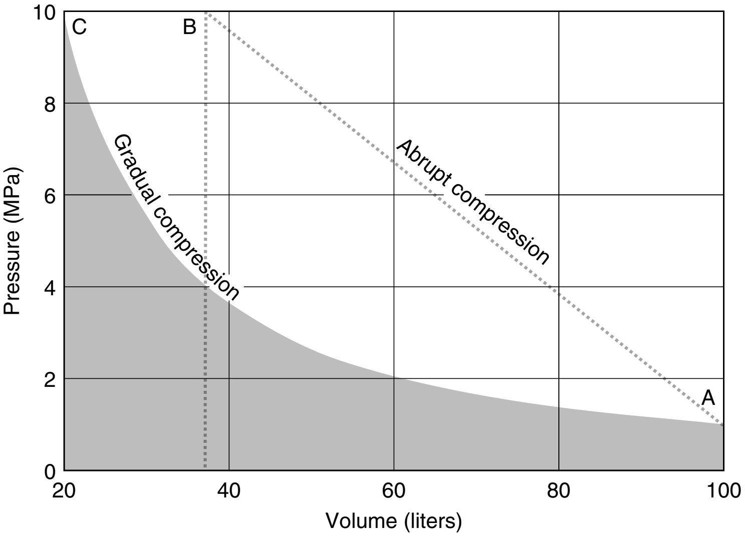


Figure 2.2 For a reversible adiabatic compression, the progression from the initial state, A, to the final state, C, is smooth and predictable. For an abrupt compression, it is impossible to specify the path of the gas.

The compression in the experiment starts from 1 MPa (when the gas volume was 100 liters) and ends when the pressure reaches 10 MPa (at a volume of 19.3 liters).

In the case of the abrupt compression, the initial state, (A), is the same as in the previous case. The final state, (B), has the same pressure (10 MPa) as before, but since the gas is at a higher temperature, the volume is larger (35.7 liters versus 19.3 liters). However, the path from one state to the other is unknown, a fact indicated by the dotted line. During such a rapid compression, the pressure and the temperature cannot be specified because they are not uniform throughout the mass of the gas. As the piston presses down, the gas piles up in the vicinity of the piston, not having time to spread out uniformly.

The terms *gradual* and *abrupt* are relative. In most machines, even at high-speed operation, the compression (or expansion) can be taken (with little error) as gradual.

2.9.4 Polytropic Law

Assume that the cylinder with piston used in our discussion of the pressure-volume work is insulated so that no heat can be exchanged between the gas inside and the environment outside. We have adiabatic conditions, and the heat exchanged is

$$dQ = 0. \tag{2.39}$$

Consider an infinitesimal step in the compression process. The work is

$$dW = p\ dV. \tag{2.40}$$

From the first law of thermodynamics, we have

$$dQ = dU + dW = 0. \tag{2.41}$$

and from the perfect-gas law,

$$p = \mu \frac{RT}{V}. \tag{2.42}$$

It follows that

$$dW = \mu RT \frac{dV}{V} \tag{2.43}$$

and

$$dU = \mu c_v dT, \tag{2.44}$$

$$\mu c_v \frac{dT}{T} + \mu R \frac{dV}{V} = 0, \tag{2.45}$$

$$c_v \ln T + R \ln V = constant. \tag{2.46}$$

But $R = c_p - c_v$; hence

$$\ln T + (\gamma - 1) \ln V = constant, \tag{2.47}$$

$$T\ V^{\gamma-1} = constant. \tag{2.48}$$

Since $pV = \mu RT$,

$$pV^{\gamma} = constant. \tag{2.49}$$

This is called the **polytropic** law and is the characteristic of a reversible adiabatic compression or expansion.

2.9.5 Work Done under Adiabatic Expansion

Since in an adiabatic process, $dQ = 0 = dU + dW$, then $dW = -dU$,

$$W_{0.1} = \mu c_v \Delta T, \tag{2.50}$$

because, there being no heat delivered to the system, all work must come from a reduction, $\mu c_v \Delta T$, of the internal energy.

2.10 Isothermal Processes

In the previous subsection, we discussed a particularly important thermodynamic transformation—the adiabatic process. We will now examine another equally important transformation—the isothermal process.

Under all circumstances, the work done when a gas expands from volume V_0 to V_1 (see Section 2.7) is

$$W_{0,1} = \int_{V_0}^{V_1} pdV. \tag{2.51}$$

Under isothermal conditions,

$$pV = p_o V_0 \qquad \therefore \qquad p = p_o \frac{V_0}{V} \tag{2.52}$$

$$W_{0,1} = p_0 V_0 \int_{V_0}^{V_1} \frac{dV}{V} = p_0 V_0 \ln \frac{V_1}{V_0} = p_0 V_0 \ln \frac{p_0}{p_1} = \mu RT \ln \frac{p_0}{p_1}. \tag{2.53}$$

We will re-derive this result using a more detailed procedure[†] in the hope that this will bring out more clearly the basic underlying mechanism of the process.

Assume that the cylinder is no longer thermally insulated; it is in thermal contact with a bath maintained at a constant 300 K. The frictionless piston is held in place by a weight, so that the initial pressure is p_0. We want to know how much energy can be extracted by allowing the expansion of the gas to a final pressure, p_f. This requires removing some weight from the piston. We assume that there is no outside atmospheric pressure.

Imagine that the original weight had a mass of 10 kg and that, suddenly, 9 kg are removed. The piston will shoot up lifting 1 kg. Since there is no friction, the system oscillates, damped only by the internal dissipation of the gas, which, cooled by expansion, warms up again as heat from the bath is conducted in. After a new equilibrium is established, a mass of 1 kg will have been lifted by, say, 1 meter. The work done is $W = mgh = 1 \times 9.8 \times 1 = 9.8$ J. Although the final temperature is the same as the initial one, *the process is not isothermal*, because during the expansion the temperature first decreased and then rose to its original value. If isothermal, the temperature must not change throughout the whole process.

Let us rerun the experiment, leaving this time 2 kg on the piston. The mass will rise, and the oscillations will eventually settle down. Now an additional 1 kg is removed, and the process repeats itself. The final state is the same as in the first experiment, but the work done is now $2 \times 9.8 \times 0.444 + 1 \times 9.8 \times 0.556 = 14.2$ J because the mass was raised 0.444

[†] Suggested by Professor. D. Baganoff of Stanford University.

m in the first step (see Problem 2.3). The obvious reason for this larger amount of work is that, although only 1 kg reached the 1 m height, a total of 2 kg was lifted part of the way.

Maximum work is done by using an infinite number of steps, each one removing infinitesimally small amounts of mass. Under such circumstances, the expansion is isothermal because heat from the bath flows into the gas after each infinitesimal cooling, thus keeping the temperature at a constant 300 K. What is this maximum work?

Define a decompression ratio, r

$$r \equiv \frac{p_0}{p_f}. \tag{2.54}$$

Let the decompression proceed by (geometrically) uniform steps and let n be the number of steps. Then,

$$r^{1/n} = \frac{p_{i-1}}{p_i}. \tag{2.55}$$

The work done in step i is

$$W_i = (h_i - h_{i-1})F_i, \tag{2.56}$$

where F_i is the weight lifted in step i and is

$$F_i = p_i A, \tag{2.57}$$

A being the area of the piston.

The height reached by the weight after step i is

$$h_i = \frac{V_i}{A}. \tag{2.58}$$

Hence,

$$W_i = \frac{V_i - V_{i-i}}{A} p_i A = V_i p_i - V_{i-1} p_i = V_0 p_0 - V_{i-1} p_i \tag{2.59}$$

because after each step the temperature is returned to its original value, T_0, and consequently $pV = p_0 V_0$.

$$W_i = V_0 p_0 \left(1 - \frac{p_i}{p_{i-1}}\right) = V_0 p_0 \left(1 - r^{-1/n}\right). \tag{2.60}$$

The total work is

$$W = V_0 p_0 \sum_{i=1}^{n} \left(1 - r^{-1/n}\right) = V_0 p_0 n \left(1 - r^{-1/n}\right), \tag{2.61}$$

$$W = V_0 p_0 \ln r. \tag{2.62}$$

This is the formula derived earlier. An expanding mass of gas does maximum work when the expansion is isothermal.

2.11 Functions of State

The **state** of a given amount of perfect gas is completely defined by specifying any two of the following three variables: pressure, volume, or temperature. The third variable can be derived from the other two (if the number of kilomoles of the gas is known) by applying the perfect gas law.

When a gas changes from one state to another, it is possible to calculate the change in its internal energy. To do this, it is sufficient to know the amount of gas, its specific heat at constant volume, and the initial and final temperatures. It is completely irrelevant what intermediate temperatures occurred during the change.

On the other hand, to calculate the amount of work done during a change of state, it is necessary to know exactly which way the change took place. Knowledge of the initial and final states is not sufficient. In the experiment described in the preceding subsection, the work to achieve the final state depended on which way the mass was raised. One cannot tell how much work is required to go from one state to another just by knowing what these states are. If one sees a person on top of a hill, one cannot know how much effort was made to climb it. The person may have taken an easy, paved path or may have traveled a longer, boulder-strewn route. In other words, energy is *not* a function of the state of gas.

2.12 Enthalpy

So far we have considered only **closed systems** in which a fixed mass of gas is involved. Many devices, such as a turbine, are **open systems** involving a flow of gas. Instead of fixing our attention on a given mass of gas, we must consider a given volume through which a fluid flows suffering some thermodynamic transformation. In order to quantify the changes in energy in such an open system, one must account for the energy that the fluid brings into the system and the energy the fluid removes from it.

To force the flow of the fluid into the open system depicted in Figure 2.3, imagine a (fictitious) piston that exerts a pressure, p, pressing a volume, V, of gas into the device. The piston exerts a force, pA, and, to push the gas a distance, L, uses an energy, $pAL = pV$. The flow of energy into the device is $p_{in}V_{in}$. Exiting, the gas carries an energy, $p_{out}V_{out}$. The net energy deposited by the flow is $p_{in}V_{in} - p_{out}V_{out}$. If the gas also changed its internal energy, then the total work that the device generates is

$$W = \Delta(pV) + \Delta U = \Delta(pV + U) \equiv \Delta H. \qquad (2.63)$$

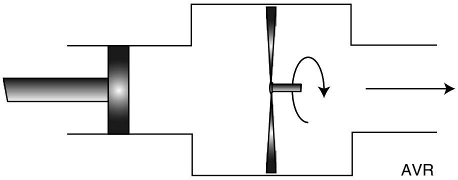


Figure 2.3 An open system in which the flow of gas does work.

This assumes that the device is adiabatic (heat is not exchanged with the environment through the walls).

The combination, $pV + U$, occurs frequently in thermodynamics, and it becomes convenient to define a quantity, called **enthalpy**,

$$H \equiv U + pV. \tag{2.64}$$

H, U, and pV, being energies, are relative; that is, they must be referred to some arbitrary level. Their magnitudes are of little importance; what is of interest is their *change*:

$$\Delta H = \Delta U + \Delta(pV) = \Delta U + p\Delta V + V\Delta p. \tag{2.65}$$

At constant pressure, ΔH is simply $\Delta U + p\Delta V$ and is therefore equal to the heat, ΔQ, added to the system:

$$\Delta H = \Delta U + W = \Delta Q \quad \text{(at constant pressure)}. \tag{2.66}$$

For this reason, enthalpy is sometimes called the **heat content**. It is a quantity commonly used by chemists because reactions are frequently carried out in open vessels—that is, at constant pressure.

From Equation 2.15 (per kilomole)

$$\Delta U = \int_{T_0}^{T} c_v dT. \tag{2.67}$$

Using the perfect-gas law (again, per kilomole), and Equation 2.65

$$\begin{aligned}\Delta H = \Delta U + \Delta(pV) = \Delta U + \Delta(RT) = \int_{T_0}^{T} c_v dT + R\int_{T_0}^{T} dT\\ = \int_{T_0}^{T} (c_v + R) dT = \int_{T_0}^{T} c_p dT.\end{aligned} \tag{2.68}$$

Compare Equations 2.67 and 2.68

2.13 Degrees of Freedom

The formulas we developed for the specific heats and for their ratio, γ, require knowledge of the number of degrees of freedom, ν, of the molecules. From our derivation, this number should be an integer, and it should be independent of temperature. What do experimental data have to say?

Consider monatomic gases whose molecules have three translational degrees of freedom. If this were all, their γ would be exactly $1+2/3 = 1.667$ (Equation 2.22). Table 2.2 shows that the measured value of γ for helium, argon, and krypton is *approximately* (but not exactly) the expected one. However, since the molecules of these gases have some volume, they must be able to spin. At least one additional degree of freedom must be assigned to this spin motion, and ν should not be smaller than 4. This would lead to a γ of $1 + 2/4 = 1.5$, a value significantly below the observed one.

Take now the diatomic molecule, H_2. It should have three translational and two rotational degrees of freedom. In addition, it can vibrate, which should contribute another two degrees of freedom (one for the kinetic energy of vibration and one for the potential). Thus, the total number of degrees of freedom should be at least seven and the value of γ should be $1 + 2/7 = 1.286$. At 2300 K, hydrogen does have a γ of 1.3, corresponding to a $\nu = 6.67$, and as can be seen from Figure 2.4, it appears that at even higher temperatures, the number of degrees of freedom might reach seven. At very low temperatures, ν tends toward three, and H_2 then behaves as a pointlike monatomic molecule.

Between these temperature extremes, ν varies smoothly, assuming fractional, noninteger values. Clearly, any single molecule can only have an integer number of degrees of freedom, but a gas, consisting of a large collection of molecules, can have an *average* ν that is fractional. At any given temperature, some molecules exercise a small number of degrees of freedom, while others exercise a larger one. In other words, the principle of equipartition of energy (which requires equal energy for all degrees of freedom of all molecules) breaks down. Thus, for real gases, the specific heats and their ratio, γ, are temperature dependent, though not extremely so. See the graphs in Free Energy Dependence on Temperature in Section 9.7.4.3 (Chapter 9).

Table 2.2 Ratio of the Specific Heats of Some Monatomic Gases

Gas	**Temp. (K)**	γ
Helium	93	1.660
Argon	298	1.668
Krypton	298	1.68

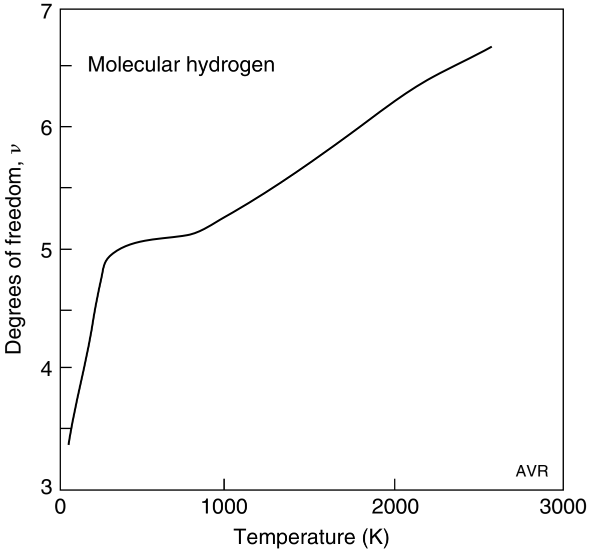


Figure 2.4 The number of degrees of freedom of molecular hydrogen as a function of temperature.

For some estimates, it is sufficient to assume that these quantities are constant. One can assume $\nu = 5$ for H_2 and O_2, and $\nu = 7$ for the more complicated molecule H_2O, all at ambient temperature. More precise calculations require looking up these values in tables. See, for instance, a listing of observed values of c_p and of γ for H_2, O_2, and H_2O in Table 9.5.

Some general trends should be remembered: More complex molecules or higher temperatures lead to a larger number of degrees of freedom and consequently larger specific heats and smaller γ.

2.14 Entropy

When one considers different forms of energy, one can intuitively rank them in order of their “nobility.” Electric energy must be quite “noble”—it can easily be transformed into any other kind of energy. The same is true of mechanical energy because it can (theoretically) be transformed into electricity and vice versa without losses. Heat, however, must be “degraded” energy. It is well known that it cannot be entirely transformed into either electric or mechanical energy (unless it is working against a heat sink at absolute zero). It turns out that chemical energy has a degree of “nobility” lower than that of electricity but higher than that of heat.

Still, intuitively, one can feel that the higher the temperature, the higher the corresponding “nobility” of the heat—that is, the more efficiently it can be transformed into some other form of energy.

Let us try to put these loose concepts on a more quantitative basis.

Consider two large adiabatic reservoirs of heat: one (which we shall call the source) at a temperature, T_H, and one (the sink) at a lower temperature, T_C. The reservoirs are interconnected by a slender metal rod forming a thermally conducting path between them. We shall assume that, for the duration of the experiment, the heat transferred from source to sink is much smaller than the energy stored in the reservoirs. Under such circumstances, the temperatures will remain unaltered.

Assume also that the rod makes thermal contact with the reservoirs but not with the environment. The amount of heat that leaves the source must then be exactly the same as that which arrives at the sink. Nevertheless, the heat loses part of its "nobility" because its arrival temperature is lower than that at the departure. "Nobility" is lost in the conduction process.

Form the Q/T ratio at both the heat source and the heat sink. Q is the amount of heat transferred. Clearly, $Q/T_H < Q/T_C$. We could use this ratio as a measure of "ignobility" (lack of "nobility"), or, alternately, the ratio $-Q/T$ as a measure of "nobility." Loosely, **entropy** is what we called "ignobility":

$$S \equiv \frac{Q}{T}. \tag{2.69}$$

It is important to realize that in the above experiment, energy was conserved but "nobility" was lost. It did not disappear from the experimental system to emerge in some other part of the universe—it was lost to the universe as a whole. There is no law of conservation of "ignobility" or entropy. In any closed system, at best, the entropy will not change, but if it does, it always increases.[†] This is a statement of the **second law of thermodynamics**.

Since there is no heat associated with electric or mechanical energy, these forms have zero entropy.

2.14.1 Changes in Entropy

Returning to the question of functions of state, it is important to know that entropy is such a function. To determine the change in entropy in any process, it is sufficient to determine the entropies of the final and the initial states and to form the difference.

Some processes can drive a system through a full cycle of changes (pressure, volume, and temperature) in such a way that, when the cycle is complete, the system is returned to the initial state. Such processes are **reversible**. To be reversible, the net heat and the net work exchanged

[†] However, a given system does not always tend toward maximum entropy. Systems may spontaneously create complicated structures such as life forms emerging from some primeval soup—a reduction of entropy.

with the environment must be zero. In any reversible process, the change in entropy of a substance owing to a change from State 1 to State 2 is

$$\Delta S = \int_{\text{“1”}}^{\text{“2”}} \frac{dQ}{T}. \tag{2.70}$$

Here, S is the entropy and ΔS is the change in entropy.

In an **adiabatic** processes, $dQ = 0$. Hence

$$S = \mathit{constant}. \tag{2.71}$$

In an **isothermal** process, $\Delta S = Q/T$ because T is constant. Notice, however, that according to the first law of thermodynamics, $\Delta U = Q - W$, but, in an isothermal change, $\Delta U = 0$, hence $W = Q$ and, since in such a change, $W = p_0 V_0 \ln p_1/p_2$,

$$\Delta S = \frac{V_0 p_0}{T} \ln \frac{p_1}{p_2} = \mu R \ln \frac{p_1}{p_2}. \tag{2.72}$$

In an **isobaric** process,

$$\Delta S = \mu \int_{\text{“1”}}^{\text{“2”}} c_p \frac{dT}{T} \tag{2.73}$$

and if c_p is constant,

$$\Delta S = \mu c_p \ln \frac{T_2}{T_1}. \tag{2.74}$$

In an **isometric** process,

$$\Delta S = \mu \int_{\text{“1”}}^{\text{“2”}} c_v \frac{dT}{T} \tag{2.75}$$

and if c_v is constant,

$$\Delta S = \mu c_v \ln \frac{T_2}{T_1}. \tag{2.76}$$

The change in entropy of μ kilomoles of a substance owing to an isobaric change of phase is

$$\Delta S = \mu \frac{Q_L}{T}, \tag{2.77}$$

where Q_L is the **latent heat of phase change** (per kilomole) and T is the temperature at which the change takes place.

We have collected all these results in Table 2.3.

Table 2.3 Changes in Entropy

Process	$\mathbf{\Delta S}$	**Equation**
Adiabatic	zero	2.71
Isothermal	$\mu R \ln \frac{p_1}{p_2}$	2.72
Isobaric	$\mu \int_{\text{“1”}}^{\text{“2”}} c_p \frac{dT}{T}$	2.73
Isobaric (const. c_p)	$\mu c_p \ln \frac{T_2}{T_1}$	2.74
Isometric	$\mu \int_{\text{“1”}}^{\text{“2”}} c_v \frac{dT}{T}$	2.75
Isometric (const. c_v)	$\mu c_v \ln \frac{T_2}{T_1}$	2.76
Phase change	$\mu \frac{Q_L}{T}$	2.77

2.15 Reversibility

In the experiment presented in Section 2.14, it is impossible to reverse the direction of heat flow without a gross change in the relative temperatures, T_H and T_C. It is an **irreversible** process.

Let us examine another case (suggested by Professor D. Baganoff of Stanford University) designed to illustrate reversibility:

Consider again an adiabatic cylinder with a frictionless piston. It contains 10 m^3 of an ideal monatomic gas ($\gamma = 1.67$) under a pressure of 100 kPa at 300 K. We will follow the behavior of the gas during compression and subsequent expansion by means of the p versus V and the T versus V diagrams in Figure 2.5. The initial state is indicated by “1” in both graphs.

When the piston is moved so as to reduce the volume to 2 m^3, the pressure will rise along the adiabatic line in such a way that the pV^γ product remains constant. When State 2 is reached, the pressure will have risen to 1470 kPa and the temperature to 882 K. The work done during the compression is $\int pdV$ between the two states and is proportional to the area under the curve between 1 and 2.

If the piston is allowed to return to its initial position, the gas will expand and cool, returning to the initial state, 1. The process is completely reversible. Now assume that inside the cylinder there is a solid object with a heat capacity equal to that of the gas. If we compress the gas rapidly and immediately expand it, there is no time for heat to be exchanged between the gas and the solid, which will remain at its original temperature of 300 K. The process is still reversible. However, if one compresses the gas and then waits until thermal equilibrium is established, then half of the heat generated will be transferred to the solid and the gas will cool to $(882 + 300)/2 = 591$ K, while its pressure falls to 985 kPa according to the perfect gas law. This is State 3.

When the gas expands again to 10 m^3, it cools down to 201 K and the pressure falls to 67 kPa (State 4). Immediately after this expansion, the solid will be at 591 K. Later, some of its heat will have been transferred

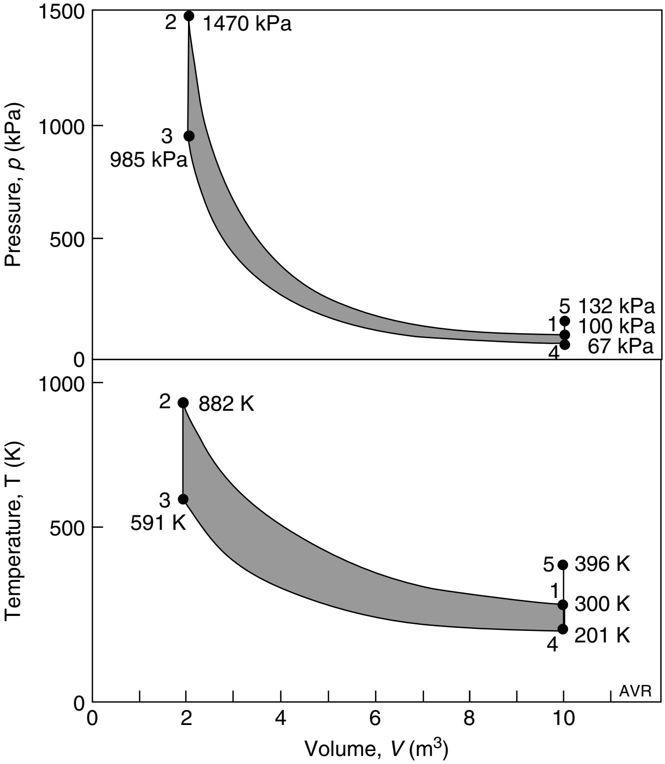


Figure 2.5 Pressure-volume and temperature-volume diagrams for the experiment described in the text.

back to the gas whose temperature rises to 396 K and the pressure will reach 132 kPa (5). The gas was carried through a full cycle but did not return to its initial state: its temperature (and internal energy) is now higher than initially. The increase in internal energy must be equal to the work done on the gas (i.e., it must be proportional to the shaded area in the upper portion of Figure 2.5, which is equal to the area under curve 1 to 2 minus that under curve 3 to 4. The process is irreversible.

What happens if the compression and the expansion are carried out infinitely slowly? Does the process become reversible? You will find that it does when you do Problem 2.2. An electric analogy may clarify the situation. When a real battery (represented by a voltage source, V, with an internal resistance, R) is used to charge an ideal energy accumulator, part of its energy will dissipate as heat through the internal I^2R losses, leaving only part to reach the accumulator. Clearly, the relative loss decreases as the current decreases—that is, as the charge time increases. If the energy is transferred from the battery to the accumulator infinitely slowly ($I \rightarrow 0$), there are no losses. The system is reversible in the sense that all the energy transferred to the accumulator can later be returned to the battery.

2.15.1 Causes of Irreversibility

Among the different phenomena that cause thermodynamic processes to become irreversible one can list the following.

2.15.1.1 Friction

Of all the causes of irreversibilities, friction is perhaps the most obvious. For example, in the cylinder-piston case, if some energy is lost by friction during compression, it is not returned during expansion; on the contrary, additional losses occur during this latter phase.

2.15.1.2 Heat Transfer across Temperature Differences

Consider a metallic wall separating a source of heat—say a flame—from the input of a heat engine. All the heat, Q, absorbed from source is transmitted without loss through the wall, yet for this heat to flow there must be a temperature difference across the wall. The source side is at T_1, the engine side is at T_2, and T_1 must be larger than T_2. The entropy on the source side is Q/T_1, and on the engine side, it is Q/T_2, which is, of course larger than Q/T_1. So, in passing through the wall, the entropy was increased—the heat became less "noble" on the engine side.

In Chapter 3, we will show that the maximum efficiency of a heat engine is $\eta = \frac{T_H - T_C}{T_H}$. If the engine could have operated without the wall, its efficiency could have reached $\frac{T_1 - T_C}{T_1}$ and would be larger than when operated on the other side of the wall when it would be limited to $\frac{T_2 - T_C}{T_2}$.

2.15.1.3 Unrestrained Compression or Expansion of a Gas

In the subsection on adiabatic processes, we dealt with an example of abrupt expansion of a gas and found that it led to irreversibilities.

2.16 Negentropy

We have stressed the idea that energy cannot be consumed. The conservation of energy is one of the laws of nature. When we use energy, we degrade it, so all energy we use is eventually degraded to heat and, one hopes, is radiated out into space in the form of long-wave infrared radiation.

Consider an example. An engine produces energy by extracting heat from the warm surface waters of the ocean (at some 300 K) and by rejecting a smaller amount of heat to the cold waters near the bottom at, say, 275 K. (See ocean thermal energy converters (OTECs) in Chapter 4.) All energy produced is used to compress adiabatically 10,000 kilomoles of a gas from 10^5 to 10^7 pascals (from about 1 to 100 atmospheres). If the initial temperature of the gas was 300 K, what is its final temperature?

Let p_0 and p_H be, respectively, the initial and the final pressures, and T_0 and T_H be the corresponding temperatures. Then,

$$\frac{T_H}{T_0} = \left(\frac{p_H}{p_0}\right)^{\frac{\gamma-1}{\gamma}}, \tag{2.78}$$

where γ is taken, in this example, as 1.4 (equivalent to 5 degrees of freedom). The corresponding specific heat at constant volume is 20.8 kJ $\text{K}^{-1}\text{kmole}^{-1}$.

For a 100:1 adiabatic compression, the temperature ratio is 3.73, and the T_H will be $3.73 \times 300 = 1118$ K. After compression, the volume is

$$V_H = \mu \frac{RT_H}{p_H} = 10,000 \frac{8314 \times 1118}{10^7} = 9178 \text{ m}^3. \tag{2.79}$$

The work required for such a compression is

$$W = 20.8 \times 10^3 \times 10,000 \times (1118 - 300) = 170 \text{ GJ}. \tag{2.80}$$

Let the gas cool to 300 K—the temperature of the ocean surface. All the thermal energy (170 GJ) is returned to the waters. If the canister with gas is towed to the beach, we will not have removed any energy from the ocean, yet we will have 10,000 kilomoles of gas at a pressure of

$$p = \mu \frac{RT}{V} = 10,000 \frac{8314 \times 300}{9178} = 2.7 \text{ MPa}. \tag{2.81}$$

The internal energy of the compressed gas is no larger than that of an equal amount of the gas at the same temperature but at a much lower pressure; the internal energy is independent of pressure. Let the gas expand through a 100% efficient turbine, allowing its pressure to fall to 10^5 Pa. The work done by the turbine must come from the internal energy of the gas; this means that the gas must cool down. The temperature ratio (Equation 2.78) will be 2.56. Thus the gas exhausted from the turbine will be at 117 K.

The energy delivered by the turbine to its load is

$$W = 20.8 \times 10,000 \times (300 - 117) = 38 \text{ GJ}. \tag{2.82}$$

We have "generated" 38 GJ that did not come from the ocean. It came from the gas itself whose internal energy was reduced by cooling to 117 K. The ambient air was cooled by the exhaust from the turbine and then reheated by the degraded energy output of the turbine. Thus, in this whole process, energy was conserved, but we were able to perform useful work (the generator output) by removing something from the ocean. What was it?

We did randomize the ocean by mixing cold bottom water with warm surface water—we increased the entropy of the ocean. The canister carried away some **negentropy**. Here we used the semantically more acceptable concept of **negative entropy**, a quantity that can be consumed.[†]

Cesare Marchetti (1976) proposed the above system as a method for "supplying energy without consuming energy." This is, of course, what is done in all cases in which energy is utilized. Although the system is technically feasible, it is commercially unattractive because the canister has but a small negentropy-carrying capacity compared with the large requirements of material for its construction. The main point to be learned from this example is that the consumable is negentropy, not energy.

2.17 How to Plot Statistics

The World Almanac lists the U.S. population distribution for 1977 by age (see Table 2.4).

In search of a meaningful way of graphically presenting the data above, one can build the histogram shown in Figure 2.6. However, owing to the disparate age intervals chosen in tabulating the data, the histogram is not very enlightening. It would be better to use uniform age intervals—the smaller, the better. In the limit, the best would be to plot $\partial N/\partial A$ versus A, where N is the number of people and A is the age. To do this, we will have to construct another table, derived from the data in Table 2.4.

The data in the last column of Table 2.5. are plotted in Figure 2.7. A continuous line joined the data points. The area under the resulting curve is proportional to the population of the country. In illustrating the energy distribution of, for instance, gas molecules, it is informative to present plots of $\partial N/\partial W$ versus W or versus v (either energy or velocity) but not a plot of N versus W, as one is sometimes tempted to do.

Table 2.4 U.S. Population Distribution by Age

Age interval (years)	**Number of people (millions)**
Under 5	15.2
5–13	32.2
14–17	16.8
18–20	12.8
21–44	72.0
45–64	43.8
65 and over	23.5

[†] Schrödinger (1943) invented the term *negative entropy*. Léon Brillouin introduced the term *negentropy*.

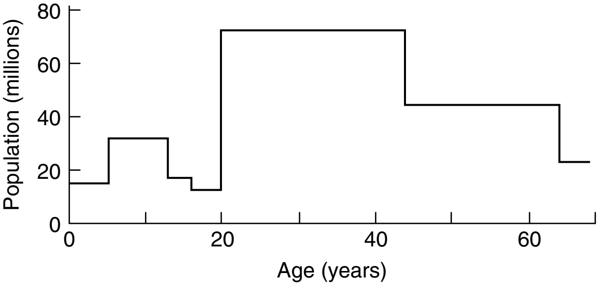


Figure 2.6 Inappropriate way of plotting population versus age. The area under the plot does not have the dimension of number of people.

Table 2.5 U.S. Population Distribution by Age

Age interval (years)	Mean age (years)	Age interval (years)	People (millions)	(people/year)
0–5	2.5	5	15.3	3.07
5–3	9.5	9	32.2	3.58
14–17	16	4	16.8	4.20
18–20	19,5	3	12.8	4.26
21–44	33.0	24	72.0	3.00
45–64	55.0	20	43.8	2.19
over 64	?	?	23.5	—

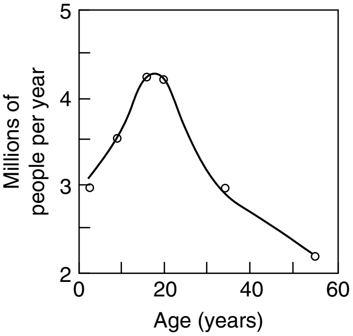


Figure 2.7 Number of people plotted in 1-year age interval.

2.18 Maxwellian Distribution

If molecules have a uniform velocity distribution, how many molecules have velocities less than a given value $|v|$? To answer this question, let us remember that at any time, each molecule is at some position $(x,\ y,\ z)$ and has some velocity $(v_x,\ v_y,\ v_z)$. As far as the energy in the gas is concerned, the exact position of the molecules is irrelevant, but their velocity is not.

Although the individual velocities are changing all the time, in a gas at constant temperature, any instant in time is statistically equivalent to any other. In other words, any instantaneous picture of the velocities is adequate to describe the statistical behavior of the gas.

Let us plot the velocities of the molecules in a system of orthogonal coordinates, v_x, v_y, and v_z—that is, in **velocity space**. Alternatively, we could plot the momenta, mv_x, mv_y, and mv_z, in the **momentum space**. Since we are assuming that molecules have uniform velocity (or momentum) distribution, the velocity (or the momentum) space is uniformly populated. Thus, the number of molecules that have less than a certain velocity, $|v|$, is proportional to the volume of a sphere of radius v (or p) in the space considered. This means that the number of molecules with velocity less than $|v|$ (or momenta less than $|p|$) must be proportional to v^3 (or p^3). Hence, the number of molecules with velocity between v and $v + dv$ (momenta between p and $p + dp$) must be proportional to $\partial v^3/\partial v$—that is, to v^2 (or p^2). In real systems, uniform velocity distribution is unusual. In common gases, a distribution that fits experimental observation is one in which the probability, f, of finding a molecule with a given energy, W, is

$$f = \exp\left(-\frac{W}{kT}\right). \tag{2.83}$$

Under such conditions, the number of molecules with velocities between v and $v + dv$ is

$$\frac{\partial N}{\partial v} = \Lambda v^2 \exp\left(-\frac{mv^2}{2kT}\right) \tag{2.84}$$

or

$$\frac{\partial N}{\partial W} = \frac{2^{1/2}}{m^{3/2}} \Lambda W^{1/2} \exp\left(-\frac{W}{kT}\right), \tag{2.85}$$

where Λ is a constant and $W = mv^2/2$.

This is the so-called **Maxwellian distribution**.

Clearly,

$$N = \int_0^\infty \frac{\partial N}{\partial v} dv = \Lambda \int_0^\infty v^2 \exp\left(-\frac{mv^2}{2kT}\right) dv, \tag{2.86}$$

where N (the total number of molecules) does not change with temperature.

It turns out that

$$\int_0^\infty v^2 \exp\left(-\frac{mv^2}{2kT}\right) dv = \frac{\pi^{1/2}}{4} \left(\frac{2kT}{m}\right)^{3/2}. \tag{2.87}$$

Thus,

$$N = \Lambda\frac{\pi^{1/2}}{4}\left(\frac{2kT}{m}\right)^{3/2} \quad \therefore \quad \Lambda = 4N\pi^{-1/2}\exp\left(\frac{m}{2kT}\right)^{3/2}, \tag{2.88}$$

$$\frac{\partial N}{\partial v} = 4N\pi^{-1/2}\left(\frac{m}{2kT}\right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2kT}\right) \tag{2.89}$$

and

$$\frac{\partial N}{\partial W} = 2N\pi^{-1/2}\frac{W^{1/2}}{(kT)^{3/2}}\exp\left(-\frac{W}{kT}\right). \tag{2.90}$$

The shape of the $\partial N/\partial v$ versus v plot depends, of course, on the temperature, as shown in Figure 2.8 where T_0 is an arbitrary reference temperature. However, the area under the curve, being a measure of the total number of molecules in the gas, is independent of temperature.

The peak value of $\partial N/\partial v$ is

$$\frac{\partial N}{\partial v} = \frac{2N}{e}\left(\frac{2m}{\pi kT}\right)^{1/2} \tag{2.91}$$

and occurs when $v = \sqrt{2kT/m}$ or, equivalently, when $W = kT$.

As T approaches 0, $\partial N/\partial v_{[max]}$ approaches ∞ and occurs for $v = 0$. The distribution becomes a delta function at $T = 0$.

This means that according to this classical theory, at absolute zero, all the molecules have zero velocity and zero energy.

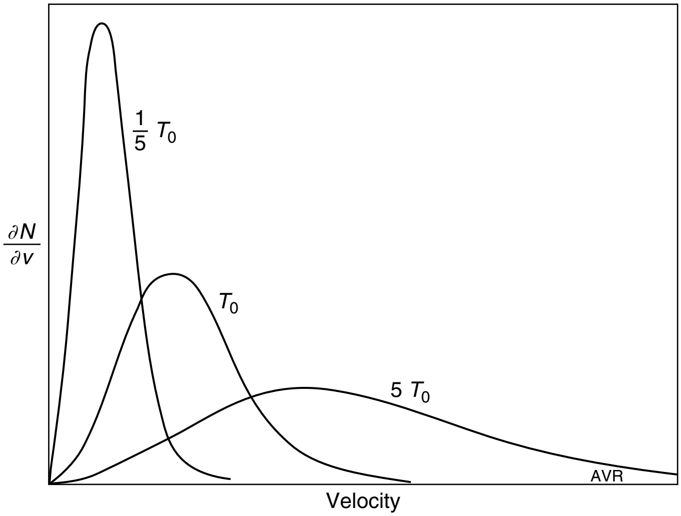


Figure 2.8 The Maxwellian velocity distribution at three different temperatures.

2.19 Fermi–Dirac Distribution

Electrons in metals do not behave in a Maxwellian way. They are governed by the **Pauli exclusion principle**, which states that in a given system, no two electrons can have the same quantum numbers, and, consequently, they cannot all have zero energy at absolute zero. Rather, at absolute zero, electrons must be distributed uniformly in energy up to a given energy level. No electron has energy above this level, called the **Fermi level**. Thus, the probability, f, of finding electrons at a given energy level is:

$$f = 1, \quad \text{for } W < W_F \quad \text{and} \quad f = 0 \quad \text{for} \quad W > W_F. \tag{2.92}$$

As before, the number of allowed states with momenta less than $|p|$ (notice the change in terminology) is proportional to the volume of a sphere with radius $|p|$ in the momentum space:

$$N \text{ (with momenta less than } |p|) = \frac{2}{h^3}\left(\frac{4}{3}\pi p^3\right).$$

We have used $2/h^3$ as a proportionality constant. Although we will not present a justification for this, it should be noticed that the factor 2 is the result of two possible spins of the electron and that the dimensions come out right: N is the number of electrons per cubic meter.

Since $p = mv$, it follows that $p^2 = 2mW$ and

$$N \text{ (with momenta less than } |p|) = \frac{2}{h^3} \times \frac{4}{3}\pi(2mW)^{3/2}. \tag{2.93}$$

Thus

$$\frac{\partial N}{\partial W} = \frac{8\sqrt{2}\pi}{h^3} m^{3/2} W^{1/2}. \tag{2.94}$$

When the temperature is larger than zero, the probability that a given state be occupied is given by

$$f = \frac{1}{1 + \exp\left(\frac{W-\mu}{kT}\right)}. \tag{2.95}$$

The quantity μ is called the **chemical potential**. When $T = 0$, $\mu = W_F$ and the function above has the property that $f = 1$ for $W < W_F$ and $f = 0$ for $W > W_F$, as required. When $T \neq 0$, μ is slightly smaller than W_F.

The **density of states**, using this probability function, becomes

$$\frac{\partial N}{\partial W} = \frac{8\sqrt{2}\pi}{h^3} m^{3/2} \frac{W^{1/2}}{1 + \exp\left(\frac{W-\mu}{kT}\right)} \tag{2.96}$$

and

$$N = \frac{8\sqrt{2}\pi}{h^3} m^{3/2} \int_0^\infty \frac{W^{1/2}}{1 + \exp\left(\frac{W-\mu}{kT}\right)} dW. \tag{2.97}$$

Since N is independent of T, the integral must itself be independent of T, which means that the chemical potential, μ, must depend on T in just the correct manner.

For $kT \ll \mu$,

$$\int_0^\infty \frac{W^{1/2}}{1 + \exp\left(\frac{W-\mu}{kT}\right)} dW \approx \frac{2}{3}\mu^{3/2} = \frac{2}{3} W_F^{3/2}, \tag{2.98}$$

that is, the integral does not depend on T and $\mu = W_F$.

At higher temperatures, to ensure the invariance of the integral, μ must change with T approximately according to

$$\mu = W_F - \frac{(\pi kT)^2}{12 W_F}. \tag{2.99}$$

The difference between μ and W_F is small. For Fermi levels of 5 eV and T as high as 2000 K, it amounts to only some 0.1%. Therefore, for most applications, the chemical potential can be taken as equal to the Fermi level. One exception to this occurs in the study of the thermoelectric effect. Figure 2.9 shows a plot of the density of states, $\partial N/\partial W$, versus W for three temperatures. Notice that the chemical potential, which coincided with the Fermi level at $T = 0$, has shifted to a slightly lower energy at the higher temperature.

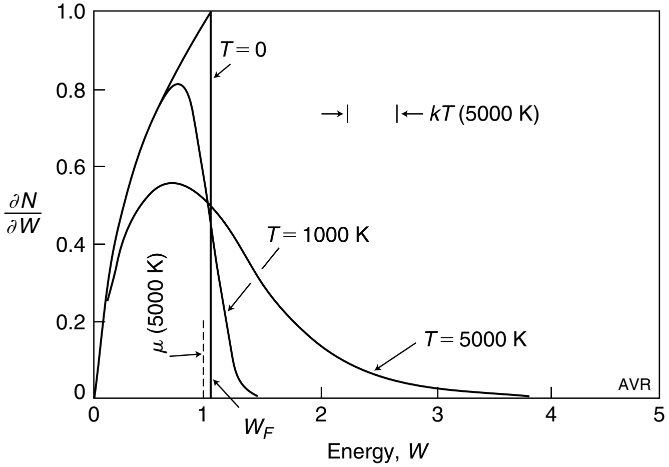


Figure 2.9 Energy distribution of electrons in a metal at three different temperatures.

2.20 Boltzmann's Law

A very useful result from statistical mechanics is **Boltzmann's law**, which describes the concentration of particles in a gas as a function of their potential energy and their temperature. This law is used in a number of chapters in this book.

Consider a force, F (derived from a potential), acting on each atom and aligned along the x-direction. nF is, of course, the total force acting on a cubic meter of gas, and, if we restrict ourselves to a rectangular prism of base area, A, and height, dx, the force is $nFdxA$, and the pressure on the base is $nFdx$. In equilibrium, this pressure must balance the gas pressure, $kTdn$.

$$Fdx = kT\frac{dn}{n}. \tag{2.100}$$

The potential energy, $W_{pot} = -\int Fdx$, is

$$W_{pot} = -kT\ln n, \tag{2.101}$$

from which, we obtain Boltzmann's equation

$$n \propto \exp\left(-\frac{W_{pot}}{kT}\right). \tag{2.102}$$

Example

Atmospheric pressure

Each molecule in air has a potential energy, $mg\Delta h$, relative to a plane Δh meters closer to the ground.

According to Boltzmann's law, the concentration of molecules must vary as

$$n = n_0 \exp\left(-\frac{mg\Delta h}{kT}\right). \tag{2.103}$$

Note that kT/mg has the dimensions of length. It is called the **scale height**, H:

$$n = n_0 \exp\left(-\frac{\Delta h}{H}\right) \tag{2.104}$$

and, if T is independent of h,

$$p = p_0 \exp\left(-\frac{\Delta h}{H}\right). \tag{2.105}$$

(*Continues*)

(*Continued*)

In an isothermal atmosphere, the air pressure falls exponentially with height.

Taking the mean mass of the molecules in air as 29 daltons, and knowing that to convert daltons to kg it suffices to divide by Avogadro's number, we find that the mass of a representative air molecule is about 48×10^{-27} kg. Consequently, the scale height of Earth's atmosphere (isothermal at 300 K) is

$$H = \frac{1.38 \times 10^{-23} \times 300}{48 \times 10^{-27} \times 9.8} = 8800 \quad \text{m.} \tag{2.106}$$

Appendix: Symbology

We will try to adopt the following convention for representing thermodynamic quantities such as

G, free energy,
H, enthalpy,
Q, heat,
S, entropy, and
U, internal energy.

1. Capital letters indicate the quantity associated with an arbitrary amount of matter or energy.
2. Lowercase letters indicate the quantity per unit. A subscript may be used to indicate the species being considered. For example, the free energy per kilomole of H_2 will be represented by $\overline{g_{H_2}}$.

g = free energy per kilogram.
$\overline{g}$ = free energy per kilomole.
g^* = free energy per kilogram, at 1 atmosphere pressure.
$\overline{g}^*$ = free energy per kilomole, at 1 atmosphere pressure.
$\overline{g}_f$ = free energy of formation per kilomole.
$\overline{g}_f^\circ$ = free energy of formation per kilomole, at 298 K, 1 atmosphere, that is, at RTP (Standard Free Energy of Formation).

For more information on some topics in this chapter, read:

Reference

Cengel, Y. A., and M. A. Boles, *Thermodynamics, An Engineering Approach*, McGraw-Hill, **1994**.

PROBLEMS

2.1 10 kg/s of steam ($\gamma = 1.29$) at 2 MPa are delivered to an adiabatic turbine (100% efficient). The exhaust steam is at 0.2 MPa and 400 K.

1. What is the intake temperature?
2. What power does the turbine deliver?

2.2 Show that the cylinder and piston experiment of Section 2.15 (with the solid object inside) is reversible, provided the compression is carried out infinitely slowly. Do this numerically. Write a computer program in which compression and expansion take place in suitably small steps and are, in each step, followed by an equalization of temperature between the gas and the solid object within the cylinder.

2.3 Refer to the experiment described in Section 2.10. Show that the work done in lifting the 1-kg mass in two steps (first 2 kg, then 1 kg) is 14.2 J. Show that the 2-kg mass rises 0.444 m. Assume that the steps occur slowly enough so that the gas cooled by the expansion returns to the original temperature after each step.

2.4 Consider 10 m^3 (V_0) of gas ($\gamma = 1.6$) at 10^5 Pa (p_0) and 300 K (T_0).

1. How many kilomoles, μ, of gas are involved?
2. The gas is compressed isothermally to a pressure, p_f, of 1 Mpa.

   2.1 What is the new volume, V_f?

   2.2 How much energy was used in the compression?

3. Now, instead of compressing the gas isothermally, start again (from V_0, p_0, and T_0) and compress the gas adiabatically to a pressure, p_2. The gas will heat up to T_2. Next, let it cool down isometrically (i.e., without changing the volume) to $T_3 = 300$ K and a pressure, p_3, of 1 MPa. In other words, let the state return to that after the isothermal compression.

   3.1 What is the pressure, p_2?

   3.2 What is the temperature, T_2, after the adiabatic compression?

   3.3 What is the work done during the adiabatic compression?

   3.4 Subtract the heat rejected during the isometric cooling from the work done during the adiabatic compression to obtain the net energy change owing to the process described in Item 3.3.

2.5 When a gas expands, it does an amount of work

$$W = \int_{V_0}^{V_1} p dV.$$

If the expansion is adiabatic, the polytropic law is observed and the integral becomes (see Chapter 6)

$$W = \frac{p_0 V_0^{\gamma}}{\gamma - 1}\left(V_1^{1-\gamma} - V_0^{1-\gamma}\right).$$

Show, by using the definitions of c_v and of γ, that this work is equal to the energy needed to raise the temperature of the gas from T_0 to T_1 under constant volume conditions.

2.6 The domains in a nonmagnetized ferromagnetic material are randomly oriented; however, when magnetized, these domains are reasonably well aligned. This means, of course, that the magnetized state has a lower total entropy than the nonmagnetized state.

There are materials (gadolinium, Gd, for example) in which this effect is large. At 290 K, polycrystalline gadolinium (atomic mass 157.25, density 7900 kg/m^3) has a total entropy of 67.6 kJ K^{-1} kmole^{-1} when unmagnetized and 65.6 kJ K^{-1} kmole^{-1} when in a 7.5 tesla field.

Assume that 10 kg of Gd are inside an adiabatic container, in a vacuum, at a temperature of 290 K. For simplicity, assume that the heat capacity of the container is negligible. The heat capacity of Gd, at 290 K, is 38.4 kJ K^{-1} kmole^{-1}.

Estimate the temperature of the gadolinium after a 7.5 T field is applied.

2.7 The French engineer, Guy Negre, invented an "eco-taxi," a low-pollution vehicle. Its energy storage system consists of compressed air tanks that, on demand, operate an engine (it could be a turbine, but in the case of this car, it is a piston device).

There are several problems to be considered. Let us limit ourselves to the turnaround efficiency of the energy storage system. For comparison, consider that a lead–acid battery has a turnaround efficiency of somewhat over 70% and flywheels, more than 90%.

A very modern compressed gas canister can operate at 500 atmospheres.

1. Calculate the energy necessary to compress 1 kilomole of air ($\gamma = 1.4$) *isothermally* from 1 to 500 atmospheres. The temperature is 300 K.
2. One could achieve the same result by compressing the air adiabatically and then allowing it to cool back to 300 K. Calculate the energy necessary to accomplish this.
3. The compressed air (at 300 K) is used to drive a turbine (in the French scheme, a piston engine). Assume that the turbine is ideal—**isentropic**—and it delivers an amount of mechanical energy equal to the change of enthalpy the gas undergoes when expanding. How much energy does 1 kilomole of air deliver when expanding under such conditions?

   To solve this problem, follow the steps suggested here.

   3.1 Write an equation for the change of enthalpy across the turbine as a function of the input temperature (300 K) and the unknown output temperature.

3.2 Using the polytropic law, find the output temperature as a function of the pressure ratio across the turbine. Assume that the output pressure is 1 atmosphere.

If you do this correctly, you will find that the temperature at the exhaust of the turbine is below the liquefaction point of the gases that make up air. This would interfere with the turbine operation, but in the present problem, disregard this fact.

3.3 Once you have the exhaust temperature, calculate the mechanical energy generated by the turbine.

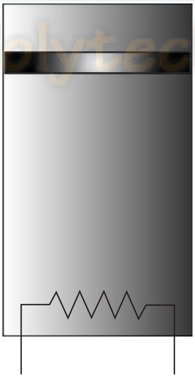
4. What is the turnaround efficiency of the compressed air energy storage system under the (optimistic) assumptions of this problem. That is, what is the ratio of the recovered energy to the one required to compress the air?

2.8 The cylinder in the picture initially, has, a 1-liter volume and is filled with a given gas at 300 K and 10^5 Pa. It is perfectly heat insulated and is in a laboratory at sea level. The frictionless piston has no mass, and the piston and cylinder, as well as the 1-ohm electric resistor, installed inside the device have negligible heat capacity.

At the beginning of the experiment, the piston is held in place, so it cannot move.

A 10-amp dc current is applied for 1 second, causing the pressure to rise to 1.5×10^5 pascals. Next, the piston is released and rises.

What is the work done by the piston?



2.9 A metallic cylinder with a 3-cm internal diameter is equipped with a perfectly gas-tight and frictionless piston massing 1 kg. It contains a gas with a $\gamma = 1.4$. When in a lab at sea level and at a room temperature of 300 K, the bottom of the piston is exactly 10 cm above the closed end of the cylinder. In other words, the gas fills a cylindrical volume 10 cm high and 3 cm in diameter.

1. An additional mass is added to the piston so that the total now masses 10 kg. Clearly, the gas inside will be compressed. Assume

that this compression occurs very rapidly but not rapidly enough to be considered an abrupt process. Make an educated guess as to the nature of the compression process (adiabatic, isothermal, etc.).

What is the height of the piston at the end of the compression?

2. A considerable time after the above compression, the piston has settled at a different height. What is that height?

3. Now, repeat 1.1 but arrange for the piston to descend very slowly. Again, make an educated guess as to the nature of the compression process (adiabatic, isothermal, etc.).

   When the piston finally settles at an unchanging height, how high will it be above the bottom of the cylinder?

4. Which, if any, of the two processes—Process 1 & 2 and Process 3—is reversible? Demonstrate mathematically the correctness of your answer. In other words, for each case, determine, if removing the excess 9 kg from the piston causes the system to return to the initial state. Again, the expansion in the reversal of Process 1 & 2 is very fast, and in Process 3, it is very slow.

Chapter 3
Mechanical Heat Engines

3.1 Heats of Combustion

The driving agent of a heat engine is a temperature differential. A heat engine must have a source and a sink of heat. The heat source may be direct solar radiation, geothermal steam, geothermal water, ocean water heated by the sun, nuclear energy (fission, fusion, or radioactivity), or the combustion of a fuel. In developed countries, over 90% of the energy is derived from combustion of fuels, almost all of which are of fossil origin.

When carbon burns completely in an oxygen atmosphere, the product is carbon dioxide, which, under all normal circumstances in this planet, is a gas. However, most fuels contain hydrogen, and, thus, when burned, they also produce water. The resulting water may leave the engine in liquid or in vapor form. In the former case it releases its latent heat of vaporization to the engine. For this reason, hydrogen-bearing fuels can be thought of as having two different heats of combustion: one, called the **higher heat of combustion**, corresponds to the production of liquid water, whereas the other, corresponding to the formation of water vapor, is called the **lower heat of combustion**.

The higher heat of combustion of hydrogen is 143 MJ/kg, while the lower is 125 MJ/kg. The heat of combustion of carbon is 32.8 MJ/kg. Thus, one could expect that the heat of combustion of a hydrocarbon, C_nH_m, is roughly

$$\Delta H \approx \frac{12n \times 32.8 + 143m}{12n + m} \tag{3.1}$$

for the case in which the product is liquid water.

Table 3.1 compares some higher heats of combustion calculated from our simplified formula with those actually measured. Naturally, our formula overestimates these heats because it does not take into account the binding energies between C and H and between C and C. However, our reasoning predicts the regular behavior of the heats as a function of molecular mass.

The higher the order of the hydrocarbon, the larger the relative amount of carbon compared with hydrogen: thus, the smaller the heat of combustion. Methane, the first of the aliphatic hydrocarbons, has the largest heat of combustion: 55.6 MJ/kg. As the order increases, the heat of combustion decreases tending toward 47 MJ/kg. This is illustrated in Figure 3.1. A similar regularity is observed as one moves from one series of hydrocarbons to the next. Table 3.2 demonstrates this effect.

Table 3.1 Comparison between Estimated and Measured Higher Heats of Combustion

	ΔH(MJ/kg) estimated	**ΔH(MJ/kg) measured**	**Error %**
CH_4	60.4	55.6	8.3
C_2H_6	54.8	52.0	5.3
C_3H_8	52.8	50.4	4.6
$C_{10}H_{22}$	49.9	47.4	5.0
$C_{20}H_{42}$	49.2	47.2	4.0

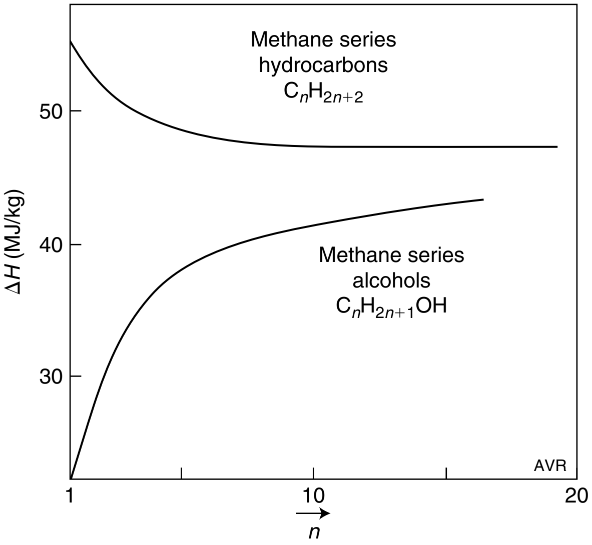


Figure 3.1 Heats of combustion of hydrocarbons and alcohols.

An alcohol results when an OH radical replaces a hydrogen in a hydrocarbon (Figure 3.2). Gaseous hydrocarbons form liquid alcohols. Thus, for vehicles (where a liquid is more useful than a gas), it is advantageous to partially oxidize hydrocarbons, transforming them into alcohols. This partial oxidation causes the alcohol to have a smaller heat of combustion than the parent hydrocarbon:

CH_4 (55.6 MJ/kg) yields CH_3OH (22.7 MJ/kg),
C_2H_6 (52.0 MJ/kg) yields C_2H_5OH (29.7 MJ/kg).

However, 1 kmole (16 kg) of methane yields 1 kmole (32 kg) of methanol. Hence, 1 kg of methane can be transformed into 2 kg of methanol with an efficiency of

$$\eta = \frac{2 \times 22.7}{55.6} = 0.82. \tag{3.2}$$

Table 3.2 Higher Heats of Combustion of Hydrocarbons of the Form C_nH_{2n+a}

n ↓	$a \to 2$	0	-2	-4	-6	-8	-10
0	143						
1	55.6						
2	52.0						
3	50.5		MJ/kg				
4	50.2						
5	49.1						
6	48.5						
7	48.2						
8	47.9						
10	47.4	47.3	45.2	44.3	44.4	42.9	41.7
16	47.4						
20	47.2						

The table shows approximate values. Exact values vary because there are, for many empirical formula, various different isomers.

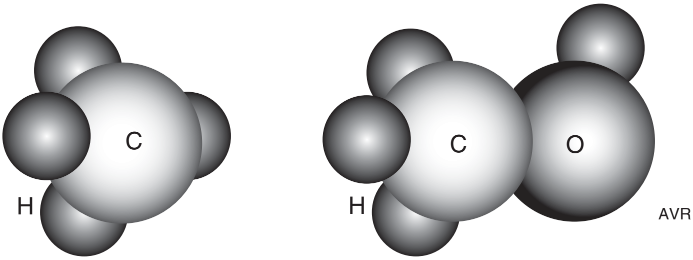


Figure 3.2 An alcohol results from the substitution of a hydrogen atom by a hydroxyl radical. The figure shows the structures of methane and its corresponding alcohol, methanol.

Table 3.3 Efficiency of Conversion of Hydrocarbons into Alcohols

Methane into methanol	0.82
Ethane into ethanol	0.88
Propane into propanol	0.91

Table 3.3 shows the *theoretical* efficiencies of converting hydrocarbons into their corresponding alcohols.

In addition to hydrocarbons and alcohols, there are a number of other fuels to be considered.

Fuels can also be derived from biomass (Chapter 13). They all contain oxygen in their molecule and, for this reason, have lower heats of combustion than hydrocarbons. Cellulose, for instance, has a heat of combustion of 17.4 MJ/kg.

3.2 Carnot Efficiency

Mechanical and electrical energy are "noble" forms of energy: no entropy is associated with them. Consequently, it is theoretically possible to transform one into another without losses—in other words, without having to reject heat. Big machines can make this transformation with over 99% efficiency.

Heat engines transform a degraded form of energy—heat—into a noble form: either electrical or mechanical. This cannot be done without rejecting part of the input heat (unless the engine works against a sink at absolute zero). Thus, even theoretically, the achievable efficiency is smaller than 1.

Figure 3.3 indicates how a heat engine operates. The input is an amount of heat, Q_{in}, the useful output is a quantity, W (of mechanical or electrical energy), and Q_{out} is an amount of heat that must be rejected. The efficiency of such an engine is the ratio of the output energy to the heat input:

$$\eta = \frac{W}{Q_{in}} = \frac{Q_{in} - Q_{out}}{Q_{in}}. \tag{3.3}$$

The entropy in the system must increase or, at best, suffer no change. Hence, the largest possible efficiency—**the Carnot efficiency**—corresponds to the situation in which the entropy at the output, Q_{out}/T_C, is equal to that at the input, Q_{in}/T_H. Of course, there is no entropy associated with the "noble" work output, W. Consequently,

$$\frac{Q_{in}}{T_H} = \frac{Q_{out}}{T_C}. \tag{3.4}$$

$$Q_{out} = \frac{T_C}{T_H} Q_{in}. \tag{3.5}$$

$$\eta_{CARNOT} = \frac{T_H - T_C}{T_H}. \tag{3.6}$$

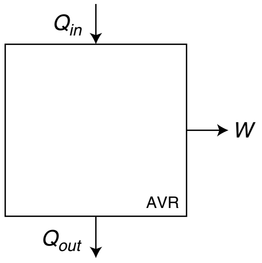


Figure 3.3 The operation of a heat engine.

Thus, the Carnot efficiency depends only on the temperatures between which the engine operates.

In steady-state conditions,

$$\frac{Q_{out}}{W} = \frac{Q_{in} - W}{W} = \frac{Q_{in}}{W} - 1 = \frac{1}{\eta} - 1. \tag{3.7}$$

As η decreases, the ratio above increases rapidly. A modern fossil-fueled steam turbine with $\eta \approx 0.6$ rejects only 0.67 joules of heat for each joule of useful energy produced. In a nuclear reactor such as all reactors in the United States, the efficiency is lower (some 28%) because of limitations in T_H: the rejected heat is then 2.6 times larger than the generated energy.[†] The cooling tower of nuclear plants must be substantially larger than that of a fossil-fueled plant.

Automotive engines have small efficiencies ($\eta \approx 0.2$) and the large amount of heat they reject can pose a serious problem when special equipment is needed for this purpose. Fortunately, Otto and Diesel engines do not need such equipment. The radiator found in most cars rejects only the heat from component inefficiency, not the thermodynamically rejected heat.

3.3 Engine Types

All mechanical heat engines involve four processes:

1. Compression (or pumping)
2. Heat addition
3. Expansion
4. Heat rejection

Figure 3.4 illustrates the four processes as they occur in a **close-cycle Rankine** engine. The expansion can be accomplished through a turbine or through a cylinder with a piston, as in the locomotives one still sees in period movies.

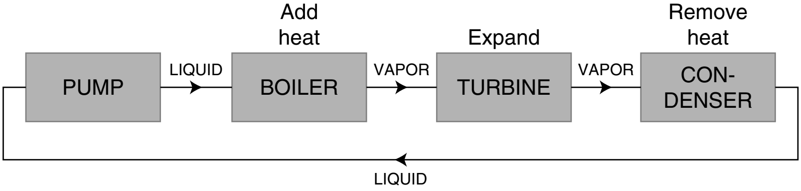


Figure 3.4 The four processes in the Rankine cycle. In the close-cycle configuration, the working fluid is condensed for reuse.

[†] This is not true of fast reactors such as the heavy-metal ones that operate at high temperatures.

Since steam engines at the beginning of the twentieth century turbine had less than 10% efficiency, the amount of heat that the condenser (in the close-cycle machines) had to reject was large: some 9 joules for each joule of mechanical energy produced (Equation 3.7).

To avoid unwieldy condensers, early locomotives released the spent vapor exhausting from the cylinder directly into the air (thus leaving to the environment the task of heat removal). See Figure 3.5.

Water could not be recovered in such open-cycle machines. The locomotive had to be equipped with a large water tank whose size limited the range of the locomotive.

Heat must be added from an external source, as the Rankine cycle is an **external heat source** (most frequently, an **external combustion**) cycle.

Another example of an open-cycle engine (in this case an **internal combustion cycle**) is the Brayton turbine used in jet and turbojet planes, illustrated in Figure 3.6.

Examples of close-cycle engines are

The Rankine or "vapor"-cycle engine and
The Stirling engine.

Examples of open-cycle engines are

The Otto (spark-ignition) engine,
The Diesel (compression-ignition) engine, and
The open-cycle Brayton engine.

The Rankine engine is distinguished by having the working fluid change phase during the cycle. Different working fluids can be employed. For operation at very low temperatures (say, boiler temperature of some

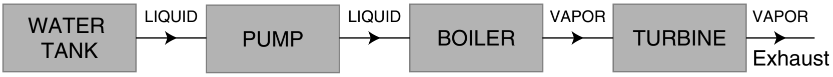


Figure 3.5 In the open-cycle Rankine, exhaust vapor is released directly into the environment. A large water tank is required.

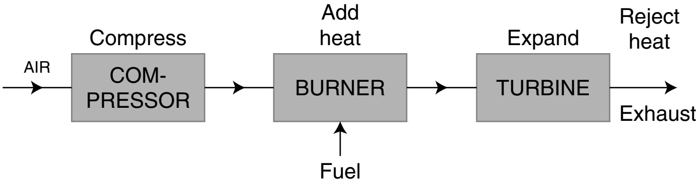


Figure 3.6 The Brayton cycle used in jet planes is an internal combustion, open-cycle engine.

Table 3.4 Nature of the Four Processes in Different Engines

Cycle	Compression	Heat addition	Expansion	Heat rejection
Carnot	adiabatic	isothermal	adiabatic	isothermal
Otto	adiabatic	isometric	adiabatic	isometric
Diesel	adiabatic	isobaric	adiabatic	isometric
Brayton	adiabatic	isobaric	adiabatic	isobaric
Stirling	isothermal	isometric	isothermal	isometric
Ericsson	isothermal	isobaric	isothermal	isobaric

25 C, as is the case in ocean thermal energy converters), the most appropriate fluid for the Rankine cycle is ammonia. For low operating temperatures such as those in some direct solar engines, one of several available freons may work best. Most fossil-fueled Rankine engines use water vapor, although mercury may be effective at higher temperatures.

Steam engines are ill suited for automobiles—they either require a large condenser or a large water tank. In part, because no heat rejection device is required, the open-cycle Brayton turbine is universally used in jet planes. Various engines that use only gases as working fluids differ in the nature of the processes used, as illustrated in Table 3.4. For good efficiency, mechanical heat engines must operate with high compression ratios.

In gas turbines, two different types of compressors can be used: radial or axial. Radial compressors can be built with compression ratios of about 3:1 per stage. Thus, for a 9:1 compression, two stages are sufficient. The disadvantage is that it is difficult to obtain high efficiencies with this type of compressor, especially when stages are ganged. In radial (centrifugal) compressors, the air is accelerated toward the rim and must be redirected to the intake at the hub of the next stage. However, channeling air around involves losses. Furthermore, part of the energy imparted to the air is rotational and therefore has to be converted to translational energy by means of (lossy) **diffusers**.

Axial compressors yield only a small compression ratio per stage, say, 1.2:1. Thus, a large number of stages must be employed. The Rolls-Royce "Tyne" aircraft turbine achieved an overall 13.5:1 compression using 15 stages, with a mean compression ratio of 1.189:1 per stage. The large number of stages is practical because of the simplicity with which the output of one stage can be fed to the input of the next. The trouble with axial compressors is that their efficiency falls rapidly with decreasing size owing mostly to blade tip leakage, something that is difficult to avoid in small machines. For this reason, Brayton turbines of relatively low power are still of the radial type.

Gas turbines may be at the verge of opening a new significant market as a bottoming cycle for solid oxide fuel cells whose exhaust consists of high-temperature gases well suited to drive this class of turbines.

In both Rankine and Brayton engines, each of the processes is carried out in a different part of the equipment. Compression and expansion are

Table 3.5 Some Characteristics of the More Common Combustion Engines

	Temperature ratio	**Component efficiency**	**Relative thermo-dynamics EFFIC**	**Heat rejection (thermo-dynamics)**
Otto	High	Fair	Poor	None
Diesel	High	Fair	Poor	None
Rankine	Low	Good	Fair	Large
Stirling	Moderate	Good	Very good	Medium
Brayton (o.c.)	Mod./high[†]	Very good	Good	None
Brayton (c.c.)	Moderate	Very good	Good	Medium

[†] *Development of more modern materials is expected to permit use of higher temperature ratios.*

accomplished by separate devices, and combustion occurs in its special chamber. This allows optimization of each part.

In Otto and Diesel engines, three processes are carried out in the same engine part—the cylinder. The fourth process—heat rejection—is carried out in the environment. The multiple functions of the cylinder require design compromises with resulting lower component efficiency.

Although Rankine and Brayton engines are of the continuous combustion types, Otto and Diesel operate intermittently and are able to tolerate higher working temperatures, since the combustion phase lasts only briefly. Also, heat is generated directly in the working fluid, thereby doing away with heat transfer problems encountered in all external combustion engines.

The Stirling engine, which will be examined in greater detail toward the end of this chapter, holds considerable promise, but, although quite old in concept, it has never achieved the popularity that it appears to deserve. It is, at least theoretically, the most efficient of the mechanical heat engines. It is less polluting than the Otto and the Diesel because it burns fuel externally. In addition, since the Stirling engine involves no explosions, it runs more smoothly and with less noise than other engines. Finally, in spite of being a close-cycle device, it requires only moderate heat rejection equipment.

The Stirling cycle also finds application in some refrigeration equipment. Since it operates with helium, hydrogen, or air, it employs no freons or other ecologically damaging fluids.

D. G. Wilson (1978) summarized the characteristics of the most common mechanical heat engines, as shown in Table 3.5.

3.4 Efficiency of an Otto Engine

In an ideal Otto cycle, a fuel/air mixture is compressed adiabatically from a volume, V_1, to a volume, V_2, during the compression phase. The gas

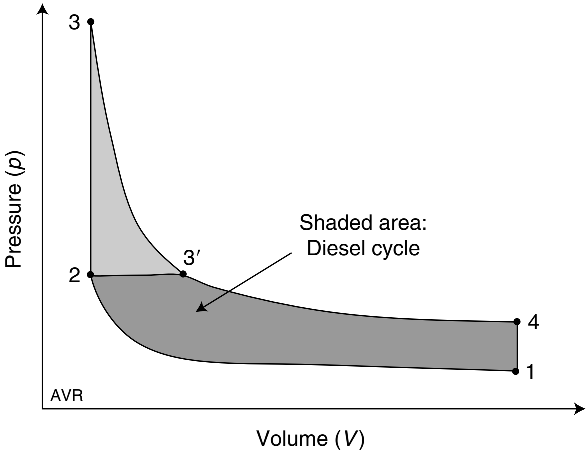


Figure 3.7 The p-V diagram of an Otto cycle.

traces out the line between Points 1 and 2 in the p-V diagram of Figure 3.7. Thus,

$$\frac{p_2}{p_1} = \left(\frac{V_1}{V_2}\right)^{\gamma}. \tag{3.8}$$

From the perfect-gas law

$$\frac{p_2 V_2}{T_2} = \frac{p_1 V_1}{T_1}. \tag{3.9}$$

$$T_2 = T_1 \frac{p_2 V_2}{p_1 V_1} = T_1 \left(\frac{V_1}{V_2}\right)^{\gamma-1} \tag{3.10}$$

$$T_2 = T_1 r^{\gamma-1}, \tag{3.11}$$

where r is the **compression ratio**, V_1/V_2. The compression is adiabatic because, owing to the rapidity with which it occurs, no heat is exchanged between the interior of the cylinder and the outside. Consequently, the work done is equal to the increase in internal energy of the gas:

$$W_{1,2} = \mu\, c_v (T_2 - T_1) = \mu\, c_v T_2 (1 - r^{1-\gamma}). \tag{3.12}$$

At the end of the compression phase, Point 2, a spark ignites the mixture, which, ideally, is assumed to burn instantaneously. Both temperature and pressure rise instantaneously, the pressure reaching the value indicated in Point 3 of the diagram. There is no change in volume during this heat addition or **combustion** phase. The amount of heat added is

$$Q_{2,3} = \mu\, c_v (T_3 - T_2). \tag{3.13}$$

The behavior during expansion is similar to that during compression:

$$W_{3,4} = \mu c_v T_3 (1 - r^{1-\gamma}). \tag{3.14}$$

The useful output of the engine is $W_{3,4} - W_{1,2}$, while the energy input is $W_{2,3}$. Hence the efficiency is

$$\eta = \frac{W_{3,4} - W_{1,2}}{Q_{2,3}} = \frac{T_3(1 - r^{1-\gamma}) - T_2(1 - r^{1-\gamma})}{T_3 - T_2} = 1 - r^{1-\gamma}. \tag{3.15}$$

Notice that from Equation 3.11, $r^{1-\gamma} = T_1/T_2$. Thus, Equation 3.15 can be written as

$$\eta = 1 - \frac{T_1}{T_2} = \frac{T_2 - T_1}{T_2} = \eta_{CARNOT}. \tag{3.16}$$

Thus, the ideal Otto cycle achieves the Carnot efficiency of an engine working between the maximum, precombustion temperature and the intake temperature. The ideal Otto cycle cannot achieve the Carnot efficiency determined by the highest and lowest temperature during the cycle.

Theoretical efficiency is plotted versus the compression ratio in Figure 3.8. The Diesel efficiency is lower than the Otto because, in the Diesel, the combustion is supposed to occur slowly, at constant pressure (see the darker shaded area in Figure 3.7), not at constant volume as in spark ignition engines. Thus, the total area enclosed by the *p*-*V* trace is smaller. In fact, the ideal efficiency of the Diesel cycle is

$$\eta = 1 - r^{1-\gamma} \left[\frac{r_c^\gamma - 1}{\gamma(r_c - 1)} \right], \tag{3.17}$$

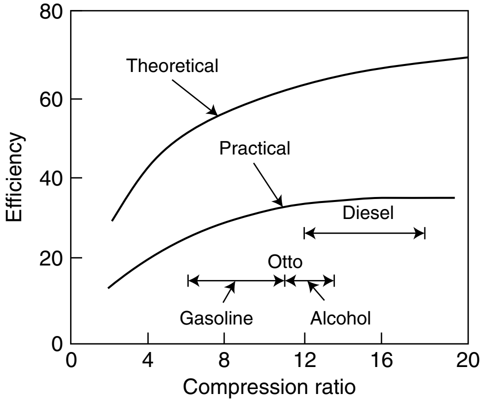


Figure 3.8 The efficiency of an Otto and a Diesel engine as a function of their compression ratios.

where $r_c \equiv V_{3'}/V_2$ (Figure 3.7) is called the **cutoff ratio** and is the expansion ratio during the combustion period.

In practice, Diesel Engines operate with higher efficiency than those based on the Otto cycle because the Otto cycle must operate at lower compression ratios to avoid **knocking** (Section 3.6).

Efficiency can be improved by

1. raising γ, and
2. raising the compression ratio.

As γ is larger for air than for fuel vapors, the leaner the mixture, the higher the efficiency. This is in part counterbalanced by the tendency of lean mixtures to burn slowly, causing a departure from the ideal Otto cycle. In addition, if mixtures become too lean, ignition becomes erratic—the engine runs "rough" and tends to backfire.

The stoichiometric air/fuel ratio for gasoline is 14.7:1. However, maximum power is achieved with a very rich mixture (12:1 to 13:1), while maximum efficiency requires lean mixtures (16:1 to 18:1). The **stratified combustion** engine achieves an interesting compromise by injecting fuel and air tangentially into the cylinder so that, owing to the resulting centrifugal force, the mixture is richer near the cylinder wall and becomes progressively leaner toward the axis. Combustion is initiated in the rich region and propagates inward. While the mean mixture is sufficiently lean to ensure high efficiency, ignition is still reliable. Incidentally, Diesel engines can operate with leaner mixtures than spark-ignition engines.

As we are going to see, the nature of the fuel used in Otto cycle engines limits their maximum usable compression ratio. Too high a ratio causes the engine to "ping," "knock," or "detonate," particularly during acceleration. Compression ratios that allow a car to accelerate without detonation are too small for efficiency at cruising speeds. This can be ameliorated by changing the ignition timing. The spark is retarded during acceleration and is advanced for cruising. An ingenious variable compression ratio engine was developed by Daimler-Benz. The cylinder has two pistons, one connected to the crankshaft as usual and an additional one that sits freely above the first. A variable volume of oil can be inserted between these two pistons regulating their spacing, thereby adjusting the compression ratio.

Consider a gasoline engine see Figure 3.9 with a 9:1 compression ratio using a fuel–air mixture with a γ of 1.3. Its ideal efficiency is about 50%.

The efficiency is reduced by departures from the ideal cycle, such as

1. Failure to burn fast enough. The combustion does not occur at constant volume. On the other hand, high-speed Diesels tend to burn not at constant pressure but at a somewhat rising pressure.
2. Heat loss through the cylinder wall and through the piston and connecting rod. Thus, the heat generated by the burning mixture does not all go into expanding the gas. Design efforts have focused on producing a more nearly adiabatic cylinder.

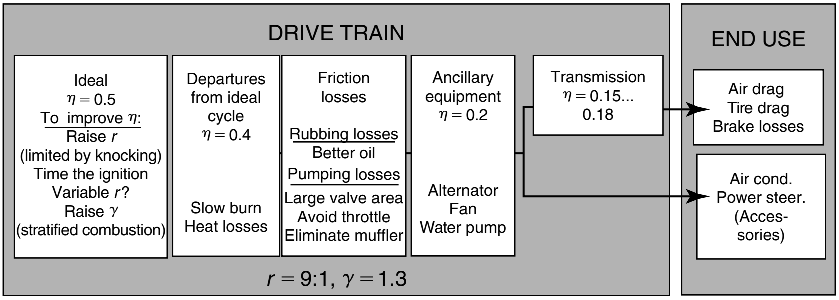


Figure 3.9 Drive train and end use in a spark-ignition vehicle. With a compression ratio of 9:1 and with an air/fuel mixture having a gamma of 1.3, the ideal engine has 50% efficiency. Assorted losses and ancillary equipment reduce the final efficiency to less than 18%.

Typically, these departures reduce the efficiency to some 80% of ideal. The engine in the example would then have a 40% efficiency.

It can be seen that for high efficiency one needs:

1. High compression ratio
2. Fast combustion
3. Lean mixture
4. Low heat conduction from cylinder to the exterior

There are numerous losses from friction between solid moving parts (**rubbing friction**) and from the flow of gases (**pumping friction**). Rubbing friction can be reduced by clever design, use of appropriate materials, and good lubricants. Pumping friction can be managed by adequate design of input and exhaust systems. Losses can be reduced by increasing the number of valves (hence the popularity of cars with four valves per cylinder). Also, part of the power control of an engine can be achieved by adjusting the duration of the intake valve opening, thereby avoiding the resistance to the flow of air caused by the throttle. Cars may soon be equipped with electronic sound cancellation systems, a technique that dispenses with the muffler and therefore permits a better flow of the exhaust gases.

An engine must use ancillary devices whose efficiency influences the overall performance. These include the alternator, the water pump, and the radiator cooling fan. In modern cars, the latter typically operates only when needed instead of continuously as in older vehicles. About half of the engine output is consumed by these devices. In the example we are considering, only some 20% of the combustion energy is available to the accessories (e.g., air conditioner and power steering) and to the transmission. The latter is about 90% efficient, so the residual power available to the propulsion of the car could be as little as 18% of the fuel energy.

Engine, transmission, and ancillary parts constitute the **drive train**. The load on the power train (e.g., tire drag, aerodynamic losses, brake losses, and accessories) constitutes the **end use load**.

3.5 Gasoline

Without a doubt, the most popular automotive fuel currently is gasoline. Gasoline is not a chemically unique substance—its composition has been continuously improved since its introduction, and it is also adjusted seasonally. It is a mixture of more than 500 components dominated by hydrocarbons with 3 to 12 carbon atoms. Most are branched (see discussion of the difference between *octane* and *isooctane* in the next section). The two main characteristics of gasoline are the following.

3.5.1 Heat of Combustion

Since the composition of gasoline is variable, its heat of combustion is not a fixed quantity. One may as well use the values for heptane or octane ($\approx$45 MJ/kg) as a representative higher heat of combustion.

3.5.2 Antiknock Characteristics

As far as energy content is concerned, gasoline has a decisive advantage over alcohol. However, there is no point in using a high-energy fuel if this leads to a low engine efficiency. As discussed previously, the efficiency is determined in part by the compression ratio, which, if too high, causes knocking (see next section). Alcohols tolerate substantially higher compressions than most gasolines and therefore lead to greater engine efficiencies. This somewhat compensates for the lower specific energy of the alcohols. Gasolines with better antiknock characteristics (higher octane rating) do not necessarily have higher energy content; in fact, they tend to have lower energy. It makes sense to use the gasoline with the lowest possible octane rating (i.e., the cheapest) compatible with the engine being fueled. On the other hand, cheaper gasolines, independent of their octane rating, may cause gum formation and other deposits in the engine and may result in more exhaust pollution.

3.6 Knocking

The efficiency of an engine increases when the compression ratio increases. In spark ignition engines, the compression must be limited to that tolerated by the fuel used. Compressions that are too high cause **detonation** or **knocking**, a condition that, in addition to being destructive to the pistons, leads to a reduction in the power of the motor.

The difference between **explosion** and **detonation** is the rapidity of combustion. Explosions propagate subsonically whereas detonations propagate supersonically. Gunpowder, for instance, will explode when confined but will only burn with a hiss when ignited in free air. On the other hand, substances such as nitrogen triiodide will decompose so fast that, even unconfined, will make a loud noise.

An explosion within a cylinder exerts a steady force on the piston analogous to the force a cyclist puts on the pedal of his bike. A detonation is as if the cyclist attempted to drive his vehicle with a succession of hammer blows.

The ability of a fuel to work with a high compression ratio without detonating is measured by its **octane rating**. A fuel is said to have an octane rating, O_f if it behaves (as far as detonation is concerned) like an isooctane/n-heptane mixture containing $O_f\%$ octane. The fuel need contain neither octane nor heptane. Pure isooctane is rated as 100% whereas n-heptane is rated as 0. This means that, by extrapolation, a given fuel may have an octane rating of less than 0 or more than 100. Experimentally, the octane rating of fuels with $O_f > 100$ is determined by comparing with isooctane to which a fraction, L (by volume), of tetraethyl lead, $(C_2H_5)_4Pb$, has been added. The octane rating is given by

$$O_f = 100 + \frac{107,000L}{(1 + 2786L) + \sqrt{1 + 5572L - 504,000L^2}}. \tag{3.18}$$

An addition of 0.17% (i.e., of a fraction of 0.0017) of tetraethyl lead to isooctane leads to a 120 octane ratio, a value common in aviation gasoline.

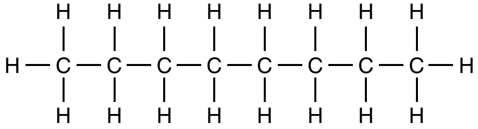
Notice that the compound used is *isooctane*. The reason is that n-octane—normal or unbranched octane—has extremely poor antiknocking behavior whereas isooctane resists knocking well.

The expression, **critical compression ratio, CCR,** actually has two different meanings. It may refer to the minimum compression ratio required to ignite the fuel in a compression ignition engine (Diesel), or, in a spark ignition engine (Otto), to the compression ratio that produces a just audible knock under a given experimental condition such as full load, 600 rpm and 450 K coolant temperature. Values for methane (13), n-pentane (3.8), n-hexane (3.3), and n-heptane (2.8) show the progressive decrease of the CCR with increasing methane series carbon number. A fairly ancient, yet interesting, article by John M. Campbell et al. (1935) lists the CCR for an extensive list of fuels.

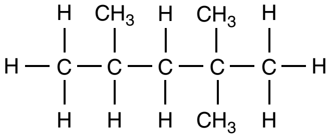
Observe that the hydrocarbons we mentioned in the preceding paragraph have all a liner chain configuration. CRCs improve if we use branched hydrocarbons. Isooctane has a CCR of 6 and benzene, the basic aromatic hydrocarbon, has a CCR of 15. For this reason, unleaded gasoline usually contains considerable amounts of aromatics to insure an acceptable octane rating.

Compare the structure of two isomers of octane (C_8H_{18}):

n-octane:



Isooctane:



The formula above shows that isooctane is technically pentane in which three hydrogens have been replaced each by a methyl (CH_3) radical. Two of the substitutions occur in position 2 and one in position 4 of the molecule. Hence isooctane is *2,2,4-trimethylpentane.*

The effective octane rating of a fuel depends on the conditions under which it operates. For this reason, more than one octane rating can be associated with any given fuel. The rating displayed on the gas station pump is usually an average of two differently measured values. A more complete discussion of this topic can be found in a book by J. B. Heywood.

Additives increase the octane rating of gasoline. Iodine can be used but is expensive. Up to a few years ago, tetraethyl lead was the standard additive in **leaded** gasolines. Environmental concerns have eliminated this type of fuel. High octane rating is now achieved by increasing the percentage of cyclic (benzene series) hydrocarbons. Thus, one avoids poisoning by increasing the risk of cancer and, incidentally, paying more for fuel.

The presence of ethanol in gasoline increases its resistance to detonation as indicated in Figure 3.10. It can be seen that the addition of 30% ethanol to low-grade gasoline raises its octane rating from 72 to 84. The

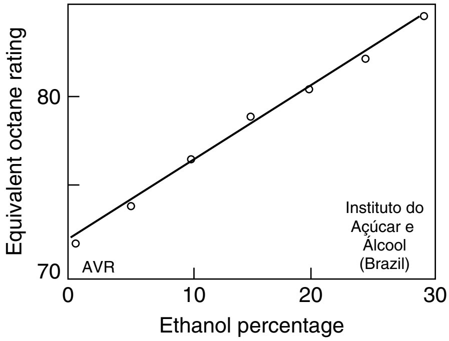


Figure 3.10 Addition of ethanol to gasoline results in a mixture with higher octane rating.

octane rating, O_m, of **gasohol** (gas/alcohol mixture) can be calculated from the octane rating, O_g, of the original gasoline and from the **blending octane value**, B, of the alcohol:

$$O_m = Bx + O_g(1 - x), \tag{3.19}$$

where x is the ratio of the additive volume to that of the gasoline. Depending on the initial quality of the fuel, the blending octane value of ethanol can be as high as 160. Methanol has a B of 130, although, when used alone, its rating is only 106. Gasohol can achieve high octane ratings without the use of lead and with only moderate addition of cyclic hydrocarbons. Thus, gasohol brings substantial public health advantages.[†]

Since 1516, Brazil has been the world's leading sugarcane producer. The widely fluctuating international price of sugar prompted Brazil to develop gasohol as a means of disposing of excess production. In years when the price was low, the alcohol percentage in Brazilian gasoline was high (typically, 24%). When sugar prices were high, much less ethanol found its way into automotive fuel (say, 5%). Starting in the 1970s, Brazil decided to sell pure (hydrated) alcohol as fuel for its fleet of specially designed cars, thus achieving a certain independence from the importation of oil.

Alcohol is more than an additive; it is itself a fuel. However, its energy content is lower than that of gasoline (Table 3.6). Per unit volume, ethanol contains only 71% of the energy of heptane, the main constituent of gasoline. Nevertheless, Brazilian alcohol-driven cars (using gasoline-free ethanol) have a kilometrage that approaches that of gasoline engines. This is due to the higher efficiency of the high-compression ethanol motors. Because water can be mixed with alcohol—inviting the "stretching" of the fuel sold at refueling stations—Brazilian pumps are equipped with densitometers permitting the consumer to check on the quality of the product.

Table 3.6 Properties of Two Important Alcohols Compared with Heptane and Octane (Higher heats of combustion for fuels at 25 C)

	Mol.	**kg/**	**MJ/**	**MJ/**	**MJ/kg**	**MJ/liter**
	mass	**liter**	**kg**	**liter**	**(rel. to**	**(rel. to**
					octane	**octane**
Methanol	32	0.791	22.7	18.0	0.475	0.534
Ethanol	46	0.789	29.7	23.4	0.621	0.697
n-Heptane	100	0.684	48.1	32.9	1.006	0.979
iso-Octane	114	0.703	47.8	33.6	1.000	1.000

[†] This may not be quite true when the additive is methanol because of the formaldehyde in the exhaust. With ethanol, the exhaust instead contains some acetaldehyde, a relatively innocuous substance.

Currently, **flex-fuel** cars are popular in Brazil where both pure ethanol (E-100) and gasoline are usually available at any refueling station. Flex-fuel engines will automatically adjust themselves to the fuel being used.

3.7 Hybrid Engines for Automobiles

Automobile emission standards are established individually by each state, but the leader is the California Air Resources Board (CARB), which has proposed the most stringent emission specifications in the country. These include a requirement that, by a given date, 2% of the vehicles sold in California be zero emission vehicles (ZEVs). This requirement was later postponed. Automobile manufacturers have spent considerable effort in the exegesis of the expression ZEV.

Clearly, a purely electric vehicle (EV) emits no noxious gases. However, it consumes electricity generated in part by fossil fuels, producing pollutants. An EV pollutes, albeit very little compared with a conventional internal combustion vehicle (ICV). Some argue that if an automobile equipped with an internal combustion engine emits the equivalent amount of pollutants (or less) as the total emission from an EV, then such an ICV should also be considered a "zero emission" car. The great popularity of the Toyota Prius, attests to the general interest in this type of vehicle.

A hybrid vehicle is an electric car equipped with an additional fuel-driven power source. Hybrids lead to a substantial lowering of emission for several reasons:

1. Whereas a normal automotive engine has to operate over a wide range of speeds, from idling to full acceleration, the engine of a hybrid is optimized for operation at, when possible, constant speed and can be fine-tuned for maximum efficiency and minimum pollution.
2. There is no waste during the frequent idling periods in normal city driving—the gas engine, instead of idling, is actually turned off.
3. Regenerative braking that returns power to the battery during deceleration can be implemented in a relatively simple manner.

There are two categories of hybrid vehicles: series and parallel. In series hybrids, the power applied to the wheels comes entirely from the electric motor(s). The fuel-driven component simply recharges the battery.

In parallel hybrids, wheel power is derived from both electric and IC motors. Clutches are used to couple these different power plants to the wheels according to the requirements of the moment.

Series hybrids are relatively simple but require large electric motors capable of delivering full acceleration power. In addition, they must have auxiliary systems to maintain battery charge. Thus series hybrids have large drive motors, a charging engine, and a generator. The sum of the powers of these three components substantially exceeds the power necessary to drive the vehicle. This can be expensive.

In parallel hybrids the electric motors can be much smaller, and the additional surge power comes from the IC power plant. However, the extra power plant in a hybrid does not have to be a heat engine. Fuel cells may prove ideal for such an application.

Many consider the hybrid car as a intermediary solution in the transition from the current IC engine to the future purely electric car whose advent still depends on battery improvements, among other things. It is easy to accept that the next step for the automotive industry will be the **plug-in hybrid**, a car that has batteries large enough for much of the daily city driving but can rely on its internal combustion engine for longer trips. For purely urban driving, no fuel is ever needed because the batteries are fully charged overnight using electricity supplied by the local utility.

3.8 The Stirling Engine

Had the early automobile developers opted for a Stirling engine rather than an Otto, present-day vehicles would be more efficient and less polluting. A quirk of history tipped the scales away from the Stirling. Nevertheless, in at least one application, there will be a revival of this old technology. As we are going to see in Chapter 5, space missions to the outer planets use as a source of power the Radioisotope Thermal Generator (RTG) because the feebleness of the sunlight in those faraway regions makes photovoltaics unpractical. RTGs use thermoelectric generators and have effective efficiencies of some 10%. By replacing thermocouples by free-piston Stirling engines, NASA plans to raise the efficiency to 30% and thus reduce both the mass and the very high cost of the plutonium heat source. Instead of the 10 kg of plutonium (millions of dollars), the radioactive fuel will be reduced to only 2 kg. Free-piston Stirling engines have longevity required by these missions. NASA expects to use them probably around 2012.

Stirlings have the following advantages:

1. They are more efficient than Otto and Diesel engines.
2. They can operate with a wide variety of fuels.
3. Being an external combustion engine, they tend to generate fewer pollutants. They still produce large amounts of carbon dioxide, but, owing to their greater efficiency, they produce less than current automotive engines of equivalent power. They can operate well with fuels having a low carbon-to-hydrogen ratio, thus producing more energy per unit amount of carbon emitted.
4. They are low-noise devices because no explosions are involved.

Table 3.7 Several Stirling Engine Configurations

Kinematic	Alpha (two cylinders, two pistons)
	Beta (one cylinder with piston and displacer)
	Gamma (one cylinder with piston, another with displacer)
Free-piston	
Ringbom	

In addition to its application to engines, the Stirling cycle can be adapted for refrigeration without needing Chlorinated Fluorocarbons (CFCs).

The Stirling cycle consists of an isothermal compression, an isometric heat addition, an isothermal expansion, and an isometric heat rejection (cf. Table 3.4). Its great efficiency results from the possibility of **heat regeneration** described in more detail later in this chapter. A number of Stirling engine configurations have been tried. See Table 3.7.

All configurations employ two pistons, in some cases a **power piston** and a **displacer**. The distinction will become clear when we examine examples of the engine.[†] **Kinematic** engines use pistons driven by the crankshaft, in general through connecting rods. In the **free-piston** configuration, the pistons are not mechanically connected to any part of the engine. The **Ringbom** uses one kinematic and one free piston.

3.8.1 The Kinematic Stirling Engine

3.8.1.1 The Alpha Stirling Engine

Since the alpha configuration is the easiest to understand, we will examine it in more detail.

Consider two cylinders interconnected by a pipe (Figure 3.11 and 3.12). One cylinder (labeled "Hot") is continuously heated by an external source—flame, radioisotopes, concentrated solar energy, and so on. The temperature of the gas in this cylinder is T_H. The other cylinder (labeled Cold) is continuously cooled by circulating water or blowing cool air or, perhaps simply by convection. The temperature of the gas in this cylinder is T_C. There is, as in any heat engine, a **source** and a **sink** of heat.

The space above the pistons is filled with a working gas (in practical engines, this may be hydrogen or helium). In order to follow the cycle, we will use a specific example. A gas with a $\gamma = 1.40$ is used. The volume of each cylinder can, by moving the piston, be changed from 10^{-3} m^3 to 0 m^3—that is, from 1 liter to 0 liters.

[†] Power pistons compress or do work on expansion. Displacers do no work. All they do is transfer fluid from one region of the machine to another.

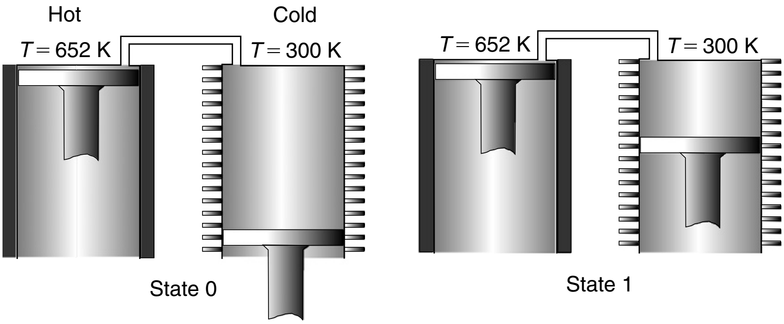


Figure 3.11 The first two states of an alpha Stirling cycle.

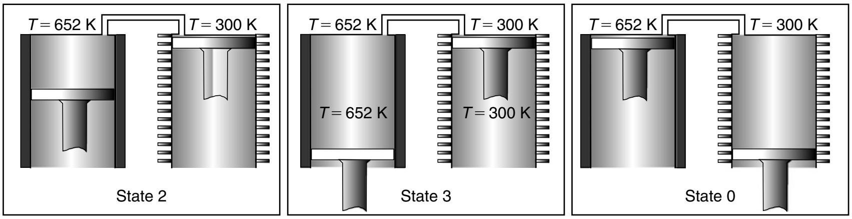


Figure 3.12 The final states of an alpha Stirling cycle.

Initially (State 0), the "cold" piston is all the way down. The volume in this cylinder is $V_{C_0} = 10^{-3}$ m^3, the temperature is $T_{C_0} = T_C = 300\,\text{K}$, and the pressure (in both cylinders) is $p_{C_0} = p_{H_0} = 10^5$ Pa or 1 atmosphere.

From the perfect gas law, $pV = \mu RT$, we calculate the amount of gas in the "cold" cylinder as 40.1×10^{-6} kilomoles. The amount of gas in the connecting pipe and in the "hot" cylinder (V_{H_0}) is assumed to be negligible.

Phase 0→1 (Isothermal compression)

The "cold" piston is moved partially up so that the gas volume is now $V_{C_1} = 10^{-4}$ m^3 (a compression ratio, $r = 10$). Since the cylinder is in contact with the heat sink, the heat generated by the compression is removed and the temperature remains unchanged. In other words, the compression is **isothermal**. The energy required is

$$W_{0\to 1} \equiv W_{compress} = \mu R T_C \ln \frac{V_{C_0}}{V_{C_1}} = 230\,\text{J}. \tag{3.20}$$

The temperature did not change while the pressure increased 10-fold. State 1 of the gas is

$$V_{C_1} = 10^{-4}\,\text{m}^3,$$

$$T_C = 300\,\text{K},$$

$$p_{C_1} = 10^6\,\text{Pa}.$$

Phase 1→2 (Gas transfer, followed by isometric heat addition)
The "cold" piston goes all the way up, and the "hot" piston goes partially down so that $V_{H_2} = 10^{-4}$, $V_{C_2} = 0$. The total volume of the gas does not change. Theoretically, there is no energy cost to this gas transfer, but the gas is now in contact with the hot source and will start heating up. Assume, arbitrarily, that the temperature rises to 652 K. For this to happen, the heat source must deliver to the gas an amount of heat, $Q_{1\to 2}$.

A gas whose $\gamma = 1.4$ has a c_v of 20.8 kJ K^{-1} kmole^{-1}. Hence, the heat necessary to raise the temperature from 300 to 652 K, while keeping the volume unchanged (**isometric** heat addition), is

$$\begin{aligned} Q_{1\to 2} \equiv Q_{add} &= \mu c_v \Delta T = \mu c_v (T_H - T_C) \\ &= 40 \times 10^{-6} \times 20.8 \times 10^3 (652 - 300) = 293\,\text{J}. \end{aligned} \tag{3.21}$$

Since the gas temperature went up without a change in volume, the pressure must have increased. State 2 of the gas is

$$\begin{aligned} V_{H_2} &= 10^{-4}\,\text{m}^3, \\ T_H &= 652\,\text{K}, \\ p_{H_2} &= \frac{652}{300} \times 10^6 = 2.17 \times 10^6 \text{ Pa}. \end{aligned}$$

Phase 2→3 (Isothermal expansion)
The high pressure pushes the "hot" piston down until the volume in the cylinder reaches 10^{-3} m^3. The corresponding 10:1 expansion would cool the gas down, but heat from the external source keeps the temperature constant—we have an **isothermal** expansion that delivers 500 J to the crankshaft:

$$\begin{aligned} W_{2\to 3} \equiv W_{expan} &= \mu R T_H \ln \frac{V_{H_3}}{V_{H_2}} \\ &= 40.1 \times 10^{-6} \times 8314 \times 652 \ln 10 = 500\,\text{J}. \end{aligned} \tag{3.22}$$

We have arbitrarily chosen a T_H of 652 K in the preceding phase, so that the energy delivered to the crankshaft comes out a round number. The heat input required is, of course, $Q_{2\to 3} \equiv Q_{expan} = 500$ J.

State 3 of the gas is

$$\begin{aligned} V_{H_3} &= 10^{-3}\,\text{m}^3, \\ T_H &= 652\,\text{K}, \\ p_{H_3} &= 2.17 \times 10^5\,\text{Pa}. \end{aligned}$$

Phase 3→0 (Isometric heat rejection)
Finally, the pistons return to their initial position. The gas volume does not change but, owing to its transfer to the "cold" cylinder, it cools

isometrically to 300 K and thus returns to State 0. This completes the cycle. The heat removed during this phase is

$$Q_{3\to 0} = \mu c_v \Delta T = 40 \times 10^{-6} \times 20.8 \times 10^3 (652 - 300) = 293\,\text{J}. \tag{3.23}$$

exactly the same as $Q_{1\to 2}$.

In one cycle, the crankshaft receives 500 J from the "hot" piston ($W_{2\to 3}$) and returns 230 J used in the compression phase ($W_{0\to 1}$). A net mechanical energy of $500 - 230 = 270$ J constitutes the output of the machine. This happens at a cost of two heat inputs, $Q_{1\to 2}$ and $Q_{2\to 3}$, amounting to 793 J.

The efficiency of the device is

$$\eta = \frac{W_{expan} - W_{compress}}{Q_{add} + Q_{expan}} = \frac{500 - 230}{293 + 500} = \frac{270}{793} = 0.34.$$

The efficiency of a Carnot cycle working between 652 and 300 K is

$$\eta_{CARNOT} = \frac{652 - 300}{652} = 0.54. \tag{3.24}$$

Thus, the Stirling engine, as described, realizes only a modest fraction of the Carnot efficiency. However, a relatively simple modification changes this picture.

We observe that $Q_{3\to 0} = Q_{1\to 2}$. The heat that has to be removed from the gas in the $3 \to 0$ phase can be stored in a **regenerator** inserted in the pipe connecting the two cylinders, and it can be used to supply $Q_{1\to 2}$. This means that the only heat required from the heat source is $Q_{2\to 3}$ (500 J). Using a perfect, ideal regenerator, the efficiency of the Stirling cycle is

$$\eta = \frac{270}{500} = 0.54, \tag{3.25}$$

which is exactly the Carnot efficiency.

In practice, regenerators can be realized by using, for example, steel wool whose large surface-to-volume ratio guarantees a speedy heat exchange. Desirable characteristics of regenerators include:

- high heat capacity,
- low longitudinal heat conductance,
- low viscous losses,
- low volume.

3.8.1.2 The Beta Stirling Engine

In a beta-configured Stirling engine, a single cylinder is used. (See the schematics in Figure 3.13.) The lower piston is called the **power** piston and fits tightly in the cylinder so that gas can be compressed. The upper

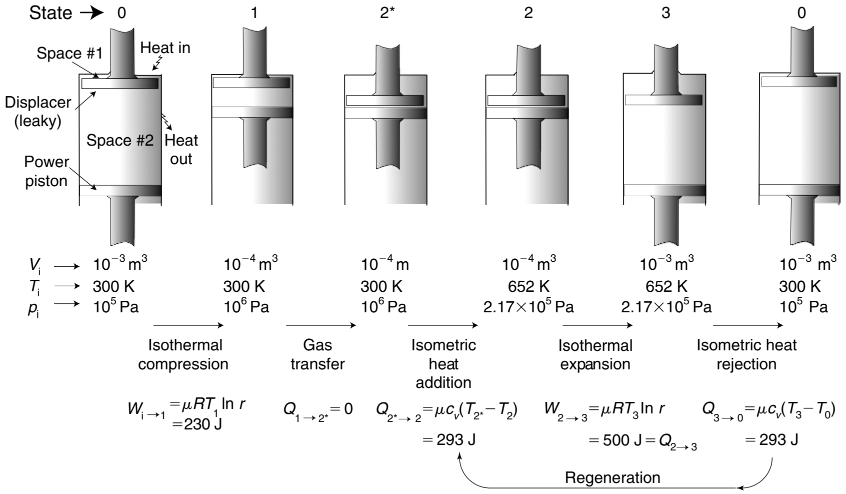


Figure 3.13 The various states of a beta Stirling cycle.

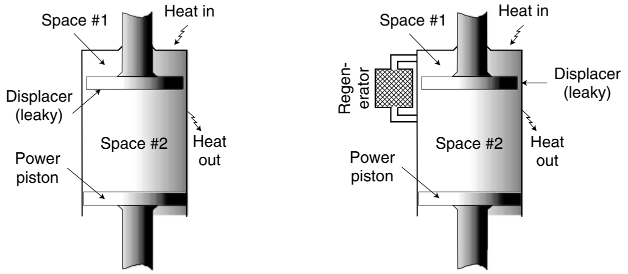


Figure 3.14 A regenerator can be fitted to a beta Stirling engine in the manner indicated above.

piston is the **displacer** and fits loosely so that it is quite leaky. The function of this displacer is to move the gas from the "cold" space (Space 2) just above the power piston to the "hot" space above the displacer (Space 1).

The phases of this cycle are the same as those of the alpha cycle so we can use the same quantitative example. In Figure 3.14, we divide the $1 \rightarrow 2$ phase into two subphases: $1 \rightarrow 2^*$ in which the displacer comes down and moves the cold gas of Space 2 into the hot region of Space 1. Again, ideally, no energy is consumed in this subphase because the gas leaks freely through the gap between displacer and cylinder. Isometric heat addition is performed in Subphase $2^* \rightarrow 2$.

The problem with this configuration is that there is no provision for the important regeneration function. This can be remedied by using an external path connecting Spaces 1 and 2, as indicated in Figure 3.14

3.8.1.3 The Implementation of the Kinematic Stirling Engine

One possible implementation of the Stirling engine is that shown in Figure 3.15. Air from a *combustion blower* is fed to a *burner* where it is mixed with fuel and ignited. The resulting hot gases are passed through an *input heat exchanger* that transfers heat to the working fluid in “Space #1.” Gases that exit the input side of this heat exchanger still contain a considerable amount of heat that can be recovered by preheating the air coming from the combustion blower, lowerings the amount of fuel required to maintain the working fluid of the engine at the required temperature.

The cooling of the working fluid in “Space #2” is accomplished through a separate *cooling heat exchanger.*

As in all types of engines, a number of factors cause the specific fuel consumption of a practical Stirling engine to be much higher than the theoretical predictions and its pollutant emissions to be higher than desirable.

Further development would be necessary if this category of engines were to become popular in the automotive world. This will probably not happen because it appears that combustion engines are approaching the end of their popularity. In all probability, fuel cells will gradually take over most areas where IC engines are now dominant. Owing to the enormous size of the automotive industry, the transition from IC to FC technology must, necessarily, proceed gradually. An important intermediate step is the use of hybrid cars.

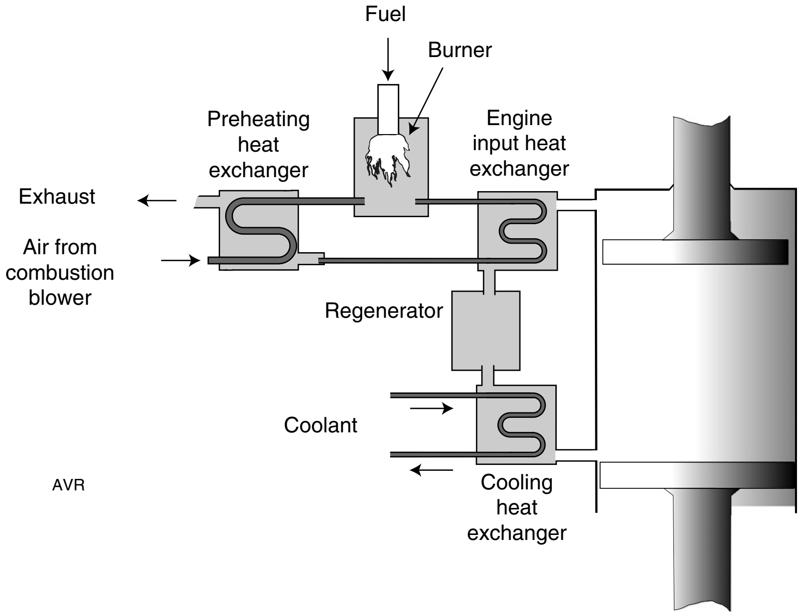

Fuel
Burner
Preheating heat exchanger
Engine input heat exchanger
Exhaust
Air from combustion blower
Regenerator
Coolant
Cooling heat exchanger
AVR

Figure 3.15 One possible implementation of a Stirling engine.

Real efficiency departs from from the theoretical because:

1. The efficiency of mechanical heat engines increases with increasing compression ratio. The Stirling engine is no exception (see Problem 3.1). The internal volume of the heat exchangers and the regenerator causes the compression ratio to be smaller than if heat could be applied directly to "Space #1."
2. The combustion gases leaving the input heat exchanger are still quite hot and carry away considerable energy. To recover part of the latter a preheater is used as explained previously.
3. The regenerator cannot operate ideally and return all the heat deposited in it during part of the cycle.
4. There are frictional losses when the working fluid is transferred from one space to the other through the heat regenerator. This is the main reason for using hydrogen as a working fluid, as it leads to minimum frictional losses.
5. Ideally, the motion of the pistons should be intermittent, but this is difficult to implement in machines operating at reasonably high rpm. One has to make a compromise in the piston programming.
6. The limited time for the heat exchanges in the working fluid leads to temperatures that never reach the desired steady state levels.

The power output of a Stirling engine can be controlled by either adding or removing working fluid. To reduce power, a compressor removes some working fluid from the engine and stores it at high pressure in a holding tank. To increase power, gas from this tank can be quickly delivered back to the engine. The input heat exchanger temperature is continuously monitored, and this information is used to control the fuel flow.

3.8.2 The Free-piston Stirling Engine

As in the case of kinematic Stirling engines, there are numerous implementations of the free-piston version. Much of the development of this type of engine is owed to Sunpower Inc. of Athens, Ohio. Figure 3.16 shows a simple form of the engine dating from 1986.

There is no rigid mechanical connection between piston and displacer. Both fit snugly in the cylinder with no intentional leakage. The motion of the displacer must lead that of the piston by more than 0° and less than 360°. The power output of the machine can be adjusted by changing this angle, which depends on the inertia of the moving parts and the rigidity of the gas coupling between them. In the design shown, the displacer moves against a **gas spring**, while the return motion of the piston is insured by the gas in the **bounce space**. In more modern designs, the displacer and piston are suspended by **flexure bearings** that also act as return

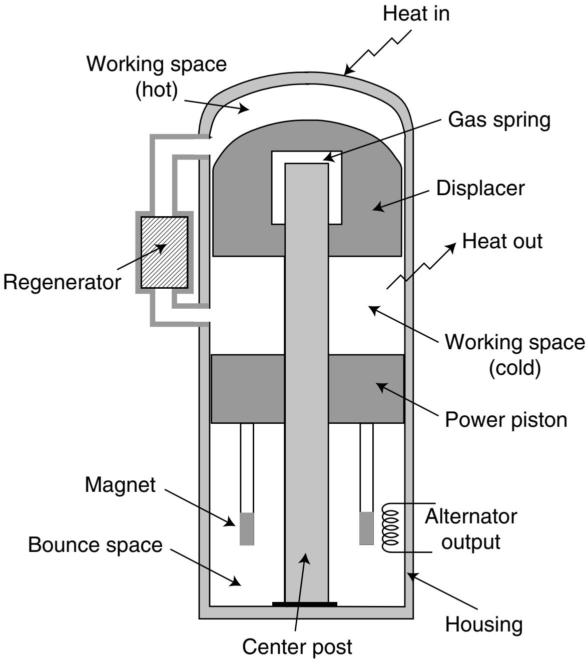


Figure 3.16 A simple free-piston Stirling engine with a linear alternator as load. The operation is somewhat similar to that of a kinematic beta Stirling.

springs. Flexure bearings are diaphragms that are quite rigid radially but allow substantial axial motion of the oscillating parts of the machine. Such bearings are of very predictable performance, are practically lossless, and have extremely long lives. Noncontact gas bearings are also used.

If all Stirling engines indeed, have, high efficiency, operate with many different heat sources, pollute substantially less than corresponding internal combustion engines, and are quite noise free, how come they have not dominated the world of heat engines? The answer is that the kinematic version of the Stirling whose development has consumed a large amount of time, effort, and funds, has some serious, nearly insurmountable problems, most of which are absent in the free-piston version developed much more recently. One area of difficulty is the need for lubrication and, consequently, the need for seals to separate the lubricated parts from the high-pressure working fluids part. This is the main reason for the limited life of the kinematic engine. Free-piston engines need no lubrication.

Kinematic engines, unlike free-piston ones, require piston rings that degrade over time. They also require costly and complicated mechanisms to permit power variations. See the subsection on implementation of kinematic Stirling engines.

Free-piston systems designed to generate electricity incorporate a (linear) alternator as an integral part of the hermetically sealed machinery, avoiding leakage problems. The frequency of the generated a.c. is determined by the mechanical resonance of the oscillating displacer/piston mass and can easily be adjusted to 50 or 60 Hz (100 Hz in some space applications).

Free-piston engines have only two moving parts, both of which are entirely inside the sealed compartment with no mechanical connection with the exterior and are, therefore, mechanically much simpler and more reliable than kinematic engines. The moving parts make no solid contact with other components, being separated from them by noncontact gas bearings. There is no life-limiting component wear. Free-piston engines are extremely easy to start and can be made essentially vibration free by operating them in opposed pairs with a common expansion space. Power modulation is achieved rapidly and in a simple manner by varying the phase angle between the displacer and piston motions. This is accomplished by varying the stiffness of the gas spring interconnecting these two components. Fast response and high efficiency are insured over a great power range. In contrast, in kinematic engines, the phase angle is determined by mechanical linkages and is, therefore, fixed.

For a more complete discussion of the advantages of the free-piston over the kinematic engine, read Lane and Beale (1997).

The great reliability, flexibility, and long life of the free-piston Stirling engines achieved in recent years are opening great possibilities for this technology. Among other applications, it not only might replace the RTG in deep space missions, but may also become the power supply of choice for the planned moon base and for many military applications as well.

3.9 Cryogenic Engines

As pointed out earlier, a heat engine depends for its operation on a heat source and a sink, which, by definition, must be at a lower temperature. Almost invariably, the sinks are at nearly the ambient temperature while the source is driven by combustion, nuclear reaction, solar energy, and so forth, to a relatively elevated temperature. However, it is possible to use a source that is at ambient temperature, while the temperature of the sink is lowered by a supply of a cryogenic substance.

Ordonez (1997) and his collaborators at the University of North Texas have proposed such cryogenic engines. Their experimental engines operate on an open Rankine cycle and use liquid nitrogen for the sink. Figure 3.17 shows a schematic representation of the setup.

The liquid nitrogen in the cryogenic reservoir is under the (adjustable) pressure, p_1. The liquid is fed to a heat exchanger to which enough heat is added (from the outside air) to produce gaseous N_2 at the pressure, p_1, and the temperature, T_1.

The exhaust valve is closed, and the intake valve is opened so that μ kilomoles of nitrogen enter the cylinder. This causes the piston to be pushed down until the volume reaches a value, v_2. The process is isobaric

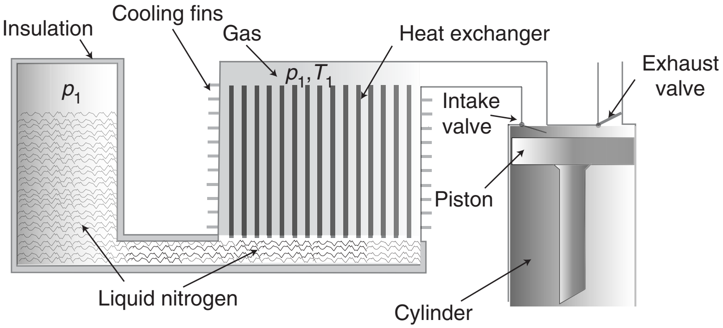


Figure 3.17 A cryogenic heat engine using an open Rankine cycle.

($p_2 = p_1$) and isothermic ($T_2 = T_1$) because sufficient heat is added to the heat exchanger.

We have

$$v_2 = \mu \frac{RT_2}{p_2} = \mu \frac{RT_1}{p_1}. \tag{3.26}$$

The work produced by the piston is

$$W_{12} = p_1 v_2 = \mu R T_1. \tag{3.27}$$

Next, the intake valve is closed. The high-pressure nitrogen continues to expand until the pressure inside the cylinder is p_3 and the volume reaches v_3. This process may be carried out isothermally ($T_3 = T_1$) or adiabatically ($T_3 < T_1$), depending on the type of hardware used. Consider the more favorable isothermal case:

$$W_{23} = \mu R T_1 \ln \frac{p_3}{p_1}. \tag{3.28}$$

The total work done by the piston during the 1-to-3 phase is

$$W_{13} = \mu R T_1 + \mu R T_1 \ln \frac{p_3}{p_1}. \tag{3.29}$$

However, this is not the available work. Part of the total work is done by the piston on the atmosphere:

$$W_{atm} = p_{atm} v_3, \tag{3.30}$$

where

$$v_3 = \mu \frac{RT_1}{p_3}. \tag{3.31}$$

Hence, assuming $T_1 = T_{atm}$,

$$W_{atm} = \mu R T_1. \tag{3.32}$$

The net useful work done is

$$W_{net} = W_{13} - W_{atm} = \mu R T_1 \ln \frac{p_3}{p_1}. \tag{3.33}$$

This means that the net work is exactly that owing to the isothermal expansion of the gas.

Example

Choose $p_1 = 1$ MPa and assume that $T_1 = T_{atm} = 298\,\mathrm{K}$ and $p_3 = p_{atm} = 0.1$ MPa. The expansion ratio, $r = p_1/p_3$, is 10:1. The net work per kilomole of "fuel" is 5.7 MJ/kmole of N_2 or 204 kJ/kg of N_2.

It may be difficult to achieve isothermal expansion; it would require an additional heat exchanger to warm up the nitrogen as it forces the piston down. However, something similar is required in a Stirling engine, so it is possible.

Let us examine the case of adiabatic expansion that is easier to achieve. If no heat is added during the expansion, the temperature will fall:

$$T_3 = T_1 \left(\frac{p_1}{p_3}\right)^{\frac{1}{\gamma}-1}. \tag{3.34}$$

Here, $\gamma = 1.4$ for nitrogen. The work during the expansion is

$$W_{23} = \mu c_v (T_1 - T_3), \tag{3.35}$$

where $c_v \approx \frac{R}{\gamma - 1} = 20.8\,\mathrm{kJ\ K^{-1}\,kmole^{-1}}$.

For the example with $p_1 = 1$ MPa, we get $T_3 = 154\,\mathrm{K}$ and $W_{23} = 3\,\mathrm{MJ/kmole}$.

When the volume, v_3, is reached, the gas will have cooled to T_3 so that

$$v_3 = \mu \frac{R T_3}{p_3}. \tag{3.36}$$

Consequently, the work done on the atmosphere is

$$W_{atm} = p_{atm} v_3 = \mu R T_3. \tag{3.37}$$

The net work is now

$$\begin{aligned} W_{net} &= \mu R T_1 - \mu R T_3 + W_{23} = \mu R (T_1 - T_3) + \mu c_v (T_1 - T_3) \\ &= \mu (T_1 - T_3) R \frac{\gamma}{\gamma - 1}. \end{aligned} \tag{3.38}$$

Still for the example being considered, the specific net energy is 4.2 MJ/kmole or 150 kJ/kg. Compare with the 5.7 MJ/kmole or 204 kJ/kg for the isothermic case and with the 47,000 kJ/kg for the typical gasoline.

Clearly, the specific energy of the cryogen will increase with increasing operating pressure. The gain, however, is logarithmic. Thus, by raising the pressure to 10 MPa (a factor of 10), the specific energy rises to 11.4 MJ/kmole, a gain of 2. Observe that 10 MPa correspond to approximately 100 atmospheres and would lead to a rather heavy (and expensive) engine.

Gasoline has to be used in an internal combustion engine with some 20% efficiency, while the pneumatic motor used in the cryogenic engine can have a very high component efficiency. This would reduce the practical specific energy advantage of gasoline to a factor of 40 over the nitrogen.

In practice, the Ordonez engine has yielded around 19 kJ/kg thus far. A demonstration car using the engine does 0.3 mile to the gallon, which is not practical. The efficiency will probably be improved.

3.9.1 Conclusions

Vehicles equipped with the cryogenic engine achieve only very modest mileage. Demonstrated mileage is 0.3 mile/gallon. The present-day cost of liquid nitrogen (2008) is \$2/gallon, when delivered in Dewars, or \$0.5/gallon when delivered in large quantities by truck. With gasoline at \$4/gallon, liquid nitrogen costs one order of magnitude less than gasoline. But since the mileage is about two orders of magnitude worse than that of an IC car, the cryogenic car, in its present configuration, has a fuel cost some 10 times that of the normal gasoline-driven car. It would take a 10-fold improvement in performance for the cryogenic car for the fuel cost to be the same for the two types of vehicle. Add to this the need to carry huge amounts of cryogenic material, which would constitute a considerable hazard in case of accident.

The advantage of the system is its low pollution (but not necessarily zero pollution because energy is needed to produce the liquid nitrogen). The question is whether these advantages compensate for the serious disadvantages that we discussed.

References

Campbell, John M., Frank K. Signaigo, Wheeler G. Lowell, and T. A. Boyd, Antiknock effect of tetraethyllead (Effectiveness of the tetraethyllead in increasing the critical compression ratio of individual hydrocarbons), *Ind. Eng. Chem.*, **27** (5), 593–597, May **1935**.

Heywood, John B., *Internal Combustion Engine Fundamentals*, McGraw-Hill, **1988**.

Lane, Neill W. and William T. Beale, Free-piston Stirling design features, Eighth International Stirling Engine Conference, University of Ancona, Italy, May 27–30, **1997**.

Ordonez, C. A., Cryogenic heat engine, *Am. J. Phys.* **64** (4), April **1996**.

Ordonez, C. A., and M. C. Plummer, Cold thermal storage and cryogenic heat engine for energy storage applications, *Energy Sources*, 19:389–396, **1997**.

Plummer, M. C., C. P. Koehler, D. R. Flanders, R. F. Reidy, and C. A. Ordonez, Cryogenic heat engine experiment, *Advances in Cryogenic Engineering* **43**, pp. 1245–1252, **1998**.

Wilson, D. G., 1978, Alternative Automobile Engines, *Scientific American,* **239** (1): 39–49.

A large number of publications on Stirling engines can be found at <http://www.sunpower.com/index.php?pg=25>

PROBLEMS

3.1

1. Demonstrate that the theoretical efficiency of a Stirling engine without regenerator is

$$\eta = \eta_{CARNOT}\left(1 + \frac{\nu \eta_{CARNOT}}{2 \ln r}\right)^{-1}$$

where η_{CARNOT} is the Carnot efficiency associated with the engine temperature differential, ν is the number of degrees of freedom of the working gas, and r is the compression ratio.

2. What gas would you suggest as a working fluid? Why?
3. In the example in the text, a compression ratio of 10 was used. What would the efficiency of that engine be if the ratio were raised to 20? What are the disadvantages of using this higher compression ratio? Is it worth the effort?

3.2 Plot a pressure versus volume diagram and a temperature versus entropy diagram for the Stirling engine in the example given in the text. What do the areas under the pV and the TS lines represent?

3.3 Consider two cylinders, A and B, equipped with pistons so that their internal volume can be changed independently. The maximum volume of either cylinder is $10\,\text{m}^3$, and the minimum is zero. The cylinders are interconnected so that the gas is at the same pressure anywhere in the system. Initially, the volume of A is $10\,\text{m}^3$ and that of B is 0. In other words, piston A is all the way up and B is all the way down. The system contains a gas with $\gamma = 1.4$.

1. If this is a perfect gas, what is the number of degrees of freedom of its molecules, and what is its specific heat at constant volume?
2. The pressure is 0.1 MPa, and the temperature is 400 K. How many kilomoles of gas are in the system?
3. Now, push piston A down reducing the volume to $1\,\text{m}^3$, but do not change the volume of cylinder B. What are the temperature and pressure of the gas assuming adiabatic conditions? What energy was expended in the compression?
4. Next, press A all the way down and, simultaneously, let B go up so that the volume in cylinder A is zero and that in cylinder B is $1\,\text{m}^3$. What are the pressure and temperature of the gas in B?
5. The next step is to add heat to the gas in B so that it expands to $10\,\text{m}^3$ at constant temperature. How much heat was added? How much work did Piston B deliver? What is the final pressure of the gas?

6. Now press B all the way down while pulling A up. This transfers gas from one cylinder to the other and (theoretically) requires no energy. Cylinder A rejects heat to the environment, and the gas cools down to 400 K. The pistons are not allowed to move. The cycle is now complete. How much heat was rejected?
7. What is the efficiency of this machine—that is, what is the ratio of the net work produced to the heat taken in?
8. What is the corresponding Carnot efficiency?
9. Sketch a pressure versus volume and a temperature versus volume diagram for the process described.
10. Derive a formula for the efficiency as a function of the compression ratio, r. Plot a curve of efficiency versus r in the range $1 < r \leq 100$.
11. If this ratio were to reach the (unrealistic) value of 10,000, what would the efficiency be? Does this exceed the Carnot efficiency? Explain.

3.4 A car is equipped with a spark ignition engine (Otto cycle). It uses gasoline (assume gasoline is pure pentane) as fuel, and, for this reason, its compression ratio is limited to 9. The highway mileage is 40 miles/gallon.

Since pure ethanol is available, the car owner had the engine modified to a compression ratio of 12 and is using this alcohol as fuel. Assuming that in either case the actual efficiency of the car itself is half of the theoretical efficiency, what is the mileage of the alcohol car?

The lower heats of combustion and the densities are

Pentane: 28.16 MJ/liter, 0.626 kg/liter, and

Ethanol: 21.15 MJ/liter, 0.789 kg/liter.

Do this problem twice, once using $\gamma = 1.67$ and once using $\gamma = 1.4$.

3.5 Consider a cylinder with a frictionless piston. At the beginning of the experiment, it contains one liter of a gas ($\gamma = 1.4$, $c_v = 20\,\text{kJ K}^{-1}$ kmole^{-1}) at 400 K and 10^5 Pa.

1. How many kmoles of gas are in the cylinder?
2. What is the pV product of the gas?

Move the piston inward reducing the volume of the gas to 0.1 liter. This compression is adiabatic.

3. What is the pressure of the gas after the above compression?
4. What is the temperature of the gas after the above compression?
5. How much work was done to compress the gas?

Now add 500 J of heat to the gas without changing the temperature.

6. After this heat addition phase, what is the volume of the gas?
7. After this heat addition phase, what is the pressure?

8. Since the gas expanded (the piston moved) during the heat addition, how much work was done?

Let the gas expand adiabatically until the volume returns to 1 liter.

9. After the expansion, what is the pressure of the gas?
10. After the adiabatic expansion, what is the temperature of the gas?
11. How much work was done by the expansion?

Now remove heat isometrically from the gas until the pressure reaches 10^5 Pa. This will bring the system to its original State 1.

12. What is the net work done by the piston on an outside load?
13. What is the total heat input to the system (the rejected heat cannot be counted)?
14. What is the efficiency of this machine?
15. What is the Carnot efficiency of this machine?
16. Sketch a pressure versus volume diagram of the cycle described in this problem statement.

3.6 Assume that gasoline is pure pentane (actually, it is a complicated mixture of hydrocarbons best represented by heptane, not pentane). Consider a 1:4 ethanol-gasoline mixture (by volume). The gasoline has an 86 octane rating. The blending octane rating of ethanol is 160. Use $\gamma = 1.4$.

1. What is the energy per liter of the mixture compared with that of the pure gasoline?
2. What is the octane rating of the mixture?

Assume that the maximum tolerable compression ratio is $r = 0.093 \times O_r$ where O_r is the octane rating.

3. What is the highest compression ratio of the gasoline motor? of the mixture motor?
4. What is the relative efficiency of the two motors?
5. What is the relative kilometrage (or mileage) of two identical cars equipped one with the mixture motor and the other with the gasoline motor?

3.7 An open-cycle piston type engine operates by admitting 23×10^{-6} kmoles of air at 300 K and 10^5 Pa. It has a 5.74 compression ratio.

Compression and expansion are adiabatic. Heat is added isobarically and rejected isometrically.

Heat addition is of 500 J.

Air has $c_v = 20,790\,\mathrm{J\,K^{-1}\,kmole^{-1}}$ and a $\gamma = 1.4$.

What is the theoretical efficiency of this engine? Compare with its Carnot efficiency.

Proceed as follows:

Calculate the initial volume of the cylinder (at end of admission).

Compress adiabatically and calculate the new V, p, T, and the work required.

Add heat and calculate new state variables.

Expand and calculate the work done.

3.8 A certain Stirling engine realizes one-half of its theoretical efficiency and operates between 1000 K and 400 K. What is its efficiency with

1. a perfect heat regenerator, argon working fluid, and 10:1 compression?
2. the same as above but with a 20:1 compression ratio?
3. the same as in (1) but without the heat regenerator?
4. the same as in (2) but without the heat regenerator?

3.9 Rich mixtures reduce the efficiency of Otto engines, but mixtures that are too lean do not ignite reliably. The solution is the "stratified combustion engine."

Consider an engine with a 9:1 compression ratio. A rich mixture may have a gamma of 1.2, while in a lean one it may be 1.6. Everything else being the same, what is the ratio of the efficiency with the lean mixture to that with the rich mixture?

3.10 Consider an Otto (spark-ignition) engine with the following specifications:

Maximum cylinder volume, V_0: 1 liter ($10^{-3}\,\mathrm{m}^3$).

Compression ratio, r: 9:1.

Pressure at the end of admission, p_0: 5×10^4 Pa.

Mixture temperature at the end of admission, T_0: 400 K.

Average ratio, γ, of specific heats of the mixture: 1.4.

Specific heat of the mixture (constant volume), c_v,: 20 kJ K^{-1} kmole^{-1}.

The calculations are to be based on the ideal cycle—no component losses. At maximum compression, the mixture is ignited and delivers 461 J to the gas.

If the engine operates at 5000 rpm, what is the power it delivers to the load?

3.11 If pentane burns stoichiometrically in air (say, 20% O_2 by volume), what is the air-to-fuel mass ratio?

Atomic masses:

H:	1 dalton.
C:	12 daltons.
N:	14 daltons.
O:	16 daltons.

Neglect argon.

3.12 The higher heat of combustion of n-heptane (at 1 atmosphere, 20 C) is 48.11 MJ/kg. What is its lower heat of combustion?

3.13 One mole of a certain gas ($\gamma = 1.6,\ c_v = 13,860\ \mathrm{JK^{-1}kmole^{-1}}$) occupies 1 liter at 300K. For each of the following steps, calculate the different variables of state, $p,\ V$ and T.

Step 1 → 2

Compress this gas adiabatically to a volume of 0.1 liter.

How much of the energy, $W_{1,2}$, is used in this compression?

Step 2 → 3

Add isothermally 10 kJ of heat to the gas.

What is the external work, if any?

Step 3 → 4

Expand this gas adiabatically by a ratio of 10:1 (the same as the compression ratio of Step 1 → 2).

Step 4 → 1

Reject heat isothermally so as to return to State 1. What is the energy involved?

What is the overall efficiency of the above cycle?

What is the Carnot efficiency of the cycle?

What power would an engine based on the above cycle deliver if it operated at 5000 rpm (5000 cycles per minute)?

3.14 The Stirling engine discussed in the example in the text uses an isothermal compression, followed by an isometric heat addition, then by an isothermal expansion, and, finally, by an isometric heat rejection.

Isothermal compression may be difficult to achieve in an engine operating at high rpm. Imagine that the engine actually operates with an adiabatic compression.

Assume that the other steps of the cycle are the same as before. Thus, the isometric heat addition still consists of absorption of 293 joules of heat. This means that the "hot" cylinder is no longer at 652 K; it is at whatever temperature will result from adding isometrically 293 J to the gas after the initial adiabatic compression.

Calculate the theoretical efficiency of the engine (no heat regeneration) and compare it with the Carnot efficiency.

Calculate the power produced by this single cylinder engine assuming that the real efficiency is exactly half of the ideal one and that the engine operates at 1800 rpm. Each full rotation of the output shaft corresponds to one full cycle of the engine. Use $\gamma = 1.4$.

3.15 Assume that an engine (Engine #1) that works between 1000 K and 500 K has an efficiency equal to the Carnot efficiency.

The heat source available delivers 100 kW at 1500 K to the above engine. A block of material is interposed between the source and the engine to lower the temperature from 1500 K to the 1000 K required. This block of material is 100% efficient: the 100 kW that enter on one side are all delivered to the engine at the other side.

What is the Carnot efficiency of the system above? What is the power output of the system?

Now, replace the block by a second engine (Engine #2) having a 10% efficiency. The heat source still delivers 100 kW. What is the overall efficiency of the two engines working together?

3.16 A boiler for a steam engine operates with the inside of its wall (the one in contact with the steam) at a temperature of 500 K, while the outside (in contact with the flame) is at 1000 K.

1 kW of heat flows through the wall for each cm^2 of wall surface. The metal of the wall has a temperature-dependent heat conductivity, λ, given by $\lambda = 355 - 0.111\,T$ in MKS units. T is in kelvins.

1. Determine the thickness of the wall.
2. Determine the temperature midway between the inner and outer surface of the wall.

3.17 A 4-cycle Otto (spark ignition) engine with a total displacement (maximum cylinder volume) of 2 liters is fueled by methane (the higher heat of combustion is 55.6 MJ/kg; however, in an IC engine, what counts is the lower heat of combustion). The compression ratio is 10:1. A fuel injection system insures that, under all operating conditions, the fuel–air mixture is stoichiometrically correct. The gamma of this mixture is 1.4.

Owing to the usual losses, the power delivered to the load is only 30% of the power output of the ideal cycle. Because of the substantial intake pumping losses, the pressure of the mixture at the beginning of the compression stroke is only 5×10^4 Pa. The temperature is 350 K.

If the engine operates at 5000 rpm, what is the power delivered to the external load?

3.18 Consider a spark-ignition engine with a 9:1 compression ratio. The gas inside the cylinder has a $\gamma = 1.5$.

At the beginning of the compression stroke, the conditions are

$$\begin{aligned} V_1 &= 1\,\text{liter}, \\ p_1 &= 1\,\text{atmosphere, and} \\ T_1 &= 300\,\text{kelvins.} \end{aligned}$$

At the end of the compression, 10 mg of gasoline are injected and ignited. Combustion is complete and essentially instantaneous.

Take gasoline as having a heat of combustion of 45 MJ/kg.

1. Calculate the ideal efficiency of this engine.
2. Calculate the Carnot efficiency of an engine working between the same temperatures as the spark-ignition engine above.

3. Prove that as the amount of fuel injected per cycle decreases, the efficiency of the Otto cycle approaches the Carnot efficiency.

3.19 In a Diesel engine, the ignition is the result of the high temperature the air reaches after compression (it is a **compression ignition** engine). At a precisely controlled moment, fuel is sprayed into the hot compressed air inside the cylinder, and ignition takes place. Fuel is sprayed in relatively slowly so that the combustion takes place, roughly, at constant pressure. The compression ratio, r, used in most Diesel engines is between 16:1 and 22:1. For Diesel fuel to ignite reliably, the air must be at 800 K or more.

Consider air as having a ratio of specific heat at constant pressure to specific heat at constant volume of 1.4 ($\gamma = 1.4$). The intake air in a cold Diesel engine may be at, say, 300 K. What is the minimum compression ratio required to start the engine?

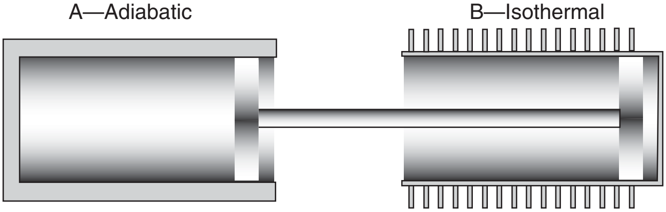
3.20 We have a machine that causes air (γ=1.4) to undergo a series of processes. At the end of each process, calculate the state of the gas (pressure, volume, and temperature) and the energy involved in the process.

The initial state (State #1) is

$$p_1 = 10^5\,\text{Pa},$$
$$V_1 = 10^{-3}\text{m}^3, \text{ and}$$
$$T_1 = 300\,\text{K}.$$

1. 1st Proc. (Step 1→2): Compress adiabatically, reducing the volume to 10^{-4} m^3.
2. 2nd Proc. (Step 2→3): Add 200 J of heat isobarically.
3. 3rd Proc. (Step 3→4): Expand adiabatically until $V_4 = 10^{-3}$ m^3.
4. List all the heat and mechanical inputs to the machine and all the mechanical outputs. From this, calculate the efficiency of the machine. (*Hint*: Don't forget to add all the processes that deliver energy to the output.)

3.21



A crazy inventor patented the following (totally useless) device: Two geometrically identical cylinders (one adiabatic and the other isothermal) have rigidly interconnected pistons as shown in the figure.

The system is completely frictionless, and at the start of the experiment (State #0), the pistons are held in place so that the gases in the cylinders are in the states described in the following:

Cylinder A	Cylinder B
(Adiabatic)	(Isothermal)
$V_{A_0} = 1$ m^3	$V_{B_0} = 0.1$ m^3
$p_{A_0} = 10^5$ Pa	$p_{B_0} = 10^6$ Pa
$T_{A_0} = 300$ K	$T_{B_0} = 300$ K

1. Now, the pistons are free to move. At equilibrium, what is the temperature of the gas in Cylinder A? The γ of the gas is 1.5.

   An external device causes the pistons to oscillate back and forth 2500 times per minute. Each oscillation causes V_B to go from $0.1\,\text{m}^3$ to $1\,\text{m}^3$ and back to 0.1 m^3.

2. How much power is necessary to sustain these oscillations?

   Consider the same oscillating system as above with the difference that in each compression and each expansion 1% of the energy is lost. This does not alter the temperature of the isothermal cylinder because it is assumed that it has perfect thermal contact with the environment at 300 K. It would heat up the gas in the adiabatic cylinder that has no means of shedding heat. However, to simplify the problem, assume that a miraculous system allows this loss-associated heat to be removed but not the heat of compression (the heat that is developed by the adiabatic processes).

3. How much power is needed to operate the system?

3.22 In a Diesel cycle one can distinguish the following different phases:

Phase 1 → 2 An adiabatic compression of pure air from Volume V_1 to Volume V_2.

Phase 2 → 3 Fuel combustion at constant pressure with an expansion from Volume V_2 to Volume V_3.

Phase 3 → 4 Adiabatic expansion from Volume V_3 to Volume V_4.

Phase 4 → 1 Isometric heat rejection causing the state of the gas to return to the initial conditions.

This cycle closely resembles the Otto cycle, with the difference that in the Otto cycle the combustion is isometric while in the Diesel it is isobaric.

Consider a cycle in which $V_1 = 10^{-3}$ m^3, $V_2 = 50 \times 10^{-6}$ m^3, $V_3 = 100 \times 10^{-6}$ m^3, $p_1 = 10^5$ Pa, $T_1 = 300$ K, and (for all phases) $\gamma = 1.4$.

1. Calculate the theoretical efficiency of the cycle by using the efficiency expression for the Diesel cycle given in Chapter 3 of the text.

2. Calculate the efficiency by evaluating all the mechanical energy (compression and expansion) and all the heat inputs. Be specially careful with what happens during the combustion phase $(2 \rightarrow 3)$ when heat from the fuel is being used and, simultaneously, some mechanical energy is being produced.

   You should, of course, get the same result from 2 and 3.

Chapter 4
Ocean Thermal Energy Converters

4.1 Introduction

The most plentiful renewable energy source in our planet by far is solar radiation: 170,000 TW fall on Earth. Harvesting this energy is difficult because of its dilute and erratic nature. Large collecting areas and large storage capacities are needed, two requirements satisfied by the tropical oceans. Oceans cover 71% of Earth's surface. In the tropics, they absorb sunlight, and the top layers heat up to some 25 C. Warm surface waters from the equatorial belt flow poleward, melting both the arctic and the antarctic ice. The resulting cold waters return to the equator at great depth, completing a huge planetary thermosyphon.

The power involved is enormous. For example, the Gulf Stream has a flow rate of 2.2×10^{12} m^3 day^{-1} of water and is some 20 K warmer than the abyssal layers. A heat engine that uses this much water and that employs as a heat sink the cool ocean bottom would be handling a heat flow of $\Delta T c \dot{V}$, where ΔT is the temperature difference, c is the heat capacity of water (about 4 MJ m^{-3}K^{-1}), and $\dot{V}$ is the flow rate.[†] This amounts to 1.8×10^{20} J day^{-1} or 2100 TW. The whole world (2008) uses energy at the rate of only $\approx$ 15 TW. These order of magnitude calculations are excessively optimistic in the sense that only a minuscule fraction of this available energy can be practically harnessed. Nevertheless, ocean thermal energy holds some promise as an auxiliary source of energy for use by humankind.

Figure 4.1 shows a typical temperature profile of a tropical ocean. For the first 50 m or so near the surface, turbulence maintains the temperature uniform at some 25 C. It then falls rapidly, reaching 4 or 5 C in deep places. Actual profiles vary from place to place and also with the seasons.

It is easier to find warm surface water than sufficiently cool abyssal waters, which are not readily available in continental shelf regions. This limits the possible sitings of ocean thermal energy converters.

4.2 OTEC Configurations

Two basic configurations have been proposed for ocean thermal energy converters (OTECs):

[†] The expression *flow rate* is redundant. The word *flow* is defined *as the volume of fluid flowing through a tube of any given section in a unit of time* Oxford English Dictionary (OED).

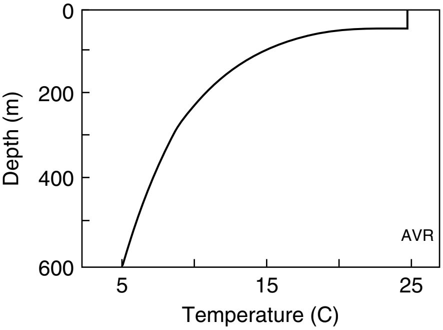


Figure 4.1 Typical ocean temperature profile in the tropics.

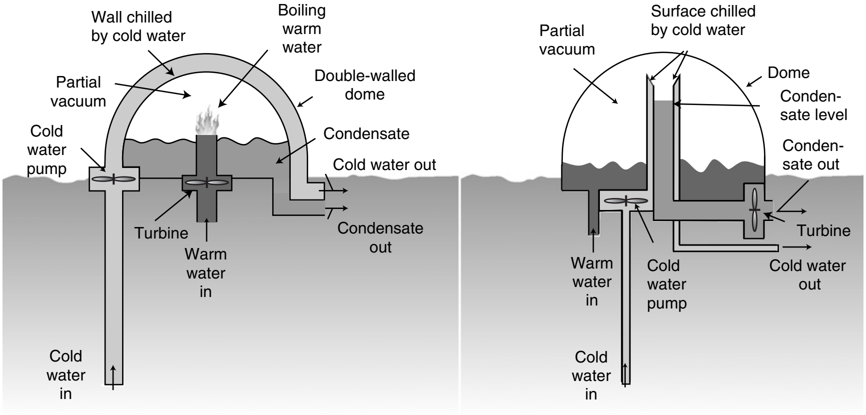


Figure 4.2 Hydraulic OTECs.

1. Those using hydraulic turbines
2. Those using vapor turbines.

The first uses the temperature difference between the surface and bottom waters to create a hydraulic head that drives a conventional water turbine. The advantages of this proposal include the absence of heat exchangers.

Consider a hemispherical canister as depicted in the left-hand side of Figure 4.2. A long pipe admits cold water, while a short one admits warm water. The canister is evacuated so that, in the ideal case, only low-pressure water vapor occupies the volume above the liquid surface. In practice, gases dissolved in the ocean would also share this volume and must be removed. This configuration was proposed by Beck (1978).

At a temperature of 15 C, the pressure inside the canister is about 15 kPa (0.017 atmospheres). At this pressure, warm water at 25 C will boil, and the resulting vapor will condense on the parts of the dome refrigerated by the cold water. The condensate runs off into the ocean, establishing a continuous flow of warm water into the canister. The incoming warm water drives a turbine from which useful power can be extracted. The equivalent hydraulic head is small, and turbines of large dimensions would be required.

To increase the hydraulic head, Zener and Fetkovich (1975) proposed the arrangement of Figure 4.2 (right). The warm surface water admitted to the partially evacuated dome starts boiling. The resulting vapor condenses on a funnel-like surface that seals one of the two concentric cylinders in the center of the dome. This cylinder receives cold water pumped from the ocean depths, which chills the steam-condensing surface. The collected condensed water subsequently flows into the central pipe, creating a head that drives the turbine. The efficiency of the device is substantially enhanced by the foaming that aids in raising the liquid.

OTECs developed in the 1980s were of the vapor turbine type. They can use open cycles (Figures 4.3a and b), close cycles (Figure 4.3c), or

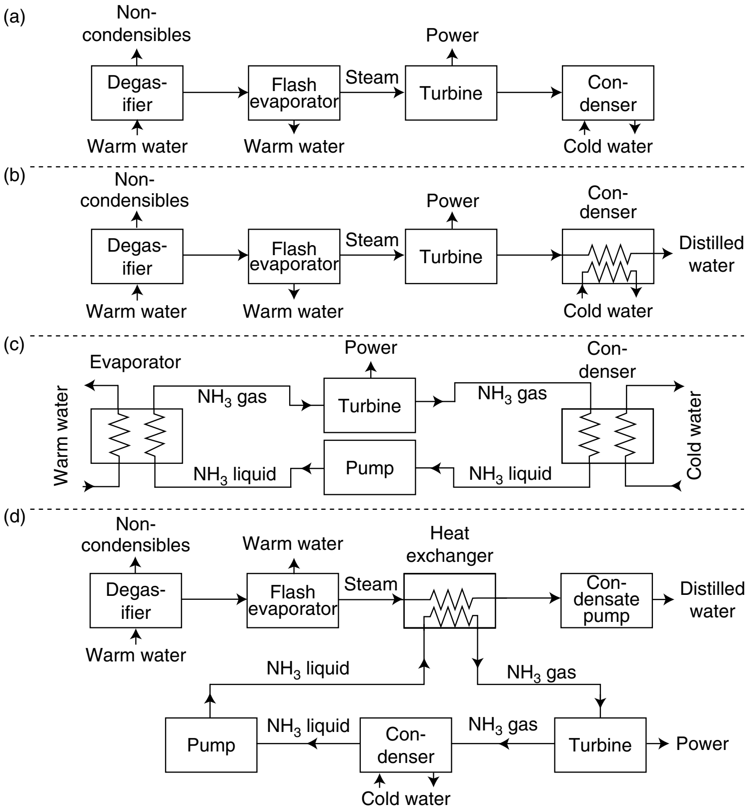


Figure 4.3 OTEC configurations include the open-cycle type without distilled water production (a), the open-cycle type with distilled water recovery (b), the close-cycle (c), and the hybrid-cycle (d).

hybrid cycles (Figure 4.3d). The open cycle avoids heat exchangers (or, if fresh water is desired, it requires only a single heat exchanger). However, the low pressure of the steam generated demands very large diameter turbines. This difficulty is overcome by using a close (or a hybrid) cycle with ammonia as a working fluid. Most work has been done on the close-cycle configuration, which is regarded as more economical. However, the costs of the two versions may turn out to be comparable.

4.3 Turbines

A turbine (Figure 4.4) generates mechanical energy from a difference in pressure. Usually, the state of the gas at the inlet and the pressure of the gas at the exhaust are specified.

Let p_{in} and T_{in} be the pressure and the temperature at the inlet of the turbine and p_{out}, T_{out} the corresponding quantities at the exhaust.

The output of the turbine is the mechanical work, W. The heat, Q, is exchanged with the environment by some means other than the circulating gases. Most practical turbines are sufficiently well insulated to be assumed **adiabatic**—that is, a condition in which $Q = 0$.

The inlet gas carries an enthalpy, H_{in}, into the turbine, while the exhaust removes H_{out} from the device. Conservation of energy requires that[†]

$$W = H_{in} - H_{out}. \tag{4.1}$$

Expressing the quantities on a per kilomole basis (quantities per kilomole are represented by lower case letters), we can write

$$W = \mu_{in} h_{in} - \mu_{out} h_{out} = \mu (h_{in} - h_{out}) \tag{4.2}$$

because, under steady-state conditions, $\mu_{in} = \mu_{out} \equiv \mu$.

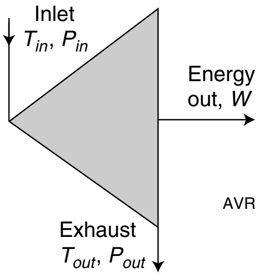


Figure 4.4 A turbine.

[†] Provided there is no appreciable change in kinetic, potential, magnetic, and other forms of energy.

In a perfect gas,

$$h_{in} - h_{out} = \int_{T_{out}}^{T_{in}} c_p dT. \tag{4.3}$$

Assuming a constant specific heat,

$$h_{in} - h_{out} = c_p(T_{in} - T_{out}), \tag{4.4}$$

and

$$W = \mu c_p(T_{in} - T_{out}). \tag{4.5}$$

$$W = \mu c_p T_{in} \frac{T_{in} - T_{out}}{T_{in}} = \mu c_p T_{in} \eta_{CARNOT} \tag{4.6}$$

Equation 4.6 looks similar to that which describes the behavior of a heat engine. However, the quantity, $\mu c_p T_{in}$, although having the dimensions of energy, is not the heat input to the device; rather, it is the enthalpy input. For a given input state and a given exhaust pressure, the mechanical energy output increases with decreasing exhaust temperature. The lowest possible value of T_{out} is limited by the second law of thermodynamics that requires that the entropy of the exhaust gases be equal or larger than that of the inlet gases. The lowest exhaust temperature (highest output) is achieved by a turbine operating **isentropically**, one in which the entropy is not changed. Any deviation from this condition is due to irreversibilities (losses) in the device. These losses will generate heat and thus increase T_{out}.

Isentropic Processes

If there is no change in entropy in the gas that flows through the turbine, then we have an isentropic process.

From the first law of thermodynamics:

$$dQ = dU + pdV \tag{4.7}$$

and from the second law:

$$dQ = TdS \tag{4.8}$$

$$TdS = dU + pdV \tag{4.9}$$

From the definition of enthalpy, $H = U + pV$,

$$dU = dH - Vdp - pdV. \tag{4.10}$$

(*Continues*)

(*Continued*)

Hence,

$$TdS = dH - Vdp, \tag{4.11}$$

$$dS = \frac{dH}{T} - \frac{V}{T}dp. \tag{4.12}$$

But $dH = c_p dT$ and $V/T = R/p$; hence

$$dS = c_p \frac{dT}{T} - R\frac{dp}{p}. \tag{4.13}$$

If the process is isentropic, then $dS = 0$; thus,

$$c_p \frac{dT}{T} = R\frac{dp}{p} = (c_p - c_v)\frac{dp}{p} \tag{4.14}$$

$$\frac{dT}{T} = \left(1 - \frac{1}{\gamma}\right)\frac{dp}{p}. \tag{4.15}$$

$$\ln T + \left(\frac{1-\gamma}{\gamma}\right)\ln p = constant \tag{4.16}$$

$$Tp^{\frac{1-\gamma}{\gamma}} = constant \tag{4.17}$$

And, finally, from the perfect-gas law,

$$pV^{\gamma} = constant. \tag{4.18}$$

Thus, the polytropic law derived for the case of adiabatic compression (see Chapter 2) applies to any isentropic process.

What is the exhaust temperature, $T_{out_{min}}$, in an isentropic turbine? Using the polytropic law,

$$p_{in}V_{in}^{\gamma} = p_{out}V_{out}^{\gamma}. \tag{4.19}$$

Applying the perfect gas law, we can eliminate the volumes:

$$p_{in}^{1-\gamma}\mu RT_{in}^{\gamma} = p_{out}^{1-\gamma}\mu RT_{out_{min}}^{\gamma}, \tag{4.20}$$

$$T_{out_{min}} = T_{in}\left(\frac{p_{out}}{p_{in}}\right)^{\frac{\gamma-1}{\gamma}}. \tag{4.21}$$

The energy delivered by the isentropic turbine is

$$W = \mu c_p T_{in} \left[1 - \left(\frac{p_{out}}{p_{in}} \right)^{\frac{\gamma-1}{\gamma}} \right]. \tag{4.22}$$

A turbine may be considered adiabatic in the sense that it does not exchange heat with the environment except through the flowing gas. However, it may exhibit internal losses that cause the exhaust temperature to be larger than that calculated from Equation 4.21.

The **isentropic efficiency** of a turbine is the ratio between the actual work produced by the turbine to the work it would produce if the input and output had the same entropy.

4.4 OTEC Efficiency

The Carnot efficiency of an OTEC is low owing to the small temperature difference that drives it. OTECs must abstract a large quantity of heat from the warm surface waters and reject most of it to the cold bottom waters. They handle great volumes of water. How do such volumes compare with those handled by a hydroelectric plant of the same capacity?

Let the temperature difference between the warm and the cold water be $\Delta T \equiv T_H - T_C$. Professor A. L. London of Stanford University has shown that the minimum water consumption occurs when the temperature difference across the turbine (in a close-cycle system) is $\Delta T/2$, leaving the remaining $\Delta T/2$ as the temperature drop across the two heat exchangers. If half of this is the drop across the warm water exchanger, and if $\dot{V}$ is the flow rate of warm water, then the power abstracted is $\frac{1}{4}\Delta T\ c\ \dot{V}$. The Carnot efficiency is $\Delta T/2T_H$. Assuming that the cold water flow is equal to that of the warm water, that is, that $\dot{V}_{TOT} = 2\dot{V}$, the power generated (with ideal turbines) is

$$P_{OTEC} = \frac{c}{16T_H} \Delta T^2 \dot{V}_{TOT} = 850\ \Delta T^2 \dot{V}_{TOT}. \tag{4.23}$$

We took $T_H = 296\,\text{K}$. Notice that the power is proportional to ΔT^2. The power of a hydroelectric plant is

$$P_{HYDRO} = \delta g \Delta h \dot{V}, \tag{4.24}$$

where δ is the density of water ($1000\,\text{kg/m}^3$), g is the acceleration of gravity, and h is the height difference between the input and output water levels.

We ask now how large Δh must be for an hydroelectric plant to produce the same power as an OTEC that handles the same water flow.

$$\Delta h = \frac{\Delta T^2 c}{16 T_H \delta g} = 0.085\ \Delta T^2 \tag{4.25}$$

An OTEC operating with a 20-K temperature difference delivers the same power as a hydroelectric plant with identical flow rate and a (moderate) 34-m head. Thus, the volumes of water required by an OTEC are not exorbitant.

These calculations are quite optimistic; they failed, for instance, to account for the considerable amount of energy required by the various pumps, especially the cold water pump. Nevertheless, the main problem with OTECs is not the large volume of water but rather one of heat transfer. Compared with fossil-fueled plants of the same capacity, the heat exchangers of an OTEC are enormous. Up to half the cost of a close-cycle OTEC lies in the heat exchangers.

4.5 Example of OTEC Design

OTECs are not designed for minimum water consumption, but rather, for minimum cost. This alters the temperature distribution in the system. Figure 4.5 shows the temperatures in a Lockheed project. It is of the close-cycle type, using ammonia as the working fluid. The overall temperature difference is 18.46 K. The warm water flow rate is 341.6 m^3/s.[†] It enters the heat exchanger at 26.53 C and exits at 25.17 C, having been cooled by 1.36 K. This warm water delivers to the OTEC a total thermal power of

$$P_{in} = 341.6 \times 4.04 \times 10^6 \times 1.36 = 1876.9.\ \text{MWt}^{\dagger\dagger} \qquad (4.26)$$

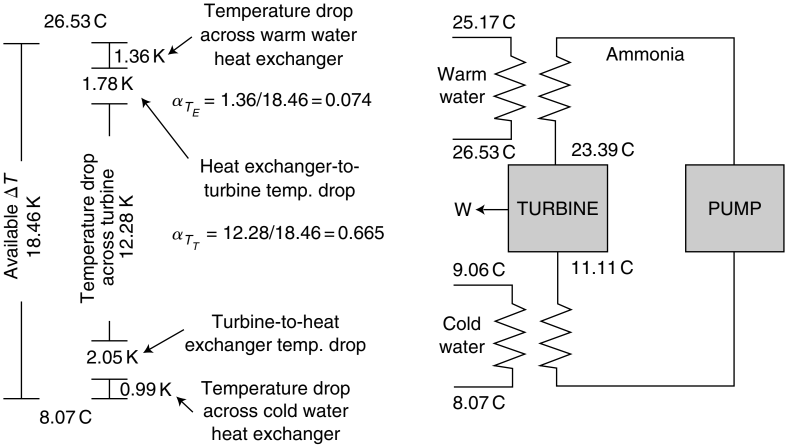

26.53 C
1.36 K
Temperature drop across warm water heat exchanger
1.78 K
α_{T_E} = 1.36/18.46 = 0.074
Heat exchanger-to-turbine temp. drop
Available ΔT 18.46 K
Temperature drop across turbine 12.28 K
α_{T_T} = 12.28/18.46 = 0.665
Turbine-to-heat exchanger temp. drop
2.05 K
0.99 K
Temperature drop across cold water heat exchanger
8.07 C
25.17 C
Ammonia
Warm water
26.53 C
23.39 C
W
TURBINE
PUMP
9.06 C
11.11 C
Cold water
8.07 C

Figure 4.5 Temperatures in an OTEC designed by Lockheed.

[†] To gain an idea of how much water is pumped, consider a 25 × 12 m competition swimming pool. The warm water pump of the Lockheed OTEC under discussion would be able to fill such a pool in less than 2 seconds!

[††] MWt stands for thermal power, whereas MWe stands for electric power.

Table 4.1 Internal Power Use in the Lockheed OTEC

Condensate pump(recirculates ammonia)	2.54 MW
Reflux pump (recirculates ammonia that failed to evaporate)	0.04 MW
Warm water pump	4.83 MW
Cold water pump	12.14 MW
TOTAL	19.55 MW

The ammonia at the turbine inlet is at 23.39 C, while at the outlet it is at 11.11 C, which leads to a Carnot efficiency of

$$\eta_{CARNOT} = \frac{23.39 - 11.11}{273.3 + 23.39} = 0.041. \qquad (4.27)$$

Electricity is produced at 90.2% of the Carnot efficiency. However, the different pumps use a great-deal of the generated power, as illustrated in Table 4.1 In fact, in this example, 28% of the total power generated is used internally just to run the system. In typical steam engines, pumps are mechanically coupled to the turbine, not electrically as in this OTEC.

The main power consumer is the cold water pump not only because it handles the largest amount of water but also because it has to overcome the friction on the very long cold water pipe.

The electric power generated is

$$P_{gen} = 1876.9 \times 0.041 \times 0.902 = 69.4\,\text{MWe}. \qquad (4.28)$$

However, the electric power available at the output bus is only

$$P_{bus} = 69.4 - 19.6 = 49.8\,\text{MWe}. \qquad (4.29)$$

Thus, the overall efficiency of this OTEC is

$$\eta = \frac{49.8}{1876.9} = 0.0265. \qquad (4.30)$$

This 2.65% efficiency is what can be expected of any well-designed OTEC.

The pressure of the ammonia at the inlet side of the turbine is 0.96 MPa (9.7 atmos) and, at the outlet, 0.64 MPa (6.5 atmos). An amount of heat equal to $1876.9 - 69.4 = 1807.5$ MW must be rejected to the cold water by the heat exchanger. This is done by taking in $451.7\,\text{m}^3\text{s}^{-1}$ of water at 8.07 C and heating it up by 0.99 K to 9.06 C.

It is necessary to make the distance between the cold water outlet and the warm water inlet sufficiently large to avoid mixing. This gives the ocean currents opportunity to sweep the cold water away. In the absence of currents, OTECs may have to move around "grazing" fresh warm water.

Propulsive power for this grazing can easily be obtained from the reaction to the outlet water flow.

4.6 Heat Exchangers

The overall efficiency of an OTEC is small. The Lockheed OTEC of the example converts 1877 MWt into 49.8 MW of salable electricity—an efficiency of 2.6%. However, the "fuel" is completely free, so the overall efficiency is of no crucial importance. What counts is the investment cost, which greatly depends on the cost of the heat exchangers.

The power transferred through a heat exchanger is

$$P_{therm} = \gamma A \Delta T_{EXB}, \tag{4.31}$$

where γ is the heat transfer coefficient of the exchanger, A is its area, and ΔT_{EXB} is the mean exchanger-to-boiler temperature difference.

Lockheed hoped to achieve a $\gamma = 2800\,\mathrm{W\ m^{-2}K^{-1}}$. With a ΔT_{EXB} of approximately 2.5 K and a P_{therm} of 1877 MW, the required area is about 270,000 m^2—that is, over 500 × 500 m.

Even minor fouling will considerably lower the value of γ, and thus it is important to keep the heat exchanger surfaces clean and free from algae. It is possible to electrolyze a small fraction of the incoming water to liberate algae-killing chlorine.

Technically, the ideal material for OTEC heat exchangers seems to be titanium, owing to its stability in seawater. Aluminum, being less expensive was also considered.

4.7 Siting

We saw that the economics of OTECs depend critically on ΔT. Consequently, a site must be found where a (comparatively) large ΔT is available. There is, in general, no difficulty in finding warm surface waters in tropical seas; the problem is to find cold water because this requires depths uncommon in the vicinity of land. For this reason, land-based OTECs may be less common than those on floating platforms.

A large ΔT (some 17 K or more) is not the only siting requirement. The depth of a cold water layer at 6 C or less should be moderate, certainly less than 1000 m; otherwise, the cost of the cold water pipe would become excessive. Anchoring problems suggest placing OTECs in regions where the total depth is less than some 1500 m. Surface currents should be moderate (less than 2 m/s).

Total ocean depth is, of course, of no importance if dynamically positioned OTECs are contemplated. This can be accomplished by taking advantage of the thrust generated by the seawater exhaust of the plant.

As explained previously, if there are no currents, such thrusting may be necessary to keep the cooler exhaust water from mixing with the warm intake.

Another siting consideration is distance from the shore if one contemplates bringing in the generated electricity by means of electric cables. OTECs may be used as a self-contained industrial complex operating as floating factories producing energy-intensive materials. Ammonia, for instance, requires for its synthesis only water, air, and electricity and is almost the ideal product for OTEC manufacture. It is easier to ship the ammonia than to transmit electric power to shore. Once on shore, ammonia can be used as fertilizer, or it can be converted back into energy by means of fuel cells (see Chapter 9).

One OTEC arrangement involves the use of only the cold ocean bottom water. Such water is first pumped through heat exchangers and then into shallow ponds, where it is heated by the sun. In this manner, it can reach temperatures well above those of the ocean, leading to larger Carnot efficiencies.

OTECs using solar-heated ponds can be combined with mariculture. Deep ocean waters tend to be laden with nutrients and, when heated by the sun, will permit the flourishing of many species of microscopic algae. The alga-rich water flows into a second pond where filter feeding mollusks are raised. Oysters, clams, and scallops are produced. The larger of these animals are either kept for reproduction or sold in the market. The smaller ones are destroyed and thrown into a third pond where crustaceans (shrimps, lobsters, etc.) feed on them. The effluent of this pond should not be returned directly to the ocean because the animal waste in it is a source of pollution. A fourth and final pond is used to grow seaweed that clean up the water and serve as a source of agar or carrageen or, alternatively, as a feedstock for methane-producing digesters (see Chapter 13). The warm water from this pond is used to drive the OTEC.

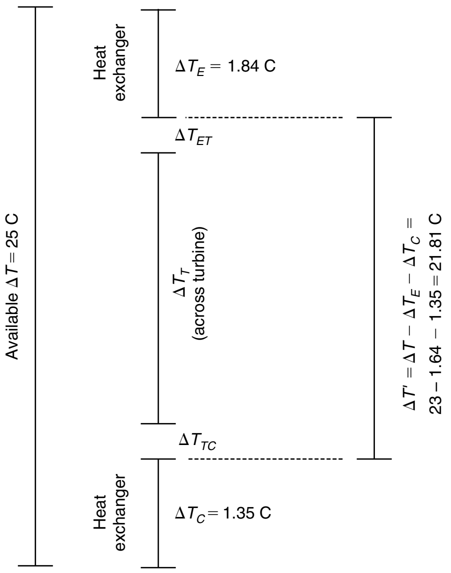
Cold water pumped up by the OTEC can be used directly for air conditioning without the need of generating electricity. Chilled-soil agriculture can allow the growing of temperate climate plants in tropical areas where sunlight is abundant. Finally, the technological development resulting from work to perfect OTECs may one day bring a payoff in the form of an economic desalination process.

References

Beck, E. J., Ocean thermal gradient hydraulic power plant, *Science*, **189**, p.293, **1975**.

Zener, C., and J. Fetkovich, Foam solar sea power plant, *Science*, **189**, p. 294, **1975**.

PROBLEMS


Available $\Delta T = 25$ C
Heat exchanger
$\Delta T_E = 1.84$ C
ΔT_{ET}
ΔT_T (across turbine)
ΔT_{TC}
Heat exchanger
$\Delta T_C = 1.35$ C
$\Delta T' = \Delta T - \Delta T_E - \Delta T_C =$
$23 - 1.64 - 1.35 = 21.81$ C

4.1 An OTEC is to deliver 100 MW to the bus bar. Its warm water comes from a solar-heated pond that is kept at 33 C. The water exhausted from the heat exchanger is returned to this same pond at a temperature of 31 C. (There is a slight heat loss in the pipes.) To reestablish the operating temperature, the pond must absorb heat from the sun. Assume an average (day and night) insolation of 250 W/m^2 and an 80% absorption of solar energy by the water.

Cold water is pumped from the nearby abyss at a temperature of 8 C. Refer to the figure for further information on the temperatures involved.

The warm water loses 1.84 K in going through the heat exchanger, while the cold water has its temperature raised by 1.35 K in its exchanger.

Assume that 80% of the remaining temperature difference appears across the turbine and that the rest is equally distributed as temperature differentials between the colder side of the warm heat exchanger (whose secondary side acts as an evaporator) and the turbine inlet and between the warmer side of the cold heat exchanger (whose secondary side acts as a condenser) and the turbine outlet (ΔT_{ET} and ΔT_{TC}, in the figure).

Internal power for pumping and other ends is 40 MW. The efficiency of the turbine-generator combination is 90%. Estimate the rates of flow of warm and cold water.

What is the required surface of the heating pond, assuming no evaporation?

If the residence time of the water in the pond is three days, what depth must it have?

4.2 The Gulf Stream flows at a rate of 2.2×10^{12} m^3/day. Its waters have a temperature of 25 C. Make a rough estimate of the area of the ocean that collects enough solar energy to permit this flow.

4.3 Assume that ammonia vaporizes in the evaporator of an OTEC at constant temperature (is this strictly true?). If the warm water enters the heat exchanger with a temperature ΔT_1 higher than that of the boiling ammonia and leaves with a ΔT_2, what is the mean ΔT? *To check your results:* If $\Delta T_1 = 4$ K and $\Delta T_2 = 2$ K, then $<\Delta T> = 2.88$ K.

4.4 A 1.2-GWe nuclear power plant is installed near a river whose waters are used for cooling. The efficiency of the system is 20%. This is the ratio of electric output to heat input.

Technical reasons require that the coolant water exit the heat exchangers at a temperature of 80 C. It is proposed to use the warm coolant water to drive an OTEC-like plant. Assume:

The river water is at 20 C,
The OTEC efficiency is one-half of the Carnot efficiency, and
Half of the available ΔT is dropped across the turbine.

1. What is the flow rate of this water?
2. What is the maximum electric power that can be generated by such a plant?

4.5 An OTEC pumps 200 cubic meters of warm water per second through a heat exchanger in which the temperature drops by 1%. All the heat extracted is delivered to the ammonia boiler. The ammonia temperature at the turbine inlet is equal to the mean temperature of the water in the warm water heat exchanger minus 1 K. The condenser temperature is kept at 10 C by the cooling effect of 250 cubic meters per second of cold water. The efficiency of the turbine/generator system is 90%, and 12 MW of the produced electricity is used for pumping.

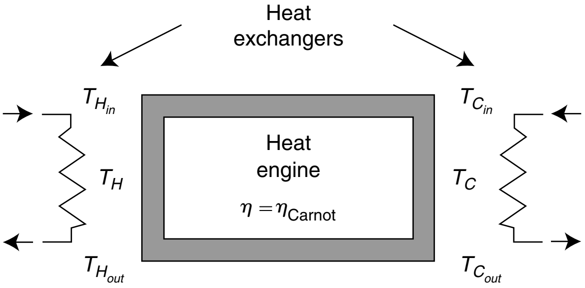
What must the intake temperature of the warm water be, so that a total of 20 MW of electricity is available for sale?

What must the intake temperature be so that the OTEC produces only exactly the amount of power needed for pumping?

4.6 Consider an OTEC whose turbine/generator has 100% mechanical efficiency. In other words, the system operates at the Carnot efficiency. Input temperature to the turbine is T_H, and the output is at T_C.

The input heat comes from a heat exchanger accepting water at $T_{H_{in}}$ and discharging it at a lower temperature, $T_{H_{out}}$. The flow rate of warm water through this heat exchanger is $\dot{V}_H$.

The heat sink for the turbine is another heat exchanger taking in water at $T_{C_{in}}$ and discharging it at a higher temperature, $T_{C_{out}}$. The flow rate of cold water through this heat exchanger is $\dot{V}_C$. Refer to the following figure.



All the heat extracted from the warm water by the heat exchanger is transferred to the input of the turbine. All the heat rejected by the turbine is absorbed by the cold water heat exchanger and removed.

The following information is supplied:

$T_{H_{in}} = 25$C.
$T_{C_{in}} = 8$C.
T_H is the mean of $T_{H_{in}}$ and $T_{H_{out}}$.
T_C is the mean of $T_{C_{in}}$ and $T_{C_{out}}$.
$\dot{V}_H = \dot{V}_C \equiv \dot{V} = 420 m^3/s.$
$T_{H_{out}}$ and $T_{C_{out}}$ are not given.

Clearly, if $T_{H_{out}}$ is made equal to $T_{H_{in}}$, no heat is extracted from the warm water, and the power output is zero. On the other hand, if $T_{H_{out}}$ is made equal to $T_{C_{in}}$, maximum power is extracted from the warm water but the output power is again zero because (as you are going to find out) the Carnot efficiency goes to zero.

Between these extremes, there must be a value of $T_{H_{out}}$ that maximizes power output. Determine this value, and determine the value of $T_{H_{out}}$ that maximizes the Carnot efficiency.

Chapter 5
Thermoelectricity

So far we have discussed heat engines in general and have examined in some detail the heat engine used in ocean thermal energy converters. All these engines convert heat into mechanical energy.

In this chapter and in the next two, we consider engines that transform heat directly into electricity: the thermoelectric, the thermionic, and the radio-noise converters. In the chapter on photovoltaic cells, we will discuss the thermophotovoltaic converter that transforms heat into radiant energy and then into electricity. Other engines such as the magnetohydrodynamic engine that converts heat into kinetic energy of a plasma and then into electricity will not be covered here.

In most of the chapters of this book, we have adopted the strategy of first introducing a simplified model that explains the phenomena that underlie the operation of the device being studied, and; from that, we deduce the way the device behaves. In this chapter, we will change the approach, describing first the behavior of thermocouples and only later examining the underlying science. The reason for this is that, although the behavior of thermoelectric devices is easy to describe, there is no simple way of explaining how these effects come about. In fact, were we to use classical mechanics with Maxwellian electron energy distribution, we would prove that there is no Peltier effect and that the Thompson effect in metals should be two orders of magnitude larger than what it really is. The Peltier and Thompson effects are two of the effects associated with thermoelectricity.

5.1 Experimental Observations

Consider a heat-conducting bar whose ends are at different temperatures, T_H and T_C. Clearly, a certain amount of heat power, P_{H_F}, will enter through one face and an amount, P_{C_F}, will leave through the opposite face. If the bar has its remaining sides perfectly insulated so that no heat can exit through them, then

$$P_{H_F} = P_{C_F} = \Lambda(T_H - T_C). \tag{5.1}$$

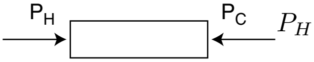
Λ is the **heat conductance** of the bar, which in the SI is measured in W/K. The subscript F is used because heat conduction is sometimes referred to as the **Fourier** effect in honor of Jean Baptiste Joseph Fourier's (1768–1830) noted contributions to the study of heat diffusion.

If, on the other hand, the bar is at uniform temperature but is heated (by means of a current) to a higher temperature than that of the bodies in contact with the two end faces, then heat must flow out. The heat power generated by the current is I^2R, where R is the resistance of the bar. Half of this generated heat will flow out one end and half the other.

$$P_{H_J} = P_{C_J} = \tfrac{1}{2}RI^2 \tag{5.2}$$

We used the subscript, J, to indicate that the heat is the result of the **Joule** effect (James Prescott Joule, 1818–1889). The arrow directions for heat flow are different in Figure 5.1 and Figure 5.2. Had we chosen to keep the same direction as in the former figure, then P_{H_J} would be negative.

If the bar is simultaneously submitted to a temperature differential and a current, then the Fourier and Joule effects are superposed:



$$P_H = \Lambda(T_H - T_C) - \tfrac{1}{2}RI^2, \tag{5.3}$$

$$P_C = \Lambda(T_H - T_C) + \tfrac{1}{2}RI^2. \tag{5.4}$$

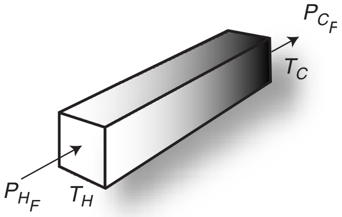


Figure 5.1 A bar with a temperature gradient.

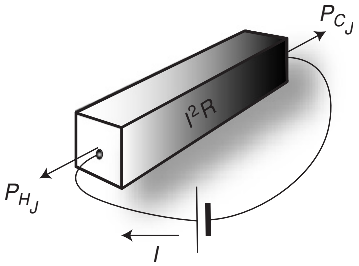


Figure 5.2 A bar heated by a current.

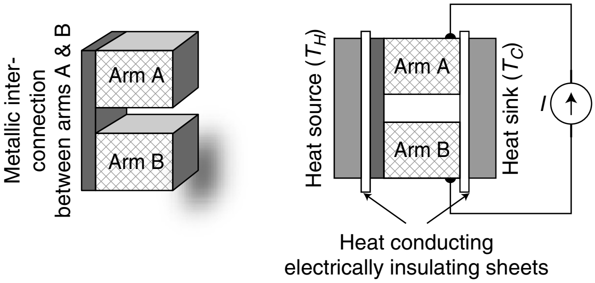


Figure 5.3 A simple thermocouple (left) and a test setup (right).

Here we used the directions of heat power flow indicated in the inset that accompanies Equations 5.3 and 5.4. However, in a more complicated structure, the heat power flow can surprisingly depart from expectation.

Consider a **thermocouple** consisting of two *dissimilar* materials (conductors or semiconductors) joined together at one end. The materials (arms A and B in Figure 5.3) may touch one another directly or may be joined by a metallic strip as indicated. As long as this metallic strip is at uniform temperature, it has no influence on the performance of the thermocouple (provided the strip has negligible electric resistance and essentially infinite heat conductivity). The free ends of arms A and B are connected to a current source.

Again, if the connecting wires are at uniform temperature, they exert no influence. Two blocks maintained at uniform temperature are thermally connected, respectively, to the junction and to the free ends. These blocks are electrically insulated from the thermocouple.

The block in contact with the junction is the **heat source** and is at the temperature T_H. The other block is the **heat sink** and is at T_C. The rate of heat flow, P_H, from the source to the sink is measured as explained in the box at the end of this subsection.

Assume the thermocouple is carefully insulated so that it can only exchange heat with the source and with the sink. If we measure P_H as a function of $T_H - T_C$ with no current through the thermocouple, we find that P_H is proportional to the temperature difference (as in Equation 5.1):

$$P_H = \Lambda(T_H - T_C). \tag{5.5}$$

As an example, take $\Lambda = 4.18$ W/K; then

$$P_H = 4.18(T_H - T_C). \tag{5.6}$$

If we force a current, I, through a thermocouple with an internal resistance, R, we expect, as explained, P_H to be given by (cf. Equation 5.3)

$$P_H = \Lambda(T_H - T_C) - \tfrac{1}{2}RI^2, \tag{5.7}$$

or, if the resistance of the thermocouple is 2.6×10^{-4} ohms,

$$P_H = 2090 - 1.3 \times 10^{-4} I^2, \tag{5.8}$$

where we used $T_H = 1500\,\text{K}$ and $T_C = 1000\,\text{K}$, as an example.

The expected plot of P_H versus I appears as a dotted line in Figure 5.4. It turns out that a change in P_H, as I varies, is in fact observed, but it is not independent of the sign of I. The empirically determined relationship between P_H and I is plotted, for a particular thermocouple, as a solid line. A seconds-order regression fits the data well:

$$P_H = 2090 + 1.8I - 1.3 \times 10^{-4} I^2. \tag{5.9}$$

In equation 5.9, we recognize the heat conduction term because it is independent of I. We also recognize the Joule heating term, $1.3 \times 10^{-4}\ I^2$.

In addition to these two terms, there is one linear in I.

This means that if the current is in one direction, heat is transported from the source to the sink and, if inverted, so is the heat transport. Evidently, heat energy is carried by the electric current. This reversible transport is called the **Peltier** effect (Jean Charles Athanase Peltier, 1785–1845).

From empirical evidence, the Peltier heat transported is proportional to the current. We can therefore write

$$P_{Peltier} = \pi I, \tag{5.10}$$

where π is the **Peltier coefficient.**

If we connect an infinite impedance voltmeter to the thermocouple instead of a current generator, we observe a voltage, V, that is dependent on the temperature difference, $\Delta T \equiv T_H - T_C$. The dependence is nonlinear as illustrated in Figure 5.5, which shows the relationship of the open-circuit

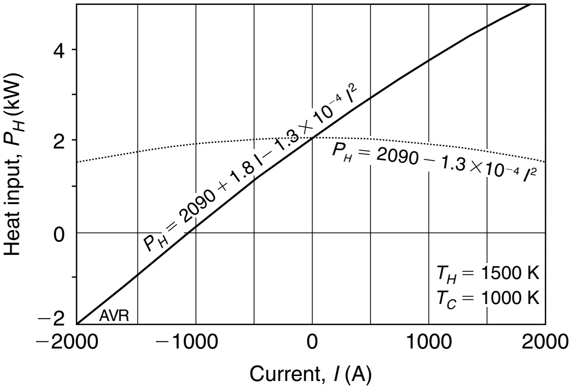


Figure 5.4 Heat input vs. current characteristics of a thermocouple.

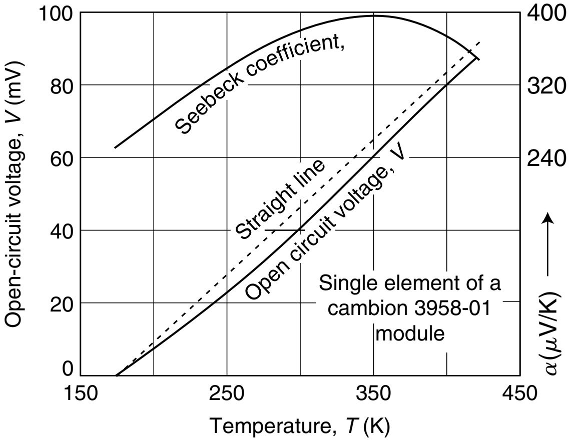


Figure 5.5 Open-circuit voltage and Seebeck coefficient.

voltage of a thermocouple to the temperature, T_H. In this case, T_C is held at a constant 173.3 K.[†]

The **Seebeck coefficient**, α, is the slope of the V versus T_H plot and depends somewhat on the temperature:

$$\alpha \equiv \frac{dV}{dT}. \tag{5.11}$$

Later we are going to show that there is a relationship between the Peltier and the Seebeck coefficients:

$$\pi = \alpha T. \tag{5.12}$$

The Peltier coefficient is a strong function of the temperature.

The open-circuit voltage developed by the thermocouple is

$$V = \int_{T_C}^{T_H} \alpha dT, \tag{5.13}$$

or, if a **mean Seebeck coefficient**, $<\alpha>$, is used,

$$V = <\alpha> (T_H - T_C). \tag{5.14}$$

In the balance of this chapter, although we will work with $<\alpha>$, we will represent it simply by α. Use of the mean value of α allows us to correctly describe the thermocouple performance in terms of only the four effects

[†] The dependence of V on ΔT can be represented quite accurately by a power series, $V = a_0 \Delta T + a_1 \Delta T^2 + a_2 \Delta T^3 + \cdots$ The values of the different coefficients, a_i, are slightly dependent on the reference temperature, T_C.

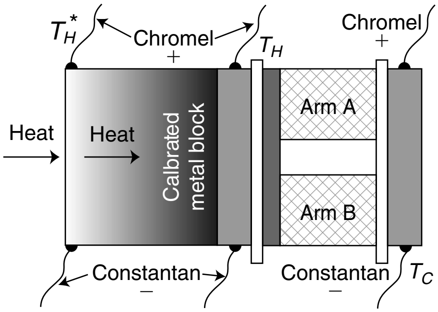


Figure 5.6 The measurement of heat flow in a thermocouple.

mentioned so far—Fourier, Joule, Peltier, and Seebeck—completely ignoring the heat convected by the carriers, that is, the **Thomson effect**. In the end of this chapter, we will justify this omission.

Thermocouples find use as

1. thermometers,
2. generators capable of transforming heat directly into electricity, and
3. heat pumps and refrigerators

Temperatures as well as electric quantities can be easily measured with good accuracy (a moderately inexpensive voltmeter can have accuracies of 0.1% of full scale). On the other hand, it is difficult to determine the precise flow of heat.

Conceptually, the heat flow can be measured by insulating the heat source except at its contact with the thermocouple. Heat can be supplied to this source by an electric resistor. The electric power necessary to keep T_H constant is a measure of the heat flow from source to couple.

More commonly, the flow of heat is measured by inserting a metallic block (whose conductivity as a function of temperature has been carefully measured) between the heat source and the thermocouple. If no heat leaks out through the side walls of the block, then the heat flow through the thermocouple can be determined from the temperature drop, $T_H^* - T_H$, along the metallic block.

T_H^* and T_H, as well as T_C, are each measured by attaching thermocouple wires to appropriate regions of the device.

5.2 Thermoelectric Thermometers

Since the open-circuit (Seebeck) voltage is a monotonic function of the temperature, thermocouples are an obvious choice as thermometers. When carefully calibrated, they will function with surprising accuracy over the temperature range from some 20 K to over 1700 K. However, to cover this full range, different units are required.

Because the Seebeck voltage is not a linear function of temperature. this voltage must be translated to temperature by means of look-up tables.

A very large number of combinations of materials have been considered for thermometry, but only a few have been standardized. Nine of these combinations are designated by identifying letters as shown in Table 5.1.

The first material in each pair in the table is the positive leg; the second is the negative leg. Composition is by weight.

Many of the alloys are better known by their commercial name:

55%Cu+45%Ni: Constantan, Cupron, Advance, ThermoKanthal JN.
90%Ni+10%Cr: Chromel, Tophel, ThermoKanthal KP, T-1.
99.5%Fe: ThermoKanthal JP.
95%Ni+2%Al+2%Mn+1%Si: Alumel, Nial, ThermoKanthal KN, T-2.
84%Ni+14%Cr+1.5%Si: Nicrosil.
95%Ni+4.5%Si+0.1 Mg: Nisil.

To identify a single-leg thermoelement, a suffix (P or N) is attached to the type letter to indicate the polarity of the material. Thus, for instance, EN—usually constantan—is the negative leg of Type E thermocouples. Materials must be selected taking into account a number of characteristics, including:

1. *Stability.* The properties of the material should not change significantly with their use. The materials must be chemically stable in the environment in which they are to operate. This is particularly true for high-temperature devices operating in an oxidizing atmosphere and is the reason for using noble metals instead of base ones.

Table 5.1 Standardized Thermocouple Pairs

Type	Material	Recommended range (K)
B	Pt+30%Rh vs. Pt+6%Rh	1640–1970
C	W+5%Re vs. W+26%Re	1920–2590
E	90%Ni+10%Cr vs. 55%Cu+45%Ni	370–1170
J	99.5%Fe vs. 55%Cu+45%Ni	370–1030
K	90%Ni+10%Cr vs. 95%Ni+2%Al+2%Mn+1%Si	370–1530
N	84%Ni+14%Cr+1.5%Si vs. 95%Ni+4.5%Si+0.1 Mg	920–1530
R	Pt+13%Rh vs. Pt	1140–1720
S	Pt+30%Rh vs. Pt	1250–1720
T	Cu vs. 55%Cu+45%Ni	70–620

The material must be physically stable. It should not experience phase changes (especially in cryogenic applications). It should not be altered by mechanical handling (for example, it must not lose ductility).

The material should be insensitive to magnetic fields.

2. *Homogeneity.* It is important that the material be homogeneous not only along a given sample but also from sample to sample so that a single calibration be is valid for an entire batch.
3. *Good thermoelectric power.* The lower temperature limit for practical use of thermocouple thermometers (about 20 K) is the result of insufficient-values of α as absolute zero is approached.
4. *Low thermal conductivity.* This is important at cryogenic temperatures.

The choice of the type of thermocouple depends mainly on the temperature range desired, as indicated in Table 5.1.

To achieve the highest precision, it is essential to perform very careful calibration. The extremely meticulous procedures required are described in detail (with ample references) by Burns and Scroger (1969). As an example, a NIST (National Institute of Standards and Technology) publication reports the calibration of a thermocouple (submitted by a customer) as having uncertainties not exceeding 3 μV in the range of 0 C to 1450 C. Since the Seebeck voltage for this particular sample was 14,940 μV at the highest temperature mentioned, the uncertainty at that temperature was less than 0.02%.

Very complete data of thermoelectric voltages versus temperature, tabulated at intervals of 1 Celsius for various types of thermocouples are found in NIST ITS-90 Thermocouple Database at <http://srdata.nist.gov/its90/main/>. An example of such data appears in Figure 5.7. Observe that the voltages developed by a thermocouple thermometer are small compared to those developed by a thermoelectric generator. See Section 5.3.

5.3 The Thermoelectric Generator

The ability of a thermocouple to generate a voltage when there is a temperature difference across it suggests its use as a heat engine capable of producing electricity directly. As a heat engine, its efficiency is limited by the Carnot efficiency, and therefore it must be of the form

$$\eta = \frac{T_H - T_C}{T_H}\eta^*, \qquad (5.15)$$

where η^* depends on the geometry of the device, on the properties of the materials used, and on the matching of the generator to the load.

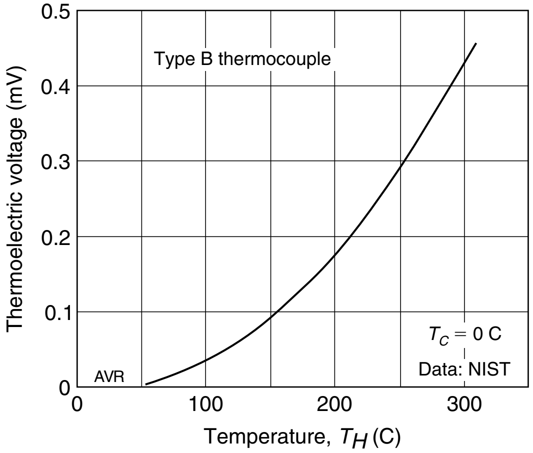


Figure 5.7 Thermoelectric voltage for a Type B thermocouple.

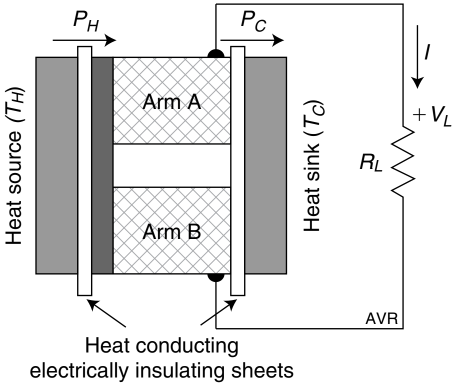


Figure 5.8 The thermocouple as a generator.

Consider a thermocouple represented by the simplified sketch in Figure 5.8. The electric resistance, R, and the heat conductance, Λ, are

$$R = \frac{\ell_A}{A_A\sigma_A} + \frac{\ell_B}{A_B\sigma_B} \tag{5.16}$$

$$\Lambda = \frac{A_A\lambda_A}{\ell_A} + \frac{A_B\lambda_B}{\ell_B}, \tag{5.17}$$

where A is the area of the cross section of each arm, ℓ is the length of each arm, σ is the electric conductivity, and λ is the thermal conductivity. Note that the two arms of the device are in parallel insofar as heat conduction is concerned but are electrically in series.

In the presence of a current, I, the heat power supplied by the heat source is, as we have seen from our experiment,

$$P_H = \Lambda(T_H - T_C) + \pi I - \tfrac{1}{2} R I^2. \tag{5.18}$$

π can be eliminated by using Equation 5.12. The Peltier term at the source is proportional to the temperature, T_H, at that point. Thus,

$$P_H = \Lambda(T_H - T_C) + \alpha T_H I - \tfrac{1}{2} R I^2. \tag{5.19}$$

The current through the load is

$$I = \frac{\alpha(T_H - T_C)}{R + R_L}, \tag{5.20}$$

and, consequently, the power delivered to the load is

$$P_L = \frac{\alpha^2 (T_H - T_C)^2}{(R + R_L)^2} R_L. \tag{5.21}$$

The efficiency of the device is

$$\begin{aligned} \eta = & \frac{P_L}{P_H} = \frac{T_H - T_C}{T_H} \times \\ & \left[\frac{(R + R_L)^2}{R_L} \frac{\Lambda}{\alpha^2} \frac{1}{T_H} + \frac{1}{2} \frac{R}{R_L} + 1 + \frac{1}{2} \frac{R}{R_L} \frac{T_C}{T_H} \right]^{-1}. \end{aligned} \tag{5.22}$$

It is convenient to write $R_L = m\ R$. The efficiency formula then becomes

$$\eta = \eta_{CARNOT} \times \left[1 + \frac{1}{2m} \left(1 + \frac{T_C}{T_H} \right) + \frac{(m+1)^2}{m} \frac{1}{T_H Z} \right]^{-1} \equiv \eta_{CARNOT} \times \eta^*, \tag{5.23}$$

where a **figure of merit**, Z, of the thermocouple is defined as

$$Z \equiv \frac{\alpha^2}{\Lambda R}. \tag{5.24}$$

The dimension of Z is [temperature^{-1}]; therefore, in the SI, the unit of measurement is K^{-1}. In the expression for η^*, all parameters other than Z are externally adjustable. The characteristics of the thermocouple are all contained in Z. *The larger Z, the greater the efficiency.*

In order to obtain a large Z, one must choose materials for the thermocouple arms that have large Seebeck coefficients. One must also make the ΛR product as small as possible. This can be achieved, again, by proper choice of materials and of the geometry of the device. If the arms are short and have a large cross section, then the resistance, R, tends to be small, but the heat conductance, Λ, tends to be correspondingly large. Similarly, if the arms are long and have a small cross section, the heat conductance tends to be small, but the resistance, R, tends to be correspondingly large. As it happens, there is a geometry that minimizes the ΛR product. This minimum occurs when the length, ℓ, and the cross-sectional area, A, of arms A and B satisfy the relationship (see derivation in the Appendix):

$$\frac{\ell_A A_B}{\ell_B A_A} = \sqrt{\frac{\lambda_A \sigma_A}{\lambda_B \sigma_B}}. \tag{5.25}$$

Under these conditions, the value of the ΛR product is

$$\Lambda R = \left[\left(\frac{\lambda_A}{\sigma_A}\right)^{1/2} + \left(\frac{\lambda_B}{\sigma_B}\right)^{1/2}\right]^2. \tag{5.26}$$

If one wants to maximize the efficiency of a thermoelectric generator, one has to choose the appropriate value for the load resistance, R_L. Remembering that we wrote $R_L = m\ R$, this means that we have to choose the appropriate value of m:

$$\frac{d}{dm}\left[\frac{(m+1)^2}{m}\frac{1}{ZT_H} + \frac{1}{2m}\left(1 + \frac{T_C}{T_H}\right) + 1\right] = 0, \tag{5.27}$$

which leads to

$$m = \sqrt{1 + <T> Z}, \tag{5.28}$$

where

$$<T> = \frac{T_H + T_C}{2}. \tag{5.29}$$

When this value of m is introduced into the expression for η^*, one obtains

$$\eta^* = \frac{(1 + <T> Z)^{1/2} - 1}{(1 + <T> Z)^{1/2} + T_C/T_H} = \frac{m - 1}{m + T_C/T_H}. \tag{5.30}$$

Summing up, there are three different considerations in optimizing the efficiency of a thermocouple:

1. Choice of appropriate materials in order to maximize Z.
2. Choice of the best geometry in order to minimize ΛR.
3. Choice of the proper value of the load resistance relative to the internal resistance of the device, that is, selection of the best value for m.

Equation 5.23 shows that for $Z \to \infty$,

$$\eta^* = \frac{1}{1 + \frac{1}{2m}\left(1 + \frac{T_C}{T_H}\right)}. \tag{5.31}$$

Maximum η^* is obtained by using $m = \infty$. Having an infinite m means that there is infinitely more resistance in the load than in the thermocouple. In other words, the couple must have zero resistance, which can only be achieved by the use of superconductors. Superconductors, unfortunately, have inherently zero Seebeck coefficients. Consequently, $\eta^* = 1$ cannot be achieved even theoretically. Indeed, with present-day technology, it is difficult to achieve Zs in excess of $0.004\,\mathrm{K}^{-1}$. This explains why thermocouples have substantially less efficiency than thermomechanical engines.

Figure 5.9 plots η^* versus T_H (for $T_C = 300\,\mathrm{K}$), using two different values of Z, in each case for optimum m. From this graph one can see that thermocouples created with existing technology can achieve (theoretically) some 30% of the Carnot efficiency. Compare this with the General Electric combined cycle "H" system that attains 60% overall efficiency. The temperature of the first-stage nozzle outlet is 1430 C. Assume a final output temperature of 300 C. The Carnot efficiency of a machine working between these two temperatures is 66%. Thus, the GE system realizes 90% of the Carnot efficiency.

5.4 Figure of Merit of a Material

The figure of merit, Z, that we have used so far refers to a pair of materials working against each other. It would be convenient if a figure of merit could be assigned to a material by itself. This would help in ranking its performance in a thermocouple. To be able to develop such a figure of merit, we need to extend the definition of the Seebeck coefficient.

Measurements and theoretical predictions show that the Seebeck effect of any junction of superconductors is zero. This permits the definition of an absolute Seebeck coefficient of a normal conductor: it is the coefficient of the material working against a superconductor.

The absolute coefficient can, of course, be measured only at temperatures low enough to allow superconductivity. The coefficient of lead was measured between 7.2 and 18 K. Below 7.2 K, lead itself becomes a superconductor, and the effect disappears; above 18 K, there were no superconductors available when these measurements were made.

Seebeck coefficients for lead above 18 K were calculated from accurate measurements of the Thompson effect (see Section 5.14.3) using the Equation 5.139,

$$\frac{d\alpha}{dT} = \frac{\tau}{T}. \tag{5.139}$$

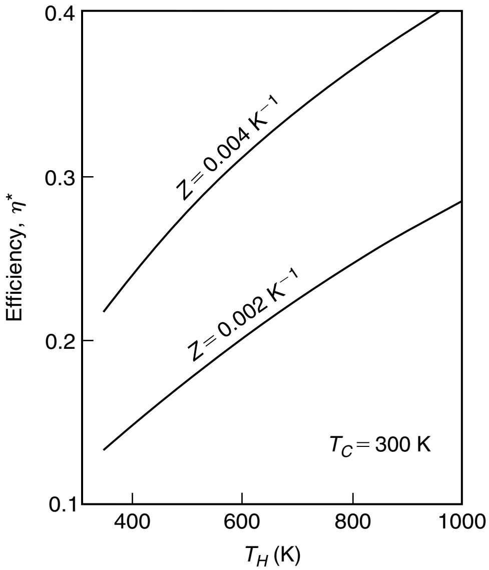


Figure 5.9 Efficiency of thermocouples vs. temperature.

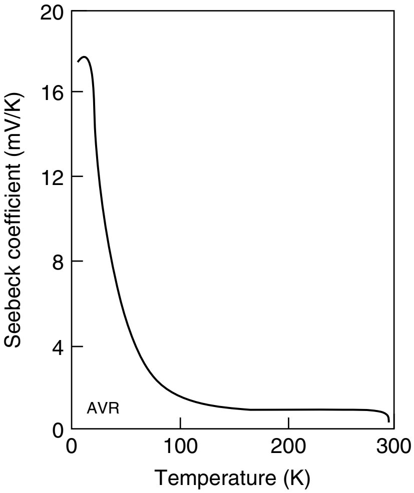


Figure 5.10 The Seebeck coefficient of germanium is large at low temperatures.

The absolute α's for other materials were determined by measuring their thermoelectric voltage against lead. Once the Seebeck coefficient is known, the figure of merit can be written as

$$Z = \frac{\alpha^2}{\Lambda R} = \alpha^2 \frac{\sigma A}{\ell} \frac{\ell}{\lambda A} = \alpha^2 \frac{\sigma}{\lambda}. \tag{5.32}$$

In this formula, all parameters refer to a single material.

It is clear that to maximize the figure of merit, one has to choose a material with the highest α and the smallest possible λ/σ. Unfortunately, this ratio is approximately the same for all metals. We will examine this question in some detail in the next section.

5.5 The Wiedemann–Franz–Lorenz Law

In the nineteenth century, physicists had difficulties measuring the thermal conductivity of materials. Gustave Heinrich Wiedemann (1826–1899) observed in 1853 that, at least for metals, the ratio, λ/σ, appeared to be constant. If so, the thermal conductivity could be inferred from the easily measure electric conductivity. Eventually, a "law" was formulated in collaboration with Rudolf Franz (1827–1902) and Ludwig Valentine Lorenz (1829–1891) expressing the relationship between thermal conductivity, electric conductivity and absolute temperature. This is the Wiedemann–Franz–Lorenz law, which can be justified based on a simple classical model of electric conduction.

Consider the heat conduction in a unidimensional gas along which there is a temperature gradient (Figure 5.11). A surface, A, at the origin of coordinates is normal to the molecular motion. We assume that there is no net mass flow. Then,

$$n_1 v_1 + n_2 v_2 = 0. \tag{5.33}$$

We now use the symbol, ℓ, as the mean free path (not as the length of the arms, as before). The molecules that cross A coming from the left originate, on the average, from a point of coordinate $-\ell/2$. Their kinetic energy is $U(-\ell/2)$. Those that come from the right have an energy $U(\ell/2)$. Half the molecules move to the left and half to the right. The net energy flux, that

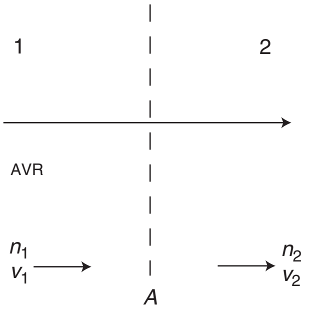


Figure 5.11 Thermal conductivity in a unidimensional gas.

is, the power density, is

$$P = \frac{nv}{2}\left[U(-\ell/2) - U(\ell/2)\right]$$
$$= -\frac{nv}{2}\frac{\partial U}{\partial x}\ell \quad \mathrm{W/m^2}. \tag{5.34}$$

The energy of each molecule is $\frac{1}{2}kT$, therefore

$$P = -\frac{nv}{4}\ell k\frac{\partial T}{\partial x} \equiv -\lambda\frac{\partial T}{\partial x}. \tag{5.35}$$

Here, the quantity, λ, is the thermal conductivity as before.

If we assume that the electrons are the only conveyers of heat in a metal (a reasonable assumption) and that they act as a gas with each electron carrying $\frac{3}{2}$kT units of energy, then the heat conductivity should be

$$\lambda = \frac{3nv\ell k}{4}. \tag{5.36}$$

The factor 3 was included to account for the three degrees of freedom of electrons in a three-dimensional gas. Actually, this overestimates the thermal conductivity because we did not correctly consider the statistical number of electrons of a three-dimensional gas that cross a given surface per unit time. Although the numerical results are incorrect, the influence of the different physical parameters on the conductivity is correctly represented.

Let us now examine the electric conductivity, σ,

$$\sigma = qn\mu, \tag{5.37}$$

where μ, the mobility, is the velocity a **carrier** (a mobile charge) attains under the influence of a unit electric field; it is the ratio of the **drift velocity**, v_d, to the electric field, E:

$$\mu = \frac{v_d}{E}. \tag{5.38}$$

Under the usual assumption that collisions are isotropic, after each collision, the velocity of the electron is statistically zero (because it has equal probability of going in any direction). This being the case, the average drift velocity of an electron is $\frac{1}{2}at$, where a is the acceleration, qE/m, and t is the mean free time, ℓ/v. Remember that v is the *thermal* velocity of the electron and is generally much larger than its *drift* velocity, v_d.

$$\mu = \frac{q\ell}{2mv} \tag{5.39}$$

and

$$\sigma = \frac{q^2 n\ell}{2mv}. \tag{5.40}$$

The λ/σ ratio becomes

$$\frac{\lambda}{\sigma} = \frac{3mv^2 k}{2q^2} = \frac{3k^2 T}{2q^2} \tag{5.41}$$

because

$$mv^2 = kT. \tag{5.42}$$

The correct ratio is

$$\frac{\lambda}{\sigma} = \frac{\pi^2}{3}\frac{k^2 T}{q^2} \equiv LT = 2.44 \times 10^{-8} T. \tag{5.43}$$

This expression is called the **Wiedemann–Franz–Lorenz** law, and the constant, $L = 2.44 \times 10^{-8}$ $(\mathrm{V/K})^2$, is the **Lorenz number**.

In the above derivation, we assumed that the only mechanism of heat conduction is electronic. This is approximately true for metals. We are going to show that in semiconductors the major heat transport is not by electrons but, rather, by phonons. In such materials, the heat conductivity is the sum of λ_C, the conductivity owing to conduction electrons plus λ_L, the conductivity owing to lattice vibrations or phonons. However, the Lorenz number is related only to λ_C:

$$L = \frac{1}{T}\frac{\lambda_C}{\sigma}. \tag{5.44}$$

If one attempted to reduce the heat conductance by using materials of low thermal conductivity, one would automatically increase the electric resistance of the thermocouple because of the proportionality between σ and λ. Thus, no improvement can be expected from manipulating these properties, and therefore, in most metallic conductors, the figure of merit depends only on the Seebeck coefficient and the temperature:

$$Z = \frac{\alpha^2}{L}\frac{1}{T} = 4.1 \times 10^7 \frac{\alpha^2}{T}. \quad \mathrm{K}^{-1} \tag{5.45}$$

Observe that the larger L, the smaller Z.

The Wiedemann–Franz–Lorenz law is not actually a law; it is rather an approximation that applies (roughly) to many metals but not to materials in general. Similarly, the Lorenz number is not constant as predicted by the theory but varies from metal to metal, as shown in Table 5.2.

The precision of the numbers in Table 5.2 is questionable as they were calculated from separate measurements of the thermal conductivity and

Table 5.2 The Lorenz Number for Some Metals

Metal	L $(\mathrm{V/K})^2 \times 10^8$	Metal	L $(\mathrm{V/K})^2 \times 10^8$
Ag	2.29	Na	2.18
Al	2.10	Ni	2.03
Au	2.53	Os	3.00
Be	1.60	Pb	2.51
Cd	2.44	Pd	2.62
Co	2.11	Pt	2.57
Cr	4.56	Sn	2.75
Cu	2.13	Ta	2.37
Fe	2.68	Ti	3.45
Gd	5.07	W	3.24
Hg	2.82	Zn	2.32
Ir	2.65	Zr	3.10
K	2.33		
Mg	1.71		
Mo	2.65		

the electric resistivity published in the *Handbook of Chemistry and Physics* (CRC). Because different samples were likely used in the two measurements and because the conductivities are very sensitive to the degree of impurity, one cannot have great confidence in the ratios displayed.

In addition, the measurement of thermal conductivity is probably not accurate beyond the second significant figure. More importantly, the assumption that electrons are the only carriers of heat in a solid is only an approximation for metals and is not valid for most other materials, as we will discuss in more detail later in this text.

5.6 Thermal Conductivity in Solids

In solids, heat propagates via two different mechanisms:

1. It is conducted by the same carriers that transport electric charges. The corresponding thermal conductivity, λ_C, is called **carrier conductivity**. It is related to the electric conductivity by the Wiedemann–Franz–Lorenz law.
2. It is also conducted by thermal vibrations of the crystalline lattice (i.e., by phonons) propagating along the material. This conductivity is called the **lattice conductivity**, λ_L.

The first mechanism is dominant in metals as they have an abundance of carriers and have relatively soft lattices. However, the opposite

Table 5.3 Thermal Conductivities Some Semiconductors

W/(m K) **Room temperature** **Doping corresponding to** $\boldsymbol{\alpha}$ **= 200** $\boldsymbol{\mu}$**V/K**			
Material	$\boldsymbol{\lambda_L}$	$\boldsymbol{\lambda_C}$	$\boldsymbol{\lambda_L/\lambda_C}$
Silicon	113	0.3	377
Germanium	63	0.6	105
InAs	30	1.5	20
InSb	16	1	16
BiTe	1.6	0.4	4

phenomenon characterizes semiconductors. For example, diamond has negligible carrier concentration and consequently has negligible electric conductivity. Nevertheless, diamond exhibits a thermal conductivity 11 times greater than that of aluminum and 30 times that of iron. In fact, it is the best heat conductor of all naturally occurring substances.[†]

The total thermal conductivity is the sum of the carrier and the lattice conductivities,

$$\lambda = \lambda_C + \lambda_L. \tag{5.46}$$

Therefore, the figure of merit becomes

$$Z = \frac{\alpha^2 \sigma}{\lambda_C + \lambda_L} = \frac{\alpha^2}{\lambda_C/\sigma + \lambda_L/\sigma} = \frac{\alpha^2}{LT + \lambda_L/\sigma}. \tag{5.47}$$

As a consequence, the value of Z in Equation 5.45 is an upper limit.

It can be seen from Table 5.3 that mechanically hard semiconductors (e.g., silicon and germanium) have lattice conductivities that are orders of magnitude larger than their respective carrier conductivities. For example, the lattice conductivity for silicon is 400 times larger than its carrier conductivity. On the other hand, the ratio of these conductivities in soft semiconductors tends to be smaller (e.g., approximately 4, for BiTe).

Compared with metals, semiconductors have an unfavorable λ/σ ratio, but have such a large advantage in α that they are the material invariably used in thermoelectric generators, refrigerators, and heat pumps. Metals are exclusively used in thermometry.

[†] Synthetic crystals, such as silicon nitride (SiN) and aluminum nitride (AlN), when carefully prepared, may have thermal conductivities exceeding that of diamond. Diamonds prepared by chemical vapor deposition (CVD), available from Fraunhoff IAF in wafers of up to 15 cm diameter and more than 2 mm thickness, have a conductivity of 5300 W/(m K) at 118 K and of 2200 W/(m K) at 273 K. Compare with copper, which, over this same range of temperatures, has about 380 W/(m K).

Table 5.4 Seebeck Coefficient for Most Metals μV/K, Temperature: 300 K

Ag	1.51	Eu	24.5	Nb	−0.44	Sr	1.1
Al	−1.66	Fe	15	Nd	−2.3	Ta	−1.9
Au	1.94	Gd	−1.6	Ni	−19.5	Tb	−1
Ba	12.1	Hf	5.5	Np	−3.1	Th	−3.2
Be	1.7	Ho	−1.6	Os	−4.4	Ti	9.1
Ca	10.3	In	1.68	Pb	−1.05	Tl	0.3
Cd	2.55	Ir	0.86	Pd	−10.7	Tm	1.9
Ce	6.2	K	−13.7	Rb	−10	U	7.1
Co	−30.8	La	1.7	Re	−5.9	V	0.23
Cr	21.8	Lu	−4.3	Rh	0.6	W	0.9
Cs	−0.9	Mg	−1.46	Ru	−1.4	Y	−0.7
Cu	1.83	Mn	−9.8	Sc	−19	Yb	30
Dy	−1.8	Mo	5.6	Sm	1.2	Zn	2.4
Er	−0.1	Na	−6.3	Sn	−1	Zr	8.9

5.7 Seebeck Coefficient of Semiconductors

No metal has a Seebeck coefficient larger than 100 μV/K. The great majority has coefficients much smaller than 10 μV/K, as can be seen from Table 5.4. Some semiconductors have coefficients of some 300 μV/K at usable temperatures. Since the figure of merit depends on the square of the Seebeck coefficient, one can see that some semiconductors have an order of magnitude (or more) advantage over metals.

The data in Table 5.4 were taken from the CRC *Handbook of Thermoelectronics*, which, in turn, obtained the values from different sources. All elements in the table are metals. The Seebeck coefficient for semiconductors depends critically on the doping level. The polarity of the Seebeck effect depends on whether the semiconductor is of the p or n type. Intrinsic semiconductors have zero Seebeck coefficients. For small doping concentrations, the Seebeck effect grows rapidly, with the doping level reaching a peak and then decreasing again as depicted in Figure 5.12.

5.8 Performance of Thermoelectric Materials

One should not select thermoelectric materials based solely on their figure of merit. A given material with good Z may be of little value because it has too low a melting point. Table 5.5 shows the figure of merit, Z, the maximum operating temperature, T_H, the ZT_H product and the "efficiency," η, for several materials. The "efficiency" was calculated from

$$\eta = \frac{(1 + Z < T >)^{1/2} - 1}{(1 + Z < T >)^{1/2} + T_C/T_H} \frac{T_H - T_C}{T_H}, \tag{5.48}$$

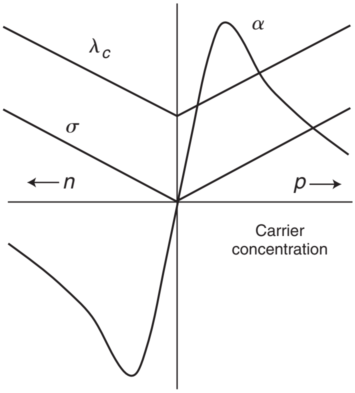


Figure 5.12 Seebeck coefficient and electric and heat conductivities of a semiconductor as a function of carrier concentration.

Table 5.5 Performance of Some Thermoelectric Materials

Material	**Z ($K^{-1} \times 10^3$)**	**T_H (K)**	**ZT_H**	**η**
Bi_2Te_3	2.0	450	0.9	5.4%
$BiSb_4Te_{7.5}$	3.3	450	1.5	7.6%
Bi_2Te_2Se	2.3	600	1.38	11.1%
PbTe	1.2	900	1.08	12.6%
CeS(+Ba)	0.8	1300	1.04	14.3%

where T_C was 300 K. Although $BiSb_4Te_{7.5}$ has the highest figure of merit of the materials in the table, its low maximum operating temperature makes it relatively inefficient. On the other hand, lead telluride, with a much smaller figure of merit, can operate at 900 K and can, therefore, yield 12.6% efficiency when its cold side is at 300 K.

The product, ZT, is the **dimensionless figure of merit**. Each semiconductor has a range of temperature over which its ZT is best. For low temperatures around 100 C, Bi_2Te_3 is normally used. At much higher temperatures around 500 C, PbTe, popular in the 1960s, was the choice. For this range, the so-called TAGS (tellurium–antimony–germanium–silver) is now preferred, as is the Michigan State University synthesized family of lead–antimony–silver–tellurium alloy (LAST) with ZT=1.4 at some 450 C. The Jet Propulsion Laboratory, JPL, (2003) developed Zn_4Sb_3, an excellent solution for the intermediary range of around 350 C. For the very high operating temperature of the radioisotope thermal generators used in space missions to the outer planets (where there is insufficient sunlight to operate photovoltaic panels), the thermoelectric material used is the SiGe alloy.

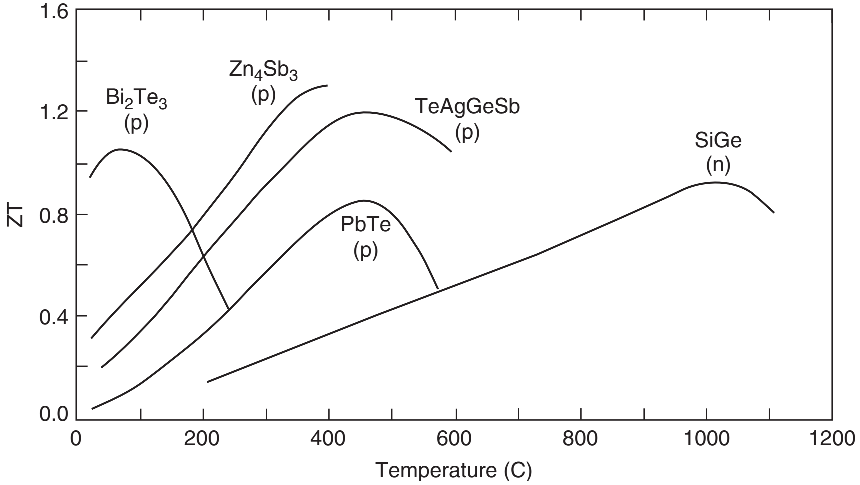


Figure 5.13 For each semiconductor, ZT is best over a given temperature range.

The data in Figure 5.13 are mostly for p-type semiconductors. PbTe, for instance, when doped with sodium, forms a p-type semiconductor, and when doped with lead iodide, it forms an n-type semiconductor. The data for an SiGe material is for the n-type variety. Of course, to build a thermocouple, one needs materials of both polarities.

In the last 50 years, by tinkering with the composition of alloys, it has been possible to come up with thermoelectric materials of acceptable performance such as the currently popular bismuth antimony tellurium alloys. In the 1990s, the idea of using skutterudides was launched.

Skutterudite is the generic name of a cobalt and nickel ore found near Skuterud, in Norway, exemplified by (Co, Ni, Fe)As_3. It has a somewhat peculiar type of crystallization: each unit cell has two relatively large voids. The carriers in the crystal have high mobility and high effective mass leading to a large α^2/ρ ratio. Unfortunately the material also has a large lattice heat conductivity. It was suggested that filling the voids with weakly bound atoms might reduce this heat conductivity because the phonons would be scattered by the "rattling" atoms in their oversized cages. This indeed happens, leading, when "filled" with some of the lanthanides, to attractive thermoelectric materials having a respectable ZT at 500 C or so. Skutterrudites approach the ideal of "electron crystal and phonon glass" in the sense that they can have a well-ordered lattice as far as carrier mobility is concerned and yet, for heat, act more like an amorphous substance such as glass, a notoriously bad heat conductor.

Shopping for appropriate bulk materials has resulted in a ZT $\approx$ 1 but seems now to have come to a dead end. Experimenters have turned to more sophisticated solutions, such as the use of superlattices. Much larger values of ZT at room temperature have been demonstrated in the laboratory. Indeed, Venkatasubramanian et al. (2001) have reported a p-type Bi_2Te_3/Sb_2Te_3 superlattice device with a ZT of 2.4 at room temperature.

Superlattices consist of alternating layers of two different semiconductors such as the Bi_2Te_3/Sb_2Te_3 or $PbSe_{0.98}Te_{0.2}/PbTe$ combinations. Superlattices and other confined conduction devices exploit the fact that frequently the mean free path of carriers is substantially smaller than that of phonons. In heavily doped silicon, for instance, the mean free path of electrons is typically 100 nm, while that of the phonons is 300 nm. If the layer spacing is made smaller than the mean free path of phonons, and larger than the mean free path of carriers, the latter can move relatively unimpeded while the phonons are scattered by the interlayer boundaries and have their propagation interfered with. Consequently, electric conductivity and carrier heat conductivity are (relatively) undisturbed, but lattice heat conductivity is reduced. If such low heat conduction lattices can be made with αs of 350 μV/K, then thermocouples would reach the golden target of a $ZT \approx 5$, which, according to experts, would open an immense field of practical applications for these devices.

Superlattice technology may be expensive to implement. An inexpensive way of achieving the same effect, at least partially, has been proposed by Poudel et al. (2008). They took one of the currently popular thermoelectric alloys, BiSbTe, and, using a ball mill, ground it to a very fine powder with average particle size of about 20 nm (some particles were smaller than 5 nm). This powder was hot pressed, forming bulk samples whose properties were then determined. As expected, owing to the small dimensions of the individual grains, phonon propagation was measurably reduced. Somewhat unexpectedly, the electric conductivity was *raised.* Observe that this requires that the grinding take place in an inert atmosphere; otherwise an electrically resistant oxide layer is created covering each particle. The Seebeck coefficient was also modified (it became somewhat smaller than that of the original alloy at temperatures below 200 C and somewhat larger than the original, at higher temperatures). The technique was effective: a value of ZT of 1.2 was measured at room temperature; it rose to 1.4 at 100 C and was still 0.8 at 250 C, a significant improvement over the original commercial alloy used in the manufacture of the material.

That nanotechnology may be the key to better thermoelectric materials is illustrated by the efforts of Boukai et al. (2008) and of Hochbaum et al. (2008), who have been trying to transform silicon (a terrible thermoelectric material due to its very large lattice heat conduction) into an acceptable material. The motivation is the cheapness of the raw silicon and the enormous technical knowhow accumulated in industrially handling it. By using nanowires, both authors have greatly reduced the undesirably high lattice heat conductivity. Unfortunately, this also reduced (by much less) the electric conductivity, partially counteracting the improvement resulting from lower heat conductivity. More significantly, it also increased α, which, being squared in the expression for Z, has great influence on the performance of the device. This boost in α is the result of the phonon-carrier drag that becomes important under the confined conditions in nano structures. Essentially, the phonon flux impels the carriers toward the colder end of the device, reinforcing the effect of the temperature on the carrier density distribution as happens in germanium of low temperature, Figure 5.10. See the explanation of the Seebeck effect in Subsection 5.14.1.

5.9 Some Applications of Thermoelectric Generators

The unparalleled reliability and simplicity of the thermoelectric generators make them the preferred device in applications in which unattended operation is more important than efficiency. These applications include:

1. Power supplies for spacecraft that operate too far from the sun to take advantage of photovoltaics.
2. Topping cycles for stationary power plants (potentially).
3. Generators for oil-producing installations, including ocean platforms.
4. Electric power providers for air-circulating fans in residential heating systems that otherwise would not operate during periods of electric power failures.
5. Power supplies for automotive use that take advantage of the heat that the engines shed.
6. Generators that produce the energy necessary to open the main valve in gas heating systems. The heat of the pilot flame activates the generator. The main gas valve will not open unless the pilot is ignited.

Thermoelectric devices are totally silent—a virtue in many cases where noises would be distracting or unacceptable, as aboard submarines, for example.

Thermoelectric generators are much less efficient than mechanical heat engines—up to a point. As pointed out by Vining (2008b), engines, unlike thermocouples, tend to become less efficient as their output power decreases, and below some 100 W, the advantage falls to thermoelectrics.

The heat necessary to drive thermoelectric generators may come from any number of sources. For example, it can come from the burning of fuel, from radioactive decay, or from reject heat, such as exhaust gases in an automobile. It appears that the recovery of reject heat in automobiles can make an important contribution to the overall efficiency of automobiles, a topic of growing importance.

Radionuclide decay is a heat source for generating electricity in space and in remote locations. Table 5.6 from a University of New York at

Table 5.6 Radionuclides Used in RTGs

Element	Half-life (years)	Specific power (kW/kg) (thermal)	Specific cost $/watt
Cesium-144	0.781	25	15
Curium-242	0.445	120	495
Plutonium-238	86.8	0.55	3000
Polonium-210	0.378	141	570
Strontium-90	28.0	0.93	250

Table 5.7 Specifications for the Radioisotope Thermal Generator (RTG) Used in the 1981 "Galileo" Mission to Jupiter

	BOL	EOM	
Heat furnished by isotope source	2460	2332	W
Heat into converter	2251	2129	W
Heat into thermocouples	2068	1951	W
Hot junction temperature	1133	1090	K
Cold junction temperature	433	410	K
Thermoelectric efficiency	11.1	10.8	%
Generator efficiency	9.4	8.6	%
Power output	230	201	W
Weight	41.7	41.7	kg
Output voltage	30	30	V
Specific power	5.52	4.82	W/kg

Stony Brook Web page (Mechanical Engineering) lists some of the radionuclides used.

Long-duration space missions invariably use plutonium-238 owing to its long half-life, although this is an extremely high-cost fuel amounting to many million dollars per RTG. For ground use, strontium-90 is preferred. Strontium-90 is, indeed, the radionuclide that powers the controversial 500 or so RTGs installed by the former Soviet Union along the coast of the Kola Peninsula (bordering on Finland and Norway).

The Radioisotope Thermal Generator (RTG) that powered the "Galileo" missions to the outer planets represented the state of the art for thermoelectric power sources in 1978. It used the then novel selenium-based semiconductors. The specifications for this RTG listed in Table 5.7 correspond to both beginning-of-life (BOL) and end-of-mission (EOM) conditions. The EOM conditions are not the same as end of life, which is many years longer. In this particular case, BOL is 1000 hours after fueling, and EOM is 59,000 hours (almost seven years) after fueling.

A ΔT of about 700 K was kept throughout the mission. The thermoelectric efficiency degraded only slightly in the seven years of operation.

The advantage of an RTG is that it is extremely light if one considers that it includes both the electrical generator and the fuel for many years of operation. Even the lightest possible gasoline engine would be orders of magnitude heavier. A large airplane gasoline engine may deliver 1500 W/kg, but one must add the mass of the fuel and oxygen needed for longtime operation. The specific consumption of a gasoline engine is about 0.2 kg $\text{hp}^{-1}\text{h}^{-1}$.[†] For each kilogram of gasoline, the engine uses 3.1 kg of O_2. The specific consumption of *fuel plus oxygen* is 0.8 kg $\text{hp}^{-1}\text{h}^{-1}$. Since 200 W correspond to 0.27 hp, the hourly consumption of a gasoline-driven generator that delivers 200 W is 0.24 kg of consumables. During the 59,000 hours of the mission, 14,000 kg of consumables would be used up. Thus, these consumables alone would mass over 3000 times more than the whole RTG. The latter, having no moving parts, requires no maintenance, while, on the other hand, it is inconceivable that a gasoline engine could possibly operate unattended for seven long years.

The low efficiency of thermocouples (less than 10%) has prompted NASA to investigate other ways of converting the heat of radioactive decay into electrical energy. One solution being actively pursued is the use of a free-piston Sterling engine (see Chapter 3) that can be made to run for decades without maintenance, thanks to hermetically sealed, lubrication-free arrangements. The oscillating piston of the engine is directly attached to a linear alternator. Efficiency of 24% (500 watts of heat input converted to 120 W of electrical output) has been demonstrated. This is important because for a given amount of electric energy required by the spacecraft, the mass of the generator becomes much smaller, as does the (very high) cost of the fuel.

The device is, of course, not an RTG; it is an SRG—a **Stirling Radioisotope Generator**. Whereas the real RTG inherently is totally silent and vibrationless, the SRG has to be designed with great care to avoid unacceptable levels of mechanical noise.

5.10 Design of a Thermoelectric Generator

Example

A thermoelectric generator is to furnish 100 kW at 115 V. Input temperature is 1500 K, while the output is at 1000 K. This output temperature is high enough to drive a steam plant—the thermoelectric generator is to serve as a **topping cycle**. See Chapter 3.

(*Continues*)

[†] One of the most economical piston aircraft engines ever built was a "turbo compound" engine that powered the Lockheed Constellation. Its specific consumption was 0.175 kg $\text{hp}^{-1}\text{h}^{-1}$.

(*Continued*)

The characteristics of materials of the thermocouple are:

Seebeck coefficient, averaged over the temperature range of interest:	0.0005 V/K
Electric resistivity of arm A:	$0.002\,\Omega\,\text{cm}$
Electric resistivity of arm B:	$0.003\,\Omega\,\text{cm}$
Thermal conductivity of arm A:	$0.032\,\text{W}\,\text{cm}^{-1}\,\text{K}^{-1}$
Thermal conductivity of arm B:	$0.021\,\text{W}\,\text{cm}^{-1}\,\text{K}^{-1}$
Maximum allowable current density:	$100\,\text{A}\,\text{cm}^{-2}$

To simplify the construction, arms A and B must have equal length (but not necessarily equal cross section). Calculate:

1. the maximum thermal efficiency,
2. the number of thermocouples in series,
3. the dimensions of the arms,
4. the open-circuit voltage, and
5. the heat input and the rejected heat at
    - 5.1. full load, and
    - 5.2. no load.

Solution:
$V_{oc} \equiv$ open-circuit voltage per thermocouple.

$$V_{oc} = \alpha(T_H - T_C) = 0.0005 \times (1500 - 1000) = 0.25\,\text{V}. \qquad (5.49)$$

$I \equiv$ current through each thermocouple (same as the total current through the battery because all elements are in series).

$$I = \frac{100,000\,\text{W}}{115\,\text{V}} = 870\,\text{A}. \qquad (5.50)$$

If there are n thermocouples, each with a resistance, R, then

$$nV_{oc} - nRI = 115\,\text{V}. \qquad (5.51)$$

To find n, we must know R. For maximum efficiency, the load resistance, R_L must be equal to mR_{batt}, or $R_L = mnR$, where $m = \sqrt{1 + < T > Z}$ (see Equations 5.28 and 5.29). Here, R is the resistance of each thermocouple, and R_{batt} is the resistance of the whole battery—that is, it is nR.

$$R_L = \frac{115\text{V}}{870\text{A}} = 0.132\,\Omega, \qquad (5.52)$$

(*Continues*)

(*Continued*)

$$Z = \frac{\alpha^2}{\Lambda R}, \tag{5.53}$$

$$\Lambda R = \left[\sqrt{\lambda_A \rho_A} + \sqrt{\lambda_B \rho_B}\right]^2 = \left[\sqrt{0.032 \times 0.002} + \sqrt{0.021 \times 0.003}\right]^2$$

$$= 254 \times 10^{-6}\,\mathrm{V}^2/\mathrm{K}, \tag{5.54}$$

$$Z = \frac{0.0005^2}{254 \times 10^{-6}} = 980 \times 10^{-6}\ \mathrm{K}^{-1}, \tag{5.55}$$

$$< T >= \frac{1500 + 1000}{2} = 1250\,\mathrm{K}, \tag{5.56}$$

$$m = \sqrt{1 + 980 \times 10^{-6} \times 1250} = 1.49, \tag{5.57}$$

$$nR = \frac{R_L}{m} = \frac{0.132}{1.49} = 0.0886\,\Omega, \tag{5.58}$$

$$n = \frac{115 + nRI}{V_{oc}} = \frac{115 + 0.0886 \times 870}{0.25} = 768.3. \tag{5.59}$$

We will need a total of 768 thermocouples.

$$V_{OC} \equiv \text{open-circuit voltage of the battery}$$

$$= nV_{oc} = 0.25 \times 768 = 192\,\mathrm{V}, \tag{5.60}$$

$$P_{H_{no\ load}} = \Lambda_{batt}(T_H - T_C). \tag{5.61}$$

$\Lambda_{batt} = n\Lambda$ because all the thermocouples are thermally in parallel.

$$\Lambda = \frac{\Lambda R}{R} = \frac{254 \times 10^{-6}}{0.0886/768} = 2.20\mathrm{W/K}, \tag{5.62}$$

$$\Lambda_{batt} = 768 \times 2.20 = 1690\,\mathrm{W/K}, \tag{5.63}$$

$$P_{H_{no\,load}} = 1690(1500 - 1000) = 846\,\mathrm{kW}, \tag{5.64}$$

$$P_{C_{no\,load}} = 846\,\mathrm{kW}, \tag{5.65}$$

$$P_{H_{full\,load}} = 846 + n\alpha T_H I - \frac{1}{2} I^2 nR$$

$$= 846 + 768 \times 0.0005 \times 1500 \times 870 \times 10^{-3}$$

$$-\frac{1}{2} \times 870^2 \times 0.0886 \times 10^{-3} = 1310\,\mathrm{kW}, \tag{5.66}$$

(*Continues*)

(*Continued*)

$$\eta = \frac{100}{P_H} = \frac{100}{1310} = 0.076 \tag{5.67}$$

$$P_{C_{full\,load}} = P_H - 100 = 1310 - 100 = 1210 \text{ kW}. \tag{5.68}$$

Since the length of the two arms is the same, Equation 5.25 simplifies to

$$\frac{A_B}{A_A} = \sqrt{\frac{\lambda_A \rho_B}{\lambda_B \rho_A}} = \frac{\sqrt{0.032 \times 0.003}}{0.021 \times 0.002} = 1.51. \tag{5.69}$$

For $J_{max} = 100$ A cm^{-2}, the smaller of the two cross sections, A_A, must be equal to $870/100 = 8.7$ cm^2. The larger cross section must be $A_B = 1.51 \times 8.7 = 13.1$ cm^2. The resistance of each individual thermocouple is

$$R = \frac{nR}{n} = \frac{0.0886}{768} = 0.000115\,\Omega, \tag{5.70}$$

$$0.000115 = \rho_A \frac{\ell}{A_A} + \rho_B \frac{\ell}{A_B} = \left(\frac{0.002}{8.7} + \frac{0.003}{13.2}\right)\ell, \tag{5.71}$$

which leads to $\ell = 0.36$ cm.

The two arms have a rather squat shape.

If the heat rejected at 1000 K is used to drive a steam turbine having 30% efficiency, the electric power generated by the latter will be $0.3 \times 1210 = 363$ kW. Adding to this the 100 kW from the thermocouple, we will have a total of 463 kW and an overall efficiency of

$$\eta = \frac{463}{1310} = 0.35. \tag{5.72}$$

It can be seen that thermocouples can be used as acceptable topping engines.

5.11 Thermoelectric Refrigerators and Heat Pumps

The Peltier effect is reversible: the direction of the heat transport depends on the direction of the current. Heat can be transported from the cold to the hot side of the thermocouple, which, consequently, can act as a heat pump or a refrigerator. We will investigate how much heat can be transported. For the sake of simplicity we will make the (not completely realistic) assumption that α, R, and Λ are all temperature independent.

5.11.1 Design Using an Existing Thermocouple

If a given thermocouple battery is available, then, presumably, the values of α, R, and Λ are known. Assume, as an example, that $\alpha = 0.055$ V/K, $R = 4.2\ \Omega$, and $\Lambda = 0.25$ W/K. Assume also that heat is to be pumped from $T_C = 278\,\mathrm{K}$ to $T_H = 338\,\mathrm{K}$, a ΔT of 60 K.

Let P_C be the heat power transported from the cold source to the cold end of the thermocouple:

$$P_C = -\Lambda\Delta T + \alpha T_C I - \frac{1}{2}RI^2. \tag{5.73}$$

For the current example,

$$P_C = -15.0 + 15.29I - 2.1I^2. \tag{5.74}$$

The electrical energy required to do this pumping is

$$P_E = \alpha\Delta TI + RI^2. \tag{5.75}$$

The ratio between the pumped heat and the required electric power is called the **coefficient of performance**, ϕ_C, of the heat pump,

$$\phi_C = \frac{-\Lambda\Delta T + \alpha T_C I - \frac{1}{2}RI^2}{\alpha\Delta TI + RI^2}. \tag{5.76}$$

In a lossless thermocouple ($R = 0$ and $\Lambda = 0$), $\phi_C = T_C/\Delta T$, which is the **Carnot efficiency**, $\phi_{C_{Carnot}}$, of the heat pump. See Problem 5.34. For our example $\phi_{C_{Carnot}} = 4.63$. The actual thermocouple does not come anywhere close to this value.

Figure 5.14 shows how the power, P_C, pumped from the cold source varies with the current. If $I < 0$ (not shown in the figure), the heat is being pumped *into* the cold source. At $I = 0$ (also not shown), there is no Peltier effect and heat still flows *into* the cold side by conduction. As I increases, some Peltier pumping starts to counteract this heat conduction, and, eventually (in a properly designed device), heat will actually begin flowing *from* the cold side to the hot side. This amount of heat will initially increase as I increases, but eventually Joule losses will begin to generate so much heat that the Peltier pumping is overwhelmed. Further increases in I will result in a reduction of the heat extracted from the cold side.

It is easy to calculate what current causes maximum heat pumping:

$$\frac{dP_C}{dI} = \alpha T_C - RI = 0 \tag{5.77}$$

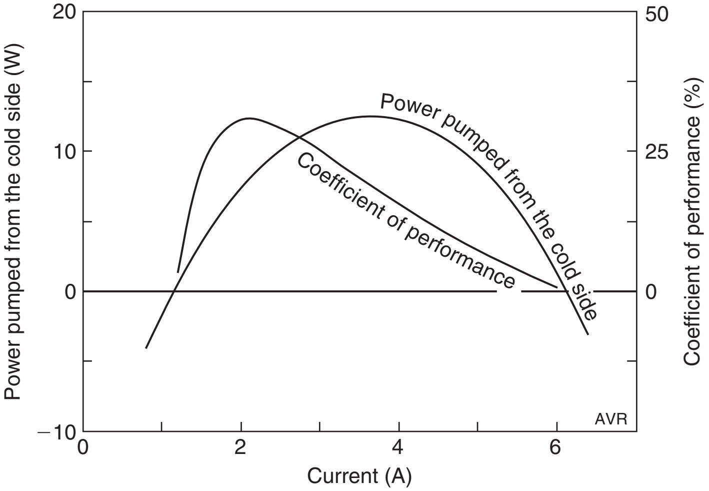


Figure 5.14 Pumped cold power and coefficient of performance as a function of current for the thermocouple of the example. Observe that the current that maximizes the pumped power is not the same as the one that maximizes the coefficient of performance.

from which

$$I_{max\,cooling} = \frac{\alpha T_C}{R}. \tag{5.78}$$

The cold power pumped when this current is used is

$$P_{C_{\max}} = -\Lambda \Delta T + \frac{\alpha^2 T_C^2}{2R}. \tag{5.79}$$

In our example, the current that maximizes the pumping is 3.64 A, and the maximum power pumped is 12.83 W.

The lowest temperature that can be reached is that at which heat just ceases to be pumped—that is, $P_C = 0$:

$$\frac{\alpha^2 T_C^2}{2R} = \Lambda(T_H - T_C) \tag{5.80}$$

from which

$$T_{C_{\min}} = \frac{-1 + \sqrt{1 + 2ZT_H}}{Z} \tag{5.81}$$

$$Z = \frac{\alpha^2}{\Lambda R} \tag{5.82}$$

which leads, for our example, to $Z = 0.00288\,\mathrm{K}^{-1}$, and $T_{C_{min}} = 249\,\mathrm{K}$. The current to achieve this is 3.26 A, but the pumped power is zero. However, any temperature above 249 K can be achieved.

With slightly more complicated math, one can find the current that maximizes the coefficient of performance:

$$\frac{d\phi_C}{dI} = (-\Lambda\Delta T + \alpha T_C I - \frac{1}{2}RI^2)(-1)(\alpha\Delta TI + RI^2)^{-2}(\alpha\Delta T + 2RI)$$
$$+ (\alpha\Delta TI + RI^2)^{-1}(\alpha T_C - RI) = 0. \tag{5.83}$$

The current that maximizes ϕ_C is

$$I = \frac{\Lambda\Delta T}{\alpha < T >}(m+1). \tag{5.84}$$

This can also be written as

$$I = \frac{a\Delta T}{R(m-1)}; \tag{5.85}$$

see Problem 5.39.

Introducing the value of I into Equation 5.76 (and after considerable algebra), one finds that the maximum value for the coefficient of performance for the thermoelectric refrigerator is

$$\phi_{C_{opt}} = \frac{T_C}{\Delta T}\left(\frac{m - T_H/T_C}{m+1}\right), \tag{5.86}$$

where $m = \sqrt{1 + Z < T >}$, as before, and $T_C/\Delta T$ is, as stated, the Carnot efficiency of the refrigerator.

Applying this to our example,

$$m = \sqrt{1 + Z < T >} = \sqrt{1 + 0.00288 \times \left(\frac{338 + 278}{2}\right)} = 1.374 \tag{5.87}$$

$$\phi_{C_{opt}} = \frac{278}{338 - 278}\left(\frac{1.374 - 338/278}{1.374 + 1}\right) = 0.308. \tag{5.88}$$

To obtain this coefficient of performance, one must use a current of

$$I = \frac{a\Delta T}{R(m-1)} = \frac{0.055 \times (338 - 278)}{4.2 \times (1.374 - 1)} = 2.10\,A. \tag{5.89}$$

Table 5.8 compares two batteries of identical thermocouples, both pumping 100 W of heat from 258 K to 323 K. One battery is adjusted to pump this heat with a minimum number of cells—that is, it operates with current that maximizes P_C. The other battery operates with the current that maximizes the coefficient of performance. The substantially larger

Table 5.8 Thermocouples Operated at Maximum P_C and at Optimum $\phi_C P_C = 100$ W

Point of operation	Number of cells	P_E (W)	P_H (W)	ϕ_C
Max P_C	100	540	640	18.5%
Opt. ϕ_C	161	336	436	29.7%

efficiency of the second battery comes at a cost of the larger number of cells required.

Commonly, the characteristics of thermocouples used as heat pumps are displayed in graphs like the one in Figure 5.15, which corresponds to the unit in our example and is roughly similar to the Tellurex CZ1-1.0-127-1.27 unit—a battery consisting of 127 cells in series. This explains the large value of α—each cell has an α of $0.055/127 = 0.000433\,\mathrm{V/K}$. For an example of how to use such graphs, see Problem 5.33.

5.11.2 Design Based on Given Semiconductors

If semiconducting materials have been selected, but the exact dimensions of the thermocouples have not yet been determined, then although the values of α, ρ, and λ are known, those of R and Λ are not. Presumably, the dimensions of the arms of the thermocouple will be optimized,

$$\Lambda R = \left[\sqrt{\lambda_A \rho_A} + \sqrt{\lambda_B \rho_B}\right]^2 \equiv \beta, \tag{5.90}$$

thus establishing a known relationship between Λ and R.

Equation 5.73 becomes

$$P_C = -\beta \Delta T \frac{1}{R} + \alpha T_C I - \frac{1}{2} R I^2. \tag{5.91}$$

If a given cooling power is desired, what is the value of R that maximizes the coefficient of performance?

$$\phi_C = \frac{P_C}{P_E} = \frac{P_C}{\alpha \Delta T I + R I^2}. \tag{5.92}$$

Solving Equation 5.91 for I,

$$I = \frac{\alpha T_C - \sqrt{\alpha^2 T_C^2 - 2\beta \Delta T - 2 P_C R}}{R}. \tag{5.93}$$

We have selected the negative sign preceding the square root because we are searching for the *least* current.

$$\phi_C = \frac{P_C R}{\alpha \Delta T \left(\alpha T_C - \sqrt{\alpha^2 T_C^2 - 2\beta \Delta T - 2 P_C R}\right) + \left(\alpha T_C - \sqrt{\alpha^2 T_C^2 - 2\beta \Delta T - 2 P_C R}\right)^2} \tag{5.94}$$

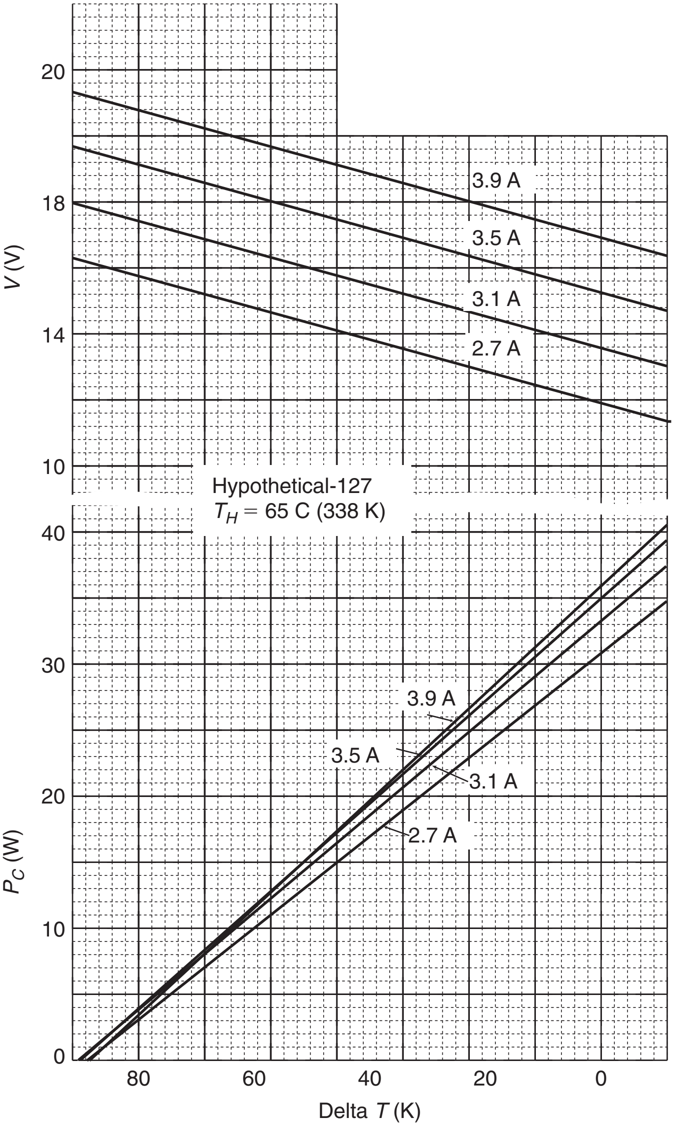


Figure 5.15 The characteristics of a given thermocouple are frequently displayed as shown.

Taking the derivative, $d\phi_C/dR$, setting it to zero, and solving for R,

$$R = \frac{-2\Delta T^2\beta(2\beta + \alpha^2 T_A) + B(\alpha^2\beta^{1/2}\Delta T T_C T_A - 2\beta^{3/2}\Delta T^2)}{\alpha^2 T_A^2 P_C}, \quad (5.95)$$

where

$$T_A \equiv \Delta T + 2T_C, \quad (5.96)$$

and

$$B \equiv \sqrt{4\beta + 2\alpha^2 T_A}. \tag{5.97}$$

As this result is sufficiently complicated to derive, it may be easier to solve the problem by trial and error using a spreadsheet.

Example

We want a refrigerator capable of removing 10 W from a cold box at −5 C, rejecting the heat to the environment at 30 C.

Owing to the temperature drops across the heat exchangers, the cold junction must be at −15 C and the hot one at 40 C.

The thermocouple materials have the following characteristics:

$$\begin{aligned} \alpha &= 0.0006\,\mathrm{V/K}, \\ \lambda_A &= 0.015\,\mathrm{W\,cm^{-1}\,K^{-1}}, \\ \rho_A &= 0.002\,\Omega\,\mathrm{cm}, \\ \lambda_B &= 0.010\,\mathrm{W\,cm^{-1}\,K^{-1}}, \text{and} \\ \rho_B &= 0.003\,\Omega\,\mathrm{cm}. \end{aligned}$$

The temperatures are

$$\begin{aligned} T_H &= 313\,\mathrm{K}\ (40\,\mathrm{C}) \text{ and} \\ T_C &= 258\,\mathrm{K}(-15\,\mathrm{C}). \end{aligned}$$

For optimum geometry,

$$\Lambda R \equiv \beta = \left[\sqrt{0.015 \times 0.002} + \sqrt{0.010 \times 0.003}\right]^2 = 120 \times 10^{-6}\,\mathrm{V^2/K}. \tag{5.98}$$

Applying Equations 5.95, 5.96, and 5.97,

$$T_A = 55 + 2 \times 258 = 571 \text{ kelvins.} \qquad \text{(from 96)}$$

$$B = \sqrt{4 \times 120 \times 10^{-6} + 2 \times 0.0006^2 \times 571} = 0.02985 \text{ V K}^{-1/2} \qquad \text{(from 97)}$$

$$\begin{aligned} R = \Big\{&-2 \times 55^2 \times 120 \times 10^{-6}\left(2 \times 120 \times 10^{-6} + 0.0006^2 \times 571\right) \\ &+ 0.02985\Big[0.0006^2 \times (120 \times 10^{-6})^{1/2} \times 55 \times 258 \times 571 \\ &-2 \times (120 \times 10^{-6})^{3/2} \times 55^2\Big]\Big\}\Big/\left(0.0006^2 \times 571^2 P_C\right) = \frac{0.00335}{P_C} \end{aligned} \qquad \text{(from 95)}$$

(*Continues*)

(*Continued*)

For this application, a single thermocouple will, draw too much current and require an inconveniently low voltage. A better strategy would be to use 100 thermocouples, connected electrically in series and thermally in parallel. Hence, we want to pump 0.1 W per thermocouple. ($P_C = 0.1$ W),

$$R = 0.0335\ \Omega. \tag{5.99}$$

The corresponding heat conductance, from Equation 5.98, is

$$\Lambda = \frac{\beta}{R} = \frac{120 \times 10^{-6}}{0.0335} = 0.00358\,\mathrm{W/K}. \tag{5.100}$$

The required current can be found from Equation 5.93:

$$I = \frac{0.0006 \times 258 - \sqrt{0.0006^2 \times 258^2 - 2 \times 120 \times 10^{-6} \times 55 - 2 \times 0.1 \times 0.0335}}{0.0335}$$

$$= 2.72\,\mathrm{A}. \tag{5.101}$$

The electric input power is

$$P_E = \alpha \Delta T I + R I^2 = 0.0006 \times 55 \times 2.72 + 0.00335 \times 2.72^2 = 0.337\,\mathrm{W}. \tag{5.102}$$

And the coefficient of performance is

$$\phi_C = \frac{0.1}{0.337} = 0.296. \tag{5.103}$$

We can obtain this same value using Equation 5.94.

We now have the required values of R and of Λ. We must determine the geometry of the two arms. It facilitates the assembly of the thermocouple if both arms have the same length, ℓ—that is, if $\ell_A = \ell_B \equiv \ell$.

$$R = \rho_A \frac{\ell}{A_A} + \rho_B \frac{\ell}{A_B}. \tag{5.104}$$

Using the values in our example,

$$\ell = \frac{0.0335}{\dfrac{0.002}{A_A} + \dfrac{0.003}{A_B}}, \tag{5.105}$$

$$\Lambda = \lambda_A \frac{A_A}{\ell} + \lambda_B \frac{A_B}{\ell}, \tag{5.106}$$

(*Continues*)

(*Continued*)

and

$$\ell = \frac{0.015A_A + 0.01A_B}{0.003580}. \tag{5.107}$$

Equating equations 5.105 to 5.107, we obtain

$$A_A = \frac{3}{2}A_B. \tag{5.108}$$

Next, we need to determine the maximum allowable current density, J_{max}. We can assume that $J_{max} = 300\,\mathrm{A/cm^2}$ and that the maximum allowable current though the thermocouple is 4 A. (It is supposed to operate at 2.7 A.) This sets an approximate area for $A_A = 4/300 = 0.013\,\mathrm{cm}^2$. The value of A_B is $0.02\,\mathrm{cm}^2$, and the length of each arm, from Equation 5.105, is 0.11 cm.

The required voltage to pump 10 W is

$$V = \frac{100P_E}{I} = \frac{100 \times 0.337}{2.72} = 12.4\,\mathrm{V}. \tag{5.109}$$

5.12 Temperature Dependence

In Section 5.11, we made the assumption that α, Λ, and R are independent of temperature. In reality, this is not the case. If we take this dependence into account, we will greatly complicate the solution of the various design problems. We will not do this in this book and shall be satisfied with the approximate results of the preceding section. Nevertheless, it is useful to acknowledge that the parameters mentioned do vary when the temperature changes. This is illustrated in Figures 5.16 and 5.17.

5.13 Battery Architecture

As we have seen, a number of characteristics of a thermocouple have to be optimized for best performance: choice of material, proper geometry of the arms, and matching the load to the generator. There is an additional factor that needs to be considered in optimizing batteries of cooling devices. T_H and T_C are typically the same for all cells of a battery. However, the coefficient of performance can be improved by making different arrangements for the heat flow. Some of these alternative arrangements are described in an article by Bell (2002).

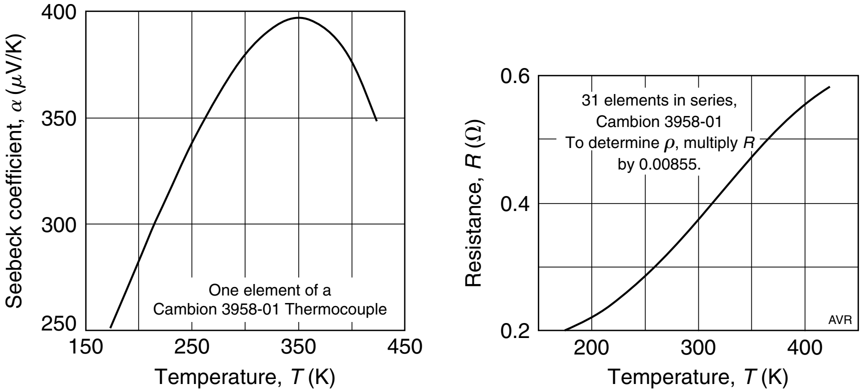


Figure 5.16 Representative behavior of the Seebeck coefficient and the resistance of a 1970 thermocouple as a function of temperature.

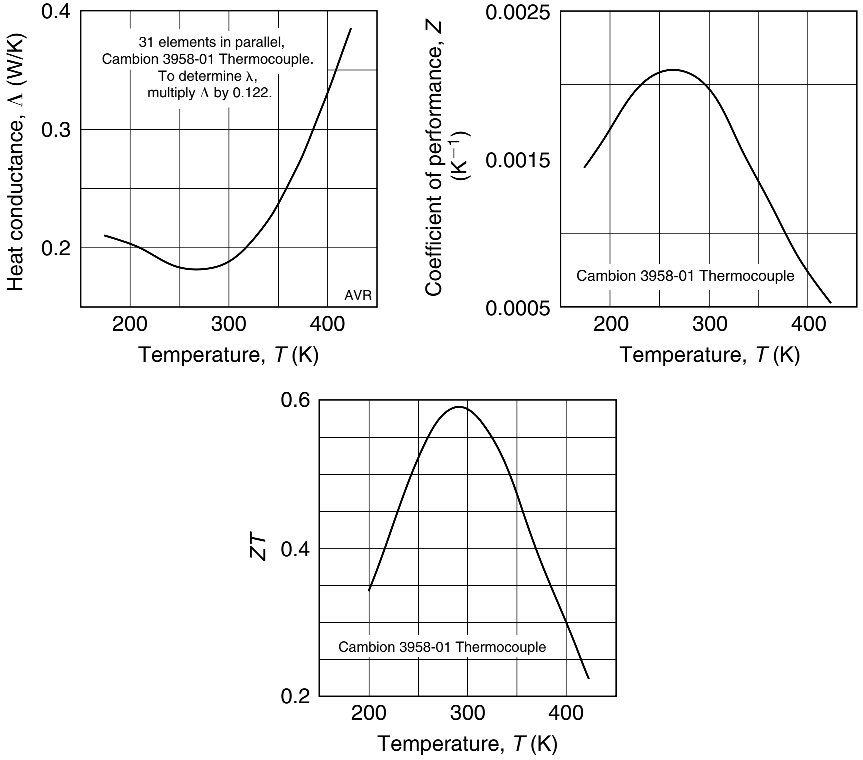


Figure 5.17 Representative behavior of the heat conductance, the coefficient of performance, and the ZT product of a thermocouple as a function of temperature. This corresponds to the state of the art in 1970. Since then, much progress has been made. The ZT values of modern materials comfortably exceeds unity, for example.

5.14 The Physics of Thermoelectricity[†]

In investigating the behavior of thermocouples, we considered four different mechanisms:

1. **Heat conduction**, a topic thoroughly familiar to most readers. It was shown that heat is conducted by both the motion of carriers and by the vibration of the crystalline lattice. Heat conduction introduces losses in the performance of thermocouple.
2. **Joule losses**. As the current flows through the device, it encounters resistance and, consequently, generates heat. This constitutes the second loss mechanism in thermocouples. Joule losses result from the scattering of carriers by lattice imperfections (thermal vibrations, impurities, dislocations, etc.). Again, it is assumed that the reader is familiar with it.
3. **Seebeck effect**, the development of a voltage in a conductor as a result of a temperature differential. It is caused by increased carrier concentration in the cold regions of a conductor. The mechanism for this is examined in Subsection 5.14.1.
4. **Peltier effect**, the absorption or the release of heat at a junction of dissimilar conductors owing to the change in heat capacity of carriers when they leave one medium and enter a different one. In Section 5.1, we stated that there is a relationship between the Seebeck and the Peltier effects. We will derive this relationship in the present section. The mechanism for the Peltier effect itself will be discussed in Subsection 5.14.2.

   So far, we have completely disregarded a fifth important effect—the Thomson effect. There is a very good reason for this omission as shall be explained in this section. We will also show how the Seebeck, Peltier, and Thomson effects are interrelated.
5. **Thomson effect**, the convection of heat by the flux of drifting carriers, will be discussed in Subsection 5.14.3.

5.14.1 The Seebeck Effect

Consider a length of pipe filled with a gas that is at uniform temperature. Clearly, both pressure and concentration are also uniform. However, if one end of the pipe is heated to a higher temperature than the other, the higher pressure on the hotter side will cause some flow of gas toward the colder end. When steady state is reestablished, the flow ceases and the pressure is again uniform. According to the perfect-gas law, $p = nkT$, constant pressure means the nT product is also constant. Consequently, the concentration of the gas will be higher at the colder end than at the hotter end of the pipe.

[†] A fuller treatment of this subject is found in Goldsmid (1986).

Conduction electrons also behave as a gas. If there is a temperature gradient, their concentration in a conducting bar will be higher in the cold side, which, as a consequence, becomes negatively charged with respect to the hotter. The resulting electric potential is the **Seebeck voltage**, a quantity that depends both on the temperature difference and on the nature of the conductor. If the carriers in the bar are holes, then the colder side will become positive with respect to the hotter. Thus, the polarity of the Seebeck voltage depends on the material of the conductor.[†]

An external connection is needed to tap into the voltage developed in the conducting bar. However, the external wire is submitted to the same temperature differential as the bar itself and will develop its own Seebeck voltage. If the external wire is made of the same material as the bar, the two voltages exactly cancel one another. If, however, the connection is made with a different material, then a net voltage may become available.

A thermocouple must always consist of two *dissimilar* materials, most often of opposing polarity so that the individual Seebeck voltages add up. Since the external connections wires are attached to the open ends of the thermocouple, which presumably are at the same temperature, these wires contribute no additional thermoelectric voltage.

Thermocouples are low-impedance devices (low voltage and large current) and, for many applications, must be connected in series forming a thermoelectric **battery**. See Figures 5.18 and 5.19.

In a thermocouple battery, individual cells are electrically in series and thermally in parallel. Thus, if Λ is the heat conductance of one single cell, then $n\Lambda$ is the heat conductance of a battery of n cells. By the same token, the resistance of the battery is nR when R is the resistance of one

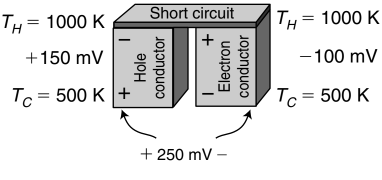


Figure 5.18 To obtain a useful output, two arms must be paired, forming a thermocouple.

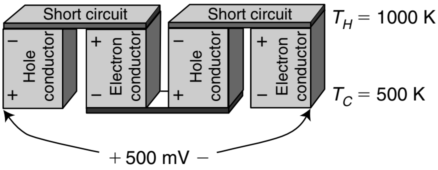


Figure 5.19 Thermocouples can be connected in series forming a battery.

[†] By convention, $\alpha < 0$ when the cold end is negative.

cell. The voltage generated per unit temperature difference (equivalent to α in a single cell) is $n\alpha$. Consequently, the figure of merit of a battery of n identical cells is

$$Z = \frac{(n\alpha)^2}{n\Lambda \times nR} = \frac{\alpha^2}{\Lambda R}, \tag{5.110}$$

the same as the figure of merit of each cell.

The above explanation of the Seebeck effect is oversimplified. It serves as a first-order model to facilitate understanding of the gross behavior of thermoelectric devices.

The migration of carriers to the cold side of a conductor creates an electric field that forces the electrons to drift back toward the hot side. A dynamic equilibrium is established when just as many carriers move under the influence of the pressure gradient as do (in the opposite direction) under the influence of the electric field.

At equilibrium, there is no *net* charge transfer from one end of the conductor to the other—the flux, nv, is the same in both directions. However, carriers moving down the temperature gradient, being more energetic, carry more heat than those moving in the opposite direction. Thus, even in the absence of a net particle flow, there is a net heat flow in the material. This explains the metallic heat conductivity.

Our simple model predicts that all electronic conductors have negative thermoelectric power—that is, a negative Seebeck coefficient. By the same token, p-type semiconductors (in which holes are the carriers) must have a positive Seebeck coefficient. The model is inadequate because, although most metals have negative αs, some, such as copper, do not.

To improve the theory, we have to consider the scattering of electrons as they move through the conductor. If the scattering cross section is temperature independent, the conclusions above hold because both fluxes (up and down the temperature gradient) are equally perturbed. However, if some mechanism causes hot electrons to be more severely scattered than cooler ones, then the flux of hot electrons is diminished and the negative thermoelectric power is reduced or even reversed. On the other hand, if the hot electrons are less scattered than the cold ones, then the negative thermoelectric power is enhanced.

Some materials exhibit a large Seebeck effect at low temperatures, as illustrated in Figure 5.10. This is due to phonon–electron interaction. When there is a temperature gradient in a material with sufficiently rigid crystalline lattice, heat is conducted as lattice waves, as discussed in Section 5.4. Such waves can be interpreted as a flux of quasiparticles called **phonons**. Here, again, we use the duality of waves and particles.

(*Continues*)

(*Continued*)

Phonons can interact with other phonons and also with electrons. At higher temperatures, phonon–phonon interaction is dominant, but at lower temperatures, phonon–electron interaction may become important. When this happens, the phonon flux (from the hot to the cold side) simply sweeps electrons along with it, causing a large charge accumulation at the cold end enhancing the negative thermoelectric power. If the material is a p-type semiconductor, then it is holes that are swept along, enhancing the positive thermoelectric power.

5.14.2 The Peltier Effect

Two different conductors, a and b, of identical cross section, A, are connected end-to-end as suggested in Figure 5.20. The surface, S, is the interface between them. Both conductors are at the same temperature, T, and a current, I, flows through them. Let the heat capacity of a typical conduction electron be c_a in conductor a and c_b in conductor b. The thermal energy associated with each (typical) electron is, respectively, c_aT and c_bT. For instance, if the electron gas obeys the Maxwellian distribution, then $c_aT = c_bT = \frac{3}{2}kT$.

The current in either side of the interface is

$$I = qnv\,A. \tag{5.111}$$

In conductor a, the current convects thermal energy toward the interface at a rate

$$P_a = nvAc_aT \tag{5.112}$$

and in conductor b, it convects thermal energy away from the interface at a rate

$$P_b = nvAc_bT. \tag{5.113}$$

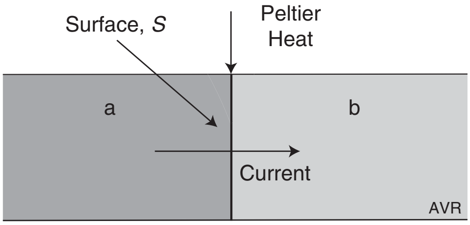


Figure 5.20 Heat convected by an electric current.

If $P_a > P_b$, then at the interface, energy must be rejected from the conductors into the environment at the rate

$$P = P_a - P_b = nvAT(c_a - c_v) = \frac{T}{q}(c_a - c_b)I \equiv \pi I. \tag{5.114}$$

For a Maxwellian electron gas, this model predicts a Peltier coefficient, $\pi = 0$, because, for such a gas, $c_a = c_b$. However, since there is a nonzero Peltier effect, one must conclude that in thermocouple materials, the electron gas does not behave in a Maxwellian manner. This agrees with the accepted non-Maxwellian model for electrons in metals. However, in lightly doped semiconductors, the conduction electrons are Maxwellian. Additional sophistication of our model is required to explain the Peltier effect in such materials.

5.14.3 The Thomson Effect

Consider again the unidimensional gas discussed in our derivation of thermal conductivity in Section 5.5. We want to derive a formula for the convective transport of heat. We will disregard heat conduction; its effect can simply be superposed on the results obtained here.

We assume that there is a net flux, nv, of molecules and that the temperature is not uniform along the gas column. Take three neighboring points: 1, 2, and 3. Each molecule that moves from 1 to 2 carries an energy cT_1. Here, c is the mean heat capacity of the molecule (i.e., c is $1/N$ of the heat capacity of N molecules). For each molecule that arrives at 2 coming from 1, another leaves 2 toward 3 carrying cT_2 units of energy. Thus, the increase in energy at 2 owing to the flow of gas must be $c(T_1 - T_2)nv$ joules per second per unit area.

If 1 and 2 are separated by an *infinitesimal distance*, then $T_1 - T_2 = -dT$ and the energy is transported at a rate

$$dP^* = -cnvdT \,\mathrm{W/m^2}, \tag{5.115}$$

where P^* is the power *density.* If instead of a gas column, we have a free-electron conductor, then heat is convected by electrons and since $J = qnv$,

$$dP^* = -\frac{J}{q}c\,dT \,\mathrm{W/m^2} \tag{5.116}$$

or

$$dP = -\frac{I}{q}c\,dT \,\mathrm{W}. \tag{5.117}$$

The above expression can be rewritten as

$$dP = \tau I dT, \tag{5.118}$$

where τ is the **Thomson coefficient** and has the dimensions of V/K. Clearly,

$$\tau = -\frac{c}{q}. \tag{5.119}$$

For conductors in which the carrier distribution is Maxwellian, $c = \frac{3}{2}k$, and the Thomson coefficient is

$$\tau = -\frac{3}{2}\frac{k}{q} = -129\mu V/K. \tag{5.120}$$

Many semiconductors do have a Thomson coefficient of about $-100\,\mu$V/K. However, a more accurate prediction of the coefficient requires the inclusion of holes in the analysis.

Electrons in metals do not obey Maxwellian statistics. As discussed in Chapter 2, they follow the Fermi–Dirac statistics: only a few electrons at the high energy end of the distribution can absorb heat. Those that do absorb energy of the order of kT units, but they represent only a fraction of about kT/W_F of the total population. Hence, the mean heat capacity of the electrons is roughly

$$c = \frac{\partial}{\partial T}\left(\frac{kT}{W_F}kT\right) = \frac{2k^2T}{W_F}. \tag{5.121}$$

The ratio of the Fermi–Dirac heat capacity to the Maxwellian is about kT/W_F.

At room temperature, kT is some 25 meV, whereas a representative value for W_F is 2.5 eV. Therefore, the quantum heat capacity is approximately 100 times smaller than the classical one. For this reason, the Thomson coefficient of metals is small compared with that of semiconductors.

5.14.4 Kelvin's Relations[†]

When we derived the formulas for the performance of thermocouples, we stated that the Peltier coefficient, π, was equal to αT. We will now prove this assertion. In addition, in developing the thermocouple formulas, we failed to introduce the Thomson effect discussed in the preceding subsection. Here, we will justify this omission. Heat flows from the hot source to the couple at a rate

$$P_H = \Lambda(T_H - T_C) + \pi_H I + I\int_{T_H}^{T_C} \tau_A dT - \frac{1}{2}I^2R. \tag{5.122}$$

[†] William Thomson was knighted Lord Kelvin in 1866, mainly in recognition of his work in transatlantic telegraphy.

Notice that in this case we have included the Thomson heat convection. The Thomson coefficients of arms A and B are, respectively, τ_A and τ_B.

We will consider a hypothetical thermocouple that has neither electric resistance nor heat conductance. It is a reversible device with no losses. Since both Λ and R are zero,

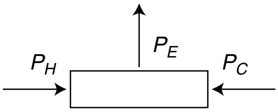
$$P_H = \pi_H I + I \int_{T_H}^{T_C} \tau_A dT. \tag{5.123}$$

In the same manner, heat flows from the sink to the couple at a rate

$$P_C = -\pi_C I + I \int_{T_C}^{T_H} \tau_B dT. \tag{5.124}$$

As there are no losses, the power, $V_L I$, delivered to the load is equal to the total heat power input, $P_H + P_C$. Moreover, because of the absence of resistance in the thermocouple, the load voltage, V_L, is equal to the open-circuit voltage, V_{oc}.

$$V_{oc} I = \pi_H I - \pi_C I + I \int_{T_C}^{T_H} (\tau_B - \tau_A) dT, \tag{5.125}$$



$$V_{oc} = \pi_H - \pi_C + \int_{T_C}^{T_H} (\tau_B - \tau_A) dT. \tag{5.126}$$

We want to find the Seebeck coefficient, α, at a given temperature. Let us hold T_C constant and see how V_{oc} varies with T_H:

$$\begin{aligned} \frac{\partial V_{oc}}{\partial T_H} &= \frac{\partial \pi_H}{\partial T_H} + \frac{\partial}{\partial T_H} \int_{T_C}^{T_H} (\tau_B - \tau_A) dT \\ &= \frac{\partial \pi_H}{\partial T_H} + \tau_B(T_H) - \tau_A(T_H). \end{aligned} \tag{5.127}$$

Since the above equation is valid for any T_H, we can replace T_H by the general temperature, T:

$$\alpha \equiv \frac{\partial V_{oc}}{\partial T} = \frac{\partial \pi}{\partial T} + \tau_B - \tau_A. \tag{5.128}$$

The entropy entering the thermocouple is

$$S_{in} = \frac{P_H}{T_H}, \tag{5.129}$$

and that leaving the thermocouple is

$$S_{out} = -\frac{P_C}{T_C}, \tag{5.130}$$

hence, the entropy change in the device is

$$\Delta S = \frac{P_H}{T_H} + \frac{P_C}{T_C} \geq 0. \tag{5.131}$$

The inequality is the result of the second law of thermodynamics. However, the thermocouple we are considering is a lossless one (isentropic); thus

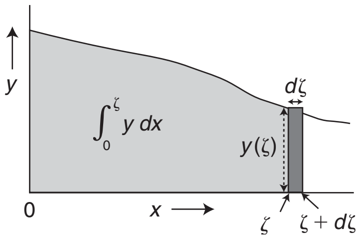
$$\Delta S = \frac{P_H}{T_H} + \frac{P_C}{T_C} = 0, \tag{5.132}$$

and $\frac{d}{dT}\Delta S = 0$, because Equation 5.132 holds for any value of T.

$$\begin{aligned} \frac{d\Delta S}{dT_H} &= \frac{d}{dT_H}\left[\frac{\pi_H I}{T_H} - \frac{\pi_C I}{T_C} + I\int_{T_C}^{T_H} \frac{\tau_B}{T} dT - I\int_{T_C}^{T_H} \frac{\tau_A}{T} dT\right] \\ &= I\left[\frac{1}{T_H}\frac{\partial \pi_H}{\partial T_H} - \frac{\pi_H}{T_H^2} + \frac{\tau_B(T_H) - \tau_A(T_H)}{T_H}\right] = 0. \end{aligned} \tag{5.133}$$

We replaced the differential of the integral with respect to its upper limit by the value of the argument of the integral at this upper limit. (See the accompanying box below. A Bit of Math.)

A Bit of Math



For those who have forgotten some of their math, here is a simple derivation of what happens when one differentiates an integral with respect to one or both limits. Consider

$$Int = \int_0^\zeta y\, dx, \tag{5.134}$$

(*Continues*)

(*Continued*)

where y is any well-behaved function of x. Refer to the figure. The integral corresponds to the light gray area and the upper limit is $x = \zeta$. We ask what happens when the limit is changed by a infinitesimal amount, $d\zeta$. Clearly, the area—the integral—increases by $y(\zeta)\,d\zeta$, which is represented by the small dark rectangle in the figure. $y(\zeta)$ is the value the function, y, assumes when $x = \zeta$ (the value of y at the upper limit). The change in the integral is

$$dInt = y(\zeta)\ d\zeta, \tag{5.135}$$

and the rate of change is

$$\frac{dInt}{d\zeta} = y(\zeta). \tag{5.136}$$

The value of the differential of an integral with respect to its upper limit is simply the value of the function at this upper limit.

Eliminating I from Equation 5.133, simplifying, and recognizing again that the result holds for any value of T_H,

$$\frac{\partial \pi}{\partial T} + \tau_B - \tau_A = \alpha = \frac{\pi}{T}. \tag{5.137}$$

The relationship between α and π used in the beginning of this chapter is thus proven correct.

The dependence of α on T is of interest:

$$\frac{\partial \alpha}{\partial T} = \frac{1}{T}\left(\frac{\partial \pi}{\partial T} - \frac{\pi}{T}\right) = \frac{1}{T}\left(\frac{\partial \pi}{\partial T} - \frac{\partial \pi}{\partial T} + \tau_A - \tau_B\right), \tag{5.138}$$

$$\frac{\partial \alpha}{\partial T} = \frac{\tau_A - \tau_B}{T}. \tag{5.139}$$

The above Expression 5.139 shows that when the two arms of a thermocouple have the same Thomson coefficients, the Seebeck coefficient is independent of temperature. This does not occur often.

The contribution of the Thomson effect to the thermocouple voltage is

$$V_\tau = \int_{T_C}^{T_H} (\tau_B - \tau_A) dT. \tag{5.140}$$

Using the relationship of Equation 5.139 (and ignoring the sign),

$$V_\tau = \int_{T_C}^{T_H} T\frac{d\alpha}{dT}dT = \int_{T_C}^{T_H} Td\alpha. \tag{5.141}$$

Integrating by parts,

$$V_\tau = (\alpha T)\Big|_{T_C}^{T_H} - \int_{T_C}^{T_H} \alpha dT. \tag{5.142}$$

The mean value of α in the temperature interval, T_C to T_H, is

$$<\alpha>= \frac{\int_{T_C}^{T_H} \alpha dT}{T_H - T_C}. \tag{5.143}$$

Hence,

$$V_\tau = \alpha_H T_H - \alpha_C T_C - <\alpha> (T_H - T_C), \tag{5.144}$$

$$V_\tau = \alpha(T_H - T_C) - <\alpha> (T_H - T_C). \tag{5.145}$$

If we use an average value for α—that is, if $\alpha_H = \alpha_C =<\alpha>$—then $V_\tau = 0$. Thus, it is possible to disregard the Thomson heat transport simply by using the **mean** Seebeck coefficient.

5.15 Directions and Signs

In the various examples discussed in this text, there has been no consistent definition of the direction in which the different powers flow into and out of a thermocouple. The directions, and the corresponding signs in the equations for P_H, P_C, and P_E, have been chosen in each case so as to best suit the problem being discussed.

Although the choice of direction for the flow of different powers is entirely arbitrary it must be consistent throughout a problem. When the thermocouple is used as a generator, the most intuitive direction is the conventional one shown in Figure 5.21, in which heat from a source flows into the couple and part of it is rejected to a sink, while electric power flows out of the device. This will cause all the powers to be larger than zero. For refrigerators and heat pumps, it is simpler to use the directions indicated in Figure 5.22 in which some heat is pumped away from a cold source by the thermocouple and is then rejected to a hotter sink. Naturally, to make heat flow from a colder to a hotter region, a certain amount of electric power is required to flow into the device.

Let us examine the equations for P_H and P_C. They consist of three terms: a conduction (Fourier) term, a Joule heating term, and a term that represents the Peltier transport of heat.

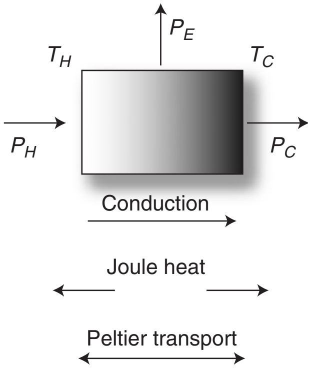


Figure 5.21 Directions of power flow in a thermocouple.

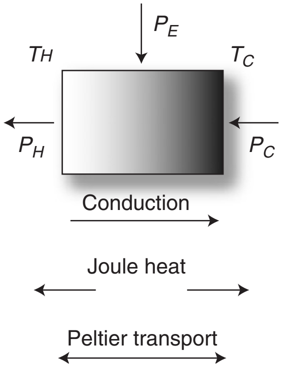


Figure 5.22 Directions of power flow in a thermocouple.

Clearly, the conduction term is positive when the flow is from T_H to T_C. In Figure 5.21, this flow is in the same direction as P_H and P_C and is therefore a positive term in the equations. In Figure 5.22, the conduction term opposes both P_H and P_C and is a negative term in the corresponding equations.

Half of the Joule heat flows out of the hot end and half out of the cold end. It is a negative term for P_H and a positive term for P_C in Figure 5.21, and it is positive for P_H and negative for P_C in Figure 5.22.

The Peltier heat can flow in either direction depending on the sign of the current applied to the thermocouple. Thus, in one of the equations (say in the equation for P_H) the sign of the Peltier term is arbitrary in both Figures 5.21 and 5.22. Once the sign is chosen in the P_H equation, then the same sign must be used in the P_C equation, provided P_H and P_C are in the same direction. The reason for this becomes obvious if we look at the equation for $P_E = P_H - P_C$ in both figures.

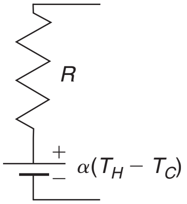


Figure 5.23 Electric model of a thermocouple.

In Figure 5.21,

$$P_H = \Lambda(T_H - T_C) + \alpha T_H I - \frac{1}{2}RI^2, \tag{5.146}$$

$$P_C = \Lambda(T_H - T_C) + \alpha T_C I + \frac{1}{2}RI^2, \tag{5.147}$$

$$P_E = \alpha(T_H - T_C)I - RI^2. \tag{5.148}$$

In Figure 5.22,

$$P_H = -\Lambda(T_H - T_C) - \alpha T_H I + \frac{1}{2}RI^2, \tag{5.149}$$

$$P_C = -\Lambda(T_H - T_C) - \alpha T_C I - \frac{1}{2}RI^2, \tag{5.150}$$

$$P_E = -\alpha(T_H - T_C)I + RI^2. \tag{5.151}$$

Notice that P_E of Figure 5.21 is the negative of that in Figure 5.22 (the arrows are inverted). Notice also that, to make the currents in the two figures consistent with one another, those in the latter figure are negative (they are in the direction of P_E, as are the currents in the former figure).

The thermocouple can be modeled (see Figure 5.23) as a voltage generator with an open-circuit voltage $\alpha(T_H - T_C)$ resulting from the Seebeck effect and an internal resistance, R. It can be seen that either equation for P_E represents such a model.

APPENDIX

The figure of merit of a thermocouple is $Z = \alpha^2/\Lambda R$ and is maximized when the ΛR product is minimum. What is the corresponding thermocouple geometry?

$$\Lambda = \frac{A_A \lambda_A}{\ell_A} + \frac{A_B \lambda_B}{\ell_B},$$

$$R = \frac{\ell_A}{A_A \sigma_A} + \frac{\ell_B}{A_B \sigma_B},$$

$$\Lambda R = \frac{A_A \lambda_A}{\ell_A}\frac{\ell_A}{A_A \sigma_A} + \frac{A_B \lambda_B}{\ell_B}\frac{\ell_A}{A_A \sigma_A} + \frac{A_A \lambda_A}{\ell_A}\frac{\ell_B}{A_B \sigma_B} + \frac{A_B \lambda_B}{\ell_B}\frac{\ell_B}{A_B \sigma_B};$$

hence

$$\Lambda R = \frac{\lambda_A}{\sigma_A} + \frac{\lambda_B}{\sigma_B} + \frac{\ell_A A_B}{\ell_B A_A}\frac{\lambda_B}{\sigma_A} + \frac{\ell_B A_A}{\ell_A A_B}\frac{\lambda_A}{\sigma_B}.$$

For a given choice of materials, $\frac{\lambda_A}{\sigma_A} + \frac{\lambda_B}{\sigma_B}$ is fixed, and a minimum ΛR occurs when $\frac{\ell_A A_B}{\ell_B A_A}\frac{\lambda_B}{\sigma_A} + \frac{\ell_B A_A}{\ell_A A_B}\frac{\lambda_A}{\sigma_B}$ is minimum. Let

$$x \equiv \frac{\ell_A A_B}{\ell_B A_A}, \quad a_1 \equiv \frac{\lambda_A}{\sigma_B}, \quad a_2 \equiv \frac{\lambda_B}{\sigma_A}.$$

Then, the minimum occurs when $\frac{d}{dx}\left(a_2 x + \frac{a_1}{x}\right) = 0$ or $a_2 = \frac{a_1}{x^2}$ which leads to

$$\frac{\ell_A A_B}{\ell_B A_A} = \sqrt{\frac{\lambda_A \sigma_A}{\lambda_B \sigma_B}}. \qquad [5.25]$$

Putting Equation 5.25 into the expression for ΛR,

$$\Lambda R_{min} = \left(\sqrt{\frac{\lambda_A}{\sigma_A}} + \sqrt{\frac{\lambda_B}{\sigma_B}}\right)^2. \qquad [5.26]$$

References

Bell, Lon E., Use of thermal insolation to improve thermoelectric system operating efficiency. International Thermoelectric Society Conference (ITC2002), Long Beach, CA, August 26–29, **2002**.

Boukai, Akram I., et al., Silicon nanowires as efficient thermoelectric materials, *Nature 451* (10), pp. 168–171, January **2008**.

Burns, G. W., and M. G. Scroger, The Calibration of Thermocouples and Thermocouple materials, National Institute of Standards and Technology (NIST), Special Publication 250–35, April **1969**.

Goldsmid, H. J., *Electronic Refrigeration*, Pion Limited, London, **1986**.

Hochbaum, Allon I., Enhanced thermoelectric performance of rough silicon nanowires, *Nature 451* (10), pp. 163–167, January **2008**.

Poudel, Bed, et al., High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys, *Science* 320, pp. 634–638, May 2, **2008**.

Venkatasubramanian, R., Edward Siivola, Thomas Colpitts, and Brooks O'Quinn, Thin-film thermoelectric devices with high room-temperature figures of merit, *Nature 413* (11) pp. 597–602, October **2001**.

Vining, Cronin B., Desperately seeking silicon, *Nature 451* (10), p. 132–133, January **2008a**.

Vining, Cronin B. (2), The limited role for thermoelectrics in the climate crisis, Solution Summit Panel on Nanotechnology and New Materials, New York City, May 1, **2008b**.

PROBLEMS

5.1 The Russians are quite advanced in thermoelectrics and have recently developed a secret material for such use. Data are naturally, hard do obtain, but the CIA discovered that a thermoelectric cooler is being built capable of depressing the temperature by 100 K when the hot junction is at 300 K. What is the figure of merit of the thermocouple?

5.2 The Seebeck coefficient of a junction is

$$\alpha = 100 + T - 10^{-3}T^2 \,\mu\text{V/K}.$$

1. At what temperature is the Peltier coefficient maximum?
2. What is the coefficient at this temperature?
3. At what temperature are the Thomson coefficients of the two arms equal?

5.3 A thermoelectric cell has a figure of merit of 0.002 K^{-1} and an internal resistance of 100 $\mu\Omega$. The average Seebeck coefficient is 200 μV/K. Heat flow meters are used to measure the flow of heat, P_H, from the hot source to the cell and the flow, P_C, from the cell to the cold sink. There is no other heat exchange between the cell and the environment.

The source is at 600 K and the sink at 300 K.

1. With no electric current flowing in the cell, what is the value of P_H and P_C?
2. A current, I, is now made to circulate. This modifies P_H. What currents cause P_H to go to zero?
3. What are the corresponding values of P_C?
4. Over what range of currents does the cell act as a power generator?

5.4 A block of metal, maintained at 850 K, is mounted on a pedestal above a platform at 350 K.

The experiment being conducted requires that there be absolutely no heat transfer from the metal block through the pedestal. To achieve this, the pedestal was made into a thermocouple and the appropriate current was driven through it. The thermocouple has a heat conductance of 1 W/K and a resistance of 1 milliohm. Its figure of merit is 0.001 K^{-1}. All these data are average values over the temperature range of interest.

What current(s) must be driven through the thermocouple? What is the voltage across its terminals? What is the electric power required?

5.5 We want to pump 100 W of heat from 210 K to 300 K. Two stages of thermocouples must be used. The *second* stage pumps 100 W from 210 K to T_{H_2}. The first stage pumps the necessary energy from $T_{C_1} = T_{H_2}$ to 300 K.

All thermocouples are made of the same materials whose combined characteristics are:

$$\alpha = 0.001\,\mathrm{V\,K^{-1}},$$
$$\Lambda\mathrm{R} = 0.0005\,\mathrm{K\,V^{2}}.$$

The geometry of each thermocouple is optimized. The current through the thermoelectric pair is adjusted for maximum cooling. The current through the first stage is not necessarily the same as that through the second.

Within each stage, all elements are connected electrically in series.

The total electric power input to the system depends critically on the choice of T_{H_2}. What value of T_{H_2} minimizes the electric power consumption, P_E? What is the value of P_E? What is the voltage applied to each stage?

5.6 A battery of thermocouples delivers 5 kW at 24.0 V to a load. The thermocouples were designed to operate at maximum efficiency under the above conditions.

The hot side of the battery is maintained at 1100 K and the cold side, at 400 K.

The figure of merit of each individual thermocouple cell is $0.0015\,\mathrm{K^{-1}}$.

What is the efficiency of the system?

5.7 What is the heat conductance of a metal bar at uniform temperature ($T = 400\,\mathrm{K}$) if the bar has a resistance of $4\,\Omega$?

5.8 To drive a current of 100 A through a thermocouple at uniform temperature (300 K), a power of 50 W is required.

The same thermocouple, when open-circuited and under a temperature differential from 800 K to 300 K, has 0.5 V between the open terminals, on the cold side.

Assume that the resistance is independent of T.

How many watts are necessary to drive a current of 100 A through the thermocouple when the differential above is maintained? Is there a unique answer to this question?

5.9 A thermocouple works between 500 K and 300 K. Its resistance is $0.0005\,\Omega$, and its heat conductance is 0.2 W/K. The mean Seebeck coefficient (between 500 and 300 K) is 0.001 V/K.

What is the open-circuit voltage generated by the thermocouple?

When there is no current, heat flows, of course, from the hot side to the cold side. Is it possible to make the heat that flows from the hot source to the thermocouple equal to zero? If so, what is the heat flow from the cold sink to the thermocouple? What is the electric power involved? What is the voltage across the couple? Does the electric power flow into the thermocouple or out of it (i.e., does the couple act as a generator or a load)?

5.10 A thermocouple operates between 900 K and 300 K. When short-circuited, it delivers 212 A, and when open-circuited, 0.237 V. The dimensions of the arm have been optimized for the mean temperature of 600 K. The material in the arms are semiconductors that have negligible lattice heat conductivity.

1. Calculate the heat taken from the hot source when open-circuited?
2. What is the power delivered to a 500 $\mu\Omega$ load?
3. What is the efficiency of the device in Question 2?

5.11 A small thermoelectric generator (single pair) is equipped with two thermal sensors, one measuring T_H (in the hotter side) and one measuring T_C (in the colder side). An electric heater warms up the hotter side, and all the heat thus generated is delivered to the thermocouple. A feedback system assures that T_H is kept at exactly 1000 K. The amount of electric power delivered to the heater can be measured.

On the colder side, a corresponding feedback system assures that T_C is kept at a constant 500 K.

The current, I, forced through the thermocouple can be adjusted to a desired value.

When $I = 0$, it takes 10 W of electric power to operate the heater at the hotter side of the thermocouple. Under such conditions, the thermocouple develops 0.50 V at its open-circuited terminals.

Now, a given current I is forced through the thermocouple so that the electric heater can be disconnected, while T_H still remains at 1000 K.

Next, the current I, above, is reversed, and it is observed that it takes 18.3 W to maintain $T_H = 1000$ K.

1. What is the value of I?
2. What voltages are necessary to drive the currents I and $-I$.

5.12 A thermocouple is connected to an adjustable current source. Disregard the resistance of the wires connecting the current source to the thermocouple. These wires and the cold side of the thermocouple are at 300 K. Under all circumstances, the current delivered has an absolute value of 100 A. Assume that the resistance of the thermocouple is temperature independent.

1. When $T_H = T_C = 300$ K, the absolute value of the voltage across the current source is 0.20 V. What is the voltage if the direction of the current is inverted?
2. When $T_H = 600$ K and $T_C = 300$ K, the absolute value of the voltage is 0.59 V. What are the voltages if the direction of the current is inverted?

5.13 Assume that the materials in the arms of a thermocouple obey strictly the Wiedemann–Franz–Lorenz law. The Seebeck coefficient, α, of the thermocouple is $150\,\mu$V/K independently of temperature, and the electric conductivities, σ (the same for the two arms), are also temperature independent. Such unrealistic behavior has been specified to make the problem more tractable; both α and σ usually do vary with temperature.

The device is to operate between T_H and T_C. For the calculation of any temperature dependent quantity, use the arithmetic mean, T_m, of these two temperatures.

The geometry of the thermocouple has been optimized, and the load connected to it always has the value that maximizes the efficiency of the system.

Show that, under the above circumstances, the electric power delivered to the load is independent of the choice of T_H provided that $\Delta T \equiv T_H - T_C$ is always the same.

At 800 K, what is the value of the ΛR product for this thermocouple?

5.14 A thermoelectric device, consisting of 100 thermocouples electrically in series and thermally in parallel, is being tested as a heat pump. One side is placed in contact with a cold surface so that it cools down to -3 C; the other side is maintained at 27 C.

The open-circuit voltage is measured by means of a high-impedance voltmeter and is found to be 900 mV.

Next, the electric output is shorted out, and it is observed that a current of 9 A flows through the short.

The device is now removed from the cold surface, and its cold end is insulated thermally so that absolutely no heat can flow in. A current of 50 A generated by an external source is forced through the device in such a direction that heat is pumped from the cold end to the warm end. A thermometer monitors the final temperature of the cold side. After steady state is reached, the temperature is 260 K. The hot side is still at 27 C.

Is this the lowest temperature that can be achieved? If not, what is the lowest temperature, and what is the necessary current?

5.15 A thermoelectric device is being tested in a laboratory. It consists of a single thermocouple and has its hot side in intimate thermal contact with an electric heater whose total heat output is transferred to the thermocouple. In other words, the power delivered to the heater is equal to the heat power, P_H, that flows into the thermocouple. Under all circumstances, the hot side is maintained at 1000 K and the cold side, at 300 K.

The two temperatures, T_H and T_C, are monitored by thermometers.

The first step in the test reveals that when the device is open-circuited, it draws 14 watts of heat from the hot source and delivers a voltage of 0.28 V. When short-circuited, it delivers a current of 35 A.

1. What is the heat power input when the device is short-circuited?
2. What is the heat power input when 0.4 V are applied in opposition to the open-circuit voltage?

5.16 A Radioisotope Thermal Generator is used as an electric power source aboard a spacecraft. It delivers 500 W at 30 V to a load optimally matched to the thermoelectric generator. Under such circumstances, this generator operates at a 12.6% efficiency.

The hot side of the generator (T_H) is at 1300 K, and the cold side (T_C) is at 400 K.

Assume that all the characteristics of the thermocouple (α, R, and Λ) are temperature independent.

1. What would be the efficiency of the generator if the load resistance were altered so that the power delivered fell to 250 W?

   Assume now that the heat power source is a radionuclide that delivers a constant heat power independent of the demands of the load. Thus, the temperature becomes a function of the load power. If the radioactive material were inside an adiabatic container, the temperature would rise until the container was destroyed—a steady heat leak must be provided to limit the temperature.

   The radionuclides in this problem release heat at a rate of 4984 W. The container, by itself, radiates enough energy to keep its outer skin at constant $T_L = 1000$ K under all circumstances, including when the thermoelectric generator is not installed, and, consequently, only the leakage path removes heat from the source. The cold side of the thermoelectric generator is always at $T_C =$ 400 K.

   When the thermoelectric generator is attached to the heat source and generates 500 W to a matched electric load, the temperature, T_S, falls to 1300 K (as in Question 1, this is T_H of the thermoelectric generator). In other words, the heat source delivers heat to two parallel paths: the leakage path and the thermoelectric generator.
2. Calculate the temperature of the heat source when no electric energy is being drawn from the thermoelectric generator.
3. When 250 W dc are drawn from the generator, what is the source temperature? There is more than one answer. Use the current with the smallest absolute value. Set up your equations and use a trial-and-error method.

5.17 Demonstrate that the voltage required to drive a thermoelectric heat pump is independent of the amount of power pumped and of the cold temperature, provided the current has been adjusted for maximum pumping.

5.18 Tungsten has an electric resistivity that (between 1000 and 3600 K) is given with acceptable precision by

$$\rho = -1.23 \times 10^{-7} + 3.49 \times 10^{-10} T,$$

where ρ is in Ω m.

Give me your best estimate for the thermal conductivity of tungsten at 1100 K and 1600 K.

5.19 A perfectly heat-insulated box is equipped with an electric heater, which allows the introduction of heat at an accurately measured rate. The only way heat can be removed is through a Peltier heat pump whose hot side is maintained at a constant 300 K.

The current, I, through the heat pump is controlled by a computer that senses the temperature inside the box and is set to keep it, if possible, at 280 K.

The experimenter chooses the amount of heat dissipated by the electric heater and tabulates the current the computer delivers to the heat pump. Here are the results:

Heat input (W)	**Current (A)**
0.50	7.382
1.00	13.964
1.25	21.225
1.00	32.702

For each case in the table, what is the coefficient of performance (COP) of the heat pump?

5.20 At 300 K, the electric resistivity of a sample is 0.002 Ωcm, and its heat conductivity is $0.03\,\mathrm{W\,K^{-1}\ cm^{-1}}$. From these data, determine if the material is a metal or a semiconductor. Explain.

5.21 A thermoelectric battery consists of 1000 identical thermocouples electrically in series and thermally in parallel. It works between 1000 K and 500 K.

Each thermocouple has the following characteristics:

Heat conductance: 3 W/K.

Electric resistance: 200 $\mu\Omega$.

Seebeck coefficient: 0.0007 V/K.

How much power does this battery deliver to a 0.3 Ω load?

5.22 A thermoelectric generator consists of a number of series-connected thermocouples. It operates between 1000 and 400 K.

When a 0.1-Ω load is used, the current is 266.7 A, and the heat rejected to the cold sink is 48 kW. When open-circuited, the voltage is 48 V.

What is the figure of merit (Z) of the device?

5.23 A prismatic block of pure sodium measures 1 by 1 by 10 cm. The two smaller faces (1 by 1 cm) are kept at a uniform temperature, one at 370 K and the other at 300 K.

Sodium has resistivity that can be represented (with fair accuracy) by

$$\sigma = 3.9 \times 10^7 - 6.6 \times 10^4 T \text{ S/m},$$

where T is in kelvins.

What is the heat power conducted by the sodium block?

Solve this problem twice:

First make some (drastic) simplifying assumption (use an average heat conductivity) to obtain an estimate of the heat power conducted.

Next, solve it by more realistically taking into account the fact that the heat conductivity varies along the length of the block.

5.24 A thermoelectric device, consisting of n series-connected thermocouples, is tested as a heat pump. A certain current, I, is chosen, and then the lowest attainable temperature, $T_{C_{min}}$, is determined. This is done with a completely heat-insulated cold end.

Next, another value of I is used and another, different, $T_{C_{min}}$ is found. In this manner, a tabulation of $T_{C_{min}}$ versus I is made. Throughout the whole experiment, T_H is maintained at a constant 320 K. The combination that resulted in the lowest $T_{C_{min}}$ is

$$I = 32.08 \text{ A for a } T_{C_{min}} = 244 \text{ K}.$$

The voltage necessary to drive this current through the thermoelectric device is 16.0 V.

What are the resistance, the heat conductance, and the α of the thermocouple battery?

5.25 The hot side of a thermocouple has, as a heat sink, a volume of water kept at 373 K under a pressure of 1 atmosphere. The water can only lose heat by evaporation (no conduction, except, of course, through the thermocouple itself). The heat of vaporization of water is 40 MJ/kmole. There is enough water, so that during the experiment the reservoir never runs dry. There is enough heat output from the hot side of the thermocouple to maintain the water at, at least, 373 K.

Assume that, under all temperatures considered in this problem, the device has

$$\Lambda R = 300 \times 10^{-6} V^2/\text{K}.$$

$$\alpha = 0.002 V/K.$$

Assume also that Λ does not change over the temperature range of interest.

Observe that this material has an extraordinarily high α. It does not exist at the moment. It will be invented in the year 2043. Just to emphasize a point, I specified this unrealistic Seebeck coefficient.

The cold side of the thermocouple is also in contact with a volume of liquid (not water), initially at 350 K. Again, the liquid (a total of 10 cubic centimeters) is in an adiabatic container, and the heat loss by evaporation is negligible.

There is no temperature drop between the water and the hot side of the thermocouple or between the cold side of the thermocouple and the cold liquid.

The current, $I = 11.67$ A, forced through the device has been chosen so that, under the initial conditions, it is the one that provides maximum heat pumping.

1. What is the heat power pumped initially from the cold liquid?
2. If the experiment runs long enough ($I = 11.67$ A), how cold will the liquid in the cold side get, and how hot will the water on the hot side get?
3. What is the rate of hot water evaporation (in kg/s) when $T_C = 350$? What is the rate when the lowest temperature of the cold liquid is reached?
4. Is the electric power used in the preceding item larger, equal, or smaller than the heat power required to evaporate the water? Do this for the cool liquid temperature of 350 K as well as for the lowest achievable temperature.
5. What is the voltage necessary to drive the 11.67 A current? Calculate this for each of the two coolant temperatures.

5.26 Usually the Seebeck coefficient, α, is a nonlinear function of the temperature, T. Under what circumstances is the coefficient independent of temperature?

5.27 Refer to the accompanying graph describing the performance of a Peltier cooler, **Hypothetical-127**, which consists of 127 thermocouples electrically in series and thermally in parallel. These data were generated under the simplifying assumptions that the Seebeck coefficient, α, the resistance, R, and the thermal conductance, Λ, are all temperature independent. In real life, the parameters actually vary with temperature.

1. Estimate the open-circuit voltage when $T_H = 65$ C and $T_C = -15$ C.
2. What is the figure of merit, Z, of the device when operating under the above conditions?

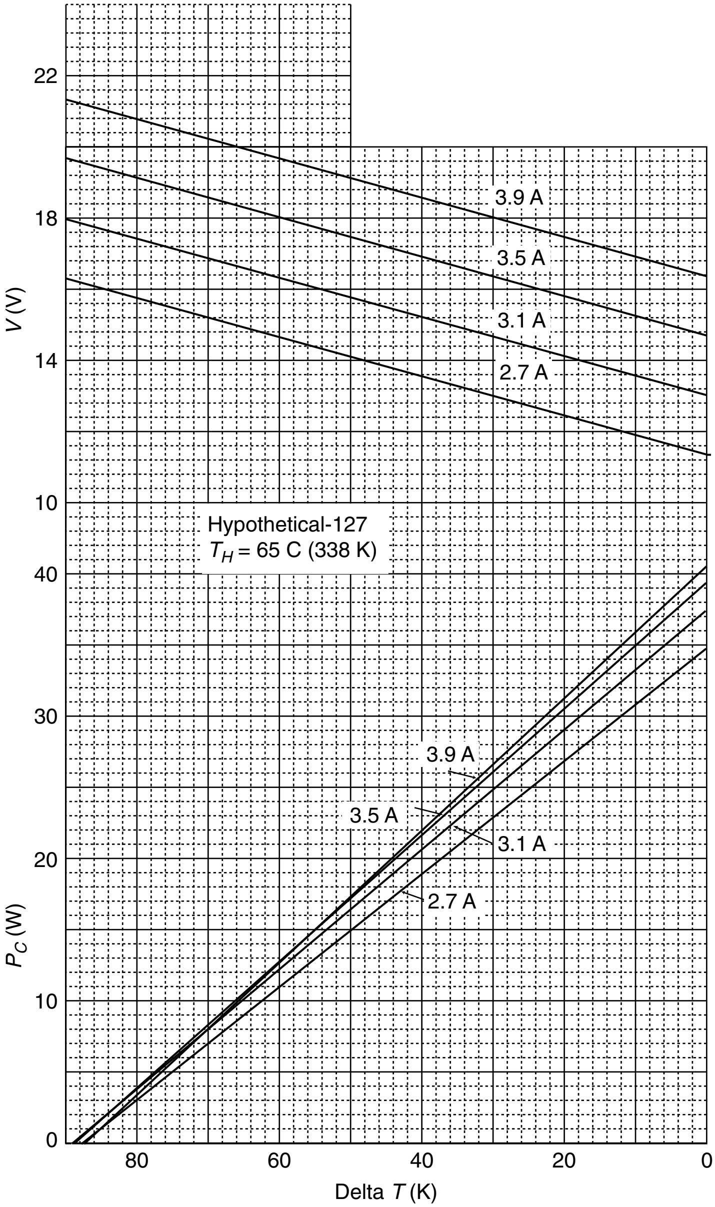
3. The figure of merit above is for the whole unit (127 thermocouples). What is the figure of merit for each individual thermocouple. Do not guess; show the result rigorously.
4. When the Peltier cooler, above, is pumping 22 W of heat from 298 K to 338 K (a ΔT of 40 K), how much heat is rejected at the hot side?
5. For the Peltier cooler, above, pumping heat from 298 K to 338 K, what would be the optimum current?
6. Assuming that α, Λ, and R are temperature independent, how much power does such a unit (used as generator) deliver to a load optimized for maximum efficiency? $T_H = 500\,\text{K}$, $T_C = 300\,\text{K}$.

5.28 A semiconductor can be doped so that it can have either p or n conductivity. Its properties (all independent of temperature) are:

		$\boldsymbol{p}$	$\boldsymbol{n}$
α	400	-400	μV/K
ρ	0.005	0.005	Ωcm
λ	0.02	0.02	W cm^{-1}K^{-1}
J_{max}	10	10	A cm^{-2}

Ammonia boils at -33.2 C under atmospheric pressure. The flask in which the ammonia is kept (at 1 atmosphere), though insulated, allows 5 W of heat to leak in. To keep the ammonia from boiling away, we must pump 5 W of heat from the flask. We propose to do this with thermocouples using the materials described above. One single stage will be insufficient to depress the temperature to the desired point; use two stages. The T_H of the colder stage (Stage 2) is 261.2 K. The T_H of Stage 1 is 40 C. The current through the two stages must be the same (they are connected in series) and must be the maximum allowable current through the material. Each stage must use the current that optimizes heat pumping. *The constraint on the current of the two stages may appear to lead to an incompatibility. Actually it does not.* Stage 2 consists of 10 thermocouples.

1. What is the electric power consumed? What voltage must be applied to the system assuming that all couples are electrically in series?
2. If power fails, the thermocouple will, unfortunately, provide a good path for heat leakage. Assuming the outside surface of the cooling system is still at 40 C, how many watts of heat leak in (as long as the ammonia is still liquid)? The heat of vaporization of ammonia is 1.38 MJ/kg. How long will it take to evaporate 1 kg of ammonia?

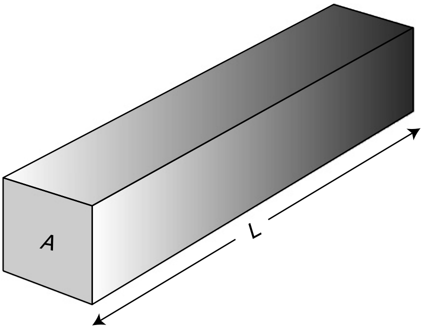


5.29 Three different materials, A, B, and C, were cut into rectangular prisms L cm long and $A\,\mathrm{cm}^2$ in cross section and were sent to a lab to be tested.

The test consists in accurately measuring the heat power flow, P_H, between the square faces (the ones with $1\,\mathrm{cm}^2$ area), when a temperature difference of exactly 1 K is maintained. The average temperature of each sample is T.

In addition, a current, I, is forced through the sample (it flows from one square face to the other). The voltage, V, developed across the sample is measured. The results are as follows.

Sample	V (V)	I (A)	P_H (W)	T (K)	A (cm^2)	L (cm)
A	0.00971	100	0.0257	500	10	1
B	0.00420	100	0.0290	50	2	5
C	22.20	0.001	4.17E-5	1000	8	3



The results for one or more of the samples above *may* be impossible. If so, identify which and explain why.

5.30 A thermoelectric unit consists of n cells. When there is no temperature difference across the leads, the resistance is 4 Ω. Λ is 0.3 W/K. When the unit is submitted to a 80 K temperature differential, it generates a 1.71 V across a 3-ohm load.

To simplify the problem, make the unrealistic assumption that α, R, and Λ are temperature independent.

1. What current pumps 10 watts of heat from 300 K to 350 K?

   Define the coefficient of performance (COP) of a thermoelectric heating system as the ratio of the delivered heat power to the electric power required.

2. What is the coefficient of performance of the device of the previous problem when extracting heat from an environment at 300 K and delivering heat at 350 K. Express this as a function of I. Determine the current that optimizes the COP. Do not waste time and risk errors by taking derivatives. Use trial-and-error numerical solutions. How much heat is pumped under such circumstances?

5.31 A thermoelectric generator works between the temperatures 1000 K and 400 K. The open-circuit voltage of the generator is 70.0 V. The geometry of the individual cells has been optimized.

Measurements show that maximum efficiency is obtained when the generator delivers 5.22 A at 38.68 V.

1. What is the efficiency of the system?
2. What is the efficiency of the system when only 100 W are delivered to the load?
3. Why is the efficiency of the generator so much smaller when delivering 100 W compared to its efficiency at 200 W?

5.32 A single thermocouple is equipped with two heat exchangers through which water is forced to flow. The input water is at 300 K for both heat exchangers. The flow is adjusted so that the water exits the hot heat exchanger at 310 K and the cold heat exchanger at 290 K. For simplicity, these are also the temperatures of the two ends of the thermocouple. The current through the thermocouple is such that exactly 10 W of heat power are removed from the water in the cold heat exchanger. This current is the smallest current capable of pumping the 10 W from the cold side. The characteristics of the thermocouple are:

$$\alpha = 0.0006 \text{V/K}.$$

$$\Lambda R = 120 \times 10^{-6} \text{V}^2/\text{K}.$$

1. What is the flow rate of the water in the cold heat exchanger?
2. What is the current through the thermocouple?
3. What is the flow rate of the water through the hot heat exchanger?
4. What is the coefficient of performance of this refrigerator? The coefficient of performance is defined as $\phi_C = P_C/P_E$.
5. Using this same thermocouple and the same T_H and T_C, it is possible to operate at a higher coefficient of performance. Clearly, the amount of heat pumped from the cold side would no longer be 10 W. What is the best coefficient of performance achievable?

5.33 Prove that the current that maximizes the coefficient of performance, ϕ_C,

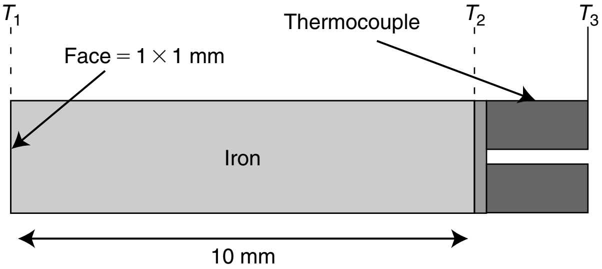
$$I = \frac{\Lambda \Delta T}{\alpha < T >}(m+1).$$

can also be written

$$I = \frac{\alpha \Delta T}{R(m-1)}.$$

5.34 Demonstrate that the Carnot efficiency of a heat pump is $T_C/\Delta T$, where T_C is the temperature of the cold side and ΔT is the temperature difference between the hot and the cold side.

5.35 In the text, there is a description of the RTG used in the Galileo mission to Jupiter in 1981. From the data given, calculate the cost in dollars of the fuel used in this RTG.



The figure illustrates an experimental setup used to determine the characteristics of a thermocouple. It consists of a rectangular prism made of pure iron (conductivity = 80 $\mathrm{Wm^{-1}K^{-1}}$) to which the thermocouple to be tested is attached. The iron slug is well insulated laterally so that heat can only enter and leave through the small faces. The temperatures of the two faces as well as that of the open end of the thermocouple can be accurately measured. All properties of the different elements are temperature independent,

Test #	**T_1**	**T_2**	**T_3**	**Applied**	**Observed**
				current	**voltage**
	(K)	(K)	(K)	(A)	(V)
1		300	300	5	0.1
2	440.6	400	300	0	0.05

This table above lists two different tests carried out.

1. From the above, calculate the figure of merit of the thermocouple.
2. What was the temperature, T_1, in Test 1?

5.37 You have 2 liters of water in an adiabatic styrofoam container. You want to cool the water by means of a battery of thermocouples consisting of 10 units, each one of which has the following (temperature independent) characteristics:

$$\alpha = 0.055\,\mathrm{V/K}.$$
$$R = 4.2\,\Omega.$$
$$\Lambda = 0.25\,\mathrm{W/K}.$$

The temperature of the hot side of the thermocouples is maintained at a constant 300 K.

1. How long does it takes to cool the water from 300 K to 285 K when you drive a 4 A constant current through each thermocouple unit.
2. Can you change the operating point of the thermocouple so as to shorten the cooling time? What is the shortest achievable cooling time?
3. How much electric energy is consumed in cooling the water when the current is that of the preceding question? Compare this with the heat energy extracted from the water. What is the coefficient of performance of the system?
4. Assume now that you adjusted the current to optimize the coefficient of performance. What is the current? What is the coefficient of performance?
5. What would be the coefficient of performance of an ideal (Carnot) heat pump cooling this amount of water?

5.38 We need to pump 15 W of heat from a CPU that works at 300 K. The heat removed has to be rejected at 340 K. In order to provide an adequate heat flow from the CPU to the cool side of the heat pump, the latter must be kept at 295 K.

Thermocouples will have to be connected electrically in series and thermally in parallel, as is usual. Each thermocouple has the following characteristics:

$$R = 0.033\,\Omega.$$

$$\Lambda = 0.002\,\mathrm{W/K}.$$

$$\alpha = 0.00043\,\mathrm{V/K}.$$

Calculate the number of thermocouples required, the total electric power, and the voltage needed to drive the system considering each of the alternatives below:

a. The current is adjusted for maximum cooling power.

b. The current is adjusted for maximum coefficient of performance.

5.39 There are several different ways one can improve the efficiency of a thermoelectric heat pump. One is, of course, by choosing a better type of thermocouple. However, if you are constrained to use a given type of thermocouple, you may want to test different architectures.

You are given thermocouples whose characteristics include:

$$R = 4.2\,\Omega.$$

$$\Lambda = 0.25\,\text{W/K}.$$

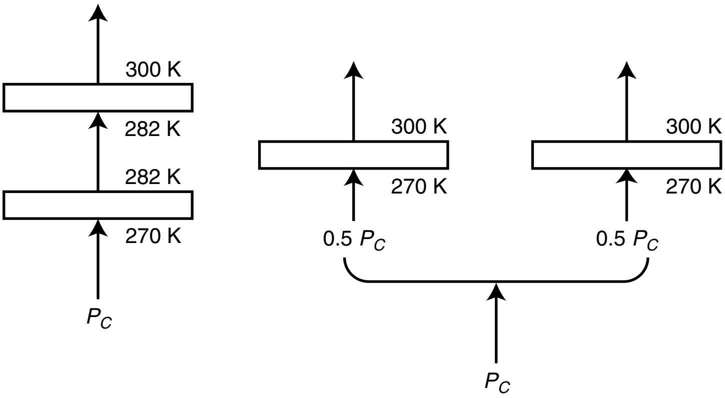
$$\alpha = 0.055\,\text{V/K}.$$

They can be used in three different ways:

a. They can be used in parallel, as shown in the right-hand side of the figure below,
b. They can be used thermally in series (left-hand side), in which case the current through one thermocouple may differ from that through the other.
c. You can ignore one of the thermocouples and use a single one.

The temperatures are indicated in the figure. You are to pump 5 watts from 270 to 300 K, that is, $P_C = 5\,\text{W}$.

If you find more than one mathematical solution for a given case, select the one that leads to a lower electric power use. Explain how, in a practical setup, you would make sure the system will operate in the more economical way.



1. Which arrangement leads to the best coefficient of performance?
2. What is the Carnot efficiency of the system?
3. Using a single thermocouple of the type described, what is the lowest temperature that can be achieved? Heat is rejected at 300 K.
4. Under the conditions of Question 3, what is the necessary current?

5. Using a single thermocouple of the type described, what is the best coefficient of performance when pumping heat from 270 K to 300 K?
6. What is the current that yields the COP of Question 5?
7. What amount of heat is pumped in Question 5?
8. What voltage has to be applied to operate in the conditions of Question 5?

Chapter 6
Thermionics

6.1 Introduction

Discovered in 1885 by Thomas Edison, thermionic emission achieved enormous industrial importance in the twentieth century, but only in 1956 was the first thermionic heat-to-electricity converter demonstrated.

The operation of a vacuum thermionic generator is easy to understand, but the device is impractical owing to the space charge created by the electrons in the interelectrode space. Introducing a plasma containing positive ions into this space overcame the space charge problem but created additional complications. It became necessary to use plasma, not only to cancel space charge but also as a source of additional electrons. Only this latter solution—the ***ignited*** *plasma device—holds promise of becoming a useful thermionic generator.*

In this text, we describe the fundamentals of the emission process, the transport of electrons through a vacuum, the operation of the vacuum diode converter, the creation of a cesium plasma, and the behavior of the plasma when present in concentrations low enough to allow disregarding collision ionization, and, finally, a few words on the converter in which the collision ionization is of fundamental importance. The behavior of the plasma in the latter is too intricate to be discussed in detail in this book. Those who need a more complete explanation of the operation of plasma diodes, especially of the only practically important class operating in the ignited mode, should refer to an extensive article by Ned Razor (1991).

The terminology used in this chapter needs some clarification. Most electronic engineers think of “anode” as the positive electrode and “cathode” as the negative one. This is misleading. In Greek, anodos *and* kathodos *mean, respectively, “the way up” and “the way down.” In electrical engineering, these words mean simply* the way in *and* the way out *of a device (using the conventional direction of current flow). Thus, in a diode, the anode is the electrode through which electricity enters the device, and the cathode is the exit electrode. Here the anode is, indeed, the electrode connected to the positive terminal of the power supply. In a battery, however, electricity (conventionally) exits from the positive electrode, which is consequently called the cathode. To simplify the terminology, we will, when possible, use self-describing terms*

(*Continues*)

(*Continued*)

such as emitter *(of electrons) for the cathode of a thermionic device and* collector *for the anode.*

Thermionic devices, once the very soul of electronics, still constitute a sizable market. Thermionics are the basis of **radio tubes** used in most high-power radio and TV transmitters and of **cathode-ray tubes** employed in oscilloscopes. Only recently have they lost their dominance in TV and computer monitors, applications in which they have been displaced by other high-resolution devices such as the much more compact light emitting diode, LCD and plasma panels. Thermionics also play an essential role in most microwave tubes such as **klystrons**, **magnetrons**, and **traveling-wave tubes**.

The majority of the classical thermionic implements were **vacuum state** devices in which electrons flow through vacuum. Nevertheless, occasionally plasmas were used, as in the once popular mercury vapor rectifiers. Vacuum state devices are undergoing a revival in microelectronics. However they do not use thermionic emission; instead they take advantage of field emission phenomena.[†] This chapter, on the other hand, concerns itself with thermionic heat-to-electricity converters and excludes field emission.

In the simplest terms, a thermionic converter is a heat engine in which electrons are boiled off a hot surface (**emitter**) and are collected by a colder one (**collector**). There is an energy cost in evaporating electrons from a solid, just as there is a somewhat similar cost in evaporating a liquid. The *minimum* energy necessary to remove an electron from a metal or a semiconductor is called the **work function**. Dividing the work function by the electron charge defines a potential (voltage) that has the same numerical value as the work function itself when the latter is expressed in eV. In this text, we will use the symbol, ϕ, to represent the potential, and, consequently, the work function is $q\phi$. Though somewhat confusing, we will use the term *work function* to designate both the energy and the potential, as do most workers in thermionics. The reader will be able to distinguish the two meanings from the context.

The work function depends on the nature of the emitting substance; if the work function of the hot emitter surface is larger than that of the cold collector, then the difference may become available as electric energy in the load. Thus, if the emitter has a work function of 3 eV and the collector has one of 2 eV, a net potential of 1 volt could be available to the load.

Electrons boiled off from the emitter accumulate on the collector unless there is a provision to drain them off. This can be done by establishing

[†] See the survey of vacuum microelectronics by Iannazzo (1993).

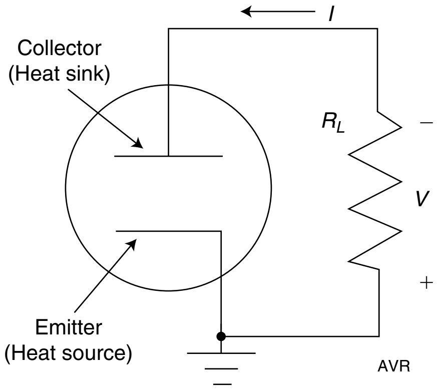


Figure 6.1 A thermionic generator.

an external path from collector to emitter. This path can include a load resistance, R_L, as depicted in Figure 6.1 through which a current, I, circulates. The external voltage drop creates an interelectrode electric field that opposes the electron flow. Notice that the collector or anode is the negative terminal of the thermionic generator. In the figure, the diode is represented by its conventional symbol in electronics.

The energy dissipated in the load comes from the emitter heat input. The required high temperature leads to substantial heat radiation losses. It is difficult to drive large currents across the vacuum gap in the emitter-to-collector space. To circumvent this difficulty, the vacuum can be replaced by a positive ion gas. Unfortunately, this introduces additional losses owing to collisions between electrons and ions.

Although thermionic converters are heat engines and are limited by the Carnot efficiency, they offer important advantages:

1. They yield electric energy directly, not mechanical energy as do mechanical heat engines.
2. They operate at high temperature, with corresponding high Carnot efficiencies.
3. They reject heat at high temperature, simplifying the design of heat sinks in space applications and making them useful as topping engines.
4. They operate at high power densities, leading to compact devices that use high-priced materials sparingly.
5. They operate at low pressures and have no moving parts.

In order to better understand the operation of a thermionic engine, it is useful to consider separately two of the basic processes involved: the thermal emission of electrons from a solid surface and the transport of such electrons across the interelectrode space.

6.2 Thermionic Emission

Qualitatively, it is easy to understand thermionic emission. In metals or semiconductors, at any temperature above absolute zero, free electrons move around in a random manner, their velocity distribution being a function of temperature. When the temperature is sufficiently high, some electrons have enough energy to overcome the forces that normally cause them to stay within the solid. However, it is not sufficient that the electrons have more than the escape energy; they must also have this excess kinetic energy associated with a velocity component normal to the emitting surface.

The thermionically emitted electron current is exponentially dependent on ϕ and is

$$J_0 = q\frac{4\pi}{h^3}mk^2T^2\exp\left(-\frac{q\phi}{kT}\right) = A_{th}T^2\exp\left(-\frac{q\phi}{kT}\right), \tag{6.1}$$

where

$$A_{th} \equiv \frac{4\pi mqk^2}{h^3}. \tag{6.2}$$

A_{th} is the **theoretical emission constant**. If the mass of the electron is taken as 9.1×10^{-31} kg, A_{th} has the value of 1.20×10^6 amperes m^{-2} K^{-2}. The actual mass of the electron may vary somewhat from material to material. Equation 6.1 is known as **Richardson's Equation** in honor of Owen Williams Richardson, who, for his work in thermionic emission, received the 1928 Nobel Prize in Physics.

The work function is a characteristic of the emitting surface. Since it is a temperature-dependent quantity, one can make the reasonable assumption that it might be a well-behaved function of T, and, as such, the corresponding potential can be expressed as

$$\phi = \phi_0 + a_1T + a_2T^2 + \ldots \text{eV}. \tag{6.3}$$

The higher order terms are, presumably, small. Truncating after the linear term, Richardson's equation becomes

$$\begin{aligned} J_0 &= A_{th}T^2\exp\left[-\frac{q}{kT}(\phi_0 + a_1T)\right] = A_{th}\exp\left(-\frac{q}{k}a_1\right)T^2\exp\left(-\frac{q\phi_0}{kT}\right) \\ &= AT^2\exp\left(-\frac{q\phi_0}{kT}\right), \end{aligned} \tag{6.4}$$

where

$$A \equiv A_{th}\exp\left(-\frac{q}{k}a_1\right). \tag{6.5}$$

One has the choice of writing Richardson's equation (Equation 6.1) using a theoretical emission constant, A_{th}, which is the same for all emitters, together with a work function, ϕ, which not only depends on the nature of the emitting surface but is, in addition, dependent on the temperature of the emitter. Or one can opt for a Richardson's equation (Equation 6.4) in which the work function does not vary with temperature (but is still dependent on the nature of the emitter), together with an emission constant that, although temperature independent, also depends on the nature of the emitter. The first option better describes the physical nature of the emission phenomenon, but the second is easier to use. Experimental thermionic emission data are usually tabulated as values of A and ϕ_0. It is customary to drop the subscript "0," with the understanding that when using the second option, one must use the value of the work function at absolute zero.

Table 6.1 shows properties of some materials typically used in thermionic devices. Work functions vary from 1 eV to a little over 5 eV, while the apparent emission constants, A, cover a much wider range, from 100 to 600,000 $\mathrm{A\,m^{-2}\,K^{-2}}$, a result of the variation in the temperature coefficients of ϕ in the different materials.

It must be emphasized that ϕ is sensitive to the exact way in which the material was prepared and to the state of its surface. Some of the tabulated values correspond to single crystals of the material considered, having as clean a surface as possible.

Devices used in electronic engineering benefit from emitters made of materials with low work functions because this allows their operation at relatively low temperatures, saving heating energy and, sometimes, prolonging the useful life of the devices. Thus, oxide-coated emitters (BaO + SrO) were popular in small vacuum tubes. However, the emitter life may still be limited by the weakness of the interfaces between the ceramic crystals.

Each emitter has a preferred operating temperature. Tungsten filaments work best at around 2500 K, thoriated tungsten at 1900 K, and oxides at 1150 K. Thoriated tungsten emitters are still used in high power vacuum

Table 6.1 Properties of Some Thermionic Emitters

Material	**Work function**	**Emiss. Const,**	**Melting**	**Temperature**
	$\boldsymbol{\phi_0}$	$\boldsymbol{A}$	**point**	**coefficient**
	(eV)	$\mathbf{(A\,m^{-2}\,K^{-2})}$	**(K)**	**eV/K**
Pt	5.32	320,000	2045	0.000114
Ni	4.61	300,000	1726	0.000120
Cr	4.60	480,000	2130	0.000079
W	4.52	600,000	3683	0.000060
Mo	4.20	550,000	2890	0.000067
Ta	4.19	550,000	3269	0.000067
Th/W	2.63	30,000	—	
BaO + SrO	1.03	100	—	0.000318
Cs	1.81	—	302	0.000810

tubes because of their lower work function compared with pure tungsten and their longer life. When pure tungsten is heated to its usual operating temperature, there is a tendency for single crystals to grow, and it is at the resulting interfaces that breaks occur. The presence of thorium retards the growth of these crystals.

Thermionic energy converters require that the work function of the emitter be larger than that of the collector. Because most converters use a cesium atmosphere to neutralize the space charge (see Subsection 6.7.1), and because cesium tends to condense on the cooler collector surface, the work function of the collector is usually near that of cesium: 1.81 eV. Thus, emitters must have work functions of more than 1.81 eV.

Owing to the exponential dependence on ϕ, the emitter current is more sensitive to ϕ than to A. For example, the current density of a BaO + SrO emitter at its normal operating temperature of 1150 K is

$$J_0 = 100 \times 1150^2 \times \exp - \frac{1.03\ q}{1150\ k} = 4090 \text{ A m}^{-2}. \tag{6.6}$$

At the same temperature, a tungsten emitter produces a much smaller current, notwithstanding having an emission constant, A, 6000 times larger:

$$J_0 = 600{,}000 \times 1150^2 \times \exp - \frac{4.52\ q}{1150\ k} = 1.3 \times 10^{-8} \text{ A m}^{-2}. \tag{6.7}$$

For this reason, tungsten emitters operate at high temperatures.

High temperatures are not the only cause of electron emission. Several other mechanisms can accomplish this, including the impact of photons (**photoelectric emission**), the impact of subatomic particles, especially electrons themselves (**secondary emission**), and intense electric fields already alluded to in the previous section (**field emission**). Intense fields near the emitter of a thermionic device will alter the magnitude of the emitted current. In most thermionic generators, such fields are sufficiently small to allow their effect to be ignored.

6.3 Electron Transport

The simplest thermionic generator would consist of an emitting surface, (the **emitter**), heated to a sufficiently high temperature, T_H, and placed in the vicinity of a collecting surface (the **collector**), operating at a lower temperature, T_C. The space between these surfaces may be a vacuum. The heat source may be of any desired nature: a flame, a nuclear reactor, the heat from nuclear decay, concentrated sunlight, and so on.

The geometry of the device plays a role in its performance, but here we are only going to consider two parallel plate electrodes because this is the easiest configuration to analyze.

Since electrons leave the emitter and travel to the collector through the interelectrode space, the conventional direction of an external current is, under all circumstances, out of the emitter into the collector. In other words, the collector is the anode, and the emitter is the cathode of the device.

When operated as a generator, the device is connected to a load, as indicated in Figure 6.2. The voltage drop across the load causes the collector to become negatively polarized with respect to the emitter; there is a **retarding interelectrode potential**. In electronic applications, devices are usually operated with an externally applied bias that causes the collector to be positive with respect to the emitter. An **accelerating interelectrode potential** is established. See Figure 6.3

One might conclude that with an accelerating bias, all the emitted electrons should reach the collector, that is, that the interelectrode current density, J, should be equal to the emitted current, J_0. However, frequently J is smaller than J_0—the device is **space charge limited** or **unsaturated**, while when $J = J_0$, it is said to be **emission limited** or **saturated**.

Most thermionic *amplifiers* operate under space charge limitation; the current through the device depends on the applied voltage and is given by the **Child–Langmuir law** discussed later on. For maximum efficiency, thermionic *generators* must operate at the maximum possible current density and must, therefore, be emission limited.

In a vacuum, such as may exist in the interelectrode space of some devices, there is no mechanism for scattering the emitted electrons. As a consequence, the electron motion is dictated in a simple manner by the local electric field, or, in other words, by the interelectrode potential.

In the absence of charges in the interelectrode space of a parallel plate device—that is, if there are no emitted electrons between the plates—the

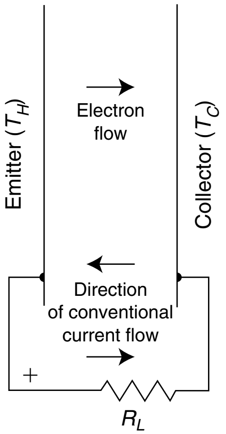


Figure 6.2 An elementary thermionic generator.

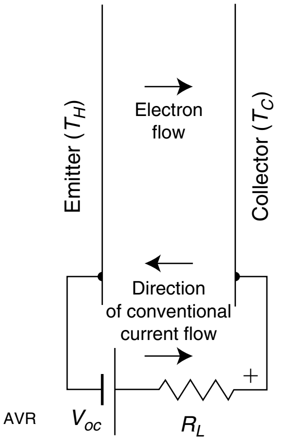


Figure 6.3 A forward-biased thermionic diode.

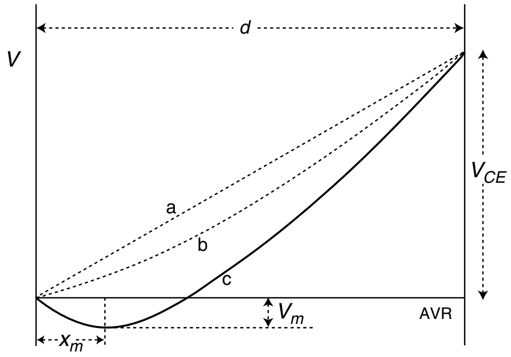


Figure 6.4 Potential across a planar thermionic diode.

electric field is constant and the potential varies linearly with the distance from the reference electrode (the emitter, in Figure 6.3). See curve a in Figure 6.4. A single electron injected into this space will suffer a constant acceleration. However, if the number of electrons is large, their collective charge will alter the potential profile causing it to sag—curve b.

If the number of electrons in the interelectrode space is sufficiently large, the potential profile may sag so much that the electric field near the emitter becomes negative (curve c), thus applying a retarding force to the emitted particles. Only electrons that are expelled with sufficient initial velocity are able to overcome this barrier and find their way to the collector. This limits the current to a value smaller than that corresponding to the maximum emitter capacity.

6.3.1 The Child–Langmuir Law

Almost all thermionic generators operate under emission saturated conditions—that is, with no net space charges in the interelectrode space. Under such conditions the current does not depend on the voltage. However, to better understand how the presence of space charges limits the current, we will derive the equation that establishes the relationship between the applied voltage and the resulting current when the device is unsaturated.

Consider a source of electrons (emitter) consisting of a flat surface located on the $x = 0$ plane, and capable of emitting a current of density, J_0. See Figure 6.5. Another flat surface (collector) is placed parallel to the emitter at a distance, d, from it. A perfect vacuum exists in the space between the electrodes. Assume initially that the electrons are emitted with no kinetic energy—they just ooze out of the emitter.

A potential is applied to accelerate the electrons from emitter to collector. A current of density, J (where $J \leq J_0$), is established between the electrodes:

$$J = qnv, \tag{6.8}$$

where q is the charge of the electron, n is the electron concentration—the number of electrons per unit volume—and v is the velocity of the electrons. J is the flux of charges.

J must be constant anywhere in the interelectrode space, but both n and v are functions of x. We want to establish a relationship between J and the potential, V, at any plane, x. We will take $V_{(x=0)} = 0$. Then, owing to the assumed zero emission velocity of the electrons,

$$\frac{1}{2}mv^2 = qV, \tag{6.9}$$

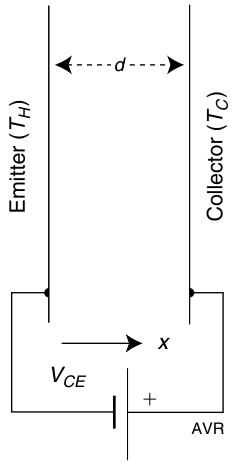


Figure 6.5 A vacuum diode.

from which

$$qn = J\sqrt{\frac{m}{2q}}\ V^{-1/2}. \tag{6.10}$$

However, the presence of electrons in transit between the electrodes establishes a space charge of density qn so that, according to Poisson's equation,

$$\frac{d^2V}{dx^2} = -\frac{nq}{\epsilon_0}, \tag{6.11}$$

where ϵ_0 is the permittivity of vacuum. Therefore,

$$\frac{d^2V}{dx^2} = -\frac{J}{\epsilon_0}\sqrt{\frac{m}{2q}}\ V^{-1/2} = -KV^{-1/2}, \tag{6.12}$$

where

$$K \equiv \frac{J}{\epsilon_0}\sqrt{\frac{m}{2q}}. \tag{6.13}$$

The solution of Equation 6.12 is

$$V = a + bx^{\alpha}. \tag{6.14}$$

The requirement that $V_{(x=0)} = 0$ forces $a = 0$. Thus,

$$\alpha(\alpha-1)bx^{\alpha-2} = -Kb^{-1/2}x^{-\alpha/2}; \tag{6.15}$$

consequently,

$$\alpha - 2 = -\frac{\alpha}{2} \quad \therefore \quad \alpha = \frac{4}{3}, \tag{6.16}$$

and

$$\frac{4}{9}b = -Kb^{-1/2} \quad \therefore \quad b^3 = \left(-\frac{9}{4}K\right)^2 = \left(\frac{9}{4}K\right)^2, \tag{6.17}$$

$$b = \left(\frac{9}{4}K\right)^{2/3}. \tag{6.18}$$

Thus

$$V = \left(\frac{9}{4}K\right)^{2/3} x^{4/3}. \tag{6.19}$$

Replacing K by its value from Equation 6.13 and solving for J,

$$J = \frac{4}{9}\epsilon_0\sqrt{\frac{2q}{m}}\ \frac{V^{3/2}}{x^2} = \frac{2.33\times10^{-6}}{x^2}\ V^{3/2} \quad \text{A m}^{-2}. \tag{6.20}$$

When $V = V_{CE}$ (the collector-to-emitter or **anode** voltage), then $x = d$, where d is the interelectrode spacing. Hence,

$$J = \frac{2.33 \times 10^{-6}}{d^2} V_{CE}^{3/2} \quad \text{A m}^{-2}. \tag{6.21}$$

More commonly, the expression is written in terms of current, not current density, and is known as the **Child–Langmuir law**:

$$I = B\, V_{CE}^{3/2}. \tag{6.22}$$

The **perveance**, B, is given by

$$B = \frac{2.33 \times 10^{-6}}{d^2} A, \tag{6.23}$$

where A is the electrode area.

The Child–Langmuir law holds for diodes of any shape; however, the perveance depends on the particular geometry of the device.

In reality, electrons do not ooze out of the emitter; they are launched with a wide spectrum of velocities. As explained before, while in transit to the collector, they constitute a space charge that causes the potential in the interelectrode region to be lower than when electrons are absent. In fact, the potential of the region near the emitter may become more negative than that of the emitter itself, creating a retarding electric field or "barrier."

The Child–Langmuir law is valid for such real diodes, but the perveance is now a function of the distance $d - x_m$ and the current is proportional to $(V + V_m)^{3/2}$. x_m is the distance from emitter to the plane where the potential reaches a minimum (see Figure 6.4). This plane represents a **virtual emitter**. V_m is the potential at $x = x_m$.

Even in the presence of space charge, a thermionic diode can become saturated if enough forward bias is applied. Consider a diode whose emitter is a tungsten plate heated to 2500 K. At this temperature, tungsten emits 3000 amperes per square meter of surface: $J_0 = 3000\ \text{A m}^{-2}$. If the collector-to-emitter voltage, V_{CE}, is small enough, then $J < J_0$ and the Child–Langmuir law is valid. At higher collector-to-emitter voltages, the law would predict currents larger than the emission capability of the cathode.

Since this is impossible, at large V_{CE}, J remains equal to J_0 independently of the value of V_{CE}. The voltage that just causes such saturation depends on the perveance of the device—that is, on the interelectrode spacing, d. The larger the spacing, the higher the saturation voltage.

The V-J characteristics of two diodes with different perveances are shown in Figure 6.6. The diode with 1-mm interelectrode spacing will operate in the space charge limited region; that is, it will obey the Child–Langmuir law as long as the collector voltage is 118 V or less. Above this voltage, it will saturate. The diode with a 0.2-mm spacing saturates at only 13.8 V.

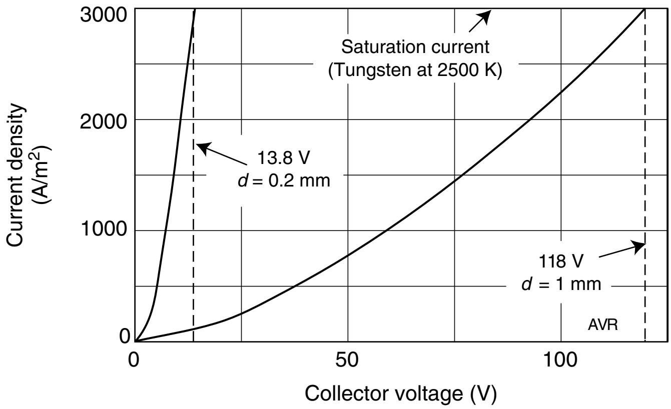


Figure 6.6 *V*-*J* characteristics of two vacuum thermionic diodes.

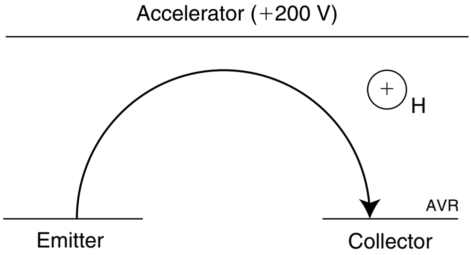


Figure 6.7 A vacuum diode with effectively zero interelectrode spacing.

Thermionic converters operate under reverse bias, and their current is severely restricted by net space charges in the interelectrode region.

To reduce the total amount of charge, the electrodes must be placed very close together, or, alternatively, a space charge neutralization scheme must be employed. Devices with spacing as small as 10 μm have been made, but they are difficult to build with large electrode areas.

It is possible to obtain a virtual zero interelectrode spacing (at least theoretically) by employing the configuration of Figure 6.7. Emitter and collector are coplanar and face a third electrode: an **accelerator**.

A voltage is applied between emitter and accelerator—say, some 200 V. A magnetic field (normal to the plane of the figure) causes the path of the electrons to bend away from the accelerator toward the collector. The effective perveance of this device is infinite; that is, the current is always saturated. Since—in the ideal case—the electrons don't reach the accelerator, no power is used in the accelerator. In practice, however, a substantial number of electrons do reach this electrode, using up a prohibitive amount of energy. Sufficiently reducing interelectrode spacing or using some scheme to circumvent the effect of space charges appears to be impractical.

In solids, the drift velocity of free electrons (or other carriers) is limited by frequent collisions with obstacles. The dissipated energy reveals itself as a resistance to the flow of current. Although there are no obstacles to scatter the electrons in a vacuum, there is nevertheless a severe impediment to the free flow of a current: the very electrons create, by their presence, a space charge that tends to oppose their motion. If instead of a vacuum the electrons move in a rarefied positive ion gas, there will be a certain scattering (with attending losses), but the space charge is neutralized and the electrostatic impediment to the flow of current is removed. The vast majority of thermionic generators use space charge neutralization. In the next section, we will derive the behavior of a thermionic diode under the assumption that there is no space charge in the interelectrode space.

6.4 Lossless Diodes with Space Charge Neutralization

6.4.1 Interelectrode Potentials

To drive large currents through a thermionic diode, interelectrode space charges must be neutralized. Thus, we will restrict our attention to the space charge-free case. Under such circumstances, in a planar diode, the potential varies linearly with distance from the emitter.

If the electrodes are externally shorted out and if thermoelectric effects are neglected, then there is no potential difference between the collector and the emitter: the energy reference level (the **Fermi level**) is the same at the two electrodes. Figure 6.8 shows the voltages in a shorted diode.

ϕ_E stands for the work function of the emitter and ϕ_C for that of the collector. Notice that the larger the ordinate, the more negative the potential (voltage). To free itself from the emitter, an electron must have (at least) an energy, $q\phi_E$; in other words, the potential just outside the emitter (referred to the Fermi level) is $-\phi_E$. The figure represents a case in which $\phi_E > \phi_C$.

Since we are accustomed to plots in which the voltage increases upward, we have redrawn Figure 6.8 upside down in Figure 6.9. It can be seen that the potential just outside the collector is more positive than that just outside the emitter, and, for that reason (in absence of space

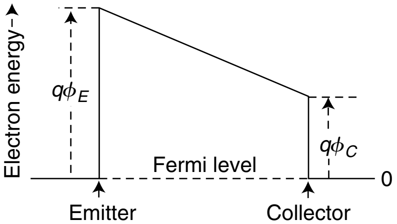


Figure 6.8 Energy levels in a short-circuited thermionic diode.

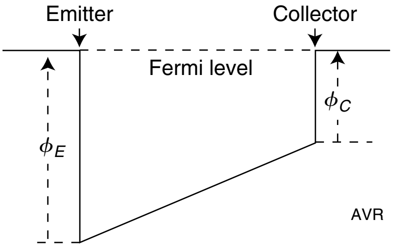


Figure 6.9 Voltages in a short-circuited thermionic diode.

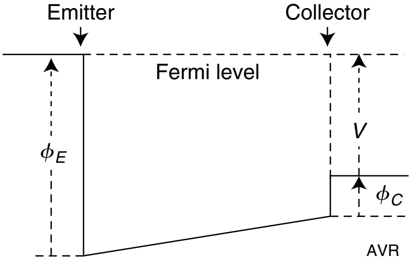


Figure 6.10 Voltages in a thermionic diode connected to a load. ($V+\phi_C < \phi_E$)

charge), the diode current, J, is saturated, that is, $J = J_0$, provided there is negligible emission from the collector:

$$J_0 = AT^2 \exp -\frac{q\phi_E}{kT}. \tag{6.24}$$

When a load is placed externally between collector and emitter as shown in Figure 6.1, the current from the generator will cause a voltage, V, to appear across this load in such a way that the terminal nearer the emitter becomes positive. The potential distribution then resembles that shown in Figure 6.10.

As long as $|V + \phi_C| < |\phi_E|$, the interelectrode potential will accelerate the electrons and the current density will remain equal to J_0.

However, if V becomes large enough to invalidate the above inequality, then a retarding potential, V_R, is developed between the electrodes, with a magnitude

$$|V_R| = |V + \phi_C - \phi_E| = |V - \Delta V|, \tag{6.25}$$

where

$$\Delta V \equiv \phi_E - \phi_C. \tag{6.26}$$

see Figure 6.11.

Under these circumstances, only electrons that, inside the emitter, have energy larger than the total barrier, $q(V+\phi_C)$ will reach the collector. The current density is still governed by Richardson's law, but the effective

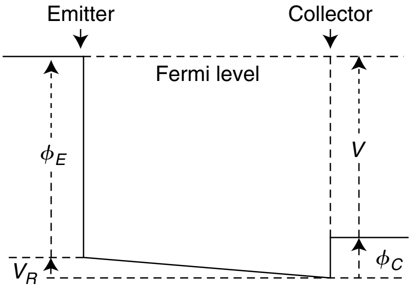


Figure 6.11 Voltages in a thermionic diode connected to a load. ($V+\phi_C > \phi_E$)

barrier is now larger than before. Electrons from the emitter now have to overcome not only the emission barrier, $q\phi_C$, but also the additional barrier, qV_R. Only electrons with more than $q(\phi_E + V_R) = q(\phi_C + V)$ joules can reach the collector:

$$J = AT^2 \exp\left[-\frac{q}{kT}(V + \phi_C)\right] = AT^2 \exp\left[-\frac{q}{kT}(\phi_E + V - \Delta V)\right]$$
$$= J_0 \exp\left[-\frac{q}{kT}(V - \Delta V)\right]. \qquad (6.27)$$

6.4.2 V-J Characteristics

The J versus V characteristic of a thermionic generator is shown in Figure 6.12.

Consider a diode with a tungsten emitter ($\phi_E = 4.52\,\text{eV}$) and a cesium plasma ($\phi_C = 1.81\,\text{eV}$). Because of the cesium condensation on the collector, the collector operates with the cesium work function.

The break point in the J versus V curve should occur at $V = 4.52 - 1.81 = 2.71$ V. Experimentally, it occurs at 2.5 V owing to, at least in part, the thermoelectric voltage created by the temperature difference between the electrodes.

6.4.3 The Open-Circuit Voltage

Under the assumed Maxwellian energy distribution of the emitted electrons (which assigns nearly infinite velocity to some electrons), it would appear that the open-circuit voltage of a thermionic diode should be infinite because only an infinite retarding potential can stop *all* electrons. In fact, the open-circuit voltage is not large because the feeble current of high-energy electrons from the emitter is counterbalanced by small photoelectric and thermionic currents from the collector.

Under open-circuit conditions (ignoring photoelectric and thermoelectric effects),

$$J_E = J_C, \qquad (6.28)$$

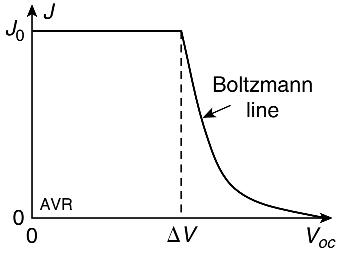


Figure 6.12 Current density versus voltage in a diode generator with neutralized space charge.

and

$$|V + \phi_C| > |\phi_E|. \tag{6.29}$$

$$A_E T_E^2 \exp\left[-\frac{q}{kT_E}(V+\phi_C)\right] = A_C T_C^2 \exp\left[-\frac{q}{kT_C}\phi_C\right], \tag{6.30}$$

$$\exp\left[-\frac{q}{kT_E}V - \frac{q}{k}\phi_C\left(\frac{1}{T_E}-\frac{1}{T_C}\right)\right] = \frac{A_C}{A_E}\left(\frac{T_C}{T_E}\right)^2, \tag{6.31}$$

$$\frac{q}{kT_E}V = -\frac{q}{k}\left(\frac{1}{T_E}-\frac{1}{T_C}\right)\phi_C - \ln\left[\frac{A_C}{A_E}\left(\frac{T_C}{T_E}\right)^2\right], \tag{6.32}$$

$$V = \phi_C\left(\frac{T_E}{T_C}-1\right) - \frac{k}{q}T_E \ln\left[\frac{A_C}{A_E}\left(\frac{T_C}{T_E}\right)^2\right]. \tag{6.33}$$

Example

A diode has an emitter made of tungsten and a collector made of thoriated tungsten. Operating temperatures are: $T_E = 2500$ K and $T_C = 500$ K. The open-circuit voltage is

$$V = 2.63\left(\frac{2500}{500}-1\right) - \frac{1.38\times 10^{-23}}{1.6\times 10^{-19}} \times 2500 \ln\left[\frac{30,000}{600,000}\left(\frac{500}{2500}\right)^2\right]$$
$$= 10.5 + 1.3 = 11.8 \text{ V}.$$

6.4.4 Maximum Power Output

The J versus V characteristics of an ideal thermionic diode with space charge neutralization can be divided into two regions:

$$\text{Region I: } |V| \le |\phi_E - \phi_C| \text{ where } J = J_0;$$

Region II: $|V| > |\phi_E - \phi_C|$ where $J = J_0 \exp\left\{\frac{q}{kT}[V - (\phi_E - \phi_C)]\right\}$.

In Region I, the diode is saturated, while in Region II, its current depends exponentially on V. The plot of J versus V in this latter region is known as the **Boltzmann line** (see Figure 6.12).

In Region I, the power density, P_{out}, delivered to the load increases linearly with the output voltage (thus increasing with the load resistance, R_L). It reaches a maximum of $J_0(\phi_E - \phi_C)$ when $V = (\phi_E - \phi_C)$. If the load resistance is increased further, there is a small increase in V coupled with an exponential decrease in J so that the output power decreases. This intuitive conclusion can be reached mathematically by determining the maximum of the function (applicable to Region II):

$$P_{out} = VJ = VJ_0 \exp\left\{\frac{q}{kT}\left[V - (\phi_E - \phi_C)\right]\right\}. \tag{6.34}$$

The maximum occurs for $V = kT/q$. Even at a temperature as high as 2500 K, kT/q is only 0.22 V, well below the break point $\phi_E - \phi_C$. This means that the maximum occurs at a voltage not valid in Region II. Thus,

$$P_{out_{max}} = J_0(\phi_E - \phi_C). \tag{6.35}$$

For the tungsten emitter, cesium plasma diode of the previous example, $\phi_E - \phi_C = 2.5$ V and, at 2500 K, $J_0 = 3000$ A m^{-2}. The maximum power output is 7500 W m^{-2}.

6.5 Losses in Vacuum Diodes with No Space Charge

6.5.1 Efficiency

If a thermionic generator had no losses of any kind and if the heat applied caused electrons to simply ooze out of the emitter, then all the heat input would be used to evaporate electrons. In reality, the evaporated electrons leave the emitter with some kinetic energy. Assuming for the moment that this kinetic energy is (unrealistically) zero, then the heat input to the generator would be simply $P_{in} = J\phi_E$ and the maximum power output would be $J_0(\phi_E - \phi_C)$. Consequently, the efficiency would be

$$\eta = \frac{P_{out}}{P_{in}} = \frac{J_0(\phi_E - \phi_C)}{J_0\phi_E} = 1 - \frac{\phi_C}{\phi_E}. \tag{6.36}$$

For the current example,

$$\eta = 1 - \frac{1.81}{4.52} = 0.60. \tag{6.37}$$

Nothing was said about the electrode temperatures, although they determine the limiting (Carnot) efficiency of the device. The implicit assumption is that $T_E >> T_C$. Otherwise, the collector itself will emit a current that may not be negligible compared with that of the emitter.

In a real generator, the heat source must supply numerous losses in addition to the energy to evaporate electrons. These losses are related to:

1. heat radiation
2. excess energy of the emitted electrons
3. heat conduction, and
4. lead resistance

and, in plasma diodes, to

5. heat convection,
6. ionization energy, and
7. internal resistance (called **plasma drop**).

6.5.2 Radiation Losses

The most serious loss mechanism is radiation from the hot emitter.

6.5.2.1 Radiation of Heat

If a *black body* is alone in space, it will radiate energy according to the **Stefan–Boltzmann** law:

$$P_r = \sigma T_E^4 \text{ W/m}^2, \tag{6.38}$$

where σ is the **Stefan–Boltzmann** constant (5.67×10^{-8} W m^{-2} K^{-4}). However, real bodies do not exactly obey this law. A fudge factor called the **emissivity**, ϵ, must be used, and the Stefan–Boltzmann law becomes

$$P_r = \sigma \epsilon T_E^4 \text{ W/m}^2. \tag{6.39}$$

Heat emissivity compares the actual heat radiated from a given material at a certain temperature with that of a black body at the same temperature. It is always smaller than 1.

To complicate matters, it turns out that ϵ depends on the frequency of the radiation. However, one can use average data taken over a wide frequency band and define a **total emissivity**, which is then, of course, frequency independent but is valid only for the frequency band considered. The body being modeled with this frequency-independent emissivity is called a **gray body**. The total emissivity is temperature-dependent as can be seen from Figure 6.13. Emissivity data valid only for a narrow band of frequencies are referred to as **spectral emissivity**, ϵ_λ.

When the emitter is in the neighborhood of another object (the collector of a thermionic generator, for instance), the net emissivity is altered. We are particularly interested in the estimation of the radiation exchanged

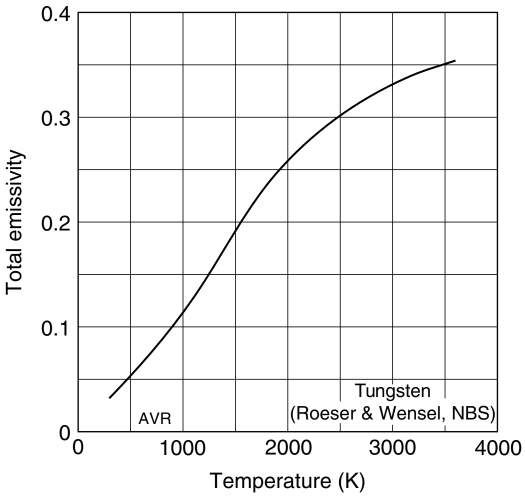


Figure 6.13 The total heat emissivity of tungsten as a function of temperature.

between a planar emitter and a planar collector that are in close enough proximity to allow us to treat them as parallel infinite planes. When radiation strikes a surface, it can be transmitted (if the material is translucent or transparent), reflected, or absorbed. In the cases of interest here, the materials are opaque and no transmission occurs.

Reflection can be specular (mirror-like) or diffuse, or a combination of both. We will consider only diffuse reflection. The part that is not reflected must be absorbed, and an **absorptivity coefficient**, α, is defined.

Based on equilibrium considerations, Kirchhoff's (thermodynamic) law concludes that the emissivity of a surface must equal its absorptivity—that is,

$$\epsilon = \alpha. \tag{6.40}$$

When the hot emitter radiates energy toward the collector, a fraction $\alpha_c = \epsilon_c$ is absorbed by the latter and a fraction, $1 - \epsilon_c$, is returned to the emitter. Here, again, part of the radiation is absorbed and part is returned to the collector. The radiation continues to bounce between the two plates, altering the apparent emissivity of the emitter. The **effective emissivity**, ϵ_{eff}, takes into account the environment around the radiating material.

Clearly, if the radiator is surrounded by a perfect mirror, its effective emissivity is zero: all radiated energy is returned to it. The effective emissivity of the emitter (in terms of the emissivities of the emitter and the collector) can be calculated by adding up the energies absorbed by the collector in all these bounces. Problem 6.1 asks you to do the math. The result is

$$\epsilon_{eff_E} = \frac{1}{1/\epsilon_E + 1/\epsilon_C - 1} \tag{6.41}$$

In the cases where either the emitter is alone in space or where it is surrounded by a black body (both situations corresponding to unity absorptivity and consequently, unit emissivity [$\epsilon_C = 1$]), the above equation reduces to

$$\epsilon_{eff_E} = \epsilon_E; \tag{6.42}$$

that is, the effective emissivity is simply the total emissivity of the emitter.

Notice also that the effective emissivity of the collector is

$$\epsilon_{eff_C} = \frac{1}{1/\epsilon_C + 1/\epsilon_E - 1} = \epsilon_{eff_E}. \tag{6.43}$$

Thus, two plates in close proximity have the same effective emissivity.

Example

Consider a tungsten plate heated to 2500 K and having a total emissivity, $\epsilon_E = 0.33$. It is mounted near another metallic plate whose total emissivity is $\epsilon_C = 0.2$. By using the formula above, one calculates an effective emissivity, for both plates, of 0.14.

The emitter radiates (from the surface facing the collector) a heat power density of

$$P_r = \sigma\epsilon_{eff}(T_E^4 - T_C^4). \tag{6.44}$$

The collector is usually at a much lower temperature than the emitter, but it may nevertheless be quite hot. Assume that $T_C = 1800$ K; then the radiated power (using the calculated $\epsilon_{eff} = 0.14$) is 226,700 W/m^2.

When $T_C << T_E$, Equation 6.44 reduces to

$$P_r = \sigma\epsilon_{eff}T_E^4, \tag{6.45}$$

and the collector temperature can be ignored.

6.5.2.2 Efficiency with Radiation Losses Only

From an examination of Equation 6.36 or from the general idea of how thermionic generators work, one could conclude that, for a fixed collector work function, ϕ_C, the larger the emitter work function, ϕ_E, the larger the efficiency. However, introducing heat radiation losses into the equation, changes this picture. In fact, there is one given value of ϕ_E that maximizes the efficiency.

Let us return to the tungsten emitter at 2500 K. At this temperature, tungsten will emit a current of 3000 A/m^2. If the collector is cesium plated, the output voltage will be about 2.5 V and the output power density will be 7.5 kW/m^2. To emit the 3000 A/m^2, the tungsten emitter consumes a heat energy of $3000\times4.52 \approx 14$ kW. For simplicity's sake, we will assume that the emitter radiates only toward the collector. The opposite side of the emitter faces a heat source that totally reflects any heat coming from the emitter. Then, as we saw in the example in the previous subsection, the emitter will radiate away into a cold collector a total of some 310 kW. This corresponds to the minuscule efficiency of 7.5/(310 + 14) = 0.023. Considering that there are really many more losses, it would seem that there is little hope of building a useful thermionic generator. Not so!

Owing to the large ϕ of tungsten, significant emitted currents require a high temperature that translates into a large radiation loss (owing to the T^4 temperature dependence). The solution is to find an emitting material capable of producing large currents even at relatively low temperatures—that is, a material with lower ϕ_E. However, although this reduces the radiation losses, it will also reduce the power output that is proportional to $\phi_E - \phi_C$. Which effect dominates?

The output power density is

$$P_L = J_0(\phi_E - \phi_C), \tag{6.46}$$

while the heat input power must be (neglecting all loss mechanisms other than heat radiation) the sum of the energy, $J_0\phi_E$, needed to evaporate the electrons plus the radiation losses, $\sigma\epsilon_{eff}T_E^4$. The efficiency is, then,

$$\eta = \frac{J_0(\phi_E - \phi_C)}{J_0\phi_E + \sigma\epsilon_{eff}T_E^4} = \frac{\phi_E - \phi_C}{\phi_E + \dfrac{\sigma}{A_E}\epsilon_{eff}T_E^2 \exp\left(\dfrac{q}{kT_E}\phi_E\right)}. \tag{6.47}$$

Here, we replaced J_0 by its Richardson's equation value.

We can now use the formula above to do some numerical experimentation. We selected an arbitrary collector work function, ϕ_C (1.81 V, in this example), and an emitter temperature, T_E. We also choose a value for the emission constant, A_E, which we took as 600,000 A m^{-2} K^{-2}.[†] We used $\epsilon_{eff} = 0.14$ as in the previous examples.

From Figure 6.14, one can see that, for maximum efficiency, an emitter at 2500 K should have a work function of about 3.3 V, and if at 2000 K, the work function should be about 2.7 V. Figure 6.15 shows the efficiency and the optimum work function depend on emitter temperature.

6.5.3 Excess Electron Energy

The tail end of the energy distribution of electrons in the emitter exhibits an exponential decay. To be emitted, the electrons need only an energy equal to $q\phi_E$. The excess appears as kinetic energy of the free electrons.

[†] The emission constant depends on the material chosen for the emitter, as does our free variable, ϕ_E. Thus, strictly speaking, it is not correct to use a constant A_E when varying ϕ_E. But this is just an exercise.

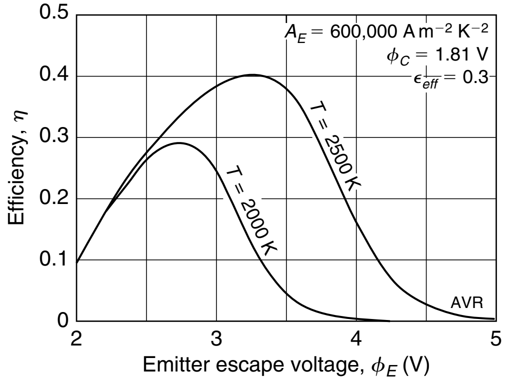


Figure 6.14 The efficiency of a thermionic generator at a given emitter temperature peaks for a given emitter work function.

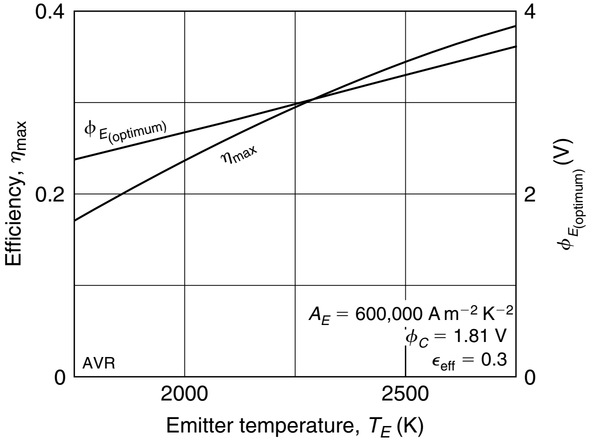


Figure 6.15 Dependence of efficiency and optimum work function on emitter temperature.

Because this energy must be supplied by the input heat source and because it is not available as electric output, it constitutes a loss. The average excess energy of the emitted electrons is between $2kT_E$ and $2.5kT_E$. Therefore, for each J_0 A/m^2, approximately $2J_0kT_e/q$ watts/m^2 is wasted and is returned as heat when the electrons impact the collector.

Including this type of loss, the efficiency of the generator is

$$\eta = \frac{\phi_E - \phi_C}{\phi_E + 2\frac{k}{q}T_E + \frac{\sigma}{A_E}\epsilon_{eff}T_E^2 \exp\left(\frac{q}{kT_E}\phi_E\right)}. \tag{6.48}$$

6.5.4 Heat Conduction

The structure that supports the emitter will unavoidably conduct away a certain amount of heat power, $\dot{Q}_E$. The heat source must supply this power. This means that the heat input must be increased by an amount $\dot{Q}_E$, and the efficiency is now

$$\eta = \frac{\phi_E - \phi_C}{\phi_E + 2\frac{k}{q}T_E + \frac{\dot{Q}_E}{J_0} + \frac{\sigma}{A_E}\epsilon_{eff}T_E^2 \exp\left(\frac{q}{kT_E}\phi_E\right)}. \tag{6.49}$$

6.5.5 Lead Resistance

The lead resistance, R_{int}, will dissipate $I^2R_{int} = J_0^2S^2R_{int}$ watts, reducing the available load power. S is the effective area of the emitter. The efficiency becomes

$$\eta = \frac{\phi_E - \phi_C - J_0S^2R_{int}}{\phi_E + 2\frac{q}{q}T_E + \frac{\dot{Q}_E}{J_0} + \frac{\sigma}{A_E}\epsilon_{eff}T_E^2 \exp\left(\frac{q}{kT_E}{}_E\right)}. \tag{6.50}$$

6.6 Real Vacuum-Diodes

The short-circuit current of a real vacuum thermionic power generator is severely limited by the space charge created by the electrons in transit from emitter to collector. The V-J characteristic of a possible generator of this type is displayed in Figure 6.16, together with the corresponding ideal characteristic of a diode of similar construction but whose space charge has magically been eliminated.

To alleviate the current limiting effect of negative space charges, the interelectrode spacing must be made extremely small. However, it is difficult

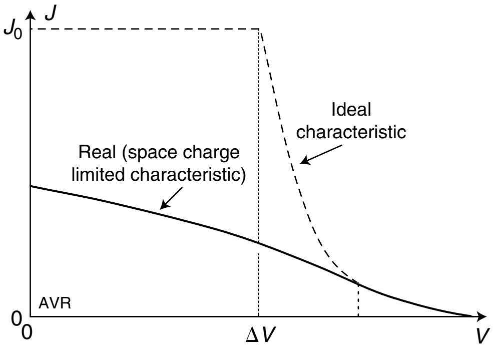


Figure 6.16 Space charge limitations degrade the output of a vacuum diode.

to make devices with spacings below some 10 μm. Even then, space charge limits the current to such an extent that the output power density can't exceed some 20 kW/m^2 (2 W/cm^2), about one order of magnitude too small for practical applications. This explains why vacuum diodes have found no practical application. Space charge neutralization arrangements discussed in the next sections are the solution.

The correct analysis of the V-J characteristic of a vacuum diode is surprisingly complicated. It was first presented by Langmuir[†] in 1913 and is reproduced in Volume II of Hatsopoulos (1973). We will skip this particular subject, especially because, as mentioned above, vacuum diodes appear to have no great future.

6.7 Vapor Diodes

Vacuum diodes with no space charge are essentially a contradiction in terms because the only practical way to eliminate space charges is to replace the vacuum by a low-pressure ion gas. To create ions, one has to fill the diode with a low-pressure neutral gas and then cause the gas to become partially ionized. The two ionization processes of interest here are **contact ionization** and **collision ionization**.

The cesium vapor used in most plasma diodes has five major effects on the performance of the diode:

1. Cesium ions can cancel the negative interelectrode space charge.
2. Cesium atoms condense on the electrodes, changing their effective work function in a predictable way.
3. Cesium atoms and ions scatter the electrons, interfering with their smooth travel from emitter to collector.
4. Inelastic electron collisions with cesium atoms can ionize them, increasing the ion current.
5. The heat convected away by the cesium vapor increases the losses of the diode.

When the cesium vapor concentration is sufficiently small, only Effects 1 and 2 are significant, and we have **low-pressure plasma diodes**; otherwise we have **high-pressure plasma diodes**. The latter are capable of high power densities (over 300 kW/m^2 or 30 W/cm^2).

Cesium vapor, from which ions will be produced, is obtained from the evaporation of the element stored in a reservoir whose temperature, T_r, can

[†] Irving Langmuir was the multitalented scientist who, working to General Electric, made vast contributions to the development of thermionics. Langmuir, besides introducing of the word "plasma" into the technical vocabulary, was also the developer of cloud seeding for producing rain. He received the 1932 Nobel Prize for Chemistry for "discoveries and inventions in surface chemistry."

be adjusted so as to permit selection of a desired cesium vapor pressure, p_{cs}. The vapor pressure of cesium was tabulated by Stull (1947), and, fitting a curve to the data he gathered, one obtains an empirical expression relating T_r, to the pressure, p_{cs}, of the resulting vapor,

$$p_{cs} = \frac{32.6 \times 10^9}{\sqrt{T_r}} \exp\left(-\frac{8910}{T_r}\right) \quad \text{Pa.} \tag{6.51}$$

The vaporization of cesium, being a heat-driven phenomenon, should obey Boltzmann's law,

$$p = p_0 \exp\left(\frac{\Delta H_{vap}}{RT_r}\right), \tag{6.52}$$

where $p_0 = 10^9$ pascals (Razor 1991).

The listed value of the enthalpy of vaporization of cesium varies considerably depending on the source consulted. The most frequently cited value is 67.74 MJ/kmole. However, using this value in the above equation leads to considerable departures from Stull's experimentally measured values. Much closer agreement with the empirically determined data is obtained when one uses $\Delta H_{vap} = 73.2$ MJ/kmole, which corresponds to the 0.75 eV suggested by Razor (1991).

The value of p_{cs} determines

1. the degree of cesium coverage of the electrodes—that is, the value of their respective ϕ; and
2. the degree of the space charge neutralization, measured by the parameter, β.

We are going to show that $\beta = 1$ corresponds to exact neutralization, $\beta < 1$ corresponds to incomplete ionization—that is, to an **electron-rich** condition, and $\beta > 1$ corresponds to overneutralization—that is, to an **ion-rich** condition.

6.7.1 Cesium Adsorption

Thermionic emission is a surface phenomenon; hence, it is not surprising that a partial deposition of cesium on the electrodes will alter their work function, ϕ, whose exact value will then depend on the **degree of coverage**, $\boldsymbol{\Theta}$.

Figure 6.17 is an example of how the work function of tungsten (**bare work function** of 4.52 eV) is influenced by the degree of cesium coverage. It is interesting to observe that in the region near $\Theta = 0.6$, the work function of the cesiated tungsten electrode is lower than that of pure cesium, which

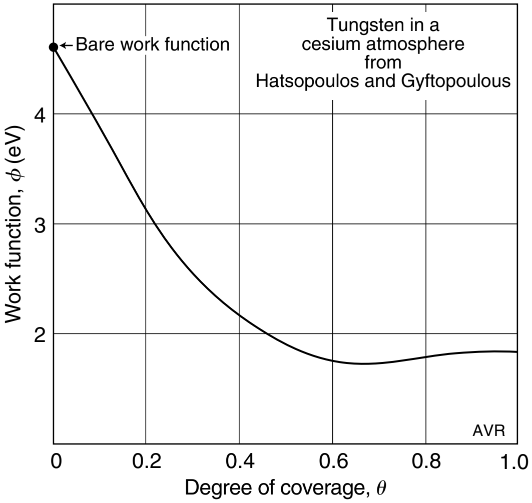


Figure 6.17 Effect of cesium adsorption on tungsten work function.

is 1.81 eV. The minimum work function of cesiated tungsten is 1.68 eV, a value lower than that of any pure metal. The tungsten-cesium combination shows sort of "eutectic-like" behavior.[†]

The degree of coverage depends on both the temperature of the emitter and on the pressure of the cesium vapor. The higher the temperature, the smaller the coverage. Consequently, Θ tends to be higher on the cooler collector than on the hotter emitter.

Obviously, the degree of coverage is quite sensitive to the cesium vapor pressure, p_{cs}. Altering this pressure is the main method of controlling ϕ, thus "tuning" this quantity to its chosen design value. As we saw, p_{cs} can be adjusted in a simple manner by varying the temperature, T_r, of the liquid cesium reservoir.

The relationship between ϕ and Θ may be of academic interest, but it is not too useful for modeling the behavior of vapor diodes because there is no simple way to measure Θ. It is better to develop a formula for ϕ based on more easily measured parameters. We used the data reported by Houston and Dederick (1973) to come up with a formula that yields the work function of a representative tungsten electrode as a function of the ratio of the electrode temperature, T, to the cesium reservoir temperature, T_r. The formula appears to be acceptably accurate for T/T_r ratios above 2 but fails to reproduce the minimum in ϕ that occurs near this point. Nevertheless, the formula is useful for general modeling work and for solving

[†] The word "eutectic" normally applies to the alloy that has the lowest melting point of all alloys of similar constituents.

homework problems.

$$\begin{aligned}\phi = 9.499 - 9.9529T/T_r + 4.2425(T/T_r)^2 \\ - 0.67592(T/T_r)^3 + 0.03709(T/T_r)^4. \qquad (6.53)\end{aligned}$$

In using this formula, one must be careful not to let ϕ exceed the bare metal value of 4.52 V. If the formula yields larger values (thus showing that its validity interval was exceeded), clamp the values of ϕ to 4.52 V, independently of the T/T_r ratio. Also clamp the value of ϕ to 1.81 V for $T/T_r < 1.9$.

The value of ϕ is a strong function of T/T_r but also depends weakly on the value of T_r. Therefore, formula 53 cannot be trusted for precision work on problems in which the latter temperature varies widely.

Symbology

A more complicated symbology is useful in the analysis of vapor diodes, where, in addition to electron currents, there are also important ion currents to consider.

In the case of low-pressure diodes, one has to deal with four different currents:

1. J_{e_E}, the electron current from emitter to collector.
2. J_{e_C}, the electron current from collector to emitter.
3. J_{i_E}, the ion current from emitter to collector resulting from contact ionization of the cesium gas at the emitter.
4. J_{i_C}, the ion current from collector to emitter resulting from contact ionization of the cesium gas at the collector.

Both electron and ion currents are thermionically emitted by the electrodes, the electron according to Richardson's law discussed previously and the ion according to a law we will derive in one of the coming sections. Only under saturation conditions will the whole emitted currents reach the opposite electrode. We shall distinguish the *emitted* or *saturation* current by appending a "0" to the appropriate symbol. For example, the ion saturation current generated by the emitter is $J_{i_{E_0}}$.

The load current, J_L is given by

$$J_L = J_{e_E} - J_{e_C} - J_{i_E} + J_{i_C}. \qquad (6.54)$$

Currents from collector to emitter are called **back emission**.

6.7.2 Contact Ionization[†]

A glance at the periodic table of elements reveals that alkali metals (Column I) have ionization potentials lower than those of elements in the other columns. It also becomes apparent that the ionization potential of any element in a given column tends to be lower the higher its atomic number. Hence, one would expect that francium, occupying the southwest tip of the table, should have the lowest ionization potential on record. However, it is cesium, the next-door neighbor to the north, that holds this honor with a meager 3.89 eV. This is one of the reasons cesium (pure or with some additives) is so popular for space charge neutralization.

Cesium vapor can undergo **contact ionization** by simply "touching" a hot metal surface provided the material of this surface has a work function larger than the ionization energy of the vapor. These conditions are satisfied by, among others, tungsten. Figure 6.18 illustrates what happens.

In a metal, there are many unoccupied levels above the Fermi level. According to classical mechanics, an electron in a gas atom can only leave the atom and transfer itself to one of the unoccupied levels in the neighboring metal if it has more energy than the ionization potential, ϕ_{ion}. In other words, the electron in cesium would have to first free itself from the parent atom. Quantum mechanics, however, predicts a finite probability that an electron, even with less than the ionization potential, will **tunnel** through and will migrate to the metal. The atom becomes "spontaneously" ionized, provided $\phi_{ion} < \phi_E$ and provided also that the atom is near enough to the metal. Remember that gas atoms in "contact" with a hot electrode are, in fact, a certain distance away from that electrode.

Do not confuse the ionization potential of cesium (3.89 eV) with its work function (1.81 eV). The ionization potential represents the energy necessary to remove the outer electron from the atom, while the work

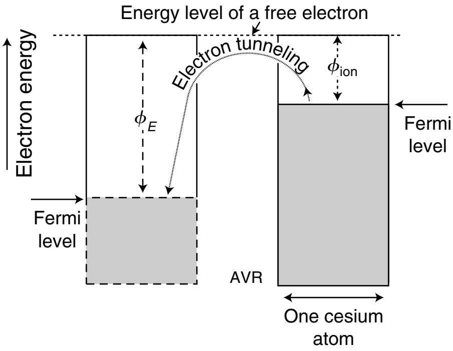


Figure 6.18 Energy levels in a gas near a metal.

[†] Also known as "thermal" or "surface" ionization.

function represents the (smaller) energy to remove a conduction electron from the solid.

6.7.3 Thermionic Ion Emission

We shall derive an expression that yields the magnitude of the thermionically emitted ion current resulting from the contact ionization of cesium gas discussed in the preceding section. Contact ionization captures an electron from a cesium atom liberating the corresponding ion.

In a one-dimensional gas, the flux, Φ, of molecules moving in either of the two possible directions is

$$\Phi = \frac{1}{2} n v. \tag{6.55}$$

The factor, 1/2, results from half the molecules moving in one direction and the other half in the opposite direction.

Since $\frac{1}{2} m_{cs} v^2 = \frac{1}{2} k T_{cs}$ (here, T_{cs} is the temperature of the cesium gas), the velocity is

$$v = \sqrt{\frac{kT_{cs}}{m_{cs}}}, \tag{6.56}$$

where m_{cs} is the mass of the cesium molecule (or atom, in this case). This results in a flux of

$$\Phi = \frac{n}{2}\sqrt{\frac{kT_{cs}}{m_{cs}}}. \tag{6.57}$$

From the perfect-gas law, $n = p_{cs}/kT_{cs}$.

$$\Phi = p_{cs}\sqrt{\frac{1}{4 m_{cs} k T_{cs}}}. \tag{6.58}$$

For a three-dimensional gas, the correct expression for the flux is

$$\Phi = p_{cs}\sqrt{\frac{1}{2\pi m_{cs} k T_{cs}}}. \tag{6.59}$$

The gas in the interelectrode space consists of electrons and two different species of larger particles: neutral atoms and ions (both with essentially the same mass, m_{cs}). Hence, the flux we calculated is the sum of the flux of ions, Φ_i, and the flux of neutrals, Φ_n. Thus, Equation 6.59 becomes

$$\Phi_i + \Phi_n = p_{cs}\sqrt{\frac{1}{2\pi m_{cs} k T_{cs}}}. \tag{6.60}$$

The ionization of neutral atoms is a heat-activated process and as such, should follow **Boltzmann's law**:

$$\frac{\Phi_i}{\Phi_n} \propto \exp\left[\frac{q}{kT}(\phi - \phi_{ion})\right]. \tag{6.61}$$

T is the temperature of the electrode, ϕ is its work function, and ϕ_{ion} is the ionization potential of cesium.

For cesium, the coefficient of proportionality is usually taken as 1/2; hence

$$\frac{\Phi_i}{\Phi_n} = \frac{1}{2\exp\left[\frac{q}{kT}(\phi_{ion} - \phi)\right]}. \tag{6.62}$$

Expression 62 is known as the **Saha–Langmuir** equation. Ionization does not stop completely when $\phi < \phi_{ion}$. However, it decreases rapidly with decreasing ϕ.

We are interested not in the ratio of Φ_i to Φ_n, but rather in the ratio of Φ_i to the total flux of particles, $\Phi_i + \Phi_n$,

$$\frac{\Phi_i}{\Phi_i + \Phi_n} = \frac{1}{1 + \frac{\Phi_n}{\Phi_i}} = \frac{1}{1 + 2\exp\left[\frac{q}{kT}(\phi_{ion} - \phi)\right]}. \tag{6.63}$$

Combining Equation 6.63 with Equation 6.60.

$$\Phi_i = \frac{p_{cs}}{\sqrt{2\pi m_{cs} kT_{cs}}\left[1 + 2\exp\left(\frac{q(\phi_{ion} - \phi)}{kT}\right)\right]}, \tag{6.64}$$

$$J_{i_0} = \frac{qp_{cs}}{\sqrt{2\pi m_{cs} kT_{cs}}\left[1 + 2\exp\left(\frac{q(\phi_{ion} - \phi)}{kT}\right)\right]}. \tag{6.65}$$

This is the ion saturation current emitted by a hot electrode exposed to cesium vapor.

6.7.4 Space Charge Neutralization Conditions

Cesium vapor was introduced into the diode in the hope of neutralizing the negative space charge resulting from the streaming cloud of electrons in transit from emitter to collector. In the preceding subsection, we showed that the cesium gas can indeed give rise to ion currents in the device. Let us investigate what magnitude these currents must have in order to cancel the electron space charge.

The average thermal velocities of ions and of electrons are

$$v_i = \sqrt{\frac{kT_i}{m_i}} \tag{6.66}$$

and

$$v_e = \sqrt{\frac{kT_e}{m_e}}. \tag{6.67}$$

The two current densities are (defining $T \equiv T_e = T_i$)

$$J_i = qn_i v_i = qn_i \sqrt{\frac{kT}{m_i}}, \tag{6.68}$$

and

$$J_e = qn_e v_e = qn_e \sqrt{\frac{kT}{m_e}}; \tag{6.69}$$

hence,

$$\frac{J_i}{J_e} = \frac{n_i}{n_e} \sqrt{\frac{m_e}{m_i}}. \tag{6.70}$$

For exact space charge neutralization, the electron concentration must equal the ion concentration—that is, $n_e = n_i$.

$$J_i = \sqrt{\frac{m_e}{m_1}} J_e = \frac{J_e}{492}. \tag{6.71}$$

Hence, the plasma will be **ion-rich** if

$$J_i > \frac{J_e}{492}. \tag{6.72}$$

If the above inequality is not satisfied, the plasma is **electron-rich** and the negative space charge is not neutralized.

It is convenient to define an **ion-richness parameter** β:

$$\beta \equiv 492 \frac{J_i}{J_e}. \tag{6.73}$$

When $\beta < 1$, negative space charge is not completely neutralized (electron-rich regimen), and when $\beta > 1$, negative space charge is completely neutralized (ion-rich regimen).

6.7.5 More V-J Characteristics

It was shown that the negative space charge caused by an electron flow, J_e, can be neutralized by the flow of an ion current, J_i, about 500 times smaller. This is, of course, the result of cesium ions being some 500 times more "sluggish" than electrons. The small ion current needed to neutralize the negative space charge has negligible influence on the output current of the device. In fact, one can ignore the influence of J_i on the output current provided $\beta < 10$—that is, provided $J_i < J_e/50$. However, the ion current is frequently much larger than that required for negative space charge neutralization, and it has then to be taken into account when computing the V-J characteristics of a diode. Most of the time, the back emission from the collector is negligible owing to the low collector temperature. Yet, this is not invariably true—sometimes the back emission current, J_{e_C}, has to be included. The same applies to the (rare) case when the ion emission current, J_{i_C}, from the collector is significant. In a general treatment of this subject, one should write the load current, J_L, as

$$J_L = J_{e_E} - J_{e_C} - J_{i_E} + J_{i_C}. \tag{6.74}$$

The values of these individual currents are

	$V < \Delta V$	$V > \Delta V$
J_{e_E}	$J_{e_{E_0}}$	$J_{e_{E_0}} \exp\left[-\frac{q}{kT_E}(V - \Delta V)\right]$
J_{e_C}	$J_{e_{C_0}} \exp\left[-\frac{q}{kT_C}(\Delta V - V)\right]$	$J_{e_{C_0}}$
J_{i_E}	$J_{i_{E_0}} \exp\left[-\frac{q}{kT_E}(\Delta V - V)\right]$	$J_{i_{E_0}}$
J_{i_C}	$J_{i_{C_0}}$	$J_{i_{C_0}} \exp\left[-\frac{q}{kT_C}(V - \Delta V)\right].$

The assumptions are as follows.

1. The interelectrode space charge is either 0 or positive—that is, $\beta \geq 1$.
2. If the interelectrode space charge is positive, it is not big enough to seriously limit the ion currents. This particular assumption is often violated, and the ion currents in the region in which $V > \Delta V$ can be much smaller than that calculated with the above formulas.

To illustrate several operating conditions, we consider a tungsten electrode diode whose emitter and collector temperatures are, respectively, 1800 K and 700 K. Different cesium reservoir temperature are used to obtain selected values of βs. The plots computed for values of β of 1, 10, and 100 appear in Figures 6.19, 6.20, and 6.21.

In Figure 6.19, the cesium reservoir temperature was adjusted to 580 K to achieve an essentially exact cancellation of the space charge. The V-J characteristic is that of an ideal vacuum diode (with no space charge). Neither ion currents nor back emission are noticeable.

When β was adjusted to 10, the ion current became large enough to influence the output. At large load voltages, this ion current is dominant, and the output current reverses its direction. Finally, at $\beta = 100$, the calculated ion current is very significant. In reality, measured results would probably show a much smaller ion current because, in our calculations, we did not take into account the large positive space charge that develops, a space charge that will severely limit the ion current.

By lowering the emitter temperature to 1200 K, while raising the collector temperature to 750 K, we create a situation in which the back emission becomes important (Figure 6.22). The collector work function under these

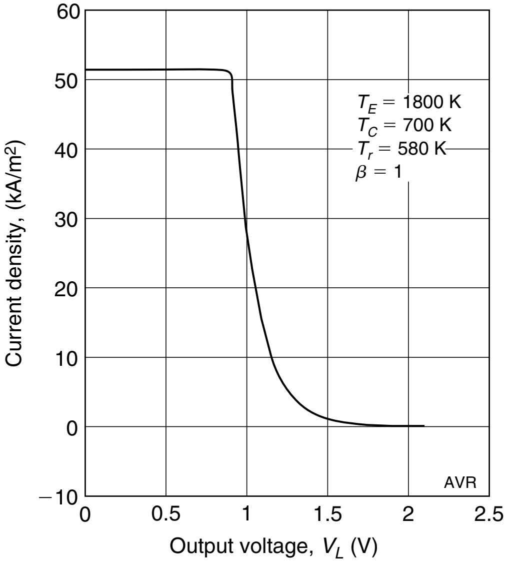


Figure 6.19 When $\beta = 1$, the only significant current is the electron current from the emitter.

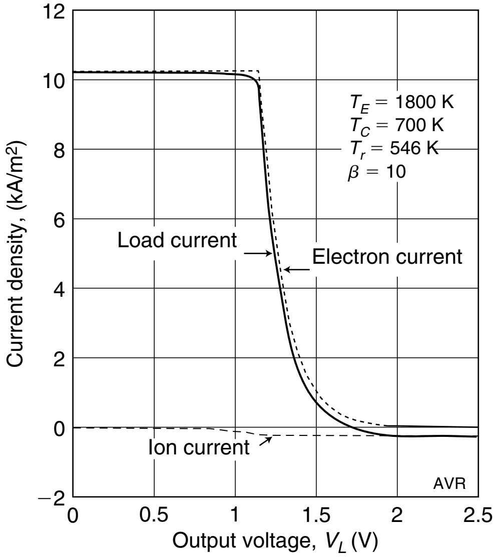


Figure 6.20 When $\beta = 10$, the ion current begins to affect the output, which is slightly negative at high output voltages.

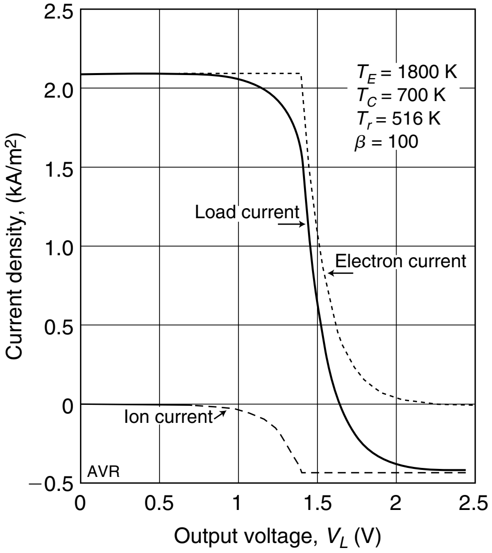


Figure 6.21 When $\beta = 100$, the ion current greatly affects the output.

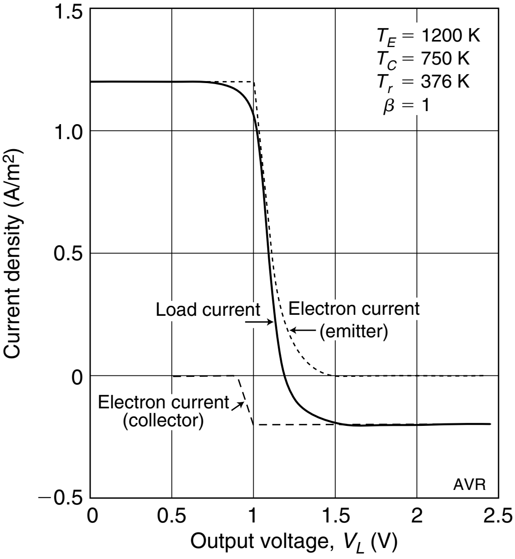


Figure 6.22 When the temperature of the emitter is not much higher than that of the collector, back emission becomes important.

conditions is 1.81 (owing to 100% cesium coverage), so that a reasonably large electron current is emitted while the hotter emitter is still operating at a relatively high work function of 2.82 V.

Those who studied the three V-J plots of Figures 6.20 to 6.21 may have spotted an apparent paradox. It would seem that increasing the cesium reservoir temperature (and consequently the cesium vapor pressure) should increase β, because a larger amount of vapor is available to be ionized and to contribute to the ion current. Yet, it was necessary to reduce T_r in order to raise β. To understand what happens, please refer to Figure 6.24.

At the relatively high emitter temperature of 2100 K, there is little cesium condensation on that electrode until the cesium vapor pressure reaches about 1 Pa. The emitter work function remains at the high bare metal value of 4.52 V, and, consequently, the emitted electron current is relatively low and unchanging (see Figure 6.23). On the other hand, the ion current is, in this low vapor pressure region, essentially proportional to the p_{cs}; in other words, it rises as p_{cs} rises. β, being proportional to J_i/J_E, also rises. Once the vapor pressure exceeds 1 Pa, ϕ_E starts decreasing, causing an exponential increase in J_E. But as ϕ_E grows, the difference, $\phi_E - \phi_{ion}$, begins to fall, and so does the rate of ionization of the gas. The ion current diminishes causing β to come down again. So the dependence of β on p_{cs} is not monotonic. In this example, β reaches a maximum at around $p_{cs} = 8$ Pa. Figures 6.20 to 6.22 correspond to a region in which β is in the descending part of its trajectory.

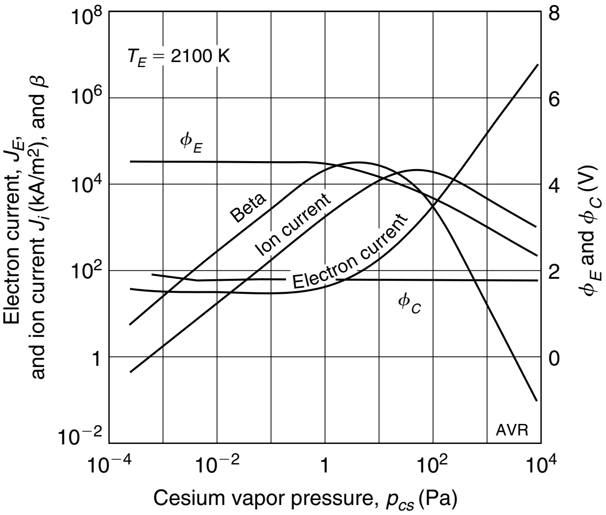


Figure 6.23 Behavior of the electron and the ion current as a function of the cesium vapor pressure.

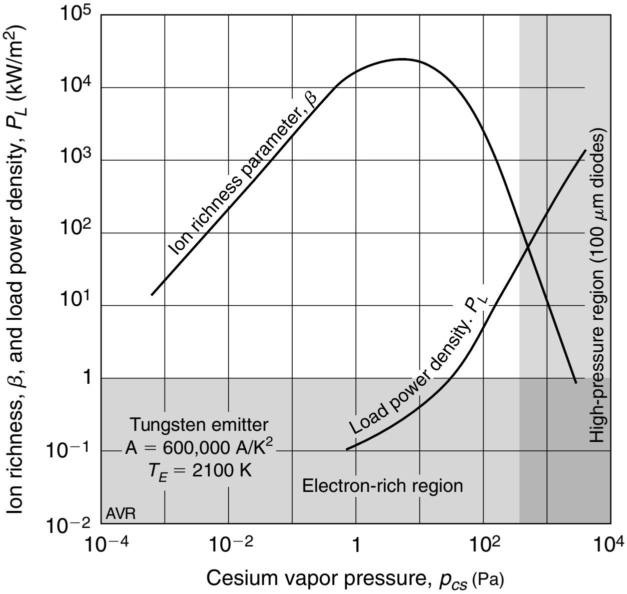


Figure 6.24 Dependence of beta and of the output power on the cesium vapor pressure.

Plasma diodes operate in this region of decreasing β,where, even at moderate temperatures, tungsten, being covered with cesium, emits current densities in the 100 kA/m^2 range. This would lead to acceptable output power densities except that the high-output power region tends to occur at such high vapor pressures that even for small interelectrode spacings, electron–cesium collisions become significant. As a rule of thumb, when $p_{cs}d > 0.0033$ N/m, collisions cannot be disregarded, and we leave the domain of low-pressure diodes to enter that of high-pressure ones. In the preceding inequality, d, is the interelectrode spacing, in meters.

The shaded region in which $\beta < 1$ corresponds to the presence of negative space charge. In this region, the performance of the diode is intermediate between that of a vacuum device and the ideal device with no space charge (see Figure 6.16). The unshaded region (where $\beta > 1$ and $p_{cs}d < 0.0033$ N/m) is the ion-rich region in which a low-pressure diode exhibits ideal characteristics. Finally, the vertically elongated rectangular region on the right is a region in which the inequality $p_{cs}d < 0.0033$ N/m is violated. The width of this shaded area is for a diode with a small interelectrode spacing of 100 μm. In this area, we no longer have a low-pressure diode. With diodes of wider spacing, the shaded area is broader, and the output limitations are correspondingly more stringent. Although diodes operating in the unshaded region can have (very nearly) ideal characteristics, their output power density is disappointing—it is comparable to that of vacuum devices. We have come back to the situation in which, similar to what happened with vacuum devices, useful devices would require prohibitively small separations between emitter and collector.

6.8 High-Pressure Diodes

The difficulty with a low-pressure diode is that it has two contradictory requirements: First, the emitter must have a *low* work function in order to produce the needed high current density. Second, it simultaneously must have a *high* work function in order to generate sufficient ionization to eliminate the negative space charge that would otherwise impede the flow of the desired current. A solution to this dilemma is the use of an ionization scheme that does not depend on a high emitter work function. Such a mechanism is collision ionization in which the very electrons whose charge is to be neutralized collide with a cesium atom with enough violence to ionize them. Although for each created ion, one additional electron is generated, there is a net gain as far as charge neutralization is concerned because the ion will linger in the interelectrode space about 500 times longer.

Inelastic collisions between the streaming electrons and the cesium atoms in the interelectrode space can, under some circumstances, cause the ionization of the cesium. This is an important process in the operation of high-pressure vapor diodes. It is intuitive that the probability of an electron colliding with a gas atom is proportional to both the gas concentration (and hence, to the pressure, p_{cs}) and to the length of the path the electron must take—that is, to the separation, d, between the electrodes. It can be shown that when the $p_{cs}d$ product is smaller than 0.0033 N/m, the electrons suffer negligible collisions. Here, the cesium pressure is measured in pascals and the interelectrode distance, in meters.

In high-pressure diodes, collisions between electrons and the interelectrode gas are numerous, so substantial scattering occurs. This causes power losses (**plasma drop**) resulting in appreciably reduced output currents and output voltages. Nevertheless, the collisions may induce so much cesium vapor ionization that the population of carriers is substantially increased and large load current can result. Current densities of more than 500,000 A/m^2 are realizable.

When the disadvantages and advantages of high-pressure operation are weighed, the latter make a compelling case for using high-pressure diodes. Indeed, practical high-pressure diodes can realize a comparatively high output power density, far beyond that achievable with their low pressure cousins. Even though load voltages tend to be smaller than those in low-pressure diodes, output power densities of some 30 W/cm^2 have been demonstrated.

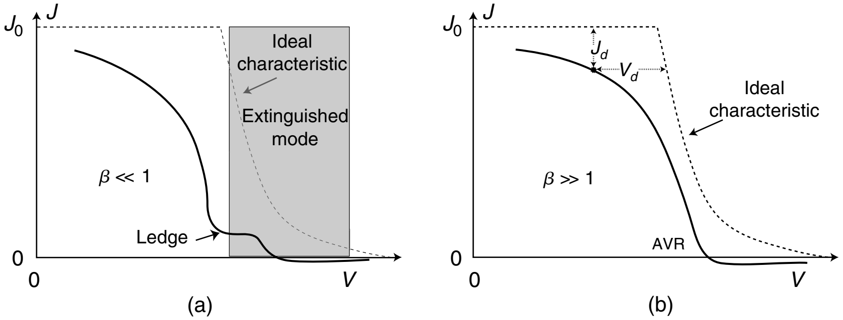


Figure 6.25 Characteristics of a high-pressure diode. Left, with very small beta; right, with very large beta.

Analysis of the performance of high-pressure diodes is very complicated owing to the interaction of many phenomena. Here, we will have to be satisfied with a simple explanation of what goes on.

In Figure 6.25a, $\beta \ll 1$ suggesting considerable negative space charge and, thus, small load currents. This is, indeed, the situation at high-output voltage where the characteristics resemble those of a space charge limited device: lowering the voltage causes the current to increase modestly and to saturate at a low level. This behavior corresponds to the ledge in the figure.

Further reduction in voltage moves the diode into a regimen in which frequent collisions generate enough ions to drive the load current up sharply. In this regimen, ions are derived mainly from inelastic collisions (ions from the emitter being negligible), and the diode is said to be in the **ignited mode**, contrasting with the initial region called the **unignited** or **extinguished** mode in which ions from the emitter are dominant.

When $\beta >> 1$ as depicted in Figure 6.25b, there is no perceptible extinguished mode—the ignited mode already starts at large load voltages.

It is this abundance of ions created by inelastic collision that leads to the large currents in high-pressure diodes and results in acceptable output power densities. But this operation comes at a cost: the real characteristics depart from the ideal by both a current and a voltage deficiency (J_d and V_d, in Figure 6.25b). V_d, called **plasma drop**, is made up of three components:

1. V_i, the voltage drop (about 0.4 V) associated with the ionization of cesium atoms by electron impact.
2. V_p, the voltage drop owing to the electron scatter by cesium atoms.
3. V_s, a potential barrier that forms next to the electrodes.

The relatively large cesium vapor pressure conducts substantial heat from emitter to collector, diverting some of the input energy. On the other hand, the heavy cesium plating that occurs in high-pressure diodes reduces the emitter work function and brings down the amount of input heat power required to evaporate the electrons from the emitter.

Although, as pointed out, the high-pressure diode is by far the most practical thermionic generator, it tends to be an expensive device owing in good part to the special materials that have to be used.

For a much more complete discussion of high-pressure thermionic diodes, read the article by Ned Razor (1991).

References

Hatsopoulos, G. N., and E. P. Gyftopoulos, *Thermionic Energy Conversion* (2 vols.), MIT Press, **1973**.

Houston, J. M., and P. K. Dederick, The electron emission of the Pt-group metals in Cs vapor, Report on the Thermionic Conversion Specialist Conference, San Diego, California, October. 25–27, **1965**, pp. 249–257.

Iannazzo, S., A survey of the present status of vacuum microelectronics, *Solid-State Electronics* (26 vol.) 3, pp. 301–320, **1993**.

Razor, N. S., Thermionic energy conversion plasmas, *IEEE Trans. on Plasma Sci.* 19, p. 191, **1991**.

Shavit, A., and G. N. Hatsopoulos, Work function of polycrystalline rhenium, *Proc IEEE* (Inst. Elec. Electron. Eng.), (54), pp. 777–781, **1996**.

Stull, D.R., *Industrial and Engineering Chemistry* 39, p. 517, **1947**.

PROBLEMS

6.1 Consider two metallic plates with linear dimensions much larger than their separation from one another (so they can be treated as infinite planes). Prove that their effective emissivity on the sides that face one another is given by

$$\epsilon_{eff} = \frac{1}{1/\epsilon_1 + 1/\epsilon_2 - 1} \quad ,$$

where ϵ_1 and ϵ_2 are the total emissivity of the two plates when taken by themselves.

6.2 Two metallic plates are held parallel to one another, separated by a small spacing. There is neither electric nor thermal conduction between them. This "sandwich" is installed in outer space, at 1 AU from the sun. One of the plates (made of pure tungsten) is exposed to concentrated sunlight, while the other (made of thoriated tungsten) is in shadow.

The tungsten plate has its outer surface (the one that receives the sunlight) treated so that its thermal emissivity and its albedo are both 10%. The inner surface (the one that faces the thoriated tungsten plate) has an emissivity of 25%. This is the absolute emissivity and does not take into account the radiation reflected by the colder plate. The colder plate has 100% emissivity on each side.

There is no radiation in space other than that from the two plates and from the concentrated sunlight. Unconcentrated sunlight, at 1 AU from the sun, has a power density of 1350 $\mathrm{W/m^{-2}}$.

The area of each plate is 1 $\mathrm{m^2}$.

There is no space charge between the plates.

The emissivities are frequency independent.

1. What is the sunlight concentration ratio that causes the hotter plate to be 1000 K above the colder plate, under steady-state conditions? What are the temperatures of the plates?
2. Now, an external electric load is connected between the plates, and the concentration ratio is adjusted so that the tungsten plate heats up to 3100 K. The load draws the maximum available power. What is the temperature of the colder plate? What is the required concentration ratio? What is the electric power generated? Assume steady state and no space charge. Consider the system as an ideal plasma diode but disregard any ion current.

6.3 A low-pressure cesium vapor diode operating under a no negative space charge condition has emitter and collector with a 2 $\mathrm{cm^2}$ area. The collector has a work function of 1.81 V. The work function of the tungsten collector can be adjusted by varying the pressure, p_{cs}, of the cesium vapor. The emission constant is unaltered by the cesium and is 600,000 A m^{-2} K^{-2}.

The temperatures of the emitter and of the collector are 2100 K and 1100 K, respectively. The effective emissivity of the emitter is 0.3 on the side facing the collector and 0 on the opposite side.

The only loss mechanisms are the cost of emitting electrons, excess energy of these electrons, and the radiation from the emitter toward the collector.

Which emitter work function maximizes the efficiency?

What is this efficiency?

6.4 Refer to the diode in Problem 6.3.

You want to operate with an emitter work function of 3.0 V.

1. Estimate the fraction, Θ, of the emitter area that has to be plated with cesium.
2. If the interelectrode spacing is large, collisions will be significant. What is the largest interelectrode spacing, d, that allows the diode to operate in the "no collision" regimen (i.e., as a low-pressure device)? Please comment.

6.5 In a low-pressure cesium vapor diode, the actual electron emitter current will be equal to the emitter saturation current only if there is no negative interelectrode space charge. Is there a negative space charge in a diode with an emitter work function of 3.3 eV, an emitter emission constant of 600,000 A m^{-2} K^{-2}, and an emitter temperature of 2350 K?

6.6 If the diode of Problem 6.5 is to operate as a low-pressure diode, what is the largest allowable interelectrode gap?

Assume that the cesium vapor pressure is 2000 Pa.

6.7 What is (approximately) the maximum power output density of the diode of Problem 6.5?

Assume that the diode operates as a low-pressure diode with no negative space charge. Assume also that there is no significant back emission.

6.8 Consider a cesium vapor thermionic generator with a 10 cm^2 emitter having an emission constant of 5×10^5 A m^{-2} K^{-2} and a work function of 3.0 eV. The collector is at 1000 K and, owing to the cesium condensation on it, has a work function of 1.8 eV. The effective emissivity of the emitter (operated at 2000 K) is 0.3 for the side that faces the collector. The other side does not radiate. Losses in the device are only through radiation and through excess kinetic energy of the emitted electrons. There are no conduction losses and no plasma drop. There is no negative space charge. Disregard any ion current.

What load resistance causes the device to furnish the highest possible output power?

Under the above conditions, what is the efficiency of the device?

What is the required input power?

6.9 Consider two solid plates: one made of tantalum and one of tungsten. Which of the two can be made to emit, in vacuum, the higher thermionic current density?

6.10 A thermionic generator has a tungsten emitter heated to 2500 K and a cesium coated collector at 1000 K. Electrodes measure 10 × 10 cm. Under all circumstances, the diode behaves as if it had an internal resistance of 10 mΩ. Disregard collector emission and thermoelectric effects. Assume there is no space charge. Calculate the power delivered to loads of 80 and 100 mΩ.

6.11 Assume the two materials being compared are at the same temperature. Which one emits thermionically, regardless of temperature, the larger current density? chromium or tantalum?

6.12 A cube (side d meters) of a highly radioactive material floats isolated in intergalactic space. Essentially, no external radiation falls on it. Owing to its radioactivity, it generates 1200 kW per cubic m. Call this generation rate Γ.

The material has the following characteristics:

Thermionic emission constant: $200,000$ A m^{-2} K^{-2}

Work function: 3.0 V

Heat emissivity (at the temperature of interest): 0.85

What is the temperature of the cube

1. if $d = 1$ m?
2. if $d = 10$ cm?

6.13 A vacuum diode is built with both emitter and collector made of pure tungsten. These electrodes are spaced 1 μm apart. The “emitter” is heated to 2200 K and the “collector” to 1800 K.

1. Estimate the maximum power that the diode can deliver to a load.
2. Introduce cesium vapor. Explain what happens when the cesium vapor pressure is progressively increased. Plot the output power as a function of the cesium vapor pressure, p_{cs}. Disregard collision ionization and all ion currents.

6.14 A thermionic diode has parallel plate electrodes 2 by 2 cm in size and separated by 1 mm. Both emitter and collector are made of tantalum.

The emitter temperature is 2300 K, and the collector is at 300 K.

If a vacuum is maintained between the plates, what is the current through the diode if a $V_{CE} = 100$ V are applied across it making the collector positive with respect to the emitter?

Repeat for $V_{CE} = 50V$.

6.15 A cube (edge length: 1 m) is placed in a circular orbit around the sun at 0.15 AU. This means that its distance to the sun is 15% of the sun–Earth distance and that the cube is way inside Mercury's orbit.

The orbit is stabilized so that a flat face always faces the sun.

The cube is made of a material that has 90% light absorptivity (i.e., has an albedo of 10%; hence it looks pretty black). Its heat emissivity is 30%.

The solar constant at the orbit of Earth is 1360 W/m^2.

The material of the cube has a low work function for thermionic emission: $\phi = 1.03\,\text{V}$. Its emission constant is $4000\,\text{A}\,\text{m}^{-2}\,\text{K}^{-2}$.

What is the equilibrium temperature of the cube?

6.16 In the text, we have presented an empirical equation relating the work function, ϕ, of tungsten to the ratio, T/T_r, of the electrode temperature to the cesium reservoir temperature. This assumes that, although ϕ is a function of T/T_r, it is not a function of T, itself. This is not quite true. ϕ is also a weak function of T. The formula value has an uncertainty of ±0.1 V, for the $1500 < T < 2000\,\text{K}$ range.

What is the corresponding uncertainty factor in the calculated value of the emitted current when $T = 1500\,\text{K}$? And when $T = 2000\,\text{K}$?

6.17 What is the output power density of a cesium plasma diode with tungsten electrodes? Assume that there are no negative space charges and that collisions between electrons and cesium atoms are negligible.

Emitter temperature, $T_E = 2100\,\text{K}$.

Collector temperature, $T_C = 1000\,\text{K}$.

Cesium reservoir temperature, $T_r = 650\,\text{K}$.

Cesium vapor temperature, $T_{cs} = T_r$.

Electrode emission constant, $A = 600,000\,\text{A}\,\text{m}^{-2}\text{K}^{-2}$.

Interelectrode spacing, $d = 0.1\,\text{mm}$.

Assume that the ion current has no effect on the load current. Nevertheless, it definitely has an effect of the space charge.

Check if your assumptions are valid. What is the nature of the space charge? Is it negative, zero, or positive? Can you neglect electron–cesium collisions?

Is the output power density that you calculated valid? If not valid, do not recalculate.

Remember that there are limitations on the validity of the expression that yields ϕ as a function of T/T_r. Use the following boolean test:

IF $T/T_r < 2.2$ THEN $\phi = 1.81$ ELSE IF $\phi > 4.52$ THEN $\phi = 4.52$ ELSE $\phi = \phi$.

6.18 If you examine the equation that yields the value of the ion current in a low-pressure thermionic generator, you will find that all else being constant, the lower the electrode temperature the higher the thermionically emitted ion current. This would suggest that the cooler collector should emit a larger ion current than the hotter emitter. Yet, in all discussions of ion currents, the only current mentioned is the one from the emitter. The ion current from the collector is usually ignored.

Give a simple and short explanation of why this makes sense. Assume that both emitter and collector are made of tungsten and that the interelectrode gas is cesium vapor.

Chapter 7
AMTEC[†]

7.1 Operating Principle

AMTEC stands for **Alkali Metal Thermal Electric Converter** and is an example of a class of devices known as **Concentration-Differential Cells**. Its operation is conceptually simple. It takes advantage of the special properties of β' alumina, which is an excellent conductor of alkali metal ions but a bad conductor of electrons. In other words, it is an electrolyte. A slab of this material constitutes **$\boldsymbol{\beta'}$ alumina solid electrolyte** or **BASE**, a somewhat unfortunate acronym because it may suggest the "base" electrode of some semiconductor devices.

If the concentration of sodium ions (see Figure 7.1) on the upper part of the slab (anode) is larger than that at the lower (cathode), sodium ions will diffuse downward through the slab, accumulating in the lower interface between it and the porous electrode that serves as cathode. Such an accumulation of ions will create an electric field that drives an upward drift of ions. In equilibrium, the downward diffusion flux equals the upward drift flux.

What is the voltage developed by a concentration cell? Let N be the concentration of the migrating ion (Na^+, in the cell described above) and x be the distance into the BASE slab, counting from the top toward the bottom. The flux, ϕ_D, owing to the diffusion of ions is

$$\phi_D = -D\frac{dN}{dx}, \tag{7.1}$$

where D is the diffusion constant. The ion migration causes their concentration at the bottom of the slab to differ from that at the top. An electric field appears driving a return flux, ϕ_E,

$$\phi_E = N_\mu E, \tag{7.2}$$

where μ is the mobility of the ions and $E = -dV/dx$ is the electric field.

In equilibrium, the two fluxes are of the same magnitude,

$$D\frac{dN}{dx} = \mu N\frac{dV}{dx}, \tag{7.3}$$

$$\frac{D}{\mu}\frac{dN}{N} = dV. \tag{7.4}$$

[†] Much of this chapter is based on the article by Cole (1983).

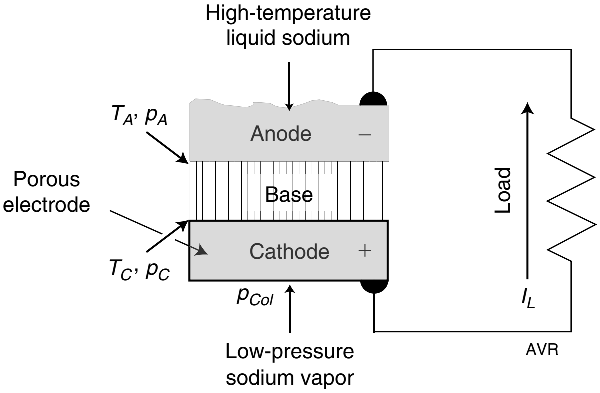


Figure 7.1 A concentration-differential cell.

Integrating from anode to cathode,

$$\frac{D}{\mu}(\ln N_A - \ln N_c) = V, \tag{7.5}$$

where N_A is the ion concentration at the anode (top) and N_C is that at the cathode (bottom). V is the potential across the cell.

According to Einstein's relation, $D/\mu = kT/q$. Thus,

$$V = \frac{kT}{q} \ln \frac{N_A}{N_c}. \tag{7.6}$$

In the usual case in which $T_A = T_C$, we have

$$V = V_{oc} = \frac{kT}{q} \ln \frac{p_A}{p_C} \tag{7.7}$$

because $p = NkT$.

Equation 7.7 gives the value of the *open-circuit* voltage of the AMTEC.

In the case of an ideal (reversible) AMTEC, the load voltage, V_L, is the same as V_{oc} because the internal resistance of the device is zero. Then, when μ kilomoles[†] of sodium ions (which are singly ionized and thus carry a positive charge equal, in magnitude, to that of an electron) pass through the BASE, the electric energy delivered to the load is

$$W_e = \mu N_{0q} V_L = \mu N_{0q} \frac{kT}{q} \ln \frac{p_A}{p_C} = \mu RT \ln \frac{p_A}{p_C}, \tag{7.8}$$

which is the energy delivered by the isothermal expansion of μ kilomoles of gas from p_A to p_C.

[†] μ now represents, the number of kilomoles, not the mobility as in the preceding page.

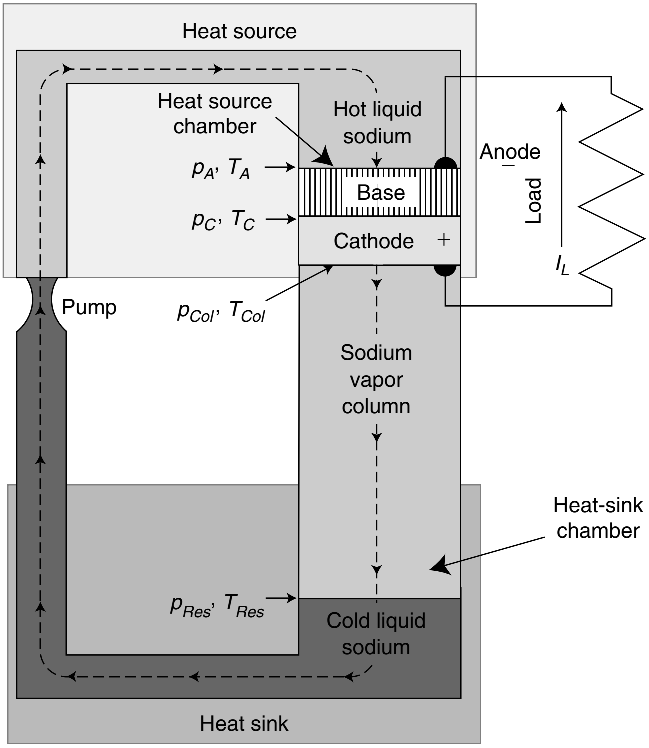


Figure 7.2 The configuration of an AMTEC cell.

An AMTEC can be built in the manner shown in Figure 7.2. A heat source raises the temperature and, consequently, the pressure of the sodium vapor in contact with the anode of the cell, while a heat sink causes the condensation of the metal into a liquid, which is circulated back to the heat source by means of a pump. Wires are attached to the electrodes in contact with the BASE.

The sodium pressure at the anode, p_A, is the vapor pressure of sodium at the temperature, T_A, while p_{Res} is the vapor pressure of the liquid sodium in the heat-sink chamber reservoir at temperature T_{Res}.

The sodium in the heat-sink chamber can be in either the liquid or the vapor state. In our examples we will consider the liquid case, the case of the **liquid-fed AMTEC**.

7.2 Vapour Pressure

From the description in the preceding subsection, it is clear that the values of the sodium vapor pressure in the hot and in the cold ends of the device play a decisive role in determining its performance.

The vapor pressure of sodium and of potassium is listed in Table 7.1, and the sodium pressure is plotted versus temperature in Figure 7.3. These elements have melting and boiling points of, respectively, 370.9 K and 1156 K for sodium and 336.8 K and 1047 K for potassium. Potassium has much less favorable characteristics than sodium. Consequently, we will concentrate on sodium.

Table 7.1 Vapor Pressure of Sodium and Potassium

Temperature	**Sodium**	**Potassium**
(K)	**(Pa)**	
400	0.0003	—
500	0.153	4
600	7	124
700	117	1235
800	1004	6962
900	5330	26832
1000	19799	78480
1100	60491	189460
1200	150368	—
1300	336309	—
1400	626219	—

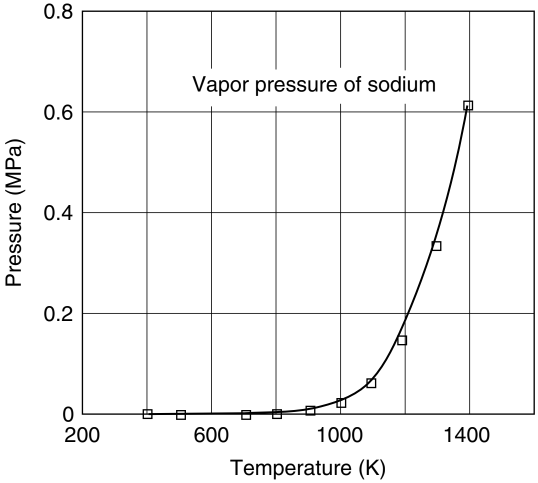


Figure 7.3 The vapor pressure of sodium.

The vapor pressure of sodium can be estimated with an accuracy adequate for modeling AMTEC performance by using the following formula:

$$\begin{aligned}\ln p = &- 62.95 + 0.2279T - 2.9014 \times 10^{-4}T^2\\ &+ 1.7563 \times 10^{-7}T^3 - 4.0624 \times 10^{-11}T^4.\end{aligned} \tag{7.9}$$

The formula yields acceptable values of p for $400 \leq T \leq 1400$.

7.3 Pressure Drop in the Sodium Vapor Column

We saw that the pressure, p_A, on the top (anode) face of the BASE is the vapor pressure of the sodium vapor at the temperature, T_A, and that the pressure, p_{Res}, just above the surface of the liquid sodium pool near the heat sink is the vapor pressure of sodium at the temperature, T_{Res}. However, the voltage across the BASE slab is determined by the pressure ratio, p_A/p_C, not p_A/p_{Res}. It is, therefore, necessary to relate p_C to p_{Res}.

If one assumes that the flow of the sodium vapor in the column between the bottom (cathode) of the BASE and the liquid sodium surface is sufficiently small not to cause, by itself, a pressure gradient and assuming a normal ideal gas behavior, then the column is isobaric—that is, $p_C = p_{Res}$. However, if the mean free path of the sodium atoms is large compared with the diameter of the column, then isobaric conditions do not hold and the column may sustain a pressure differential without a corresponding gas flow. Under such circumstances, $p_C \neq p_L$, and it becomes necessary to establish a relationship between these pressures.

Consider a gas in a horizontal pipe. If there is no net mass flow, then across any surface normal to the pipe, the flux, ϕ_1, from left to right, must equal the flow, ϕ_2, from right to left:

$$\phi_1 = \phi_2. \tag{7.10}$$

Here, we assumed that the flux is uniform over the cross section of the pipe. We will *not* impose the constraint, $p_1 = p_2$, which would, of course, lead to isobaric conditions.

Let n_1 and v_1 be, respectively, the concentration and the mean thermal velocity of the molecules that cross the surface coming from the left, and n_2 and v_2 the corresponding quantities for the flux from the right.

$$n_1 v_1 = n_2 v_2, \tag{7.11}$$

or

$$\frac{n_1}{n_2} = \frac{v_2}{v_1} = \sqrt{\frac{T_2 m_1}{T_1 m_2}} \tag{7.12}$$

because

$$mv^2 = kT. \tag{7.13}$$

Thus,

$$\frac{n_1}{n_2} = \sqrt{\frac{T_2}{T_1}}. \tag{7.14}$$

Since $p = nkT$, we get

$$\frac{p_1}{p_2} = \sqrt{\frac{T_1}{T_2}}. \tag{7.15}$$

This means that, throughout the column of gas, when the mean free path, ℓ, is much larger than the dimensions of the column, the pressure in the gas obeys the rule,

$$\frac{p}{\sqrt{T}} = constant, \tag{7.16}$$

provided there is no gas flow in the column—in other words, provided no current is withdrawn from the AMTEC.

These circumstances are known as **Knudsen's conditions**.[†]

For example, consider a situation in which $T_A = 1300\,\text{K}$ and $T_L = 400\,\text{K}$. The corresponding vapor pressures are $332,000\,\text{Pa}$ and $3 \times 10^{-4}\,\text{Pa}$. This would lead to an open-circuit voltage of

$$V_{oc} = 1300\frac{k}{q} \ln \frac{332,000}{3 \times 10^{-4}} = 2.33\ V, \tag{7.17}$$

provided isobaric conditions did prevail because then $p_C = p_L = 7\,\text{Pa}$.[††]

However, if Knudsen conditions prevail, then

$$\frac{p_C}{\sqrt{T_C}} = \frac{p_L}{\sqrt{T_L}}, \tag{7.18}$$

and

$$p_C = p_L\sqrt{\frac{T_C}{T_L}} = 3 \times 10^{-4}\sqrt{\frac{1300}{400}} = 540 \times 10^{-6}\ \text{Pa}, \tag{7.19}$$

[†] Martin Knudsen, Danish physicist (1871–1949). The Knudsen number is the ratio of the mean free path length of a molecule to some characteristic length in the system.

[††] The temperature span in this example is quite optimistic. Advanced Modular Power Systems, Inc., manufacturer of AMTEC cells, for example, quotes heat-source temperatures from 870 K to 1120 K and heat-sink temperatures from 370 K to 670 K.

and the open-circuit voltage is

$$V_{oc} = 1300\frac{k}{q}\ln\frac{332,000}{540\times10^{-6}\times100} = 2.27\,\mathrm{V}, \tag{7.20}$$

just a tiny bit lower than in the isobaric case.

It is of interest to determine which is the regimen of the pressure behavior in the sodium column. For this, we need to know the dimensions of the column and the mean free path of the sodium ions.

7.4 Mean Free Path of Sodium Ions

The mean free path, ℓ, of a gas molecule can be estimated from

$$\ell = \frac{1}{4nA}, \tag{7.21}$$

where n is the gas concentration and A is the cross-sectional area of the molecule, which, for sodium can be taken as 110×10^{-21} m^2. Thus,

$$\ell = \frac{2.28\times10^{18}}{n} \tag{7.22}$$

From the perfect-gas law, $p = nkT$, or $n = \frac{p}{kT}$; hence,

$$\ell = 2.28\times10^{18}k\frac{T}{p} = 31\times10^{-6}\frac{T}{p}. \tag{7.23}$$

If the gas is at its saturation pressure, then for each value of T, the value of pressure can be determined from Table 7.1 and a plot of ℓ as a function of T can be constructed as shown in Figure 7.4.

At sufficiently low temperatures, the mean free path becomes surprisingly large because the vapor pressure is so low and may very well exceed the dimensions of the sodium vapor column in the device. Under these circumstances, the Knudsen conditions prevail.

7.5 *V*-*I* Characteristics of an AMTEC[†]

Under open-circuit conditions, there is no net flow of sodium ions through the BASE and through the cathode slab of the device. However, if there is an external current through a load connected to the terminals of the AMTEC, then ions must flow through the base and recombine with electrons in the cathode. A net flux, Φ, of sodium atoms is established in the vapor column between the cathode and the cold liquid sodium. To drive

[†] With thanks to John Hovell.

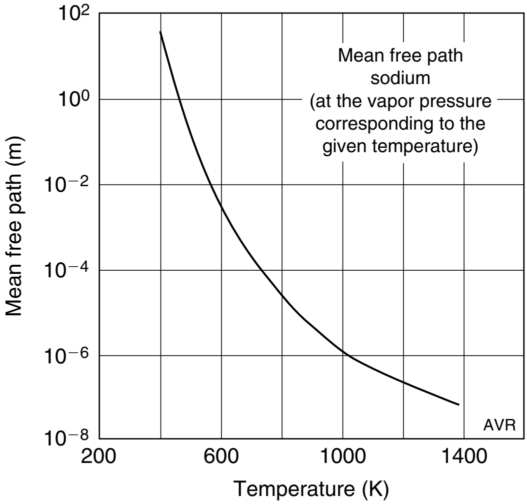


Figure 7.4 The mean free path of sodium molecules as a function of temperature. The gas is assumed to be at its vapor pressure.

this flux, a vapor pressure differential, Δp, must appear between the top of the BASE and the bottom of the cathode slab. The relationship between the flux and the pressure differential is

$$\Phi = \frac{\Delta p}{\sqrt{2\pi m k T_A}}, \tag{7.24}$$

where m is the mass of a sodium ion (22.99 daltons or 38.18×10^{-27} kg). This equation, known as the **Langmuir Assumption**, was derived in Subsection 6.7.3 of Chapter 6. We are making the assumption that $T_A = T_C$ (i.e., that the BASE is at uniform temperature).

The current density is $J = q\Phi$; hence,

$$\Delta p = \frac{\sqrt{2\pi m k T_A}}{q} J = \chi J, \tag{7.25}$$

where

$$\chi \equiv \frac{\sqrt{2\pi m k T_A}}{q}. \tag{7.26}$$

The sodium vapor pressure at the top of the sodium vapor column is now

$$p_{col} = p_{col0} + \chi J = p_{Res}\sqrt{\frac{T_A}{T_{Res}}} + \chi J, \tag{7.27}$$

and, extending Equation 7,

$$V = \frac{kT_A}{q} \ln \frac{p_A}{p_{Res}\sqrt{\frac{T_A}{T_{Res}}} + \chi J} = \frac{kT_A}{q}\left[\ln p_A - \ln\left(p_{Res}\sqrt{\frac{T_A}{T_{Res}}} + \chi J\right)\right]$$

$$= \frac{kT_A}{q}\left\{\ln p_A - \ln\left[\chi\left(\frac{p_{Res}}{\chi}\sqrt{\frac{T_A}{T_{Res}}}\right) + J\right]\right\}$$

$$= \frac{kT_A}{q}\left\{\ln p_A - \ln \chi - \ln\left(\frac{p_{Res}}{\chi}\sqrt{\frac{T_A}{T_{Res}}} + J\right)\right\}. \tag{7.28}$$

Defining

$$J_\vartheta \equiv \frac{p_{Res}}{\chi}\sqrt{\frac{T_A}{T_{Res}}}, \tag{7.29}$$

$$V = \frac{kT_A}{q}\left[\ln p_A - \ln \chi - \ln\left(J_\vartheta + J\right)\right] = \frac{kT_A}{q}\left[\ln \frac{p_A}{\chi} - \ln\left(J_\vartheta + J\right)\right]. \tag{7.30}$$

Defining

$$V_p \equiv \frac{kT_A}{q}\ln\frac{p_A}{\chi}, \tag{7.31}$$

$$V = V_p - \frac{kT_A}{q}\ln\left(J_\vartheta + J\right). \tag{7.32}$$

When $J = 0$, $V = V_{oc}$,

$$V_{oc} = V_p - \frac{kT_A}{q}\ln(J_\vartheta). \tag{7.33}$$

Using Equation 7.29, we find that Equation 7.33 is the same as Equation 7.7.

In the preceding derivation, we assumed that there is no resistance to the flow of ions in the BASE. This ideal situation is not achievable, and a voltage drop, $J\Re$, appears across that component. $\Re = AR$ is the **specific resistivity** of the cell. Here A is the cross-sectional area of the BASE, and R is its resistance. The load voltage becomes

$$V = V_p - \frac{kT_A}{q}\ln\left(J_\vartheta + J\right) - J\Re. \tag{7.34}$$

Equation 7.34 can be used for modeling the behavior of an AMTEC. Cole (1983) states that under normal operating conditions, the calculated values are within 50 mV from the experimentally determined ones.

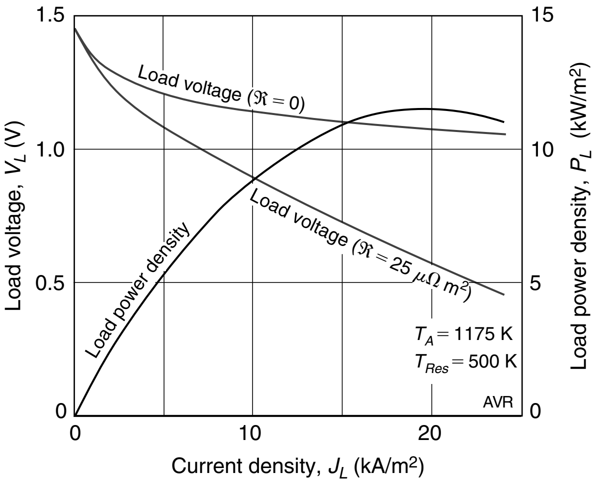


Figure 7.5 Calculated characteristics of an AMTEC.

Figure 7.5 displays the V-I characteristics of a sodium AMTEC operating between 1175 K and 500 K. The values were calculated using Equation 34. One curve corresponds to the situation in which there is no BASE resistance, and the other, to a more realistic situation in which $\Re = 25\ \mu\Omega\text{m}^2$. The load power is also plotted in the figure. It peaks at 11.6 kW/m^2. Both the load current density and the load power density are quite high—a characteristic of AMTECs.

7.6 Efficiency

The efficiency of an AMTEC is (see Cole 1983)

$$\eta = \frac{\text{Electric power delivered to the load}}{\text{Total heat input power}} \tag{7.35}$$

The AMTEC uses up heat by the following four processes:

1. The liquid sodium has to have its temperature raised from T_{Res} to T_A. This amounts to a heat power of

$$\begin{aligned}\dot{Q}_1 &= c\,(T_A - T_{Res})\,\Phi = c\,(T_A - T_{Res})\,\frac{J}{{}_qN_0}\\ &= 311.4 \times 10^{-6}\,(T_A - T_{Res})\,J,\end{aligned} \tag{7.36}$$

where $c = 30$ kJ K^{-1} kmole^{-1} is the specific heat of sodium and Φ is the sodium flux, which is, of course, proportional to the current density through the AMTEC.

2. The heat power needed to vaporize the liquid sodium. This amounts to a heat power of

$$\dot{Q}_2 = \Delta\overline{h_{vap}}\Phi = \dot{Q}_2 = \Delta\overline{h_{vap}}\frac{J}{{}_qN_0} = 0.924J, \tag{7.37}$$

where $\Delta\overline{h_{vap}} = 89$ MJ K^{-1} kmole^{-1} is the heat of vaporization of sodium.

3. The heat power needed to drive the output electric power,

$$\dot{Q}_3 = VJ. \tag{7.38}$$

4. The sum of all parasitic heat losses, $\dot{Q}_4 \equiv \dot{Q}_{loss}$. The main contributors to these parasitic losses are heat conduction through the output leads and heat radiation from the cathode.

Equation 7.35 becomes

$$\begin{aligned}\eta &= \frac{JV}{J\left[V + \frac{c(T_A - T_{Res})}{{}_qN_0} + \frac{\Delta\overline{h_{vap}}}{{}_qN_0}\right] + \dot{Q}_{loss}} \\ &= \frac{VJ}{J[V + 311.4 \times 10^{-6}(T_A - T_{Res}) + 0.924] + Q_{loss}}\end{aligned} \tag{7.39}$$

In the hypothetical case in which there are no parasitic losses ($Q_{loss} = 0$), the AMTEC would operate at its maximum efficiency, which Cole (1983) calls the **electrode efficiency**. It is

$$\eta_{electrode} = \frac{1}{1 + [311.4 \times 10^{-6}(T_A - T_{Res}) + 0.924]/V}. \tag{7.40}$$

The efficiency for given values of T_A and T_{Res} depends, of course, on the operating point. Figure 7.6 shows how the electrode efficiency of a sodium AMTEC operating between 1175 and 423 K varies with the load current, and, consequently, output power. Although the electrode efficiency exceeds 60% at low-power outputs, at the maximum power of 6.64 kW/m^{-2} ($I_L = 11.5$ kA/m^{-2}), the efficiency is down to 33.3%.

Parasitic losses considerably reduce the real efficiency of the device. For instance, consider only radiation losses from the cathode (neglecting heat conduction losses through the output leads). If both the cathode and the surface of the liquid sodium pool were to act as black body radiators, the radiation loss from the cathode would be (read about radiation losses in Chapter 6),

$$\begin{aligned}P_r &= \sigma(T_A^4 - T_{Res}^4) \approx \sigma T_A^4 \\ &= 5.67 \times 10^{-8} \times 1175^4 = 108,000 \text{ W/m}^2.\end{aligned} \tag{7.41}$$

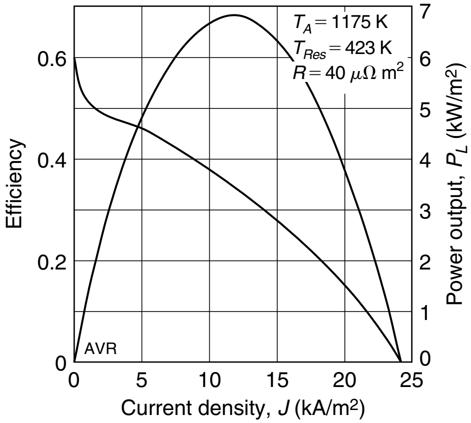


Figure 7.6 Electrode efficiency of an AMTEC.

With such large radiation losses, the efficiency of the AMTEC in our example would barely exceed 5% at maximum power. Fortunately, the surface of the liquid sodium in the heat-sink chamber is far from a black body (which would absorb 100% of the radiation falling on it). In fact, the liquid sodium will do almost the opposite: it reflects more than 98% of the infrared radiation it receives. If the reflected radiation is reabsorbed by the cathode (even if the latter were a black body radiator), the net losses would be 50 times smaller than that calculated above. This would result in an efficiency of 30% for the AMTEC in the example. For such a reduction of radiation losses to take place, it is necessary to have the reradiation from the liquid sodium focused back onto the cathode. This may not be easy to arrange.

Conduction losses are far from negligible. The output leads connect a very hot BASE to the cool world outside. These leads must have high electric conductance, which implies high heat conductance (see the Wiedemann–Franz–Lorenz law in Chapter 5). Low-conductance leads reduce heat conduction losses but increase the I^2R losses in the device. One way to reduce conduction losses is to operate a number of AMTEC cells in series so that for all but the first and the last cell, there is no temperature differential across the leads.

7.7 Thermodynamics of an AMTEC[†]

We will examine the thermodynamics of a liquid-fed sodium AMTEC cell by following the states of the working fluid through a complete cycle. Please

[†] For more detailed information on this topic, please refer to the article by Vining et al. (1993).

refer to the p-V and to the T-S diagrams in Figure 7.7. The fluid sodium will be in one of the several regions of the cell described in Figure 7.2:

1. Heat-sink chamber
2. Pump
3. Heat-source chamber
4. Base
5. Cathode
6. Sodium column

In this example, the device uses 0.001 kmole (μ = 0.001 kmole) of sodium (about 0.032 kg) and operates between the 1200 K and 600 K. We will start at State 1 when the sodium vapor is in the heat-sink chamber and is in contact with the heat sink. Its temperature is 600 K, and its vapor pressure is 7 Pa. The vapor occupies a volume,

$$V = \frac{\mu R T}{p} = \frac{1 \times 10^{-3} \times 8314 \times 600}{7} = 713 \text{ m}^3. \qquad (7.42)$$

State 1 (sodium vapor, in the heat-sink chamber).
$V = 713$ m^3,
$T = 600$ K,
$p = 7$ Pa,
$S = 150$ aeu (arbitrary entropy units).

In Step 1 → 2, the sodium vapor condenses isothermally and isobarically. When fully condensed, the 0.001 kilomole of liquid occupies 23.3×10^{-6} m^3 (assuming a liquid sodium density of 970 kg/m^3).

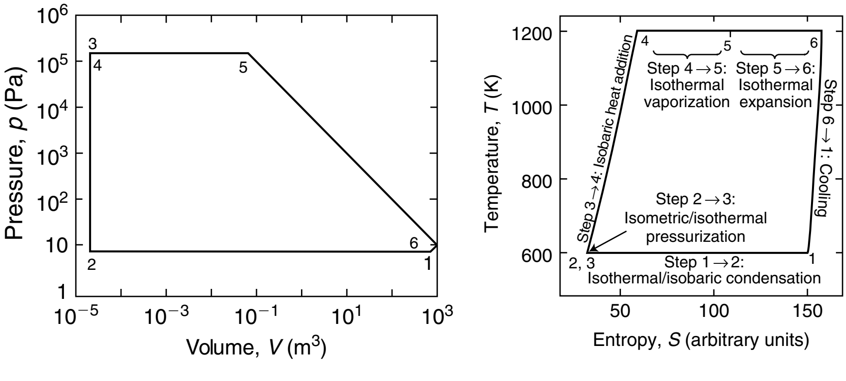


Figure 7.7 Pressure/volume and temperature/entropy graph for an AMTEC.

State 2 (liquid sodium, in the heat-sink chamber).
$V = 23.3 \times 10^{-6}$ m^3,
$T = 600$ K,
$p = 7$ Pa,
$S = 32$ aeu.

In Step 2 → 3, a pump pressurizes the liquid sodium to a pressure equal to the vapor pressure at 1200 K. In this example, it will be 150,400 Pa.[†] The pressurization is carried out at constant temperature, so points 2 and 3 cannot be distinguished in the temperature/entropy plot but are quite apart in the pressure/volume plot.

State 3 (liquid sodium moving through the pump).
$V = 23.3 \times 10^{-6}$ m^3,
$T = 600$ K,
$p = 150,400$ Pa,
$S = 32$ aeu.

In Step 3 → 4, the liquid sodium is now in the heat-source chamber and temperature rises to 1200 K. The process is isobaric and essentially isometric because of the small expansion of the liquid sodium. For this reason, points 3 and 4 cannot be distinguished in the pressure/volume plot, but owing to the large increase in temperature and a modest increase in entropy, they are widely apart in the temperature/entropy plot.

State 4 (liquid sodium, in the heat-source chamber).
$V = 23.3 \times 10^{-6}$ m^3,
$T = 1200$ K,
$p = 150,400$ Pa,
$S = 60$ aeu.

In Step 4 → 5, the liquid sodium vaporizes isothermally and isobarically.

[†] In practice, a certain overpressurization is used to make sure that the sodium remains in the liquid state when heated to 1200 K. A depressurization phase is then used to bring the pressure to the desired value.

State 5 (sodium vapor, in the heat-source chamber).
$V = 0.0663$ m^3,
$T = 1200$ K,
$p = 150,400$ Pa,
$S = 80$ aeu.

In Step 5 → 6, the sodium vapor expands isothermally until the pressure is slightly above that of the initial State 1. Say, $p = 10$ Pa.

State 6
$V = 997.7$ m^3,
$T = 1200$ K,
$p = 10$ Pa,
$S = 157$ aeu.

Finally, in Step 6 → 1, the sodium vapor cools to 600 K and is returned to the initial State 1.

State “1” (sodium vapor, in the heat-sink chamber).
$V = 713\,\mathrm{m}^3$,
$T = 600$ K,
$p = 7$ Pa,
$S = 150$ aeu.

References

Cole, Terry, Thermoelectric energy conversion with solid electrolytes, *Science* **221** 4614, p. 915 2 September 2, **1983**.

Vining, C. B., R. M. Williams, M. L. Underwood, M. A. Ryan, and J. W. Suitor, Reversible thermodynamic cycle for AMTEC power conversion, *J. Electrochem. Soc.*, 140, p. 10, October **1993**.

Chapter 8
Radio-Noise Generators

A subcommittee of the Committee on Government Operations of the House of Representatives of the Ninety-Fourth Congress of the United States heard, on June 11, 1976, testimony on "Converting Solar Energy into Electricity: A Major Breakthrough." Joseph C. Yater, the inventor, read a prepared statement. Although the implementation of the idea is improbable, we will discuss it here as an interesting intellectual exercise.

The idea is fundamentally simple: It is well known that a resistor generates electric noise owing to the random motion of electrons. The available noise power is proportional to the temperature but is independent of the value of the resistance. If two resistors are connected in parallel and maintained at different temperatures, there is a net flow of electric noise power from the hotter to the colder. This energy can be converted into direct current and used for any desired purpose. The system described converts heat into electricity directly.

There is also heat transfer by convection, conduction, and radiation. The crucial question is how much of the input heat is lost by these parasitic processes compared with what is transformed into electricity.

By taking appropriate precautions, one can, at least conceptually, eliminate both convection and conduction but not radiation.

Radiation losses are proportional to the surface area of the heated part while the generated noise power is independent of this area. By reducing the dimensions of the device, it is possible to reduce radiation losses without diminishing the useful power output. Can one build a device small enough to achieve acceptable efficiencies?

The electric output from a heated resistor is in the form of "white" noise—in other words, noise whose power density (power per unit frequency interval) is independent of frequency, up to a given upper limiting frequency or **upper cutoff** frequency, f_U.

The **available noise power** generated by a resistor is

$$P = kTB, \tag{8.1}$$

where k is Boltzmann's constant, B is the bandwidth under which the noise is observed, and T is the temperature. Available power is the power a generator can deliver to a matched load. The bandwidth is

$$B = f_U - f_L \approx f_U \tag{8.2}$$

because, in general, $f_U >> f_L$.

When two resistors of equal resistance are interconnected, the net flow of radio-noise power from the hotter to the colder is

$$P = k\left(f_{UH}T_H - f_{UC}T_C\right). \tag{8.3}$$

The subscripts H and C designate quantities associated with the hotter and the colder resistors, respectively. Notice the assumption that the bandwidth, f_{UH}, for generation of power by the hotter resistor is different from that for the colder, f_{UC}.

It must be remembered that P is the total power transferred, not the power density—the dimensions of the device play no role.

The upper cutoff frequency is a parameter of utmost importance. It is determined by the mean collision frequency of the electrons in the material of the resistors. The mean thermal velocity of the electrons is

$$v = \sqrt{\frac{k}{m}}T^{1/2}. \tag{8.4}$$

The mass to be used in the formula is the effective mass of the electrons in the conductor. If one takes this mass as 10% of that of the free electron, then $v \approx 12{,}000 T^{1/2}$.

For $T = 700$ K, $v \approx 3.3 \times 10^5$ m/s. The mean free path, ℓ, of the electron can be estimated as, roughly, one order of magnitude larger than the lattice constant of the material, which is, typically, 5×10^{-10} m. Take $\ell = 10^{-8}$ m. The collision frequency is then

$$f_U = \frac{v}{\ell} \approx 1.2 \times 10^{12} T^{1/2}. \tag{8.5}$$

Again, for $T = 700$ K, $f_U = 3 \times 10^{13}$ Hz (30 THz). The red end of the visible spectrum occurs at 400 THz. Therefore, f_U is in the infrared.

The available power from any resistor is

$$P_{avail} = 16 \times 10^{-12} T^{3/2}. \tag{8.6}$$

For a resistor at 700 K, this power amounts to 3×10^{-7} W. The power delivered to the load is

$$P_L = P_{avail,\ H} - P_{avail,\ C}. \tag{8.7}$$

In our example, $P_L = 2 \times 10^{-7}$ W. This small amount of power must be compared with the losses incurred when a resistor is heated to 700 K.

If there is a direct connection between the hot and the cold resistors, heat will be conducted through this path and will cause excessive losses. Fortunately, it is possible to transfer the noise from the hot to the cold resistor through a vacuum capacitor having essentially zero heat conductance.

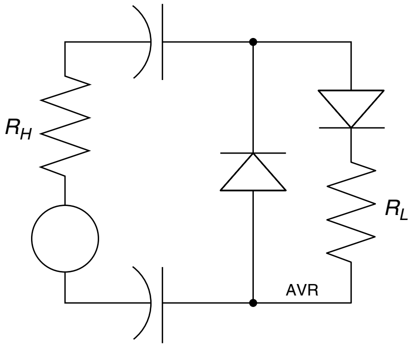


Figure 8.1 Half-wave rectifier.

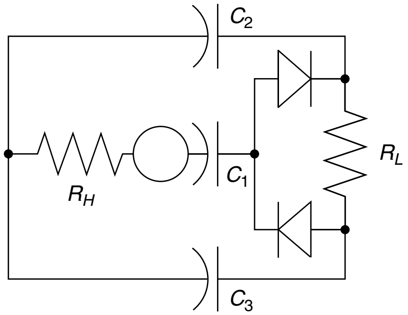


Figure 8.2 Full-wave rectifier.

Several circuits can be used. Figure 8.1 shows a simple half-wave rectifier. The diode that shunts the load provides a dc path otherwise blocked by the capacitors. The circuit in Figure 8.2 is a full-wave rectifier and consists of two of the preceding rectifiers connected back-to-back.

Figure 8.3 shows a perspective of a possible realization of the full-wave rectifier. It can be seen that in the arrangement, the major radiative losses occur across the capacitor plates—one plate is at T_H, while the opposing one is at T_C.

Such losses are

$$P_R = A\sigma\epsilon(T_H^4 - T_C^4). \tag{8.8}$$

A is the total area of the three capacitors (one side only), σ is the Stefan–Boltzmann radiation constant, and ϵ is the effective thermal emissivity (see Chapter 6).

To minimize P_R at a given temperature, it is necessary to reduce both A and ϵ. The latter depends on the nature of the material, on surface

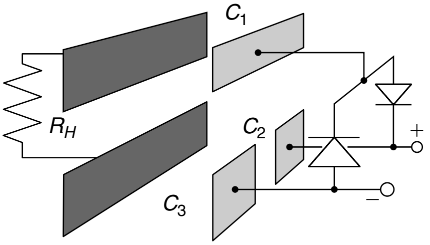


Figure 8.3 A possible geometry for the radio-noise converter.

Table 8.1 Heat Emissivity of Some Materials

Lamp black	0.95
Some white paints	0.95
Oxidized copper	0.60
Copper	0.15
Nickel	0.12
Silver	0.02
Gold	0.015

roughness, on the surroundings, and on the temperature. At low temperatures, typical values of ϵ are given in Table 8.1.

Gold seems to be a good candidate for capacitor plates in the radionoise converter because it combines low emissivity with good stability at high temperatures.

Let us assume that, to ensure good efficiency, radiation losses are to be limited to 10% of the useful power transferred to the load which is

$$P_L = 1.38 \times 10^{-23} \times 1.2 \times 10^{12}(T_H^{3/2} - T_C^{3/2}). \tag{8.9}$$

For $T_H = 700$ K and $T_C = 350$ K, $P_L = 2 \times 10^{-7}$ W. Thus, we must limit P_R to, say, 2×10^{-8} W. Then the total area of capacitors must not exceed

$$A = \frac{P_R}{\sigma\epsilon(T_H^4 - T_C^4)} \approx 10^{-10}\ \mathrm{m}^2, \tag{8.10}$$

assuming gold-plated capacitors. This amounts to 3.3×10^{-11} m^2 per capacitor, a square of about 6×10^{-6} m to the side.

If R_H occupies the same area as one capacitor, the whole device will have an area of 1.3×10^{-10} m^2, leading to a device density of 7.5×10^9 m^{-2} and a power density of about 1500 W m^{-2}.

What is the limit on power densities? The smaller the device, the larger the ratio between useful power and radiation losses. We must investigate what is the smallest size of a resistor that still produces white noise.

The mean free path of the electrons was estimated at 10^{-8} m. A resistor with 10 times this linear dimension will satisfy the statistical requirements for random noise. Thus, the resistor should have a minimum area of $10^{-7} \times 10^{-7} = 10^{-14}$ m^2, which is much less than the area used in the previous example.

If, again, we assume that the area of the capacitors is equal to that of the resistor, the device will occupy a total area of 4×10^{-14} m^2 and the power density will be 5 MW m^{-2}.

The capacitance of a capacitor with area, A_c, is

$$C = \epsilon_0 \kappa \frac{A_c}{d}, \tag{8.11}$$

where ϵ_0 is the permittivity of free space (8.9×10^{-12} F/m), d is the separation between the plates and κ is the dielectric constant ($\kappa = 1$, because there is no dielectric, the capacitor being in vacuum). With an area of 10^{-14} m^2, $C = 8.9 \times 10^{-26}/d$ farads. If the device can be built with a separation of 10^{-7} m between the plates, the capacitance will be about 10^{-18} F. The lower cutoff frequency is

$$f_L = \frac{1}{2\pi C(R_H + R_L)}. \tag{8.12}$$

Since for maximum power transfer, the load must match the generator, $R_L = R_H$. To have the lowest possible f_L, we want the largest possible R_H. High-resistivity material should be used. Possibly, some semiconductor would be appropriate. Let us, however, try a metal—gold—as an example. Assume that the resistor is a ribbon of gold 20 atoms thick and 10^{-7} by 10^{-7} m in area.

The resistivity of gold at room temperature is 2.44×10^{-8} Ωm, and its temperature coefficient†† is 3.4×10^{-11} Ωm/K. Thus, at 700 K, the resistivity is 3.8×10^{-8} Ωm. The ribbon would have a resistance of 0.4 Ω, leading to a lower cutoff frequency of 2×10^{17} Hz, which is much higher than the upper cutoff frequency. The device as planned is patently too small. If the resistor could be made of lightly doped silicon, the lower cutoff frequency would be brought down by a factor of, say, 10^5, leading to $f_L = 2 \times 10^{12}$. Such a cutoff frequency is acceptable because it would result in a bandwidth of $3 \times 10^{15} - 2 \times 10^{12} \approx 3 \times 10^{15}$.

References

Yater, Joseph C., Power conversion of energy fluctuations, *Phys. Rev. A, 10*(4), p. 1361, October **1974**.

Yater, Joseph C., Rebuttal to "Comments on Power conversion of energy fluctuation," *Phys. Rev. A, 20*, p. 623, August **1979**.

†† Temperature coefficient of resistivity is the ratio of the change in resistance to the corresponding change in temperature.

Chapter 9
Fuel Cells

9.1 Introduction

It is said that the nineteenth was the century of mechanical engineering, the twentieth, that of electronics, and the twenty-first, that of biology. In fact, the twentieth century could just as well be known as the century of the mechanical heat engine. Counting cars, trucks, and buses, the United States alone, built, from 1900 to 1999, slightly more than 600 million vehicles. If one adds the rest of the world's automotive production and includes lawn mowers, motorcycles, motorboats, railroad locomotives, airplanes, and heavy construction machinery, the production of internal combustion engines in the twentieth century probably reached the 2 billion mark!

Mechanical heat engines generally use the heat released by the reaction of a chemical substance (fuel) with oxygen (usually from air). The heat is then upgraded to mechanical energy by means of rather complicated machinery. This scheme is inherently inefficient and cumbersome. It is the final outcome of our millenarian struggle to control and use fire. Converting chemical energy directly into electricity is more straightforward, especially in view of the electric nature of the chemical bond that holds atoms in a molecule. Devices that convert chemical energy directly into electricity are called **voltaic** cells,[†] a subgroup of **electrochemical** cells, which also include devices that use an electric current to promote a chemical reaction. Such devices are called **electrolytic** cells or **electrolyzers**, and are covered in the chapter on hydrogen production.

Flashlight batteries, automobile batteries, and fuel cells are examples of voltaic cells. Because voltaic cells transform chemical energy directly into electricity without requiring an intermediate degradation into heat, they are not limited by the Carnot efficiency.

The words "cell" and "battery" are, in modern parlance, interchangeable. "Cell" suggests one single unit (although "fuel cell" most frequently consists of a number of series-connected units). "Battery" suggests a number of units, but a single 1.5-V flashlight cell is commonly called a battery.

[†] Voltaic cells are also called **galvanic** cells, in honor of the Italian physician Luigi Galvani (1737–1798).

If the battery is not worth preserving after its first discharge, it is an **expendable** (also called **primary**) battery. If the device is reusable after discharge, it may fall into one of two categories:

1. **Rechargeable** (also called **secondary**) devices, in which the activity is restored by means of an electric **charging** current, as is the case of automobile batteries.
2. **Refuelable** devices (**fuel cells**), which deliver a sustained output because their consumables are replenished. To facilitate such replenishment, these consumables are usually fluids, although some fuel cells use solid consumables as is the case of **zinc-air cells**, described later in this chapter.

$$\text{Voltaic Cells}\begin{cases}\text{Expendable}\\ \text{Nonexpendable}\begin{cases}\text{Rechargeable}\\ \text{Refuelable}\end{cases}\end{cases}$$

Although fuel cells date back to 1839 when Sir William Groves demonstrated his “gaseous voltaic battery,” until recently they remained in their technological infancy. NASA revived fuel cell research: both *Gemini* and *Apollo* used fuel cells, and so does the space shuttle. Their most important applications in the near future are as power sources for buses and automobiles, as central utility power plants, as dispersed (including residential) power suppliers, and as power sources for cell phones and other small electronic devices.

9.2 Voltaic Cells

The purpose of voltaic cells is to provide a flow of electrons in an external circuit where useful work can be done. To this end, the cells must consist of a source and a sink of electrons.

The reactions used in electrochemical cells are called **reduction-oxidation (redox)** reactions, because the buzz word for releasing electrons is **oxidation** and that for capturing electrons is **reduction**.

Numerous old scientific terms are confusing or at least not self-explanatory. The terms **reduction** and **oxidation** require explanation.

The word “oxygen” stems from *oxús* = acid or sharp and means generator of acids, a name that appears in de Morveau and Lavoisier’s *Nomenclature Chimique* in 1787, when chemists were under the wrong impression that oxygen was an essential element in acids. Actually, it is hydrogen that is essential. When an acid is dissolved in water, some of its hydrogen atoms lose their electron—the water becomes *acidic*; the hydrogen is *oxidized.* By extension, any reaction that involves the loss of electrons is called **oxidation**. The reverse reaction—gaining electrons—is called **reduction**.

The simplest way of thinking of a voltaic cell is as a combination of an electron source or anode in which some chemical is oxidized delivering

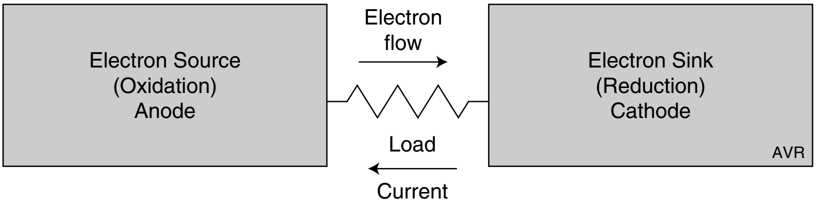


Figure 9.1 A voltaic cell must consist of a source and a sink of electrons.

a flow of electrons to an external load. The latter is connected to a sink of electrons, i.e., a cathode in which a chemical is reduced thus taking up the electrons exiting from the load. See Figure 9.1. In practical electrochemical cells, the full reaction is broken down into two **half-cell reactions** or **half-reactions** that occur in physically separate regions of the device. These regions are interconnected by an **electrolyte** that conducts ions but not electrons. The electrons, having (in voltaic cells) been released by the oxidizing half-reaction, can move to the reduction side only via an external circuit, establishing the external current that is the purpose of the cell. The *conventional* direction of this external current is from the reduction to the oxidizing side—the current exits the device from the reduction side, which thus becomes the **cathode**, and enters the device at the oxidizing side, which becomes the **anode**. As in any source of electricity, the cathode is the positive electrode and the anode the negative one, the opposite of what happens in sinks of electricity (**loads**). See the Introduction to Chapter 6 for a discussion of the words "anode" and "cathode."

As an example of an electrochemical cell, consider a membrane capable of acting as an electrolyte. Put hydrogen in contact with one side of this membrane. At ambient conditions, most of the gas will be in the form of H_2 molecules; however, a small amount will **dissociate**:

$$H_2 \rightarrow 2H, \tag{9.1}$$

and some of the resulting H will **oxidize** (ionize)—that is, lose an electron:

$$H \rightarrow H^+ + e^-. \tag{9.2}$$

Since the membrane does not conduct electrons, the electrons will remain on its surface, while the positive ions will **diffuse** through it and arrive at the other side. Because the ions carry a positive charge, the hydrogen side becomes negative owing to the excess electrons that remain on it, and the opposite side becomes positive owing to the positive ions that arrived there. The resulting electric field causes some of the positive ions to drift back to the hydrogen side. A **dynamic equilibrium** is established when the diffusion exactly equals the returning drift. It is easy to calculate the potential developed (Chapter 7, Section 7.1).

Now sprinkle a conducting powder on both sides of the membrane so as to create two porous electron-conducting layers, that is, two **electrodes**. Interconnect the electrodes externally through a load resistance, R_L. Ions

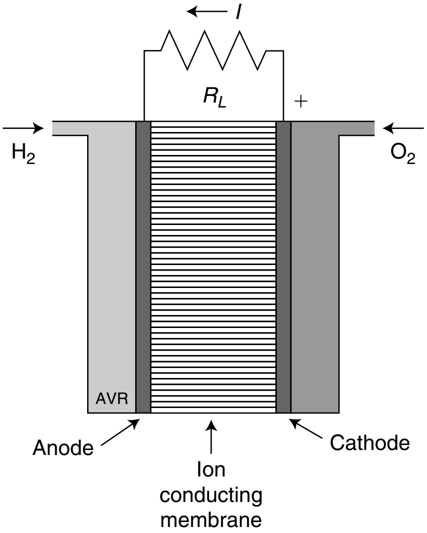


Figure 9.2 The simplest fuel cell.

cannot flow through this external connection, but electrons can and, when they do, they flow from the hydrogen side where they are abundant to the opposite side establishing an electric current as indicated in Figure 9.2. The reaction of interest that occurs at the hydrogen electrode is

$$\underline{\text{Anode reaction:}} \quad 2\text{H}_2 \rightarrow 4\text{H}^+ + 4\text{e}^-. \tag{9.3}$$

The difficulty with this picture is that it contradicts the first law of thermodynamics in that it causes an I^2R_L amount of heat to be generated in the load, while, at the cathode, the incoming electrons will combine with the H^+ that diffused through the membrane, regenerating the hydrogen atom, H, and, eventually re-creating the H_2 gas used as "fuel." We would generate heat without using any fuel.

The external circuit creates a *path* for the electrons but cannot by itself force a current to circulate, just as a pipe with one end dipped into a lake cannot cause water to flow up inside it. For the water to flow, the open end of the pipe must be lower than the surface level of the lake. Similarly, to have an external current, it is necessary to lower the (thermodynamic) potential on the cathode side. This can conveniently be done by introducing oxygen so that, combined with the electrons and the H^+, water is formed:

$$\underline{\text{Cathode reaction:}} \quad 4\text{e}^- + 4\text{H}^+ + \text{O}_2 \rightarrow 2\text{H}_2\text{O}. \tag{9.4}$$

This reaction is strongly exothermal—that is, it releases energy (although, in this case, not mostly as heat, as in case of the combustion of hydrogen, but mainly as electricity). This is, of course, the energy that powers the fuel cell.

The electrochemical cell just described is shown in Figure 9.2.

Under STP conditions, the degree of hydrogen dissociation at the anode is small. It can be increased somewhat by altering physical conditions (for example, increasing the temperature,). Remember Le Chatelier's principle. It also can be increased by the action of catalysts.

The overall cell reaction is

$$2H_2(g) + O_2(g) \rightarrow 2H_2O(g). \tag{9.5}$$

Clearly, a voltage appears across the terminals of our fuel cell. All voltaic cells rely on the potential difference that develops when an electrode is put in contact with an electrolyte. In the text box titled "Cell Potential," we discuss this phenomenon in a simple qualitative fashion, and, we also point to the mechanisms that cause oxidation and reduction.

Cell Potential

Let us gain an understanding of what gives rise to a voltage when two different metals are dipped into an electrolyte.

When *one* electrode is dipped into an electrolyte, a potential difference is established between the electrode and the electrolyte. However, this potential cannot be measured because a contact must be made with the electrolyte and, consequently, an additional electrode must be used. We can only measure the potential between *two* electrodes.

The potential difference between two identical electrodes is exactly zero because one cancels the other.

We elect one electrode as a universally accepted reference (the standard hydrogen electrode, SHE). All others are measured relative to this SHE.

But what causes a potential difference (voltage) to appear? Here is a very simplified explanation:

Remember that if the atomic number (the number of protons in the nucleus[†]) is equal to one of the values generated by progressive truncations of the series below, then the atom is very stable. (The atom frequently refuses any chemical interaction with other atoms and even with itself—the molecule is invariably monoatomic. The formula is

$$2(1^2 + 2^2 + 2^2 + 3^2 + 3^2 + 4^2 + \cdots),$$

whose progressive truncations yield the series

$$2, 10, 18, 36, 54, 86, \ldots,$$

corresponding to helium, neon, argon, krypton, xenon, and radon.

[†] In a neutral atom, the number of electrons equals to the number of protons.

(*Continues*)

(*Continued*)

Plausibly, elements that have just one electron more than the above magic number don't mind losing it because they covet a stable electron configuration. Such elements have the lowest ionization potentials in the table of elements. On the other hand, atoms that are one electron short of the magic number hang on tightly to what they have and, correspondingly, have high ionization potentials:

Atomic number	**Element**	**Ionization potential (e-volts)**	**Atomic number**	**Element**	**Ionization potential (e-volts)**
3	Lithium	5.40			
11	Sodium	5.15	9	Fluorine	17.45
19	Potassium	4.35	17	Chlorine	12.99
37	Rubidium	4.18	35	Bromine	11.83
55	Cesium	3.90	53	Iodine	10.47
87	Francium	3.94	85	Astatine	8.90

Notice the gradual decrease in ionization potentials of the alkali metals as the atomic number grows, leading to a minimum at 3.90 eV for cesium. Francium has, anomalously, a slightly higher ionization potential. The same happens in the halogen column, starting with a record ionization potential of 17.45 eV for fluorine.

These ionization potentials are for isolated atoms. When atoms are part of a chemical reaction, they behave somewhat differently, but still follow the general tendency of either making it easy for electrons to leave (oxidizing the substance) or for electrons to be captured (reducing the substance). This tendency to either shed or grab electrons is indicated by the **reduction potential** of the substance: it is negative if the substance oxidizes and positive if it is reduced. The potentials are the ones measured when a substance is put in contact with an electrolyte, as described initially.

All alkali metals have large negative reduction potentials—they are strong **reducing agents** or **reductants** because, by becoming oxidized, they provide the electrons necessary to reduce other chemical species. Halogens have positive reduction potentials and are strong **oxidizing agents** or **oxidants** because, by being reduced, they provide opportunity for electrons to leave other chemical species, which consequently become oxidized. Oxygen is, of course, an oxidizer.

(*Continues*)

(*Continued*)

Atomic number	Element	Reduction potential (volts)	Atomic number	Element	Reduction potential (volts)
3	Lithium	−3.05			
11	Sodium	−2.71	9	Fluorine	2.87
19	Potassium	−2.93	17	Chlorine	1.36
37	Rubidium	−2.98	35	Bromine	1.07
55	Cesium	−2.92	53	Iodine	0.54

In greater detail, here is what happens:

Consider zinc, which is a good reductant; having two valence electrons it does not mind giving up. Its reduction potential is −0.76 V. When dipped into an electrolyte, zinc will oxidize:

$$\mathrm{Zn(s)} \rightarrow \mathrm{Zn^{++}(aq)} + 2\mathrm{e}^{-}.$$

The neutral solid (s) zinc will split into a zinc ion and two electrons. The electrons, being insoluble, stay in the metal, while the positive ion (a **cation**) goes into solution as indicated by the (aq). Of course, the concentration, $[\mathrm{Zn^{++}}]$, of zinc ions is larger on the surface of the electrode in contact with the electrolyte than farther away in the bulk of the electrolyte. This concentration disequilibrium causes a diffusion current of zinc ions to stream from the electrode to the electrolyte.

The electrode, because it retains the electrons but loses some ions, becomes negatively charged with respect to the electrolyte that has an excess of cations. This generates an electric field that creates a drift current of cations moving from the electrolyte to the electrode. We have a dynamic equilibrium resulting of two sizable **exchange currents**—diffusion versus drift—which exactly cancel one another. When this cancellation occurs, a steady **electrode potential** is established. When at RTP and at unit molarity, the cell potential is called the **standard cell potential.**

9.3 Fuel Cell Classification

As in many technical areas, there is here a proliferation of acronyms capable of driving the uninitiated to distraction. We will use a few:

AFC	Alkaline fuel cell
DMFC	Direct methanol fuel cell

MCFC	Molten carbonate fuel cell
PAFC	Phosphoric acid fuel cell
SAFC	Solid acid fuel cell
SOFC	Solid oxide fuel cell (ceramic)
SPFC	Solid-polymer fuel cell

In the beginning of the twentieth century, at least three different technologies—steam, electric, and internal combustion—were competing for the automotive market. Now, in the beginning of the twenty first century, the different fuel cell technologies, listed above, are vying for dominance. Although AFC, MCFC, and PAFC are still in commercial production, it would appear that only SPFC and SOFC have a really good chance of emerging victorious.

Fuel cells can be classified according to several criteria, as follows.

9.3.1 Temperature of Operation

The open-circuit voltage of a fuel cell depends on the nature of the fuel used and, to a small extent, on the temperature and pressure. The maximum deliverable current density, however, rises quickly with increasing temperature because it is related to the rapidity of the chemical reaction—that is, to the chemical kinetics, which is, of course, also a function of the type of fuel and can be improved by catalysts.

The higher the temperature, the larger the current the cell can deliver. On the other hand, high temperatures may be incompatible with the electrolyte or with other materials in the cell and tend to limit lifetime.

Large power plants that work continuously may profit from high operating temperature: the kinetics are favorable, catalysts are either unnecessary, or, when necessary, are immune to CO poisoning,[†] and exhaust gases can be used for cogeneration, increasing the overall efficiency. Plants that operate at somewhat lower temperatures can produce hot water for general use, an advantage in the case of district or residential generators.

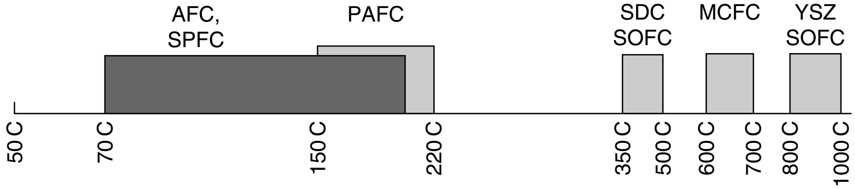


Figure 9.3 The operating temperatures of the different fuel cell types fall into relatively narrow ranges. Notice the two different types of SOFC: SDC = samarium-doped ceria, YSZ = yttria-stabilized zirconia.

[†] CO poisoning of the catalyst may be a serious problem in low-temperature cells operating with fossil fuel-derived hydrogen.

For intermittent use, especially in automobiles, it is important to use fuel cells that operate at temperatures low enough to permit short start-up times. However, low temperatures bring the disadvantage of higher vulnerability to carbon monoxide poisoning, of high catalyst loading requirements, and of the need for more complicated cooling systems.

At present, the more common SPFC operating at slightly below 100 C are too cool for optimal performance, while SOFCs, are too hot, especially for use in cars. Indeed, at the low temperature of present-day SPFC, problems of catalysis and sensitivity to CO poisoning are important, while, in SOFC, the advantages of high temperature mentioned before are counterbalanced by a number of disadvantages listed in Subsection 9.5.4. For this reason, SPFC research strives for high-temperature plastics such as the membranes developed by Celanese AG that operate at 200 C[†], while SOFC research is seeking lower temperature ceramics, such as doped lanthanum strontium gallate magnesite (LSGM) and samarium-doped ceria (SDC).

9.3.2 State of the Electrolyte

Most early fuel cells used liquid electrolytes, which can leak, require liquid level and concentration management, and may be corrosive.[††] Modern cells tend to use solid electrolytes, which, among other advantages, have a composition that does not change as a function of the current drawn as happens with liquid electrolytes. Such composition variation may lead do electrolyte depletion that contributes to transport losses at the high current end of the V-I characteristics. Solid electrolyte cells employ either ceramics for high temperatures or plastics for low temperature. Some second-generation utility-type fuel cells still use molten carbonates but seem to be on the way out.

9.3.3 Type of Fuel

At least in a laboratory, the simplest fuel to use in a fuel cell is hydrogen. However, this gas, owing to its extremely low density, is difficult to store (see Chapter 11). Especially for automotive use, efforts are being made to use easily storable liquid substances, from which hydrogen can be extracted as needed (see Chapter 10). The extraction process frequently consists of a **steam reforming** step in which a carbon-bearing feedstock is transformed into a **syngas** or **reformate** consisting mainly of H_2 and CO, followed,

[†] One of the most common membranes used as an electrolyte in low-temperature cells is some variation of Nafion, a polymer derived from the high-temperature plastic, teflon. The Celanese membrane is derived from the even higher temperature plastic, polybenzimidazole, used in spacesuits, among other applications.

[††] Liquid electrolyte fuel cells (using a potassium hydroxide, KOH, solution) are still employed by NASA in the space shuttle.

when necessary, by a **shift reaction** in which the CO in the reformate is made to react with additional water and is transformed into H_2 and CO_2. CO happens to interfere with the catalytic action of platinum, especially at low temperatures.

Higher temperature fuel cells such as MCFCs and SOFCs can use CO as a fuel so that the H_2–CO reformate can be fed directly to the cells.

Methanol is used in some fuel fells, especially in small ones for use in portable electronic equipment.

High reactivity fuels were used when the technology was quite immature. They included such combinations as hydrazine (NH_2NH_2) with hydrogen peroxide (H_2O_2) as oxidant. These fuels are, however, corrosive and expensive. Hydrazine—visualize two amines (NH_2) held together by a nitrogen–nitrogen bond—is highly toxic.

9.3.4 Chemical Nature of the Electrolyte

Electrolytes can be alkaline, acid, molten carbonates, or ceramic.

Alkaline electrolytes, such as potassium hydroxide (KOH), are commonly used in **electrolyzers** (the dual of fuel cells—instead of generating electricity from fuels, they produce fuels from electricity). Alkalis are avoided in most fuel cells because the presence of CO_2 in air (used as oxidant) causes the formation of insoluble carbonates that spoil the electrolyte. For special applications where pure oxygen is available, KOH fuel cells offer the advantage of high efficiency. This is the case of fuel cells used in space.

Acids tend to be more corrosive than alkalis, but relatively weak acids perform well in fuel cells. Phosphoric acid, in particular, was popular.[†] It tolerates carbon dioxide. For good performance, it must operate at temperatures between 150 and 220 C, and to keep the liquid from boiling away, a certain degree of pressurization is needed. At lower temperatures, the conductivity of the solution is too small, and at higher temperatures, there are problems with the stability of the materials. Solid acids (Subsection 9.5.7) have been proposed as electrolytes for fuel cells. They may contribute to the solution of the vexing methanol crossover problem (Subsection 9.5.6).

Most ceramic electrolytes, as, for instance, yttria-stabilized zirconia (YSZ) and samarium-doped ceria (SDC), are anion conductors (conductors of negative ions such as O^{--}). However, cation conductors have been proposed.

Solid polymers act, in general, as proton conductors—that is, as acids, although, as in the case of ceramics, cation conductors have been investigated.

[†] Phosphoric acid is a benign acid as attested by its daily consumption by the millions of Coca-Cola drinkers.

9.4 Fuel Cell Reactions

The chemical reaction in a fuel cell depends on both the type of fuel and the nature of the electrolyte. Some of the most common combinations are listed below and are displayed in the illustrations that follow this subsection.

9.4.1 Alkaline Electrolytes

Hydrogen-oxygen fuel cells with alkaline electrolytes (generally, KOH) use OH^- as the current-carrying ion. Because the ion contains oxygen, water is formed at the anode.

The KOH in the electrolyte dissociates:

$$KOH \rightleftharpoons K^+ + OH^-. \tag{9.6}$$

Neutral hydrogen at the anode combines with the hydroxyl ion to form water, releasing the electrons that circulate through the external load:

$$2H_2 + 4OH^- \rightarrow 4H_2O + 4e^-. \tag{9.7}$$

At the cathode, the electrons regenerate the hydroxyl ion:

$$4e^- + O_2 + 2H_2O \rightarrow 4OH^-. \tag{9.8}$$

The KOH is, of course, not consumed. The overall reaction is

$$2H_2 + O_2 \rightarrow 2H_2O. \tag{9.9}$$

9.4.2 Acid Electrolytes

When the electrolyte is acid, H^+ ions are available. These can come from the ionization of the hydrogen (as in the SPFC cells) or from the dissociation of the acid in the electrolyte. Take phosphoric acid:

$$H_3PO_4 \rightleftharpoons 3H^+ + PO_4^{---}. \tag{9.10}$$

In either case, the H^+ ion is replenished by the anode reaction:

$$2H_2 \rightarrow 4H^+ + 4e^-. \tag{9.11}$$

At the cathode, the H^+ is reduced in the presence of O_2 by the electrons that circulate through the load, forming water:

$$4H^+ + 4e^- + O_2 \rightarrow 2H_2O. \tag{9.12}$$

The overall reaction is the same as in the previous case. Water is formed at the cathode, and the active ion is hydronium.

9.4.3 Molten Carbonate Electrolytes

Molten carbonate electrolytes are not bothered by carbon oxides. They operate at relatively high temperatures, hence under more favorable kinetics.

When fueled by hydrogen, the reactions are:

$$H_2 + CO_3^{--} \rightarrow H_2O + CO_2 + 2e^-, \tag{9.13}$$

$$\frac{1}{2}O_2 + CO_2 + 2e^- \rightarrow CO_3^{--}. \tag{9.14}$$

When the fuel is CO, the anode reaction is

$$CO + CO_3^{--} \rightarrow 2CO_2 + 2e^-, \tag{9.15}$$

while the cathode reaction is the same as in the hydrogen case.

9.4.4 Ceramic Electrolytes

Ceramic electrolytes usually conduct negative ions (anions), although, as pointed out before, some cation conductors have been demonstrated.

For anion conductors:

At the cathode, oxygen is reduced by capturing the electrons that constitute the load current:

$$O_2 + 4e^- \rightarrow 2O^{--}. \tag{9.16}$$

At the anode, the fuel combines with the O^{--} ions that drifted through the electrolyte, freeing electrons. The fuel may be hydrogen,

$$2H_2 + 2O^{--} \rightarrow 2H_2O + 4e^-, \tag{9.17}$$

or carbon monoxide,

$$2CO + 2O^{--} \rightarrow 2CO_2 + 4e^-. \tag{9.18}$$

9.4.5 Methanol Fuel Cells

The anode reaction is

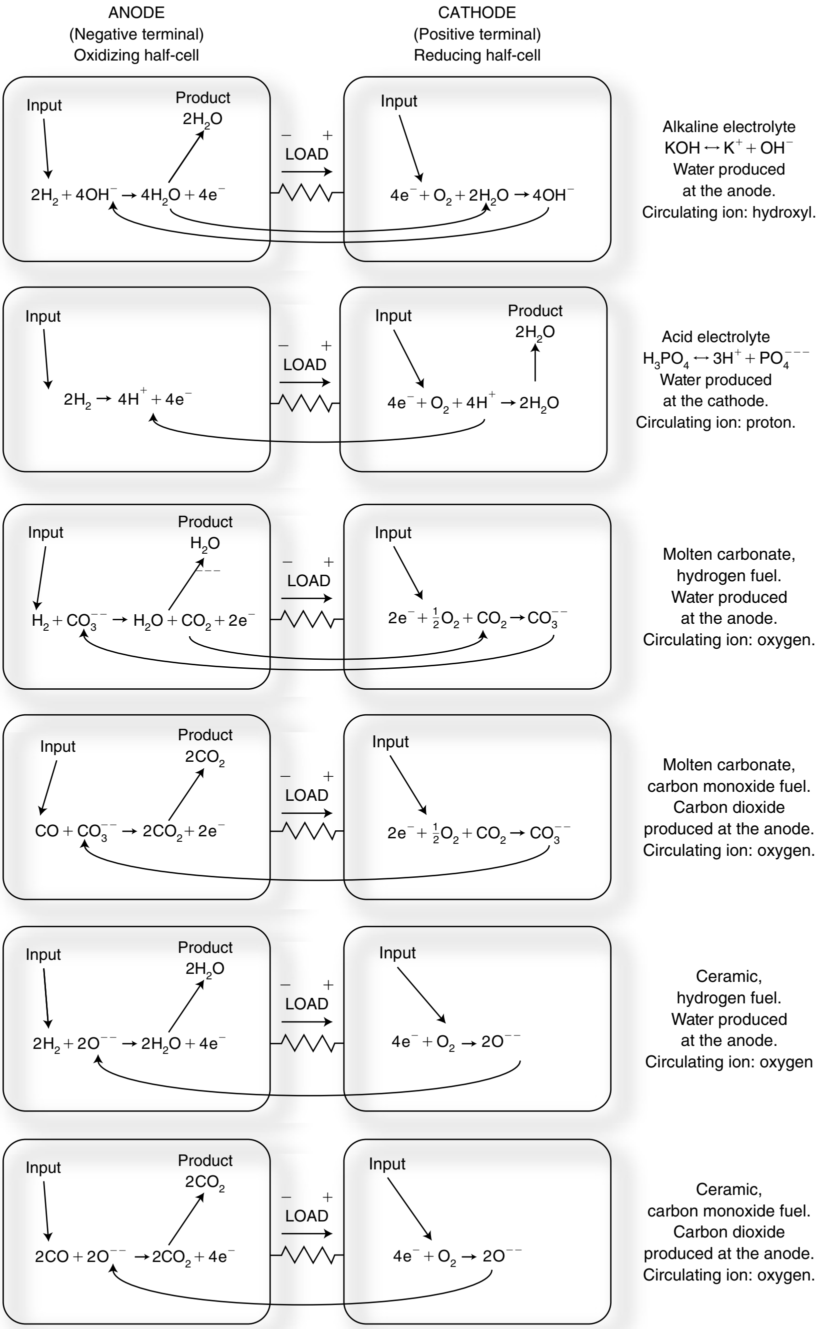
$$CH_3OH + H_2O \rightarrow CO_2 + 6H^+ + 6e^-. \tag{9.19}$$

The cathode reaction is

$$6H^+ + 6e^- + \frac{3}{2}O_2 \rightarrow 3H_2O. \tag{9.20}$$

Thus, the overall reaction is

$$CH_3OH + \frac{3}{2}O_2 \rightarrow CO_2 + 2H_2O. \tag{9.21}$$


ANODE
(Negative terminal)
Oxidizing half-cell
CATHODE
(Positive terminal)
Reducing half-cell
Input
Product
$2H_2O$
$2H_2 + 4OH^- \rightarrow 4H_2O + 4e^-$
− +
LOAD
Input
$4e^- + O_2 + 2H_2O \rightarrow 4OH^-$
Alkaline electrolyte
$KOH \leftrightarrow K^+ + OH^-$
Water produced
at the anode.
Circulating ion: hydroxyl.
Input
$2H_2 \rightarrow 4H^+ + 4e^-$
− +
LOAD
Input
Product
$2H_2O$
$4e^- + O_2 + 4H^+ \rightarrow 2H_2O$
Acid electrolyte
$H_3PO_4 \leftrightarrow 3H^+ + PO_4^{---}$
Water produced
at the cathode.
Circulating ion: proton.
Input
Product
H_2O
$H_2 + CO_3^{--} \rightarrow H_2O + CO_2 + 2e^-$
− +
LOAD
Input
$2e^- + \frac{1}{2}O_2 + CO_2 \rightarrow CO_3^{--}$
Molten carbonate,
hydrogen fuel.
Water produced
at the anode.
Circulating ion: oxygen.
Input
Product
$2CO_2$
$CO + CO_3^{--} \rightarrow 2CO_2 + 2e^-$
− +
LOAD
Input
$2e^- + \frac{1}{2}O_2 + CO_2 \rightarrow CO_3^{--}$
Molten carbonate,
carbon monoxide fuel.
Carbon dioxide
produced at the anode.
Circulating ion: oxygen.
Input
Product
$2H_2O$
$2H_2 + 2O^{--} \rightarrow 2H_2O + 4e^-$
− +
LOAD
Input
$4e^- + O_2 \rightarrow 2O^{--}$
Ceramic,
hydrogen fuel.
Water produced
at the anode.
Circulating ion: oxygen
Input
Product
$2CO_2$
$2CO + 2O^{--} \rightarrow 2CO_2 + 4e^-$
− +
LOAD
Input
$4e^- + O_2 \rightarrow 2O^{--}$
Ceramic,
carbon monoxide fuel.
Carbon dioxide
produced at the anode.
Circulating ion: oxygen.

9.4.6 Formic Acid Fuel Cells

The anode reaction is

$$\mathrm{HCOOH} \rightarrow \mathrm{CO_2} + 2\mathrm{H^+} + 2\mathrm{e^-}. \tag{9.22}$$

The cathode reaction is

$$\frac{1}{2}\mathrm{O_2} + 2\mathrm{H^+} + 2\mathrm{e^-} \rightarrow \mathrm{H_2O}. \tag{9.23}$$

Thus, the overall reaction is

$$\mathrm{HCOOH} + \frac{1}{2}\mathrm{O_2} \rightarrow \mathrm{CO_2} + \mathrm{H_2O}. \tag{9.24}$$

9.5 Typical Fuel Cell Configurations

All fuel cells are basically a sandwich of two electrodes separated by an electrolyte. Numerous variations of this theme have been tried, but as we have pointed out, the dominant variable is the nature of the electrolyte. Early fuel cells used liquid electrolytes—alkaline or acid—but the current trend is in the direction of solids—mostly polymers or ceramics. In the subsections that follow, we will briefly describe configurations that have had some success, many of which have now been abandoned.

9.5.1 Demonstration Fuel Cell (KOH)

The demonstration fuel cell described here illustrates the ideas that perhaps originated in the 1950s. The design is self-evident from Figure 9.4. The six holes near the rim of the two Lucite covers allow the passage

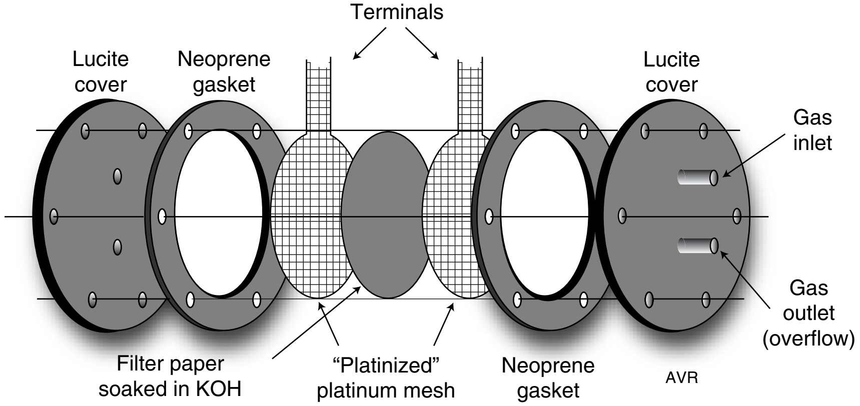


Figure 9.4 Exploded view of an early demonstration fuel cell.

of screws that hold the system together. The diameter of the device is 8 cm. Fuel sources are two toy balloons (not shown), one containing oxygen and the other hydrogen. Excess gas is vented into a beaker with water.

As in many liquid electrolyte cells, the electrolyte was immobilized by soaking it up in some matrix, in this case, a simple disc of filter paper saturated with a KOH solution.

It can be seen that the cell is symmetrical—there is no structural difference between the anode and the cathode side. The electrodes are made of platinum mesh. Prior to using the cell, platinum black is sprinkled over the mesh to enhance the catalytic action.

Open-circuit voltage is about 1 V, and the cell will deliver some 200 mA at 0.7 V, a power of 140 mW.

The design is quite primitive and predates the era of ion-exchange membranes. Much better demonstration fuel cells can currently be obtained from a number of vendors. Go to Google for "Fuel cell demonstration kits." Some of the available kits are well designed and easy to operate. They all require distilled water, but that should be no problem. Particularly instructive are kits that incorporate a photovoltaic converter, an electrolyzer, and a simple hydrogen storage system, in addition to the fuel cell. They demonstrate that electricity can be produced and stored with absolutely no CO_2 emission. The problem is, of course, that it is not yet economically feasible to do so on a large scale.

9.5.2 Phosphoric Acid Fuel Cells (PAFCs)

9.5.2.1 A Fuel Cell Battery (Engelhard)

Engelhard[†] attempted to commercially implement the technology exemplified by the demonstration fuel cell of the previous subsubsection. It offered a completely autonomous 750-W battery that included a fuel supply and all the control devices. The knotty problem of providing the hydrogen for the cell was solved by an integrated ammonia-cracking subunit. Ammonia, NH_3, is indeed a convenient way of storing hydrogen because, at room temperature under moderate pressures, it is a liquid and can, therefore, be easily handled.

The electrolyte was phosphoric acid. The cell construction can be seen in Figure 9.5.

Since fuel cells are low-voltage, high-current devices, they must be connected in series. The simplest way of doing so is to use a **bipolar** configuration: a given electrode acts as an anode for one cell and (on the other face) as a cathode for the next. The **bipolar electrode** was made of

[†] Engelhard (of New Jersey), creator of the catalytic converter for automobiles, was bought in 2006 by BASF, the large German chemical company, part of the (somewhat) infamous I. G. Farbenindustrie of the Nazi era.

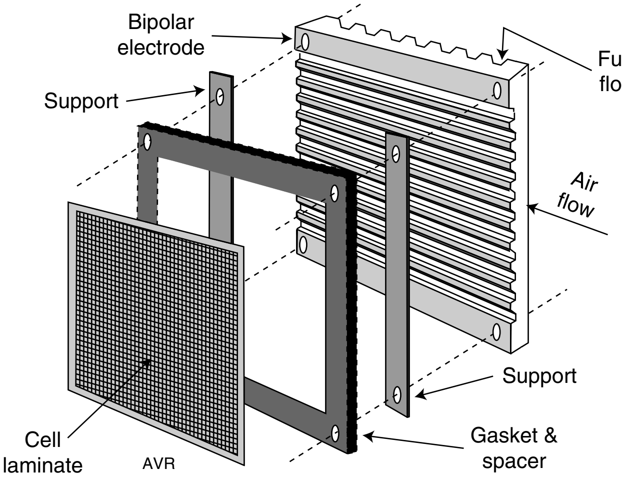


Figure 9.5 One element of the Engelhard PAFC.

either aluminum or carbon. Gold was plated on aluminum for protection and flashed on carbon to reduce contact resistance.

The plate was 3 mm thick and had grooves or channels machined on both faces. The oxidant channels ran perpendicular to those for hydrogen on the opposite face. The plate was rectangular with its smaller dimension along the air-flow direction to minimize pressure drop. The channels, in addition to leading in the gases, served also to increase the active surface of the electrodes.

The electrolyte soaked a "cell laminate" held in place by a rubber gasket. It can be seen that this type of construction does not differ fundamentally from that used in the demonstration cell described earlier.

The cell operated at 125 C. The oxidant was air, which entered at ambient temperature and left hot and moist carrying away excess heat and water (formed by the oxidation of the fuel in the cell). The air flow was adjusted to provide sufficient oxygen and to ensure proper heat and water removal.

Apparently, no catalysts were used. This, combined with the relatively low operating temperature (phosphoric acid cells frequently operate at temperatures above 150 C), resulted in somewhat adverse kinetics and, consequently, in low-voltage efficiency.

The system consisted of an ammonia source (a pressure cylinder), a dissociator or **cracker**, a scrubber, the fuel cells, and ancillary pumping and control mechanisms.

The fuel cells themselves operated at an optimistic 45% efficiency (referred to the higher heating value of ammonia). However, the system needed to divert fuel to heat the dissociator that worked at 850 C, thus reducing the overall efficiency at rated power to about 30% and to much less, at lower power levels.

9.5.2.2 First-Generation Fuel Cell Power Plant

One early effort to adapt fuel cells for dispersed utility-operated power plant use was made by United Technologies Corporation (UTC). It was a 4.8-MW module fueled by natural gas, to be installed in Manhattan.

The UTC fuel cells used phosphoric acid and delivered 0.64 V per cell. At the operating temperature of 150 to 190 C, current densities of 1600 to 3100 $\text{A}\,\text{m}^{-2}$ were possible. The life of the cells was expected to be some 40,000 hours (around five years).

The UTC project in New York suffered such lengthy delays in the approval process by the city government that when it was finally cleared for operation in 1984, the equipment had deteriorated to the point that it was uninteresting to put it in operation. The first commercial demonstration of this fuel cell application took place in Japan where a 4.5-MW unit started operation in 1983. The same company (Tokyo Electric Power Company) later operated an 11-MW PAFC facility built by Toshiba using U.S. built cells (International Fuel Cell Company, a subsidiary of UTC).

UTC manufactures a 200-kW PAFC power plant, known as PureCell 200 (formerly PC25). It is used as a large uninterruptible power supply. From 1991 until the beginning of 2008, worldwide sales approached 300 plants. They delivered both electricity (at 37% efficiency) and heat. The latter could be in the form of 260 kW of hot water (at 60 C) or 130 kW of hot water plus another 130 kW of heat at 120 C. They were fueled by natural gas or, in some cases, by biogas from anaerobic digesters. The PureCell 200 package includes the appropriate (external) reformer to extract hydrogen from the feedstock. By 2004, some PC25s could be bought in the surplus market.

Modern PAFCs use platinum catalysts and are vulnerable to CO in the fuel stream (see SPFC, in Section 9.5.5). Fortunately, the tolerance of Pt to CO increases with operating temperature: at 175 C, CO concentration can be as high as 1%, and, at 190 C, as high as 2%. This contrasts with the more stringent requirements of the cool SPFCs, which at 80 C require hydrogen with less than 10 ppm of CO unless special procedures or special catalysts are employed.[†] On the other hand, whereas the SPFCs have a very stable electrolyte, the phosphoric acid in PAFCs can be degraded by ammonia or sulfur whose concentrations in the fuel stream must be kept low.

9.5.3 Molten Carbonate Fuel Cells (MCFCs)

9.5.3.1 Second-Generation Fuel Cell Power Plant

Unlike the first-generation utility-type fuel cells described earlier, the molten carbonate systems operate at sufficiently high temperature so that

[†] Roughly, the tolerance, λ (in ppm), of platinum catalysts to CO as a function of the temperature, T (in kelvins), is given by $\lambda = 255 \times 10^{-12} \exp(T/14.5)$.

natural gas, digester-produced biogas, and other fuels can be reformed internally, yielding the H_2 consumed by the cells.

In 1988 the American Public Power Association (APPA), together with the Electric Power Research Institute (EPRI), promoted an international competition to design fuel cells tailored to urban needs. The winning project was a 2-MW MCFC developed by Energy Research Corporation (ERC), now renamed Fuel Cell Energy, Inc. These "second-generation" cells were evaluated and in 1996 appeared to have come close to their design performance. The plant achieved the excellent measured efficiency of 43.6% when delivering 1.93 MW ac to the grid.

The cells were assembled in bipolar stacks (see the Engelhard battery in Figure 9.5) bracketed by nickel-coated stainless steel end plates and separated by bipolar plates made of the same material. The cells themselves were a sandwich of an anode, an electrolyte matrix, and a cathode. The nickel-ceramic anode and the nickel oxide cathode were porous. The reactive gases were fed in through the side opposite to the one in contact with the electrolyte, which was a mixture of lithium and potassium carbonates held in lithium aluminate matrix. Note that the general configuration of the cells is still quite similar to that of the UTC PAFC.

At the high operating temperature of between 600 and 700 C, no expensive platinum-based catalyst is needed, and, as we stated, internal reforming of the fuel can be achieved. Cogeneration schemes can be implemented. On the other hand, these high temperatures limit the lifetime of the cells by slowly dissolving the cathode and by causing the carbonate to poison the reforming catalyst.

Based on the experience gained from operating the Santa Clara plant and from other development work, Fuel Cell Energy, Inc. is now offering plants of up to 2.8 MW for commercial sale. The plants operate at 47% efficiency. Exhaust gases at 370 C can be used to drive bottoming cycle engines, thus, very substantially increasing the overall efficiency. It is not clear what the lifetime of the cell modules is; what is clear is that these modules (not the whole plant) are quite massive. They generate only some

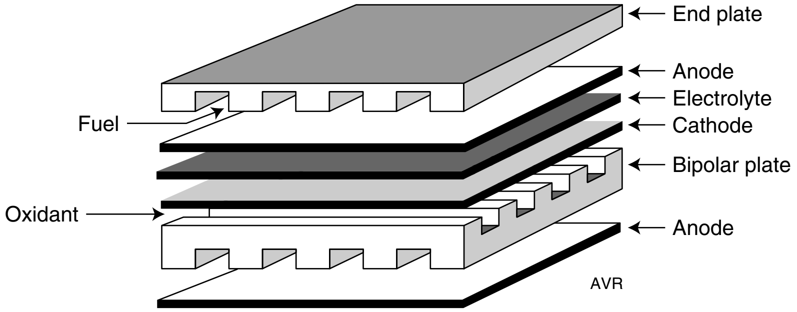


Figure 9.6 Exploded view of a MCFC unit.

20 W/kg. Compare with a modern SPFC that delivers about 700 W/kg. To be sure, the MCFC does internal reforming, while the SPFC requires pure hydrogen and has to operate with an external reforming unit.

9.5.4 Ceramic Fuel Cells (SOFCs)

9.5.4.1 Third-Generation Fuel Cell Power Plant

The quest for utility-sized fuel cell plants started with low-temperature liquid-electrolyte PAFC, then went to medium-temperature molten carbonate, MCFC, units and reached its third generation by going to high-temperature ceramic fuel cells, SOFC. All three are still in contention, but SOFC may turn out to be the final survivor.

One of the most critical components in any type of fuel cell is the electrolyte: the ceramic in SOFC is no exception. The ceramics used as electrolytes must have a much higher conductivity for ions than for electrons, a requirement that demands operation at elevated temperatures. Figure 9.7 illustrates the strong influence of the operating temperature on the achievable power density in a fuel cell using yttria-stabilized zirconia (YSZ) electrolyte. The data in the figure are from Global Thermoelectric Corp. (see Ghosh et al. 2000). No information on the thickness of the electrolyte was given. Because there are difficulties associated with high-temperature operation, the performance of the cell is frequently limited by the relatively low conductance of the electrolyte.

1. High temperatures require the use of expensive alloys.
2. Temperature cycling introduces mechanical stresses.

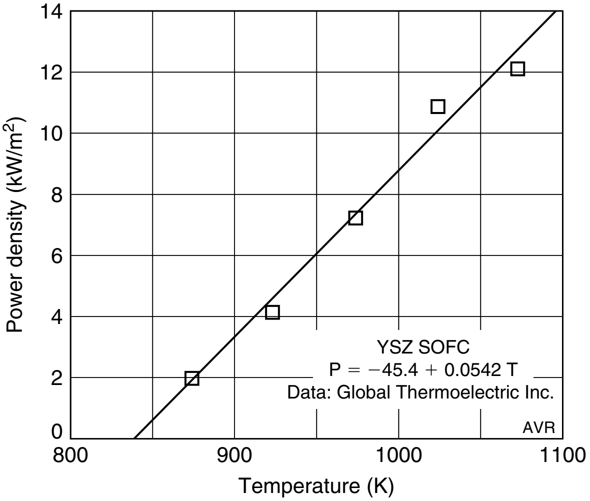


Figure 9.7 The power density of a YSZ SOFC rises sharply with operating temperature.

3. The electrodes (but not the electrolyte) must be porous; however, high temperatures promote their sintering, causing them to become progressively more impermeable to fuel and air.
4. High temperatures promote the diffusion of constituents of the electrodes into the electrolyte.

Items 2 through 4 reduce the lifetime of the fuel cell.

The conductance of a ceramic electrolyte depends on three factors:

1. The operating temperature.
2. The thickness of the electrolyte.
3. The nature of the electrolyte, which, in most current cells is YSZ, typically $(ZrO_2)_{0.9}(Y_2O_3)_{0.1}$.

In order to achieve acceptable conductances, the electrolyte must be quite thin. Global Thermoelectrics has demonstrated 5-μm-thick electrolytes (Ghosh et al. 2000) but current commercial cells use thicknesses that are almost one order of magnitude larger. The reason is that the electrolyte must be impermeable to gas, a condition difficult to satisfy when very thin layers are used—porosity will be high and small pin holes are apt to occur. Layer densities of some 98% of the theoretical density are desirable. In addition, thin electrolytes are exceedingly fragile and must be supported by either the anode or the cathode (i.e., they must be thin compact ceramic layers deposited on one of the two electrodes).

Lower temperature operations can be achieved by employing ceramic materials other than the popular YSZ. For example, Samarium doped cerial (SDC) is a ceramic electrolyte that, at any given temperature, has much higher ionic conductivity than YSZ. (See Sub-subsection 9.5.4.2.) Ceramics that conduct protons instead of negative ions have been demonstrated in the laboratory. It is expected that, at 700 C, these ceramics, working with the same power density as zirconia at 1000 C, will deliver some 10% higher efficiency. A promising “low-temperature” ceramic proton conductor is $BaCeO_3$. (See Rocky Goldstein, EPRI, Advanced Fossil Power System Business Unit.) Any ceramic used as an electrolyte must not react chemically with the materials with which it is in contact.

The most popular cathode material for YSZ fuel cells is strontium-doped lanthanum manganite, LSM ($La_xSr_{1-x}MnO_3$), which is compatible with the electrolyte. Strontium doping increases the electronic conductivity of the material, which is mainly an electron conductor. Consequently, a chemical reaction is required in which the negative ion conducted by the electrolyte must be transformed into electrons conducted by the cathode. This reaction must occur at a triple point where cathode, electrolyte, and oxygen come together. The triple point area can be increased, as shown by Yoon et al. (2000), by coating the pores of the cathode with macroporous YSZ. The cathode material must exhibit high porosity (say, 50%) and must be stable in an oxidizing atmosphere. Porosity is achieved by incorporating a pore-forming substance such as starch in the powder mixture that will form the ceramic.

The anode, being in a reducing atmosphere, could be made of porous nickel were it not for the high coefficient of thermal expansion of the metal. To correct this, the nickel is dispersed in a matrix of yttria-stabilized zirconia forming a cermet.[†] YSZ powder is mixed with NiO_2 and sintered together in an environment that reduces the nickel oxide to dispersed metallic nickel, leaving behind a porous ceramic. The fuel electrode must have good catalytic action if internal fuel reforming is used. If the temperature is much smaller than some 700 C, internal reformation may become impossible or difficult.

All three layers—anode, electrolyte, and cathode—must have matching thermal expansion to avoid delamination and must not react chemically with one another or diffuse into the neighbor.

To form a stack, interconnections between the different cells are required. They must have good mechanical properties and good electric conductivity; they must also have appropriate thermal expansion coefficients and good resistance to corrosion and chemical compatibility with the rest of the cells. Above all, they must be of low cost.

Iron alloys, rendered corrosion resistant by the addition of chromium, satisfy most of the above requirements. As in all stainless steels, their rust resistance depends on the formation of a thin scale of chromia on the surface effectively sealing the bulk of the material from the corroding atmosphere. Unfortunately, chromia, at the high operating temperature of the fuel cells, forms volatile CrO_3 and $CrO_2(OH)_2$, which contaminate the cathode, severely degrading the performance. CrO_3 is the dominant chromium vapor in cells using oxygen or very dry air as oxidant, while $CrO_2(OH)_2$ (chromium oxyhydroxide) is dominant when moist air is used. Chromium vaporization can be reduced by using dry air and operating at low temperature. The latter may not be economically attractive.

Power densities as high as $12\,\mathrm{kW/m^2}$ have been demonstrated in the laboratory (Pham et al. 2000). These levels are below the $20\,\mathrm{kW/m^2}$ reached by modern SPFC (Ballard) and refer to individual cells, not to the whole stack. SOFCs in or near production show a more modest power density of some $3\,\mathrm{kW/m^2}$ (Siemens Westinghouse). This may not be a major disadvantage in stationary applications, but may be more of a problem in automotive use when compactness is desirable.

SOFCs have high efficiencies (over 50%), especially if used in hybrid (cogeneration) as in Figure 9.8 arrangements when overall efficiencies approaching 60% have been achieved. Since no corrosives are used (such as phosphoric acid in PAFC or molten carbonates in MCFC), long life is possible.

[†] Cermets are combinations of ceramic materials with metals.

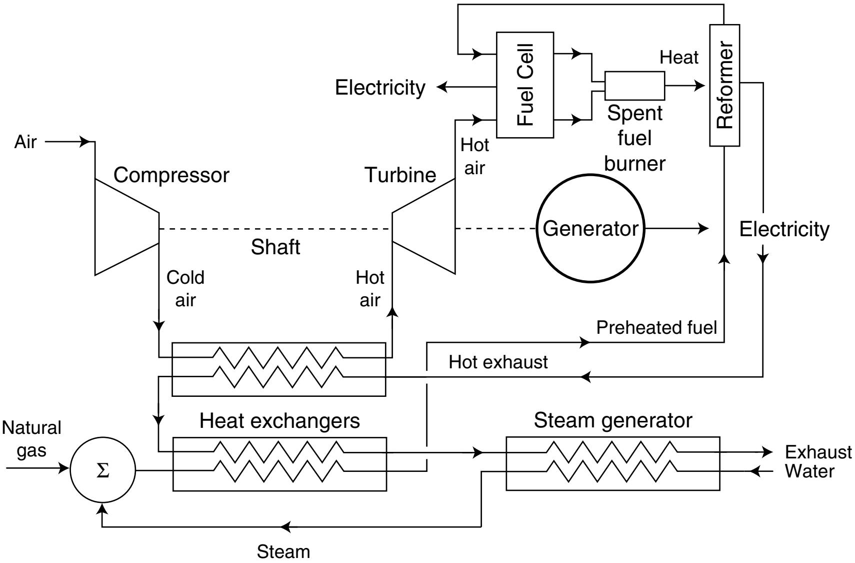


Figure 9.8 An arrangement for a SOFC/microturbine combined cycle plant (Fuller et al. 2000).

Chromium

Chromium forms compounds in which it can assume at least three different oxidation states.[†] Therefore, it has a number of different oxides: chromous oxide, CrO, also called chromium monoxide, oxidation state II, chromic oxide, Cr_2O_3, also called chromium sesquioxide or **chromia**, oxidation state III, and chromium trioxide, CrO_3, oxidation state VI.

The significance, as far as the use of iron-chromium alloys in fuel cells is concerned, is that, whereas chromia is highly refractory (melting point 2435 C), chromium trioxide is quite volatile, having a melting point of only 196 C. The conditions inside a fuel cell lead to the transformation of stable chromium oxides into volatile compounds capable of contaminating the cathode of the cell.

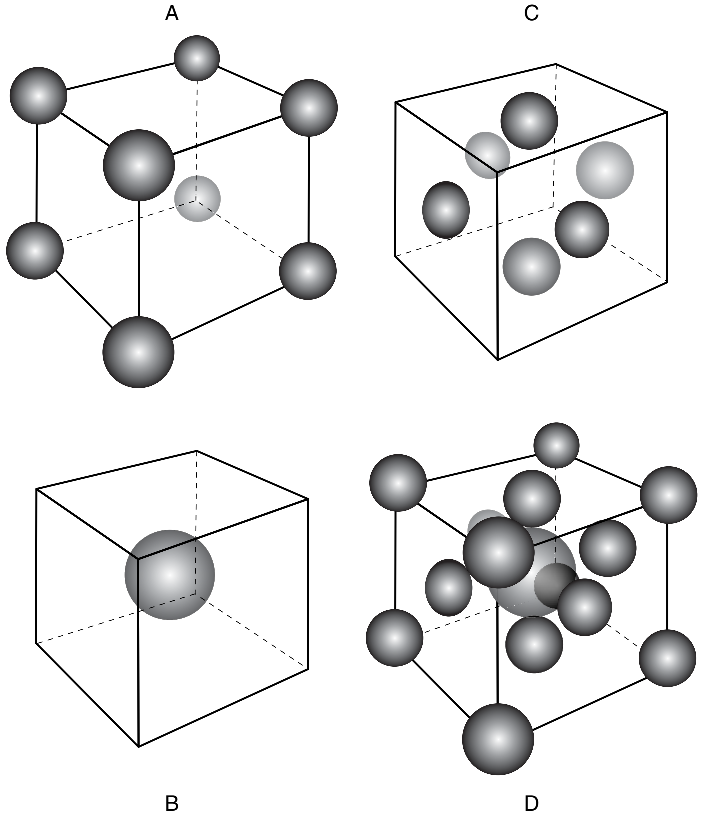
[†] The oxidation state is the difference in the number of electrons of a neutral atom and that of the corresponding ion in an ionic compound.

Continuous operation for over 35,000 hours has been demonstrated (see Bessette and Pierre, 2000). SOFCs have an unlimited shelf life. Many of the ceramics used in fuel cells have a perovskite structure; (see the box, Perovskite on the next page).

In negative-ion solid electrolyte fuel cells, the reactions are those described in Subsection 9.44.

SOFCs are usually either planar or tubular. The tubular SOFCs, as exemplified by the Siemens Westinghouse cells, have the great advantage of doing away with seals that plague planar cells.

Perovskite



Perovskite is the mineral $CaTiO_3$. By extension, the word designates any substance having the empirical formula DEF_3 (where F usually corresponds to oxygen) which crystallizes in the perovskite arrangement.

Consider a cubic crystalline unit cell having 8 atoms of element D at the eight vertices (see “A” in the figure). When these unit cells are stacked together, each of the atoms is shared among 8 different, adjacent cells. In other words, there is, on average, exactly 1 atom of element D per cell.

(*Continues*)

(*Continued*)

Now refer to “B” in the figure. There is 1 single element E atom per unit cell.

Finally, refer to “C.” Element F atoms are centered on the face of the cube. There are 6 faces, but each atom is shared by 2 adjacent unit cells. Thus, on average, there are 3 F atoms per cell. In “D,” the three arrangements are put together yielding a unit cell with the empirical formula DEF_3.

Perovskites are important in SOFC electrolytes and in high temperature superconductors.

9.5.4.2 High-Temperature Ceramic Fuel Cells

Siemens Westinghouse has done much work in the area of tubular cells. In 2001, this company had a 100-kWe plant in the Netherlands that operated unattended, delivering power to the grid. It also had a 220-kWe hybrid (operating in conjunction with a micro gas turbine) at the Southern California Edison Co., in Los Angeles, California. A 1-MWe plant with 58% efficiency was being installed in Fort Meade, Maryland, for the Environmental Protection Agency, while another plant of equal power was planned for Marbach, Germany.

When starting up and when stopping, SOFCs are subjected to large-amplitude thermal cycling, which creates serious problems with seals. Siemens Westinghouse has come up with a clever seal-free configuration using tubular cells. Each individual cell consists of a triple-layered ceramic tube, as depicted in Figure 9.9. The inner layer is the air electrode (cathode), the middle layer, the electrolyte, and the outer, the fuel electrode (anode). Manufacture of the fuel cell starts with an air electrode made of lanthanum manganite ($La_{0.9}Sr_{0.1}$). The YSZ electrolyte is built up on the

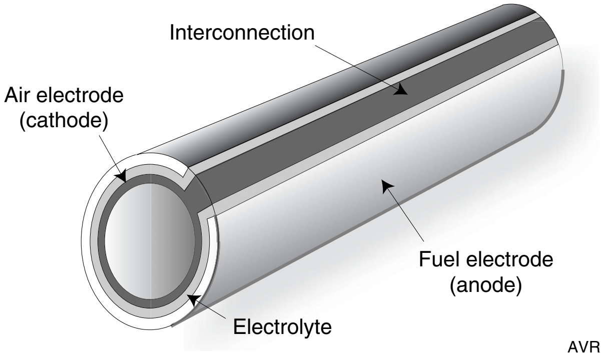


Figure 9.9 Tubular concentric electrodes of the Siemens Westinghouse SOFC.

cathode tube, and the anode is deposited on top of that. Connections to the anode are made directly to the outer electrode. To reach the cathode, an interconnecting strip is placed longitudinally on the tube and penetrates to the inner layer (see Figure 9.9). This interconnecting strip must be stable at both the oxidizing environment of the air electrode and the reducing one at the fuel electrode. It must also be impermeable to gases. These requirements are met by lanthanum chromite. To enhance conductivity, the material is doped with Ca, Mg, Sr, or other metals of low valence.

Figure 9.10 shows how bundles of tubes can be stacked in series-parallel connections forming modules. Nickel felt made of nickel fibers sintered together provides mechanically compliant interconnections. Notice that all these interconnections take place in a chemically reducing environment.

The modules consist of two chambers as indicated in Figure 9.11. The lower and longer **fuel reaction chamber** receives the fuel from openings at the bottom. The fuel flows up and most of it is used up by the anode.

Unreacted fuel, some 15% of the total, exits into the **spent fuel combustion chamber** where it is mixed with the excess air that comes out from the top of each tube and burns, producing heat used in preheating the input air and increasing the temperature (up to 900 C) of the exhaust gases in order to drive a bottoming cycle generator such as a turbine.

If an adequate amount of water vapor is mixed with fossil fuel gases, automatic reformation will take place thanks to the catalytic action of the nickel in the anode. (See subsection on modern hydrogen production in Chapter 10.)

$$CH_4 + H_2O \rightarrow CO + 3H_2.$$

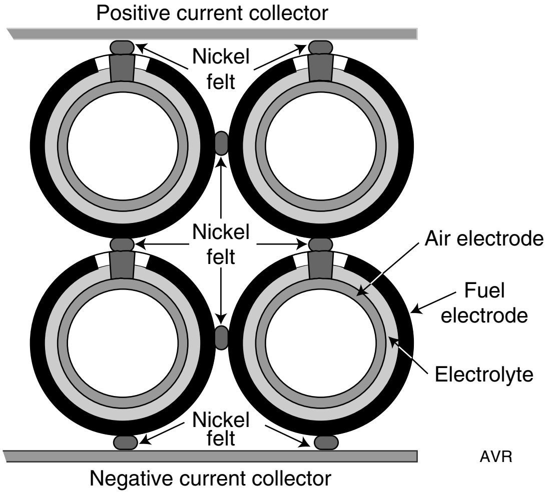


Figure 9.10 Tubular cells can easily be stacked in series-parallel connections.

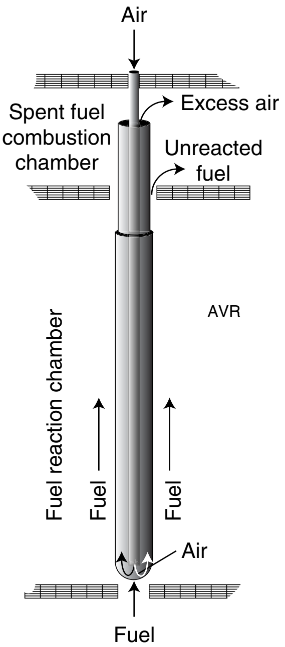


Figure 9.11 Tubular cells are stacked in seal-free modules.

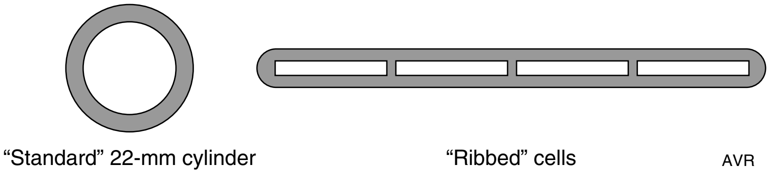


Figure 9.12 Cross section of cylindrical and flattened "ribbed" cells.

"Standard" Siemens Westinghouse cells are 22-mm diameter tubes of varying lengths, usually 1.5 m. Better performance is obtained by the use of alternate geometries. See Figure 9.12.

Flattening the cells into the "ribbed" configuration not only improves the stacking factor (so that more units fit into a given volume) but also reduces production cost. These flattened cells have incorporated improvements that substantially increase cell efficiency. See the comparison of the performances of cylindrical and "ribbed" cells shown in Figure 9.13.

9.5.4.3 Low-Temperature Ceramic Fuel Cells

It would be useful to operate SOFCs at temperatures lower than those currently used with YSZ. Among other things, the life of the stacks would presumably be longer. It would also be simpler if one could build a **single-chamber** fuel cell.

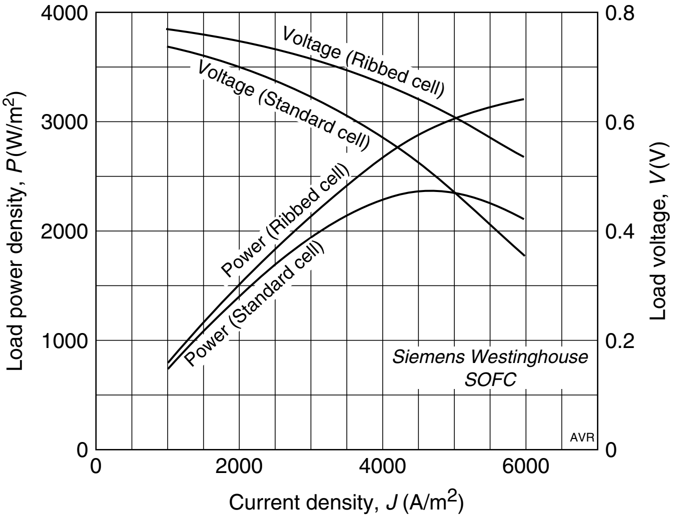


Figure 9.13 "Ribbed" cells show substantially better performance than cylindrical ones.

In the usual fuel cell, fuel and oxidizer are fed separately to the device, each gas to its own electrode. In a single chamber cell, the fuel is mixed with air in proportions that are too rich to allow an explosion. The mixture is fed simultaneously to both electrodes, one of which reacts preferentially with the fuel, the other with the oxygen.

In the cell described by Hibino, the SDC electrolyte was a ceramic disc ground to 0.15-mm thickness. The anode was a layer of nickel-enriched SDC, while the cathode consisted of $Sm_{0.5}Sr_{0.5}CoO_3$—samarium strontium cobalt oxide.

The ethane (18%)/air mixture in contact with the nickel-rich anode is reformed into a hydrogen/carbon monoxide mixture according to

$$C_2H_6 + O_2 \rightarrow 3H_2 + 2CO. \tag{9.25}$$

The two gases in the reformate are oxidized:

$$3H_2 + 3O^{--} \rightarrow 3H_2O + 6e^-, \tag{9.26}$$

$$2CO + 2O^{--} \rightarrow 2CO_2 + 4e^-. \tag{9.27}$$

Thus, each ethane molecule yields 10 electrons that circulate in the load. The five oxygen ions come from the cathode (by moving through the electrolyte). They are produced by combining the 10 electrons with 2.5 oxygen molecules from the fuel/air mixture:

$$10e^- + 2.5O_2 \rightarrow 5O^{--}. \tag{9.28}$$

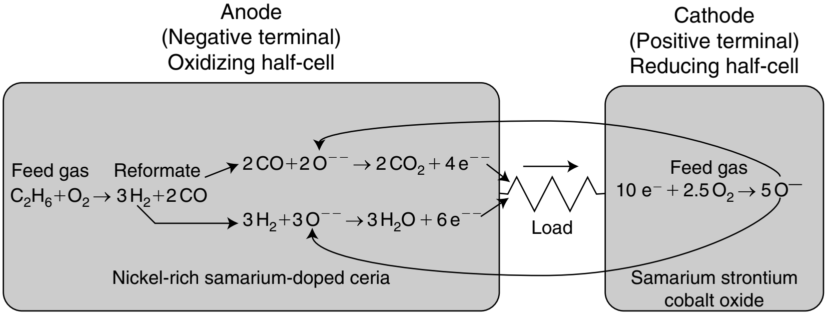

Anode
(Negative terminal)
Oxidizing half-cell
Cathode
(Positive terminal)
Reducing half-cell
Feed gas
$C_2H_6+O_2 \rightarrow 3H_2+2CO$
Reformate
$2CO+2O^{--} \rightarrow 2CO_2+4e^{--}$
$3H_2+3O^{--} \rightarrow 3H_2O+6e^{--}$
Load
Feed gas
$10\,e^- + 2.5\,O_2 \rightarrow 5\,O^{--}$
Nickel-rich samarium-doped ceria
Samarium strontium cobalt oxide

Figure 9.14 Cell reactions of the SDC low-temperature SOFC.

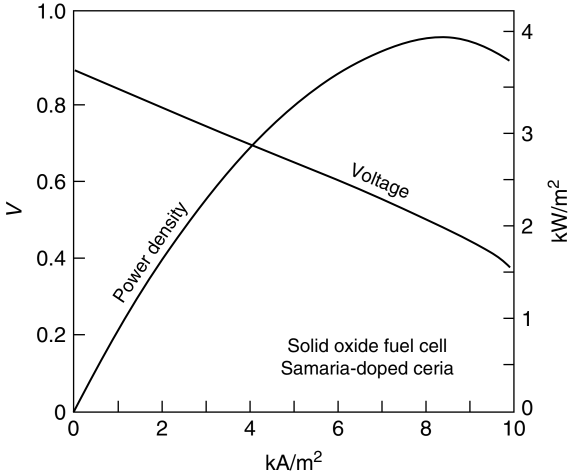


Figure 9.15 Characteristics of an experimental SDC fuel cell developed by Hibino.

The overall reaction is shown in Figure 9.14.

The single-chamber configuration simplifies the construction and makes the cell more resistant to both thermal and mechanical shocks.

The performance of some of the experimental cells prepared by Hibino is quite promising: over 4 kW/m^2 at 500 C and over 1 kW/m^2 at 350 C. Compare with the 20 kW/m^2 of the much more developed Ballard SPFC.

The V–I characteristics of one of the experimental SDC cells, with 0.15-mm electrode thickness, is shown in Figure 9.15.

9.5.5 Solid-Polymer Electrolyte Fuel Cells

The solid-polymer electrolyte is conceptually the simplest type of fuel cell and, potentially, the easiest to manufacture. As the name suggests, the electrolyte is a solid membrane made of an ion-conducting polymer. The ion can be positive (a cation), usually a proton, or negative (an anion) usually an OH^-.

SPFCs are safer than liquid or molten electrolyte cells because they use noncorrosive (and of course, nonspillable) electrolytes. They can stand fairly large differential pressures between the fuel and the oxidant sides, thus simplifying the management of these gases.

A short history of the development of SPFC cells can be found in an article by David S. Watkins. SPFCs were pioneered by GE, starting in 1959. In 1982, the technology was transferred to United Technology Corporation/Hamilton Standard. Little additional progress was made until Ballard Power Systems[†] took up further development and pushed it, with the collaboration of Daimler Benz, to the production stage. Prior to Ballard, fuel cell cars were predicted to come on the market by 2020. Now, there is hope for their introduction by 2010.

The progress made by the original GE effort is illustrated by the growth of power densities from 50 W/m^2 in 1959 to over 8 kW/m^2 in 1982, a 160-fold improvement. Many factors contributed to such progress, including better membranes (early devices used sulfonated polystyrene, later ones used Nafion), thinner membranes (from 250 μm to 123 μm), higher operating temperatures (from 25 C to 150 C), and better catalysts.

SPFCs are fast emerging as the preferred solution for automotive use and as a competitor for fixed power plants. The extremely fast advances in this type of cell, a sign of a young and immature technology, is attested to by the exponential improvement in power density of Ballard's cells illustrated in Figure 9.16. Ballard cells now exceed 20 kW per square meter of active surface (compare with 8 kW/m^2 for the best GE cells). The power-to-mass ratio (gravimetric power density) has also improved substantially—it exceeds 0.7 kW/kg, approaching the values for a modern aircraft engine.[††] Thus, power-to-volume and power-to-mass ratios of a modern SPFC are already well into the range of acceptability in automotive applications. What needs to be improved—substantially—is the power-to-cost ratio.

In 2005, the cost[†††] of Ballard automotive SPFC was 73 \$/kW, and it is expected to reach the Department of Energy (DOE) target of \$30/kW by 2010. Cost can be reduced in many areas. Most obvious is the need to create a market large enough to permit mass production. Cheaper membranes and catalysts are a must.

To be acceptable as an automobile power plant, fuel cells must be able to start up quickly under freezing conditions. By 2004, Ballard had already

[†] For Ballard's history, read Tom Koppel (1999). For current Ballard technology, go to <**http://www.ballard.com**>.

[††] To be fair, one must compare the aircraft engine weight with the **sum** of the fuel cell stack weight plus that of the electric motor.

[†††] Cost of stack alone, not of the whole system.

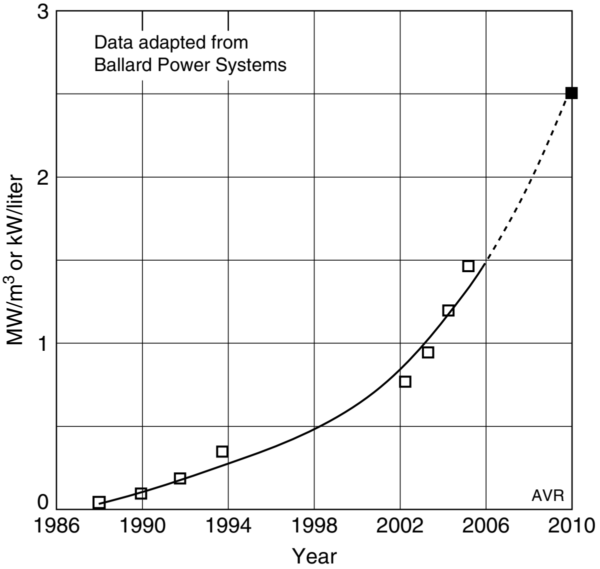


Figure 9.16 The power density of Ballard fuel cells has risen exponentially since 1986. In 2005, about 1.5-MW could be generated by each m^3 of cells (1.5 kW/liter). The black square indicates the estimated performance by 2010.

demonstrated an 8-second start-up time (the time to reach 50% of full power) when the ambient temperature was −15 C. In 2005, a 16-second start-up time was demonstrated when ambient temperature was −20 C. The hoped-for performance for 2010 is 30 seconds when the temperature is −30 C.

A modern automobile will last as much as 300,000 km, which corresponds to 6250 hours at an average speed of 48 km/h. Present cells have less than half this durability (2200 hr). The DOE target for 2010 is 5000 hr.

9.5.5.1 Cell Construction

The construction of a typical SPFC is illustrated in Figure 9.17.

The heart of the cell is the **membrane electrode assembly (MEA)**, which consists of a solid electrolyte sandwiched between two electrodes. A catalyst layer is inserted between electrodes and electrolyte. The catalyst can be bonded to either the electrode or the electrolyte. When hydrogen is used, the anode catalyst is frequently a platinum/ruthenium mixture either unsupported or supported on high surface area carbon blacks (more than 75 m^2 per gram of carbon). Ruthenium is used to improve tolerance to CO, which may be an impurity in hydrogen produced by reformation of an organic fuel. The catalyst at the cathode is either pure platinum or platinum alloyed to some stable metal (Cr, Zr, or Ti). This is necessary because of the highly oxidizing environment at the cathode.

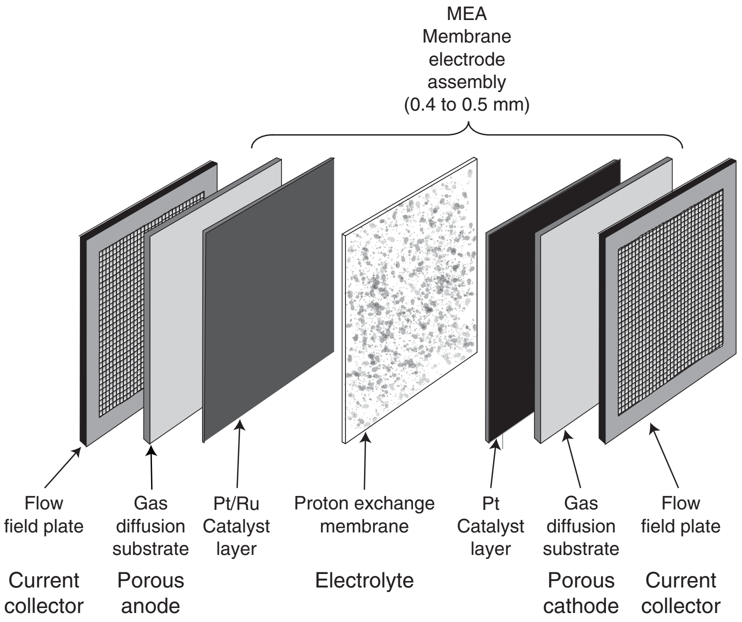


Figure 9.17 Structure of a SPFC.

The electrodes must be good electronic conductors and must be porous. They should be inexpensive to mass produce, possibly using a reel-to-reel technique. The British firm, Johnson Matthey, has developed such a technique using a tangle (not a weave) of carbon fibers embedded in teflon to produce flexible electrode structures.[†] There is a possibility of incorporating into the gas diffusion substrate a catalyst capable of promoting the oxidation of CO into CO_2, thus reducing the concentration of CO.

Contacts to the electrodes are made through gas flow plates (metal or, more commonly, graphite) on whose surface grooves have been formed to distribute the fuel and the oxidant uniformly over the whole active area of the cell. An example of the complicated pattern of grooves in the gas flow plates appears in Figure 9.18.

The plates have a number of holes near the corners constituting a manifold through which fluids flow. All but the two end plates are bipolar—one face connects to the anode, the other to the cathode. The plate depicted shows its cathode side to which oxidant is fed and from which oxygen-depleted oxidant and product water are removed. The anode side of the flow plate is fed through the smaller hydrogen holes. This is because air is delivered in larger quantities than what would be demanded by stoichiometry—the excess cools the cells and removes the product water.

[†] U.S. Patent 6949308.

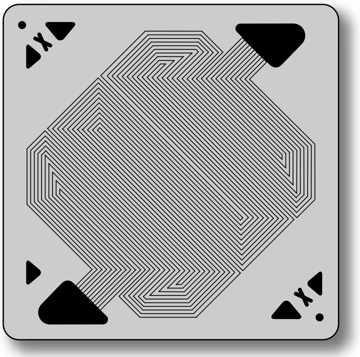


Figure 9.18 Grooves are formed on the gas flow plates to distribute fuel and oxidant. In this cathode face of the plate, oxidant enters and exits through the large, roughly triangular holes at the upper right-hand and lower left-hand corners of the plate. The remaining holes are for hydrogen and water (Ballard).

The design of the groove pattern can greatly influence the performance of the cell. The grooves should not trap any of the reaction water, thereby "drowning" areas of the cell, and must distribute the reactants uniformly. The bipolar configuration facilitates stacking the cells so that they are electrically in series.

9.5.5.2 Membrane

Although most membranes are of the proton-exchange type, anion-exchange ones have been proposed claiming a number of advantages. Among these are their immunity to CO_2 fouling, simplified water management requirements, and the possibility of using less expensive nonnoble metal catalysts.

Membranes that act as electrolytes must be excellent ionic conductors, but must not conduct electrons. They must be reasonably impermeable to gases and sufficiently strong to withstand substantial pressure differences.

Many such membranes consist of a Teflon-like backbone to which side chains containing a sulfonic (SO_3) group are attached. These acid groups are immobile and cannot be diluted or leached out.

The thinner the membrane, the smaller its resistance to the flow of protons. Typically, membranes 50 to 180 μm thick are used.[†] If they are too thin, gas crossover occurs and the performance of the cell is degraded, as witnessed by a reduction in the V_{oc} of the cell.

For a membrane to be a good proton conductor, its SO_3 concentration must be sufficiently high. The ratio between the mass of dry polymer (in kg) and the number of kilomoles of SO_3 sites is called the **equivalent weight**

[†] The paper commonly used in copying machines is roughly 100 μm thick.

(EW). Smaller EWs lead to higher ion conductivity. For membranes with the same backbone, the shorter the side chains, the smaller the EW. Thus Nafion, having a side chain,

$$-\mathrm{O}-\mathrm{CF_2}-\mathrm{CF_3CF}-\mathrm{O}-\mathrm{CF_2}-\mathrm{CF_2}-\mathrm{SO_3H},$$

has a larger EW than the Dow membrane with its shorter chain,

$$-\mathrm{O}-\mathrm{CF_2}-\mathrm{CF_2}-\mathrm{SO_3H}.$$

In fact, Nafion has a typical EW of 1100 kg per kmoles, whereas the Dow membrane has an EW of some 800 kg.

The proton conductivity of a membrane also depends on it being thoroughly hydrated—a dry membrane loses most of its conductivity, and, for this reason, careful water management is essential (see Section 9.5.5.4).

Membranes can be surprisingly costly. In the 1980s, Nafion sold at \$800 per m^2. One reason is that this material was created to last 100,000 hours of operation in a chlorine-producing plant. On the other hand, an automobile is not expected to survive much more than 300,000 km, which, even at a very modest speed of 30 km/h, represents only some 10,000 hours of operation. Nafion is, as far as automotive fuel cells are concerned, "overdesigned" by one order of magnitude.

Ballard set out to develop its own proprietary membranes, which contain much less fluorine and are much cheaper in spite of being even more efficient than Nafion and the Dow membrane. Surprisingly the life of the membrane exceeds 15,000 hours, more than enough for an automotive fuel cell.

9.5.5.3 Catalysts

Because of their low operating temperature, the kinetics of SPFC are unacceptable unless special catalysts are used. The problem is much more serious in the cathode where the oxygen reduction takes place than in the anode where the simpler hydrogen oxidation occurs. For the cathode, the catalyst is platinum (or an alloy of platinum with Cr, Zr, or Ti, stable in the reducing environment). In the anode, Pt is used alone or alloyed to Ru. A problem at the anode arises when hydrogen is extracted from a carbon-containing feed stock such as methanol or methane. This hydrogen may be contaminated by as much as 1% or 2% carbon monoxide that strongly binds to the platinum. At 80 C, even a few parts per million of CO will cover some 98% of the platinum surface interfering with the catalysis.

Hydrogen containing small amounts of CO (say 100 ppm) will drastically impair the performance of low-temperature fuel cells with pure Pt catalysts as shown in Figure 9.19. However, if a small amount of O_2 is added to the fuel, it will selectively combine with the carbon monoxide (oxidizing it to CO_2), and almost complete recovery of the cell performance is achieved. The CO vulnerability of Pt catalysts diminishes with

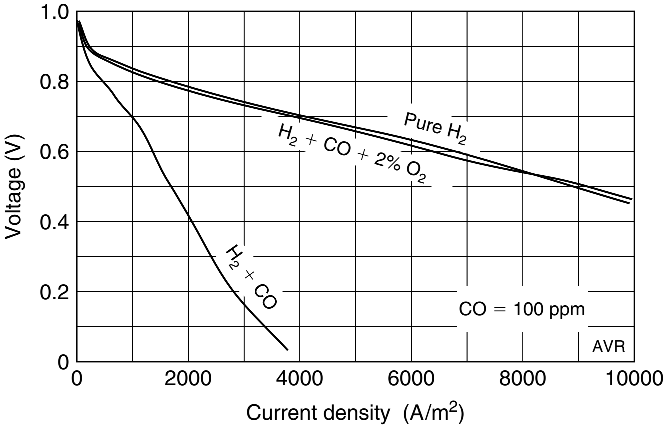


Figure 9.19 When a SPFC using pure Pt catalysts is fed hydrogen with 100 ppm of CO, its performance suffers drastically. However, if a small amount of oxygen is added to the fuel stream, it will selectively oxidize the CO, and the performance is then essentially indistinguishable from that with pure hydrogen.

increasing temperature. At 80 C, CO concentration must be kept below 10 ppm. At 90 C, where Ballard cells operate, somewhat higher concentrations are acceptable.

Another way to reduce the sensitivity of platinum to CO is to alloy it with ruthenium. $Pt_{0.5}Ru_{0.5}$ is commonly used as an anode catalyst.

The CO sensitivity can, of course, be ignored if pure hydrogen or hydrogen containing only traces of CO is used. To that end, purification processes are being actively investigated. As an example, IdaTech, of Bend, Oregon, has developed fuel processors capable of producing very pure hydrogen thanks to the use of palladium filters (see Chapter 10). The Los Alamos National Laboratory has built such a filter based on a thin tantalum sheet plated with palladium on both sides.

If the fuel cell catalysts become poisoned with CO, all is not lost—a short exposure to pure H_2 will completely rejuvenate the cell.

CO is not the only contaminant to avoid. SPFCs have extremely low tolerance for sulfur compounds (H_2S, for instance) and for ammonia (NH_3).

Since platinum is expensive and its sources are limited (the price of platinum grows when the demand increases), great effort has been made to lower the amount of catalyst used. In earlier cells, platinum loading of about 4 mg/cm^2 added more than \$500 to the cost of each square meter of active fuel cell area. Fortunately, techniques that permit a reduction of a factor of 10 in the platinum loading are now practical, and the Los Alamos National Laboratory has had success with even less catalyst. The British firm, Johnson Matthey, producers of catalytic converters for conventional

(IC) cars, is working with Ballard to reduce the cost of catalysts in SPFCs.

The quest for reduced platinum loading is partially based on the fact that in a catalyst it is only the surface that is active. By reducing the size of the grains, the surface-to-volume ratio is increased—that is, less platinum is needed for a given catalyst area. **Supported catalyst**—catalyst supported on high surface area grains of carbon black with graphitic characteristics—are employed to improve platinum usage. The grains must have a very large surface area (more than 75 m^2 per gram of carbon) and adequate porosity.

Other types of catalysts are being considered, and it may even be possible to use enzymes for this purpose. The anode reaction consists of splitting the hydrogen molecule into two protons and two electrons (see Equation 9.3), a feat performed by many anaerobic bacteria (Chapter 13) that use enzymes called **hydrogenases**.

Synthetic hydrogenases have been created by Robert Hembre of the University of Nebraska, Lincoln, using compounds containing ruthenium and iron. Ruthenium, being substantially cheaper than platinum, may have a future as a catalyst in low-temperature fuel cells.

9.5.5.4 Water Management

Present-day ion-exchange membranes are an acid electrolyte. Therefore, water forms at the cathode where it would collect and "drown" the electrode—that is, it would impede the penetration of the oxygen to the active surface. To reduce the tendency of the water to collect in the pores and interstices of the MEA, a hydrophobic material is applied (typically a Teflon emulsion).

Water is removed by the flow of oxidant, which is generally supplied well in excess of the stoichiometric requirement. Usually, the amount of air circulated is double that needed to furnish the correct amount of oxygen.

Although the anode reaction (Equation 9.3) indicates the formation of protons, the ion that actually migrates through the electrolyte is hydronium—a hydrated proton. Thus, water is consumed at the anode because it is electrically "pumped" to the cathode, tending to desiccate the membrane. This drying is compensated, in part, by water from the cathode that diffuses back through the membrane, driven by the concentration gradient. Nevertheless, there is a tendency to dehydrate the proton-exchange membrane, which tends to wrinkle and have its proton conductivity drop catastrophically. To avoid desiccation, a partial pressure of at least 50 kPa of H_2O must be maintained in both the fuel and oxidant streams. If the feed pressure is 100 kPa (1 atmos), the partial pressure of H_2 would be some 50 kPa and the output voltage would fall correspondingly. For this reason, the fuel cell is usually pressurized to 300 kPa (3 atmos), leading to a hydrogen partial pressure of 250 kPa, compatible with an efficient fuel cell operation.

Careful water management is necessary to avoid either “drowning” or “desiccation” of the membrane.

9.5.6 Direct Methanol Fuel Cells (DMFCs)

As discussed in the preceding subsection, one attractive alternative to providing a portable source of hydrogen is to use a hydrogen carrier, a substance from which the gas can be extracted as needed. This requires some sort of chemical reforming process. Methanol is a leading candidate as a hydrogen carrier for vehicular applications. Nevertheless, the need for separate fuel processing equipment in the vehicle adds to the cost and complexity of the system. An obvious solution is to develop fuel cells that can use methanol directly without the need of preprocessing.

The Jet Propulsion Laboratory has done considerable work on the development of direct methanol fuel cells (DMFCs) (see Halpert et al. 1997).

The fuel used is a low-concentration (3%) methanol in water solution with air as the oxidant.

The anode reaction is

$$CH_3OH + H_2O \rightarrow CO_2 + 6H^+ + 6e^-. \tag{9.29}$$

The cathode reaction is

$$6H^+ + 6e^- + \frac{3}{2}O_2 \rightarrow 3H_2O. \tag{9.30}$$

Thus, the overall reaction is

$$CH_3OH + \frac{3}{2}O_2 \rightarrow CO_2 + 2H_2O. \tag{9.31}$$

Halpert et al. (1997) reports a cell producing 5.0 MJ of energy per liter of methanol consumed when operated at 90 C and 2.4 atmospheres of air pressure. This can be interpreted as an efficiency of 28% if referred to the higher heat of combustion of the alcohol or 32% if referred to the lower.

The JPL DMFC is a solid-polymer fuel cell using a Nafion 117 membrane with a fairly high loading of noble-metal catalyst (2 to 3 mg of Pt-Ru per cm^2). The combination of Nafion with high catalyst loading leads to high costs. The search for better anode catalysts can be greatly accelerated by the technique described by Reddington et al. (1998) that permits massively parallel testing of a large number of catalyst samples. See also Service (1998).

In DMFC, since the anode is in direct contact with water, problems of membrane dehydration are circumvented and the constantly flowing liquid simplifies heat removal. Pollution problems are alleviated because only carbon dioxide and water are produced, whereas carbon monoxide is

Table 9.1 Thermodynamic Data for Methanol Fuel Cells at RTP

Methanol	**Water**	$\Delta \boldsymbol{H}^\circ$	$\Delta \boldsymbol{G}^\circ$
		MJ/kmole	**MJ/kmole**
liquid	gas	−638.5	−685.3
gas	gas	−676.5	−689.6
liquid	liquid	−726.5	−702.4
gas	liquid	−764.5	−706.7

Data from Dohle et al. (2000).

generated when methanol is reformed into hydrogen for use in other types of cells.

"Methanol crossover" (i.e., the transport of methanol from anode to cathode through the Nafion membrane) severely reduces cell efficiency. In the JPL cell mentioned earlier, crossover consumed 20% of the fuel. Efforts are being made to develop new low-cost membranes that are less subject to this difficulty. It appears that the figure of 10% has already been achieved. Lower crossover rates permit operation, with more concentrated methanol mixtures yielding better efficiencies. JPL feels confident that efficiencies above 40% can be attained. Crossover reduces the efficiency of the cell, not only because some fuel is diverted, but also because the methanol that reaches the cathode is prone to undergo the same reaction (Equation 9.29) that normally takes place at the anode, generating a countervoltage and, thus, reducing the voltage delivered to the load. International Fuel Cells Corporation has a patent for a catalyst that promotes the reaction of Equation 9.30, but not that of Equation 9.29.[†] If the methanol at the cathode is not consumed, perhaps it can be recovered. This can be done by condensing the methanol–water vapor mixture exhausted from the cathode.

Instead of employing the conventional "bipolar" configuration, direct methanol fuel cells designed for powering small portable equipment may employ the "flat-pack" design shown in Figure 9.20 (top) or the "twin-pack" design (bottom). This lends to the batteries (or "stacks") a shape that is more compatible with the equipment with which they are used.

For convenience, we display some thermodynamic data for methanol in Table 9.1.

9.5.7 Direct Formic Acid Fuel Cells (DFAFCs)

Fuel crossover has been a difficult problem for the DMFC designers. One obvious solution is to use a different fuel, such as formic acid. The

[†] Chu et al. (2000) at the U.S. Army Research Laboratory have developed an iron–cobalt tetraphenyl porphyrin catalyst that can promote the oxygen reduction to water (Equation 9.30) but will not catalyze methanol oxidation (Equation 9.29). Consequently, it has possibilities as a cathode catalyst in DMFC.

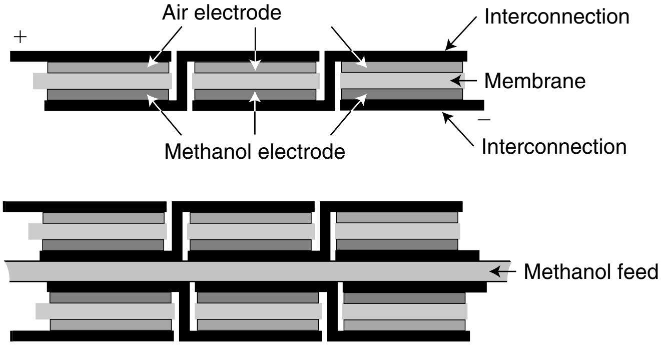


Figure 9.20 Bipolar configuration reduces the series resistance between individual cells, an advantage when high currents are involved. Flat-packs can be extremely thin, making it easy to incorporate them into small-sized equipment. In the flat-pack configuration (developed by the Jet Propulsion Laboratory; see Narayanan Valdez, and Clara (2000)) cell interconnections pierce the membrane as shown. Two flat-packs deployed back-to-back permit the sharing of the methanol feed.

crossover problem is practically eliminated, and formic acid is less toxic than methanol. The reactions involved are spelled out in Subsection 9.4.6.

Platinum does not work well with formic acid but an addition of palladium improves the situation (Rice et al. 2003). Much work on this type of fuel cell has been done at the University of Illinois in Champaign-Urbana by Richard Masel's group.

Direct formic acid fuel cells designed to provide power for laptops and cell phones are being industrialized by the Canadian company, Tekion, which has acquired the rights from the University of Illinois. It is also being pushed by the German Chemistry giant, BASF, the world's largest producer of formic acid (next to ants, themselves).

9.5.8 Solid Acid Fuel Cells (SAFCs)

Phosphoric acid and most solid polymer fuel cells are examples of cells that use hydrated acid electrolytes. Solid acids, on the other hand, can exhibit anhydrous proton conductivity and may have certain advantages over the popular SPFC. (See Boysen et al. 2004.)

H_2SO_4 is a common acid. However, it is liquid at temperatures one would expect fuel cells to operate. On the other hand, replacing one of the hydrogens by a cesium atom leads to $CsHSO_4$, an acid (known as cesium hydrogen sulfate) with a high melting point. Above 414 K, this acid becomes a good proton conductor (commonly referred to as a **superprotonic** or **superionic** conductor) and has been proposed as a fuel cell electrolyte by groups at the California Institute of Technology and the University of Washington, among others.

Since these solid acids are soluble in water and since water is produced by the fuel cell reaction, it is essential to operate the cells at temperatures high enough to ensure that any water that comes in contact with the electrolyte is in vapor form. This means that operating temperatures should be above some 150 C.

A major difficulty with $CsHSO_4$ (and with the corresponding selenate) is that they react with hydrogen forming, in the sulfate case, hydrogen sulfide. Greater stability is expected from phosphate-based compounds. CsH_2PO_4 fuel cells, for instance, were demonstrated at 235 C (see Boysen et al. 2004).

The advantage of SAFC over SPFC is that, operating at higher temperatures, they are less susceptible to CO poisoning and require fewer catalysts. In addition, since the water is always in vapor form, they do not need the careful water management required by the polymer cells.

Perhaps the most promising aspect of SAFC is that the electrolyte is not permeable to methanol, thus eliminating the serious methanol crossover problem of polymer cells. See the subsection on direct methanol fuel cells. The power densities of solid acid methanol fuel cells (Boysen et al. 2004) are within a factor of 5 of the density of the most advanced DMFCs (2004). The electrolyte thickness of these cells was 260 μm. Future cells with substantially thinner electrolytes may exhibit considerable improvement in their performance. However, one important difficulty with this type of cell is the fragility of the electrolyte.

9.5.9 Metallic Fuel Cells—Zinc–Air Fuel Cells

One of the advantages of **refuelable** cells over **rechargeable** cells is that refueling tends to be much faster than recharging (minutes versus hours). However, hydrogen fuel cells suffer from the difficulty of transporting and storing the gas. The use of certain metals as fuel may lead to simple and safe transportability and large volumetric energy density. Metallic cells are an exception to the rule of using fluids as fuel. Two of the metals considered for this purpose are aluminum and zinc. Aluminum, however, is corroded by the caustic electrolyte even when no current is being generated. Aluminum-Power, Inc. has developed a type of cell in which the electrolyte is pumped away when the cell is not in use. This apparently leads to a cumbersome system. It may be easier to use zinc. Metallic Power, Inc. has a promising zinc–air fuel cell. Small zinc pellets (about 1 mm in diameter) are one of the reactants, the other being oxygen, in general just plain air. The electrolyte is a concentrated KOH solution.

At the anode (the negative electrode) the reaction is

$$Zn + 4OH^- \rightarrow Zn(OH)_4^{--} + 2e^-. \qquad (9.32)$$

The ion on the right-hand side of the equation is a zincate ion that consists of a positively charged Zn^{++} surrounded by four hydroxyls (OH^-).

The electrons, after circulating through the load, recombine with the oxygen at the cathode and with water, regenerating half of the hydroxyl ions used at the anode,

$$2e^- + \frac{1}{2}O_2 + H_2O \rightarrow 2OH^-. \tag{9.33}$$

The electrolyte containing the zincate ion is pumped to an **Electrolyte Managing unit**—an integral part of the fuel cell—where the reaction,

$$Zn(OH)_4^{--} \rightarrow ZnO + H_2O + 2OH^-, \tag{9.34}$$

not only supplies the water consumed at the cathode and the other half of the needed hydroxyls, but also generates zinc oxide, which precipitates out and can be removed from the cell for recovery of the zinc.

The overall fuel cell reaction is

$$Zn + \frac{1}{2}O_2 \rightarrow ZnO. \tag{9.35}$$

The cell is refueled by loading its hopper with zinc pellets. As the zinc is consumed, the hopper drops additional pellets into the reaction area. The ZnO is transferred to a stand-alone **Zinc Regeneration/Refueling System** to be electrochemically reduced to metal and pelletized.

9.6 Fuel Cell Applications

Fuel cells are in a stage of rapid development and are on the verge of achieving maturity. As time goes on and its economic potential becomes well documented, this technology will occupy a growing number of niches.

At present, much of the effort is being concentrated in two distinct areas of application: stationary power plants and automotive power plants.

9.6.1 Stationary Power Plants

Stationary power plants of various types include central utility-operated power plants of large capacity (say, up to 1 GW), dispersed utility-operated power plants (perhaps in the tens of MW sizes), and on-site electrical generators (some 10 to 100 kW). For these applications, fuel cells present the following advantages over conventional heat engines:

1. Absence of noise.
2. Little pollution.
3. Ease of expansion (owing to modular construction).
4. Susceptibility to mass production (again, owing to modularity).
5. Possibility of dispersion of power plants. Owing to the low pollution and low noise, plants can be operated even in residential areas, thus economizing transmission lines.

6. Possibility of using reject heat for ambient heating because fuel cells can be near populated areas where there is a demand for hot water.
7. Possibility of cogeneration, using the high-temperature exhaust gases in some types of plants.
8. Fast response to demand changes.
9. Good efficiency at a fraction of rated power.
10. Extremely good overload characteristics.
11. Small mass/power ratio, in some types of plants.
12. Small volume/power ratio, in some types of plants.
13. Great reliability (potentially).
14. Low maintenance cost (potentially).

Owing to modular construction, plant capacity can be easily expanded as demand grows. Capital investment can be progressive, lessening the financial burden. Not all advantages can be realized simultaneously. Cogeneration can only be achieved with high-temperature fuel cells such as MCFCs and SOFCs. With cogeneration, low noise advantage may be lost.

9.6.2 Automotive Power Plants

Imagine a hypothetical country having a modern industrial base but, for some unexplained reason, lacking completely any automotive industry. Imagine also that someone wants to build a small number of automobile engines of the type that drives present-day cars and that in the United States, Germany, or Japan can be bought for roughly $4000 apiece. Since specialized mass production machines do not exist, the engines have to be built one by one, and their cost would be certainly more than an order of magnitude higher than that of the imported model.

Exactly the same situation prevails, all over the world, regarding fuel cell power plants for cars. The current technology is already quite adequate as far as efficiency, power-to-mass ratio, power-to-volume ratio, and so forth are concerned. Further improvements will be made, but the existing technology is acceptable.

To be sure, a doubling of the present-day lifetime of the cells would be welcome, and a major problem to be solved is how to provide the hydrogen to feed them. But the greatest obstacle for their adoption is definitely one of cost. Cost will come down only when a sufficient number of units can be sold; a sufficient number will only be sold when the cost comes down. This vicious circle is hard to break. One must remember that the low cost of automobiles results from millions upon millions being sold. The retooling for a new model runs to 1 or 2 billion dollars and can only be justified by large sales. Once the vicious circle is broken, a second, very attractive, area of application of fuel cells will be in transportation.

For small and medium-sized (and perhaps even for large) vehicles, the compact SPFCs are nearly ideal. Among the many advantages that can be claimed for them, one must count the extremely low pollution, the high efficiency (guaranteeing fuel economy), and the high power density (permitting compact designs). In addition, their low operating temperature permits rapid start-ups.

Although the life of current SPFC fuel cells is a bit low, their expected eventual long life may lead to an unusual situation. A cell with a plausible 50,000 hour life would, even at only 40 km/h average speed, drive a car some 2 million kilometers. The fuel cell would outlive by far the automobile body. It would, then, make sense to reuse the cell in a new car. Different vehicle models could be designed to operate with one of a small number of standardized fuel cell types. This would have the adverse effect of reducing the market for fuel cells.

Hydrogen is the usual fuel for SPFC. The gas can be used as such (compressed or liquefied as discussed in Chapter 11) or in the form of a hydrogen "carrier" such as methanol (Chapter 10) or a metallic hydride (Chapter 11). Hydrogen carriers can be derived from fossil fuels (Chapter 10) or from biomass (Chapter 13). Fuel cell cars may need energy storage devices (batteries, flywheels, or ultra-capacitors) to provide start-up power and to accumulate the energy recovered from dynamic braking. This stored energy can be used to supplement the fuel cell during fast accelerations. Hence, the needed fuel-cell power might be closer to the average power required by the vehicle than to the peak power. This would cut down the size and cost of the cell stack. Under all circumstances, the total electrically stored energy would be only a small fraction of that of a purely electric car.

The current trend is toward hybrid cars, and soon the trend will be toward plug-in hybrids. This may be the end of the line for a prolonged period of time, especially if a fuel cell hybrid catches on.

9.6.3 Other Applications

The first practical application of fuel cells was in space, where reliability far outweighs cost. The cells work with the hydrogen and the oxygen already available for other uses in the spacecraft and provide valuable drinking water as an output. GE SPFCs started the trend; at the moment AFCs are more popular—they were used in the Apollo program and currently supply energy for the space shuttle and for the International Space Station.

Small submersibles as well as full-scale submarines benefit from the clean operation of fuel cells. The German navy developed 400-kW fuel cells for their submarines. This is half the power of their standard Diesel engines. The fuel cells will probably be used in a hybrid combination with the Diesels. Some military applications benefit from the low heat signature and absence of noise. Once fuel cells become even more compact and light, they may allow the operation of "cold" aircraft invisible to heat-seeking missiles. Fuel cells are important in the propulsion of other naval vessels. Very compact power plants can be built in combination with superconducting motors. The U.S. Department of Defense is actively pursuing this development. Presumably a number of other nations are doing the same.

Micro fuel cells may be used to trickle charge portable electronics. In this application, micro fuel cells may have to face stiff competition from **nuclear batteries** using β^- emitters such as tritium or nickel-63. Such batteries take advantage of the enormous energy density of nuclear materials, (thousands of times more than chemical batteries). One nanotechnological implementation consists of a metallic target that, collecting the emitted electrons, becomes negatively charged. The resulting electrostatic attraction causes a thin silicon cantilever to bend until the target touches the source and discharges. The cantilever snaps back, and the process repeats. The energy of the vibrating cantilever is transformed into electrical energy by piezoelectric converters. Efficiencies of 4% have been demonstrated, but Lal and Blanchard (2004) expect 20% with more advanced configurations.

9.7 The Thermodynamics of Fuel Cells

Some students will profit from rereading the first few pages of Chapter 2 to reacquaint themselves with such concepts as internal energy, enthalpy, and entropy. Here we will repeat the convention (Appendix to Chapter 2) for representing thermodynamic quantities.

Symbology

G, free energy,
H, enthalpy
Q, heat,
S, entropy, and
U, internal energy:

1. Capital letters indicate the quantity associated with an arbitrary amount of matter or energy.
2. Lowercase letters indicate the quantity per unit. A subscript may be used to indicate the species being considered. For example, the free energy per kilomole of H_2 will be represented by $\overline{g}_{H_2}$.

g = free energy per kilogram.
$\overline{g}$ = free energy per kilomole.
g^* = free energy per kilogram, at 1 atmosphere pressure.
$\overline{g}^*$ = free energy per kilomole, at 1 atmosphere pressure.
$\overline{g}_f$ = free energy of formation per kilomole.
$\overline{g}_f^\circ$ = free energy of formation per kilomole, at 298 K, 1 atmosphere, i.e., at RTP (Standard Free Energy of Formation).

9.7.1 Heat of Combustion

Let us return to the reaction

$$2H_2 + O_2 \rightarrow 2H_2O \tag{[9.4]}$$

(in Section 9.2) in which hydrogen and oxygen combine to form water. In rearranging four hydrogen atoms and two oxygen atoms to form two molecules of water, some energy is left over. Although the force that binds atoms into molecules is electric in nature, when H_2 reacts directly with O_2, only heat results because electrons and ions collide and their energy is randomized.

Assume that a measured amount, μ, of hydrogen is introduced into a constant-pressure calorimeter and is made to react with sufficient oxygen. A certain amount of heat, Q, is released. The ratio, $Q/\mu = \overline{h}_{comb}$, is the **heat of combustion** of hydrogen in an atmosphere of oxygen (joules/kmole). The exact amount of heat released depends on the temperature of both the reactants and the product and on the state of the product. If the water produced is liquid, the heat of combustion is larger than when water is in the form of gas because the heat released includes the heat of condensation of water. As explained in Chapter 3, all fuels containing hydrogen have two different heats of combustion: the **higher heat of combustion** in case of liquid water and the **lower heat of combustion** in case of water vapor.

Since the reaction occurs at constant pressure, the heat released equals the change in enthalpy, ΔH, of the system (see *enthalpy* in Chapter 2).

$$Q = \Delta H \tag{9.36}$$

Q is taken as positive if it is added to the system as happens in heat input to a heat engine. In **exothermic** reactions, Q is negative and so is ΔH.

Notice the sign convention used here. To conform with Chapter 2, energies introduced into a system, $\sum W_{in}$, are taken as positive; energies rejected by the system, $\sum W_{out}$, are negative. Consequently, the energy balance reads

$$\sum W_{in} + \sum W_{out} = 0. \tag{9.37}$$

Enthalpy, as any other form of energy, has no absolute value; only changes can be measured. We can therefore arbitrarily assume any value for the enthalpy of the reactants. By convention, *at 298.15 K, the enthalpy of all elements in their natural state in this planet is taken as zero.* This means that, at 298.15 K, the enthalpies of H_2 and O_2 are zero, but those of H and O are not.

The difference between the enthalpy of a product and those of its elements is the **enthalpy of formation**, ΔHf, of the product. If both reactants and products are at the **reference temperature and pressure**

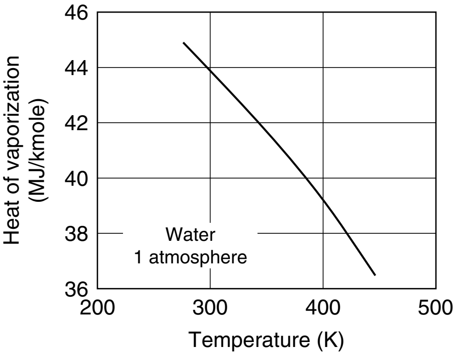
(RTP),[†] then the enthalpy of formation is called the **standard** enthalpy of formation, ΔHf°.

From the calorimeter experiment, for water (expressing the enthalpies in a per kilomole basis),

$$\text{H}_2\text{O (g)}: \quad \overline{h}^\circ_{f_g} = -241.8 \;\; \text{MJ/kmol}, \tag{9.38}$$

$$\text{H}_2\text{O } (\ell): \quad \overline{h}^\circ_{f_\ell} = -285.9 \;\; \text{MJ/kmol}. \tag{9.39}$$

The subscripts g and ℓ indicate the state of the product water. The difference of −44.1 MJ/kmol between the enthalpies of formation of liquid and gaseous water is the **latent heat of condensation**, $\overline{h}_{con}$, of water. Clearly, $\overline{h}_{con} = -\overline{h}_{vap}$, where $\overline{h}_{vap}$ is the **latent heat of vaporization**.



For coarse estimates, the latent heat of vaporization can be taken as constant. However, it is, in fact, temperature dependent, falling somewhat with rising temperature as can be seen in the figure.

9.7.2 Free Energy

If the reaction under consideration occurs not in a calorimeter but in an ideal fuel cell, part (but not all) of the energy will be released in the form of electricity. It is important to investigate how much of the ΔH of the reaction will be converted into electric energy and to understand why some of the energy must take the form of heat even in the ideal case.

Let V_{rev} be the voltage produced by the cell. Each kilomole of water (or of any other substance) contains N_0 molecules, where $N_0 = 6.022 \times 10^{26}$

[†] **Standard temperature and pressure** (STP), frequently used by chemists, corresponds to 1 atmosphere and 0 C (273.15 K). The *CRC Handbook of Chemistry and Physics*, however, lists the standard thermodynamic properties at 1 atmosphere and 298.15 K, which we will call RTP.

is **Avogadro's number**.[†] From Equation 12, Subsection 9.4.2, describing the cathode reaction of the cell,

$$4e^- + 4H^+ + O_2 \rightarrow 2H_2O \qquad [9.12]$$

we see that for each molecule of water, 2 electrons (charge q) circulate in the load. The energy delivered to the load is the product of the charge, $2qN_0$, times the voltage, V_{rev}. More generally, the electric energy produced per kilomole of product by a reversible fuel cell is

$$W_e = n_e q N_0 V_{rev} \equiv n_e F V_{rev}, \qquad (9.40)$$

where

n_e = number of kmoles of electrons released per kilomole of products.
q = charge of the electron = 1.602×10^{-19} coulombs.
N_0 = Avogadro's number in meter kilogram second (MKS) systems.
$F \equiv qN_0 \equiv$ **Faraday** constant $= 1.602 \times 10^{-19} \times 6.022 \times 10^{26}$
$= 96.47 \times 10^6$ Coulombs/kmole = charge of 1 kmole of electrons.
V_{rev} = voltage.

Consider a hypothetical experiment in which the open-circuit voltage of a reversible fuel cell is accurately measured. The voltage at RTP would be 1.185 V.[††] The voltage delivered by a reversible fuel cell is called the **reversible voltage** and is designated by V_{rev}, as was done above.[†††] Thus, the electric energy produced by a reversible fuel cell of this type is

$$|W_e| = 2 \times 96.47 \times 10^6 \times 1.185 = 228.6 \text{ MJ/kmole}. \qquad (9.41)$$

The electric energy delivered to a load by a reversible fuel cell is called the **free energy change** owing to the reaction (designated ΔG). Usually, if there is a single product, ΔG is given per kilomole of product and is represented by $\overline{g}_f^\circ$ (if at RTP). Again, since the cell *delivers* electric energy, the free energy change is negative, conforming to the convention for the sign of the enthalpy change of the reaction. As in the case of enthalpies, by convention, at RTP, *the free energy of all elements in their natural state in this planet is taken as zero.*

[†] Observe that, owing to our use of kilomoles instead of moles, Avogadro's number is three orders of magnitude larger than the value usually listed.

[††] This assumes that the product water is created as a gas. If the water is created as a liquid, the reversible voltage will be 1.229 V, and the free energy is 237.1 MJ/kmole.

[†††] Owing to irreversibilities in practical cells, such a measurement cannot be carried out accurately. However, the reversible voltage can be estimated by connecting a voltage generator to the cell and observing its voltage-versus-current characteristic as the applied voltage is varied. If this voltage is sufficiently high, current will be driven into the cell, and if it is low, current will be delivered by the cell to the generator. The characteristic should be symmetrical around the reversible voltage.

In most cases (but not always—see Problem 9.3), $|\Delta G| < |\Delta H|$. Thus the energy from the reaction usually exceeds the electric energy delivered to a load even in an ideal reversible cell. The excess energy, $\Delta H - \Delta G$, must appear as heat. Consider the entropies involved. Each substance has a certain entropy that depends on its state. Entropies are tabulated in, among other works, the *Handbook of Chemistry and Physics from* CRC Press. In the reaction we are examining, the absolute entropies, at RTP, are:

$$\begin{aligned}
&H_2(g): &&\bar{s}^\circ = 130.6\ \mathrm{kJ\,K^{-1}kmol^{-1}},\\
&O_2(g): &&\bar{s}^\circ = 205.0\ \mathrm{kJ\,K^{-1}kmol^{-1}},\\
&H_2O\ (g): &&\bar{s}^\circ = 188.7\ \mathrm{kJ\,K^{-1}kmol^{-1}}.
\end{aligned}$$

When 1 kilomole of water is formed, 1 kilomole of H_2 and 0.5 kilomole of O_2 disappear, and so do the corresponding entropies: a total of $130.6 + 205.0/2 = 233.1$ kJ/K disappear. This is, in part, compensated by the appearance of 188.7 kJ/K corresponding to the entropy of the water formed. The matter balance leads to an entropy change of $188.7 - 233.1 = -44.4$ kJ/K. In a closed system, the entropy cannot decrease (second law of thermodynamics), at best—under reversible conditions—its change is zero. Consequently, an amount of entropy, $\Delta S = Q/T$ must appear as heat:

$$Q = T\Delta S = 298 \times (-44.4 \times 10^3) = -13.2\ \text{MJ/kmol}. \tag{9.42}$$

This amount of heat must come from the enthalpy change of the reaction, leaving $-241.8 - (-13.2) = -228.6$ MJ/kmol as electricity.

Chemical energy can be thought of as consisting of two parts: an entropy-free part, called **free energy**, that can entirely be converted to electricity and a part that must appear as heat. The free energy,[†] G, is the enthalpy, H, minus the energy, TS, that must appear as heat:

$$G = H - TS \tag{9.43}$$

and

$$\Delta G = \Delta H - \Delta(TS). \tag{9.44}$$

In isothermal cases (as in the example above),

$$\Delta G = \Delta H - T\Delta S. \tag{9.45}$$

[†] The idea of free energy was first proposed by the American physicist Josiah Willard Gibbs (1839–1903), hence the symbol G.

The electric energy, W_e, delivered by the reversible fuel cell is ΔG:

$$\Delta G = -n_e q N_0 N |V_{rev}|. \tag{9.46}$$

where N is the number of kilomoles of water produced. Note that since ΔG is removed from the cell, it must be < 0. Per kilomole of water,

$$\overline{g}_f = -n_e q N_0 |V_{rev}|. \tag{9.47}$$

Table 9.2 lists some of the thermodynamic quantities used in this section.

The dual of a fuel cell is an electrolyzer (Chapter 10). This is illustrated in Figure 9.21. A fuel cell may use hydrogen and oxygen, generating electricity and producing water and heat. The electrolyzer consumes water and electricity, producing hydrogen and oxygen. In the ideal case, the electrolyzer absorbs heat from the environment, acting as a heat pump. If there is insufficient heat flow from the environment to the electrolyzer, the electrolyzer will cool down.

For a given amount of gas handled, the electric energy generated by the reversible fuel cell is precisely the same as that required by the reversible electrolyzer, and the heat produced by the reversible fuel cell is precisely the same as that absorbed by the electrolyzer. This amount of heat is reversible.

Table 9.2 Some Thermodynamic Values

	$\Delta\overline{h}f^\circ$	$\Delta\overline{g}f^\circ$	$\overline{s}^\circ$
	(MJ/kmole)	**(MJ/kmole)**	**(kJ K^{-1} kmole^{-1})**
H_2O (g)	−241.8	−228.6	188.7
H_2O (l)	−285.9	−237.2	70.0
H_2 (g)	0	0	130.6
O_2 (g)	0	0	205.0

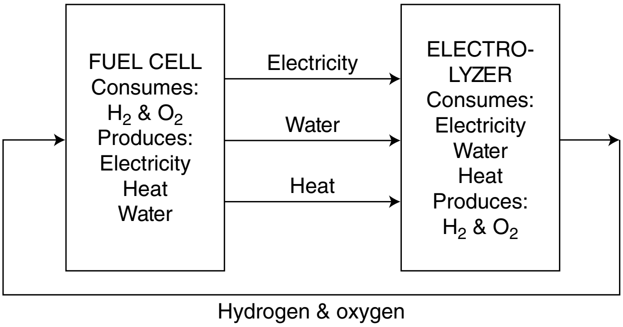


Figure 9.21 An ideal fuel cell and its dual, the ideal electrolyzer, act as a reversible system. The inputs of one are precisely the same as the output of the other. For this to be true, the input to the electrolyzer must be water vapor.

Clearly, the reversibility is destroyed if the system has losses. A lossy (read, practical) fuel cell generates more heat than $T\Delta S$, while a lossy electrolyzer will generate heat, which may (and frequently does) exceed the thermodynamically absorbed heat. Nevertheless, some realizable electrolyzers may operate with sufficient efficiency to actually cool down during operation.

9.7.3 Efficiency of Reversible Fuel Cells

The efficiency of a fuel cell is the ratio of the electric energy generated to the enthalpy change of the chemical reaction involved. In reversible cells,

$$\eta_{rev} = \frac{\Delta G}{\Delta H}. \tag{9.48}$$

It is of interest to examine how this efficiency depends on the temperature and the pressure of reactants and products, or, in other words, how the enthalpy and free energy changes of the reaction depend on such variables.

Figure 9.22 (left) depicts the manner in which ΔH and ΔG vary with temperature and (right) ΔG varies with pressure. The data assume that both reactants, H_2 and O_2, and the product, H_2O, are at identical pressures and identical temperatures.

Increasing the temperature while maintaining the pressure constant has a small effect on ΔH but appreciably reduces the available free energy, ΔG (the electric energy generated). Thus, for reversible cells, the higher the temperature, the lower the efficiency. Exactly the opposite happens

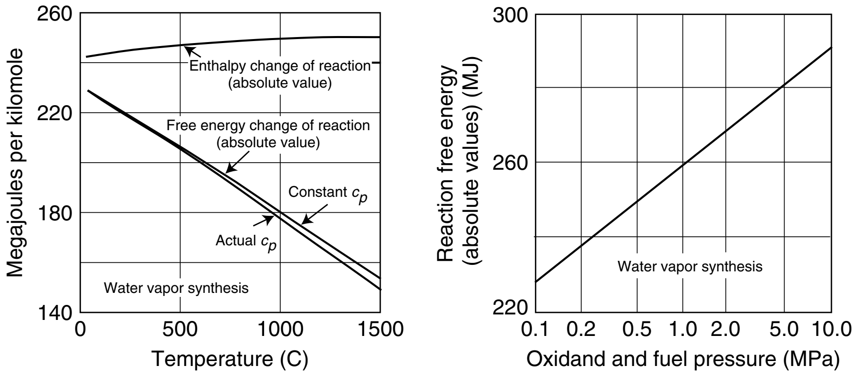


Figure 9.22 (Left) At constant pressure (0.1 MPa), increasing equally the temperature of reactants and product increases slightly the absolute value of the enthalpy change of the reaction, but reduces substantially the corresponding free energy. (Right) At constant temperature (298 K), increasing equally the pressure of reactants and product increases substantially the absolute value of the free energy change of the reaction.

with practical cells. The reason is that the improvement in the cell kinetics with temperature more than offsets the free energy loss. This point will be discussed in more detail further on.

We will now derive the dependence of the enthalpy and the free energy changes of a reaction on temperature and pressure.

9.7.4 Effects of Pressure and Temperature on the Enthalpy and Free Energy Changes of a Reaction

9.7.4.1 Enthalpy Dependence on Temperature

Let ΔH be the enthalpy change owing to a chemical reaction.

By definition, ΔH is equal to the sum of the enthalpies of the products of a reaction minus the sum of the enthalpies of all reactants. Thus, $\Delta H =$ *enthalpy of products* $-$ *enthalpy of reactants.* When there is a single product, then the enthalpy change of the reaction can be expressed per kilomole of product by dividing ΔH by the number of kilomoles of product created. An equivalent definition holds for the free energies.

In mathematical shorthand,

$$\Delta H = \sum n_{P_i} H_{P_i} - \sum n_{R_i} H_{R_i}. \tag{9.49}$$

where,

$n_{P_i} =$ number of kilomoles of the ith product,
$n_{R_i} =$ number of kilomoles of the ith reactant,
$H_{P_i} =$ enthalpy of the ith product, and
$H_{R_i} =$ enthalpy of the ith reactant.

For the water synthesis reaction,

$$\Delta H = 1 \times H_{\mathrm{H_2O}} - 1 \times H_{\mathrm{H_2}} - \frac{1}{2} \times H_{\mathrm{O_2}}, \tag{9.50}$$

where $H_{\mathrm{H_2O}}$ is the enthalpy of water, and so on.

Of course, at RTP, for water vapor, $H_{\mathrm{H_2O}} = -241.8$ MJ/kilomole, and $H_{\mathrm{H_2}}$ and $H_{\mathrm{O_2}}$ are both zero by choice.[†] Hence, $\Delta H = -241.8$ MJ/kilomole.

If the value of the enthalpy, H_0 of a given substance (whether product or reactant), is known at a given temperature, T_0, then at a different temperature, T, the enthalpy is

$$H = H_0 + \int_{T_0}^{T} c_p dT, \tag{9.51}$$

where c_p is the specific heat at constant pressure.

[†] Remember that, by convention, the enthalpy of all elements in their natural state under normal conditions on Earth is taken as zero.

c_p is somewhat temperature dependent, and its value for each substance can be found either from tables or from mathematical regressions derived from such tables. For rough estimates, c_p, can be taken as constant. In this latter case,

$$H = H_0 + c_p \Delta T, \tag{9.51a}$$

where $\Delta T \equiv T - T_0$.

Replacing the various enthalpies in Equation 9.49 by their values as expressed in Equation 9.51,

$$\begin{aligned}\Delta H =& \sum n_{P_i} H_{0_{P_i}} + \sum n_{P_i} \int_{T_{0_{P_i}}}^{T_{P_i}} c_{p_{P_i}} dT - \sum n_{R_i} H_{0_{R_i}} - \sum n_{R_i} \int_{T_{0_{R_i}}}^{T_{R_i}} c_{p_{R_i}} dT \\ &= \Delta H_0 + \sum n_{P_i} \int_{T_{0_{P_i}}}^{T_{P_i}} c_{p_{P_i}} dT - \sum n_{R_i} \int_{T_{0_{R_i}}}^{T_{R_i}} c_{p_{R_i}} dT. \end{aligned} \tag{9.52}$$

For a simple estimate of the changes in ΔH, one can use the constant c_p formula:

$$\Delta H = \Delta H_0 + \sum n_{P_i} c_{p_{P_i}} (T_{P_i} - T_{0_{P_i}}) - \sum n_{R_i} c_{p_{R_i}} (T_{R_i} - T_{0_{R_i}}). \tag{9.52a}$$

Example 1

The standard enthalpy of formation of water vapor is −241.8 MJ/kmoles. What is the enthalpy of formation when both reactants and product are at 500 K?

We need to know the specific heats at constant pressure of H_2, O_2, and $H_2O(g)$. To obtain an accurate answer, one must use Equation 9.52 together with tabulated values of the specific heat as a function of temperature. An approximate answer can be obtained from Equation 9.52a using constant values of the specific heats. We shall do the latter.

We saw in Chapter 2 that, if one can guess the number of degrees of freedom, ν, of a molecule, one can estimate the specific heat by using the formula

$$c_p = R\left(1 + \frac{\nu}{2}\right).$$

The advantage of this procedure is that it is easier to remember ν than c_p. For diatomic gases, ν can be taken as 5, yielding a $c_p = 29.1$ kJ kmole^{-1}K^{-1}, and for water vapor, ν can be taken as 7, yielding $c_p = 37.4$ kJ kmole^{-1}K^{-1}.

(*Continues*)

(*Continued*)

Using Equation 9.52a (in this problem, all ΔT are the same: 500 K −298 K),

$$\Delta H = -241.8 + (500 - 298)\left[0.0374 - (0.0291 + \frac{1}{2} \times 0.0291)\right]$$
$$= -243.1 \text{ MJ per kmole of water vapor.}$$

Since the enthalpy changes only little with temperature, these approximate results are close to the correct value of −243.7 MJ/kmole obtained through the use of Equation 9.52.

9.7.4.2 Enthalpy Dependence on Pressure

In Section 2.12, we defined enthalpy as the sum of the internal energy, U, and the pressure-volume work, pV:

$$H \equiv U + pV. \tag{9.53}$$

The internal energy of the gas is the energy stored in its molecules. Such storage can take the form of excitation, ionization, and so on. However, in this chapter, we will limit ourselves to the internal energy stored as kinetic energy of the molecules, a quantity measured by the temperature of the gas.

Changing the pressure at constant temperature does not change the average energy of the molecules, and since the mass of gas (the number of molecules) is constant, the total internal energy remains unaltered—that is, U remains constant when p is changed provided T is unchanged.

We have $pV = \mu RT$. At constant temperature, pV must also be constant.

Thus, neither U nor pV changes when the pressure is altered isothermally. Consequently, H does not depend on pressure provided the temperature, and the mass of the gas is unaltered.

9.7.4.3 Free Energy Dependence on Temperature

The free energy is

$$G = H - TS. \tag{9.54}$$

The behavior of H as a function of temperature was discussed previously. We must now investigate the behavior of the entropy, S.

From Chapter 2, we have the relationship (for isobaric processes and per kilomole of gas),

$$dS = c_p \frac{dT}{T}, \tag{9.55}$$

which integrates to

$$S = S_0 + \int_{T_0}^{T} c_p \frac{dT}{T}. \tag{9.56}$$

The change in free energy when the temperature is changed under constant pressure is

$$\begin{aligned} G - G_0 &= H - H_0 - (TS - T_0 S_0) \\ &= \int_{T_0}^{T} c_p dT - \left(TS_0 + T \int_{T_0}^{T} c_p \frac{dT}{T} - T_0 S_0 \right) \\ &= \int_{T_0}^{T} c_p dT - T \int_{T_0}^{T} c_p \frac{dT}{T} - S_0 \Delta T. \end{aligned} \tag{9.57}$$

Given a table of values for c_p as a function of T, the change, $G - G_0$, in the free energy can be numerically calculated from Equation 9.57 and the ΔG of reaction can be obtained from

$$\Delta G = \sum n_{P_i} G_{P_i} - \sum n_{R_i} G_{R_i}, \tag{9.58}$$

an equation equivalent to Equation 9.49 for ΔH.

For the case when c_p is assumed constant, Equation 9.57 reduces to

$$G - G_0 = (c_p - S_0)\Delta T - T c_p \ln \frac{T}{T_0}. \tag{9.57a}$$

Example 2

Estimate the free energy of the $H_2(g) + \frac{1}{2}O_2(g) \rightarrow H_2O\ (g)$ reaction at standard pressure and 500 K, using constant c_p.

The necessary values are

	Entropy **kJ K^{-1}kmole^{-1}** **(at RTP)**	**Specific Heat** **kJ K^{-1}kmole^{-1}**
H_2 (g)	130.6	29.1
O_2 (g)	205.0	29.1
H_2O (g)	188.7	37.4

Since the product and all reactants are at the same temperature, $\Delta T = 500 - 298 = 202$ for all gases. Let us calculate individually their change in free energy remembering that elements, in their natural state at RTP, have zero free energy.

(*Continues*)

(*Continued*)

From Equation 9.57a,

$$G_{\mathrm{H_2}} = 0 + (29.1 \times 10^3 - 130.6 \times 10^3) \times 202 - 500 \times 29.1 \times 10^3 \times \ln \frac{500}{298} = -28.03 \times 10^6 \text{ J/kmole},$$

$$G_{\mathrm{O_2}} = 0 + (29.1 \times 10^3 - 205.0 \times 10^3) \times 202 - 500 \times 29.1 \times 10^3 \times \ln \frac{500}{298} = -43.06 \times 10^6 \text{ J/kmole},$$

$$G_{\mathrm{H_2O}} = -228.6 \times 10^6 + (37.4 \times 10^3 - 188.7 \times 10^3) \times 202 - 500 \times 37.4 \times 10^3 \times \ln \frac{500}{298} = -268.8 \times 10^6 \text{ J/kmole}.$$

Thus, the ΔG of the reaction is

$$\Delta G = G_{\mathrm{H_2O}} - G_{\mathrm{H_2}} - \frac{1}{2} G_{\mathrm{O_2}} = -268.8 - (-28.0) - \frac{1}{2}(-43.1) = -219.3 \,\text{MJ/kmole}.$$

This estimate came fortuitously close to the correct value of −219.4 MJ/kmole. If we had calculated ΔG at, say, 2000 K, we could be making a substantial error.

Nevertheless, the assumption of constant c_p leads to estimates that indicate the general manner in which the free energy depends on temperature.

More accurate values can be obtained from numerical integration of Equation 9.57. The values of c_p for H_2, O_2, and H_2O can be read from experimentally determined tables (reproduced below from Haberman and John 1989).

The data of Table 9.3 are displayed in the plots of Figures 9.23 through 9.25. They show that the specific heats at constant pressure, c_p, and the γ of the three gases of interest, far from being temperature independent as suggested by simple theory (see Chapter 2), do vary substantially.

The value of c_p derived from a guess of the number of degrees of freedom are indicated by the horizontal dotted line in each figure. For hydrogen and oxygen, two diatomic molecules, a reasonable number of degrees of freedom would be $\nu = 5$, which leads to $c_p = 29.1$ kJ K^{-1} per kilomole. This matches the actual value for hydrogen at temperatures between 350 K and 600 K. At higher temperatures, the c_p of this gas seems to be

Table 9.3 Specific Heats at Constant Pressure and Gammas

		H_2	H_2	O_2	O_2	H_2O	H_2O
T	T	c_p	γ	c_p	γ	c_p	γ
(C)	(K)	(kJ/K)		(kJ/K)		(kJ/K)	
		per kmole		per kmole		per kmole	
−50	223.18	27.620	1.426	29.152	1.399	33.318	1.333
0	273.18	28.380	1.410	29.280	1.397	33.336	1.332
25	298.18	28.560	1.406	29.392	1.406	33.489	1.330
50	323.18	28.740	1.402	29.504	1.403	33.642	1.328
100	373.18	28.920	1.399	29.888	1.386	34.020	1.323
150	423.18	28.980	1.398	30.336	1.378	34.434	1.318
200	473.18	29.020	1.397	30.816	1.369	34.902	1.312
226.8	500	29.031	1.397	31.090	1.365	35.172	1.309
250	523.18	29.040	1.397	31.328	1.361	35.406	1.307
300	573.18	29.080	1.396	31.840	1.354	35.946	1.301
350	623.18	29.120	1.395	32.320	1.346	36.522	1.295
400	673.18	29.180	1.394	32.768	1.340	37.098	1.288
450	723.18	29.240	1.393	33.184	1.334	37.710	1.283
500	773.18	29.340	1.391	33.536	1.329	38.322	1.277
550	823.18	29.440	1.389	33.888	1.325	38.952	1.271
600	873.18	29.560	1.387	34.208	1.321	39.564	1.266
650	923.18	29.720	1.384	34.496	1.318	40.194	1.261
700	973.18	29.880	1.381	34.752	1.315	40.788	1.256
750	1023.2	30.040	1.378	34.976	1.312	41.382	1.251
800	1073.2	30.240	1.375	35.200	1.309	41.958	1.247
850	1123.2	30.420	1.372	35.392	1.307	42.642	1.242
900	1173.2	30.640	1.369	35.584	1.305	43.326	1.237
950	1223.2	30.840	1.365	35.744	1.303	43.920	1.233
1000	1273.2	31.060	1.362	35.904	1.301	44.514	1.229
1050	1323.2	31.280	1.358	36.064	1.300	45.072	1.226
1100	1373.2	31.500	1.355	36.224	1.298	45.630	1.223
1150	1423.2	31.720	1.351	36.352	1.296	46.170	1.219
1200	1473.2	31.940	1.348	36.480	1.295	46.674	1.216
1250	1523.2	32.140	1.345	36.608	1.294	47.178	1.214
1300	1573.2	32.360	1.342	36.736	1.292	47.664	1.211
1350	1623.2	32.560	1.339	36.864	1.291	48.114	1.209
1400	1673.2	32.760	1.337	36.992	1.290	48.564	1.206
1450	1723.2	32.960	1.344	37.120	1.289	48.978	1.204
1500	1773.2	33.160	1.331	37.248	1.287	49.392	1.202

heading toward the 37.4 kJ K^{-1} per kilomole that correspond to 7 degrees of freedom. For a somewhat more detailed discussion of these changing degrees of freedom, refer to Chapter 2.

Water, with its presumed 7 degrees of freedom, should have $c_p = 37.4$ kJ K^{-1} per kilomole. This is actually the correct value at some 700 K.

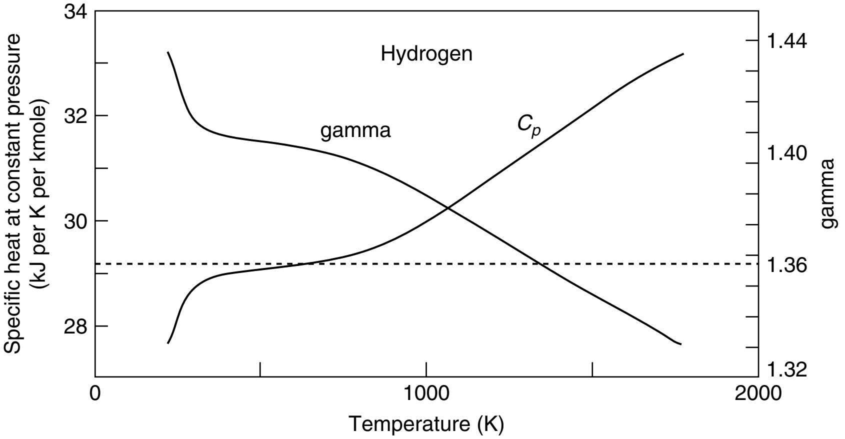


Figure 9.23 Specific heat and gamma of hydrogen.

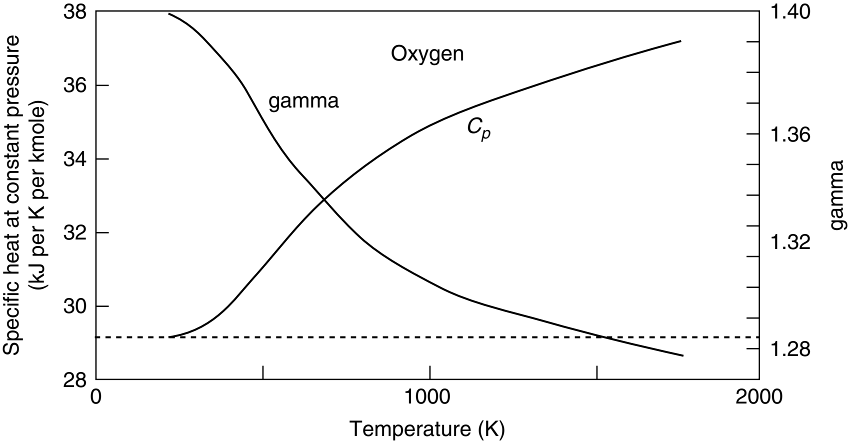


Figure 9.24 Specific heat and gamma of oxygen.

However, the c_p of water, like that of most gases, varies fairly rapidly with temperature.

We fitted a fifth-order polynomial to the data in Table 9.3, so that the values can be calculated as a function of T with reasonable accuracy:

$$c_p = a + bT + cT^2 + dT^3 + eT^4 + fT^5 \tag{9.59}$$

where the constants a through f are given in Table 9.4.

It should be noticed that these regressions must be used only in the 220 K $< T <$ 1800 K interval. Outside this range, the errors become, in some cases, unacceptably large.

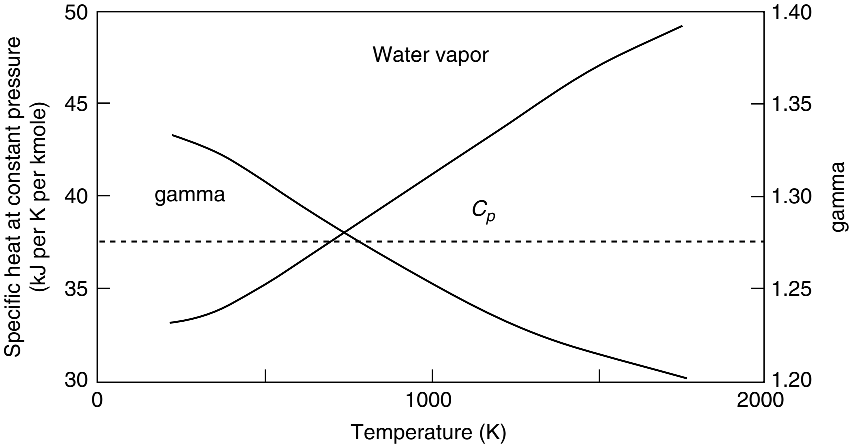


Figure 9.25 Specific heat and gamma of water vapor.

Table 9.4 Coefficients of the Polynomial Used to Calculate the Specific Heats at Constant Pressure 100,000 Pa (1 atmosphere)

Gas	a	b	c	d	e	f
H_2	22.737	37.693E-3	−85.085E-6	89.807E-9	−42.908E-12	7.6821E-15
O_2	30.737	−19.954E-3	72.554E-6	−80.005E-9	38.443E-12	−6.8611E-15
H_2O	32.262	1.2532E-3	11.285E-6	−3.7103E-9	—	—

9.7.4.4 Free Energy Dependence on Pressure

The free energy, G, is defined as

$$G = H - TS. \tag{9.60}$$

We saw that the enthalpy does not change when the pressure is altered isothermally. Thus, isothermally, pressure can only alter the free energy through its effect on the entropy, S:

$$\Delta G = -T\Delta S. \tag{9.61}$$

When an amount, Q, of heat is added to a system at constant temperature, the entropy increases by Q/T:

$$\Delta S = \frac{\Delta Q}{T}. \tag{9.62}$$

From the first law of thermodynamics,

$$\Delta U = \Delta Q - \Delta W. \tag{9.63}$$

But we saw previously that in an isothermal compression, $\Delta U = 0$, hence, $Q = W$. However, the work done per kilomole of gas isothermally compressed (see Section 2.10) is

$$W = RT \ln \frac{p_1}{p_0}. \tag{9.64}$$

Thus,

$$-\Delta G = T\Delta S = \Delta Q = RT \ln \frac{p_1}{p_0}. \tag{9.65}$$

Consequently, the energy of isothermal compression of a gas is entirely free energy. This is an important effect. It is possible to change the efficiency (and the voltage) of a fuel cell by changing the pressure of products and reactants.

Example 3

A reversible fuel cell, when fed hydrogen and oxygen at RTP, delivers a voltage of 1.185 V. Calculate the voltage delivered by the same cell if air (at RTP) replaces the oxygen.

Air contains roughly 20% of oxygen. Thus, the partial pressure of this gas is 0.2 atmosphere, a 5:1 decompression relative to the pure oxygen case.

The energy of isothermal decompression is

$$W_{decompr.} = \frac{1}{2} RT \ln \frac{1}{5} = \frac{1}{2} \times 8314 \times 298 \times \ln 0.2 = -2 \times 10^6 \text{ J/kmole}.$$

The factor, $\frac{1}{2}$, results from the stoichiometric proportion of one-half kilomole of oxygen per kilomole of water. This energy must be *subtracted* from the ΔG of the reaction.

$$\Delta G = -228.6 - (-2) = -226.6 \,\text{MJ/kmole}.$$

The voltage is now

$$V_{rev} = \frac{|\Delta G|}{n_e q N_0} = \frac{226.6}{2 \times 1.60 \times 10^{-19} \times 6.02 \times 10^{26}} = 1.174 \,\text{V}.$$

9.7.4.5 The Nernst Equation

One can generalize Equation 9.65 If there is more than one reactant and more than one product,

$$a\text{A} + b\text{B} + \ldots \rightarrow c\text{C} + d\text{D} + \ldots, \tag{9.66}$$

then the energy of isothermal compression from 1 atmos to p_i atmos is

$$\begin{aligned} W &= aRT \ln p_A + bRT \ln p_B + \cdots - cRT \ln p_C - dRT \ln p_D - \cdots \\ &= RT \ln \left(\frac{p_A^a p_B^b \cdots}{p_C^c p_D^d \cdots} \right), \end{aligned} \tag{9.67}$$

and the free energy change owing to the reaction is

$$\Delta G = \Delta G_0 - RT \ln \left(\frac{p_A^a p_B^b \cdots}{p_C^c p_D^d \cdots} \right). \tag{9.68}$$

Dividing by nF, we obtain the reversible voltage owing to the reaction,

$$V_{rev} = V_{rev_0} + \frac{RT}{nF} \ln \left(\frac{p_A^a p_B^b \cdots}{p_C^c p_D^d \cdots} \right). \tag{9.69}$$

This is the **Nernst equation** for gases. It can be generalized to other phases by replacing the pressure by the **activity** of the species. Solids and pure liquids, being incompressible, are not affected by pressure. For these species, the pressure in the formula must be replaced by the number 1; that is, the activity is 1.

In the isothermal compression with which we are dealing here, we are taking the ratio of two pressures, which, using the ideal gas law, is

$$\frac{p_1}{p_0} = \frac{\frac{n_1}{V_1} RT}{\frac{n_0}{V_0} RT} = \frac{\frac{n_1}{V_1}}{\frac{n_0}{V_0}} \tag{9.70}$$

because T is constant. Thus the ratio of pressures is equal to the ratio of concentrations. This means that for solutions the activity is equal to the concentration, n/V.

9.7.4.6 Voltage Dependence on Temperature

We have derived expressions that show how the free energy depends on both pressure and temperature. The evaluation of these expressions requires numerical integration and a look-up table of values of c_p as a function of temperature. We will now derive an equation that will give us directly the temperature dependence of the voltage, V_{rev}, of an ideal fuel cell in a rigorous manner for the constant pressure case.

By definition,

$$G = H - TS \tag{9.71}$$

and

$$H = U + pV \tag{9.72}$$

hence

$$G = U + pV - TS, \tag{9.73}$$

from which

$$dG = dU + pdV + Vdp - TdS - SdT. \tag{9.74}$$

From the combined laws of thermodynamics,

$$dU = TdS - pdV, \tag{9.75}$$

hence

$$dG = Vdp - SdT. \tag{9.76}$$

G is a function of the independent variables p and T. Thus, formally,

$$dG = \left(\frac{\partial G}{\partial p}\right)_T dp + \left(\frac{\partial G}{\partial T}\right)_p dT. \tag{9.77}$$

Comparing the last two equations, one can see that

$$\left(\frac{\partial G}{\partial T}\right)_p = -S. \tag{9.78}$$

From here on, V represents voltage, not volume, as before.

$$\sum n_{P_i} n_e q N_0 V = -\Delta G = -\left(\sum n_{P_i}\ G_{P_i} - \sum n_{R_i}\ G_{R_i}\right), \tag{9.79}$$

$$\begin{aligned}\sum n_{P_i} n_e q N_0 \left(\frac{\partial V}{\partial T}\right)_p &= -\left(\sum n_{P_i} \frac{\partial G_{P_i}}{\partial T} - \sum n_{R_i} \frac{\partial G_{R_i}}{\partial T}\right) \\ &= \sum n_{P_i}\ S_{P_i} - \sum n_{R_i}\ S_{R_i} \\ &= \frac{\sum n_{P_i}\ T\ S_{P_i} - \sum n_{R_i}\ T\ S_{R_i}}{T}. \end{aligned} \tag{9.80}$$

But $TS = H - G$, hence,

$$\begin{aligned}&\sum n_{P_i} n_e q N_0 \left(\frac{\partial V}{\partial T}\right)_p \\ &\quad = \frac{\sum n_{P_i}\ H_{P_i} - \sum n_{R_i}\ H_{R_i} - \sum n_{P_i}\ G_{P_i} + \sum n_{R_i}\ G_{R_i}}{T}. \end{aligned} \tag{9.81}$$

Therefore,

$$\left(\frac{\partial V}{\partial T}\right)_p = \frac{V + \Delta H / (\sum n_{P_i} n_e N_0 q)}{T}, \tag{9.82}$$

where $\Delta H/(\sum n_{P_i} n_e N_0 q)$ is the voltage the cell would have if all the enthalpy change of the reaction were transformed into electrical energy. Let us call this the **enthalpy voltage**.

For an H_2/O_2 fuel cell producing water vapor,

$$\frac{\Delta H}{\sum n_{P_i} n_e N_0 q} = \frac{-241.8 \times 10^6}{1 \times 2 \times 6.022 \times 10^{26} \times 1.6 \times 10^{-19}} = -1.225 \text{ V},$$

where we set $\sum n_{P_i} = 1$ because the value of ΔH used is the one for a single kilomole of water.

$$\left(\frac{\partial V}{\partial T}\right)_p = \frac{1.185 - 1.255}{298} = -2.3 \times 10^{-4} \text{ V/K}.$$

9.8 Performance of Real Fuel Cells

In examining the performance of real fuel cells, we must inquire:

1. What current can the cell deliver?
2. What is the efficiency of the cell?
3. What are the current/voltage characteristics?
4. What is the heat balance?
5. How can the excess heat be removed?

9.8.1 Current Delivered by a Fuel Cell

If $\dot{N}$ is the rate (in kilomoles/sec) at which the product is generated (water, in case of hydrogen/oxygen cells) and n_e is the number of electrons per molecule of product (2, for hydrogen/oxygen cells), then the rate at which electrons are delivered by the cell to the load is $n_e N_0 \dot{N}$. Consequently, the current is

$$I = q n_e N_0 \dot{N}. \tag{9.83}$$

One defines a **current efficiency** as the ratio of the actual load current, I_L, to the theoretical current calculated above. In many cases, one can safely assume 100% current efficiency.

9.8.2 Efficiency of Practical Fuel Cells

It was shown that the theoretical efficiency of a reversible fuel cell is

$$\eta_{rev} = \frac{\Delta G}{\Delta H}. \tag{9.84}$$

One has the choice of using for ΔH either the higher or the lower heat of combustion of the reactants, and, in stating a given efficiency, reference should be made to which was used. As to ΔG, for uniformity, we will use the one appropriate for the formation of water vapor (if, indeed, water is involved).

The efficiency of practical fuel cells is the ratio of the electric power, P_L, delivered to the load to the heat power that would be generated by combining the reactants in a calorimeter, under the same temperature and pressure used in the cell,

$$\eta_{practical} = \frac{P_L}{P_{in}} = \frac{I_L V_L}{\Delta \bar{h} \dot{N}} = \frac{q n_e N_0 V_L}{\Delta \bar{h}}. \tag{9.85}$$

Again, one has a choice of which $\Delta \bar{h}$ to use. Assuming 100% current efficiency, for hydrogen/oxygen fuel cell at RTP referred to the lower heat of combustion, the efficiency is, at RTP,

$$\eta_{pract_{RTP}} = 0.798 V_L. \tag{9.86}$$

Practical fuel cells have lower efficiency than ideal ones owing to:

1. Not all reactants used up take part in the desired reaction; some may simply escape, whereas others may take part in undesired side-reactions. Sometimes part of the fuel is consumed to operate ancillary devices such as heating catalytic crackers, and so on, or it may be "after burned" to raise the exhaust gas temperature in cogeneration arrangements.
2. Not all the current produced will go through the load; some may leak through parallel paths (a minor loss of current) or may be used to drive ancillary equipment such as compressors.
3. The voltage, V_L, that the cell delivers to a load is smaller than, V_{rev}, the **reversible voltage** (the theoretical voltage associated with the change in the free energy).
   A number of factors contribute to such voltage loss:

   3.1. Unavoidably, fuel cells have internal resistance to the flow of electrons and, in the electrolyte, to the flow of ions.
   3.2. The rate at which the chemical reactions take place—the **kinetics** of the chemical reaction—limits the rate at which electrons are liberated—that is, limits the current produced.
   3.3. Unwanted reactions generate voltages that oppose the normal potential of the cell. Some fuel may leak to the oxidizer side (**fuel crossover**), creating a worrisome problem with direct methanol fuel cells.

4. The effective electrode area may become reduced because

   4.1. Excess water may "drown" the electrodes disturbing the "triple point" contact between reactants, electrolyte, and electrodes.

4.2. Insufficient moisture may dry out the solid electrolyte membrane of SPFCs further increasing the resistance to ion flow.

Incomplete use of fuel (Effect 1), diversion of some of the generated current (Effect 2), and loss of effective area (Effect 4) are problems dealt with mostly by system design, whereas voltage loss (Effect 3) is an inherent property of the individual cell itself. We will discuss this voltage loss by examining the voltage–current (or voltage–current density) characteristics of the cell.

9.8.3 V-I Characteristics of Fuel Cells

Reversible fuel cells will deliver to a load a voltage, $V_L = V_{rev}$, which is independent of the current generated. Their V-I characteristic is a horizontal line. Such ideal cells require reaction kinetics fast enough to supply electrons at the rate demanded by the current drawn. Clearly, reversible cells cannot be realized. In practical fuel cells, two major deviations from the ideal are observed:

1. The open-circuit voltage, V_{oc}, is smaller than V_{rev}.
2. The load voltage, V_L, decreases as the load current, I_L, increases.

Frequently, the V-I characteristics approaches a straight line with some curvature at low currents as, for example, those of Figure 9.26 (left), while other cells exhibit characteristics in which the linearity can be seen only in a limited region (right).

To facilitate fuel cell performance calculations, it is useful to express the cell's characteristics in a mathematical form, that is, to describe the load voltage, V_L, as an analytical function of the load current, I_L,

$$V_L = f(I_L). \tag{9.87}$$

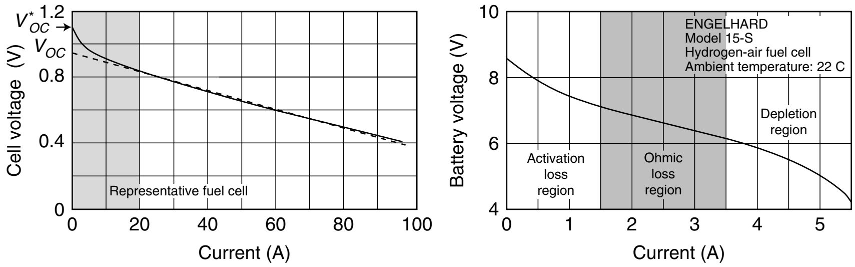


Figure 9.26 The typical modern fuel cell tends to have V-I characteristics consisting of a long stretch of apparently linear relationship between current and voltage, with a small curvature at the low current end (left, above). The small Engelhard liquid-electrolyte demonstration cell had a limited region in which voltage decreases linearly with current. At both lower and higher currents, the V-I characteristic exhibited marked curvature (right, above).

One way to accomplish this is to fit, empirically, a curve to the observed data, that is, to treat the cell as a black box without inquiring what is happening internally. It turns out that good fits can, in general, be achieved. They describe accurately what the load voltage is for a given load current, as long as such parameters as temperatures and pressures are the same as those used in obtaining the data. However, such empirically obtained mathematical expressions fail to provide adequate insight on the workings of the fuel cell, and, thus, they provided limited guidance as how to improve its design. For that, one must be able to interpret the relationship between the empirical parameters and the physical processes inside the cell.

9.8.3.1 Empirically Derived Characteristics

For many calculations, it may be sufficient to use a simple straight-line approximation relating V_L to I_L,

$$V_L = V_{oc} - R_{app} I_L. \tag{9.88}$$

Consider the fuel cell whose characteristics are depicted in Figure 9.27.

In the figure, the straight line was made tangent to the real characteristic at the point where $I_L = 50$ A. This is arbitrary; the point of tangency can be at any current, depending on the chosen region of operation. Since the real data are not linear in I_L, the farther from the point of tangency,

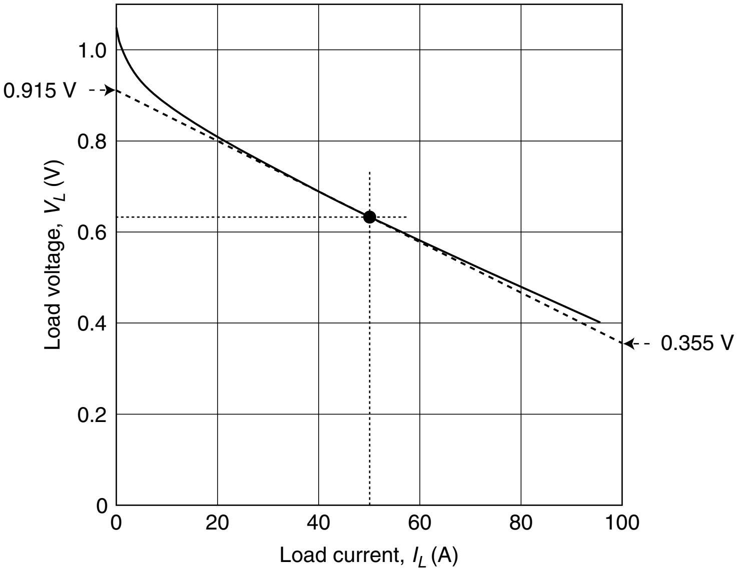


Figure 9.27 A fuel cell has the observed characteristics represented by the solid line. Roughly, over a reasonable range of load currents, the curved observed line can be replaced by the straight, dotted line.

the more the straight line departs from the observed values. Thus, even though the measured open-circuit voltage of this cell is 1.1 V, the value of V_{oc} in Equation 9.88 is only 0.915 volt: the linear characteristic does not do a very good job of describing the cell's performance when the load current is small (or too large). Notice that, although the parameter, R_{app} in Equation 9.88, has the dimensions of resistance, it is not the real internal resistance, R_{int}, of the cell.

Using a straight-line approximation allows the modeling of the cell as a voltage generator in series with an internal resistance, R_{app}, as suggested by the diagram in Figure 9.28

The two circuits of Figure 9.28 are entirely equivalent. The one on the left is obvious; in the one on the right, the open-circuit voltage, V_{oc}, is represented by two opposing voltage generators, V_{rev} and $V_{rev} - V_{oc}$. Such representation facilitates the calculation of internal heat generation, as shown later in this chapter. V_{oc} is, of course, the intercept of the V-I (or V-J) line with the ordinate axis.

From the circuit model,

$$V_L = V_{oc} - I_L R_{app} = V_{oc} - J_L A R_{app}. \tag{9.89}$$

Not uncommonly, voltages are plotted versus current densities, $J_L = I_L/A$, A being the **active area** of the electrodes:

$$V_L = V_{oc} - J_L A R_{app} = V_{oc} - \Re_{app} J_L, \tag{9.90}$$

where $\Re_{app}$, the cell's **specific resistance**, has dimensions of ohms × m^2.

Let us present a parenthesis in our explanation and say a few words about the use of load current density, J_L, in lieu of the actual load current, I_L. Current densities permit a certain generalization of the characteristics and are useful in scaling fuel cells.

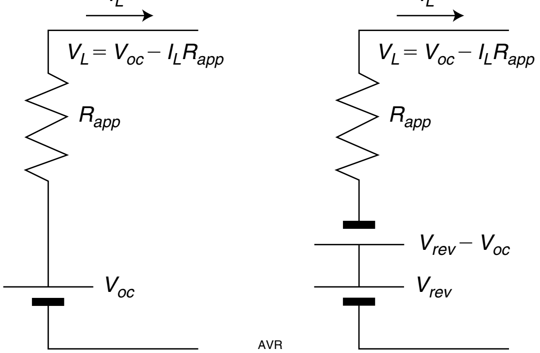


Figure 9.28 Circuit model for a fuel cell with straight-line V-I characteristics.

9.8.3.2 Scaling Fuel Cells

Some of the data on fuel cell performance are presented in the form of V-J rather than V-I characteristics. Such practice permits the scaling of the cells—estimating the performance of a larger cell based on the data from a smaller one of the same type.

Consider a fuel cell that has an active area A_0 and a V-I characteristic

$$V_L = V_{oc} - R_{int_0} I_L. \tag{9.91}$$

Since

$$J_L = \frac{I_L}{A_0}, \tag{9.92}$$

the equation can be written

$$V_L = V_{oc} - R_{int_0} A_0 J_L = V_{oc} - \Re J_L, \tag{9.93}$$

where, remember, $\Re$ is the **specific resistance** of the fuel cell.

If now another fuel cell is built with exactly the same configuration and the same materials but with a different active area, A, then, plausibly, its internal resistance, R_{int}, will be

$$R_{int} = R_{int_0} \frac{A_0}{A} \tag{9.94}$$

because $R = \rho L/A$ (assuming that the thickness of the cell does not change).

V_{oc} does not depend on the area and will therefore remain the same. Thus, the load voltage will be

$$V_L = V_{oc} - R_{int_0} \frac{A_0}{A} I_L = V_{oc} - \Re J_L. \tag{9.95}$$

In other words, the specific resistance of the larger cell is the same as that of the smaller and both have the same V-J characteristics, although they may have quite different V-I characteristics.

In practice, scaling up a particular fuel cell is not necessarily simple. Problems with heat removal and water management arise when dimensions are changed. The same occurs when cells are operated together in "stacks" to raise the overall voltage, as is almost invariably the case.

Even though linear characteristics are only a rough representation of the behavior of fuel cells, they may lead to some modest theoretical conclusions. The published data for a certain type of Ballard SPFC are displayed in Figure 9.29. Linear regressions yield,

$$\begin{aligned} \text{at } 50\ C, \quad & V_L = 0.912 - 54.4 \times 10^{-6} J_L, \\ \text{and at } 70\ C, \quad & V_L = 0.913 - 49.3 \times 10^{-6} J_L, \end{aligned}$$

where J_L is the current density in amperes per m^2.

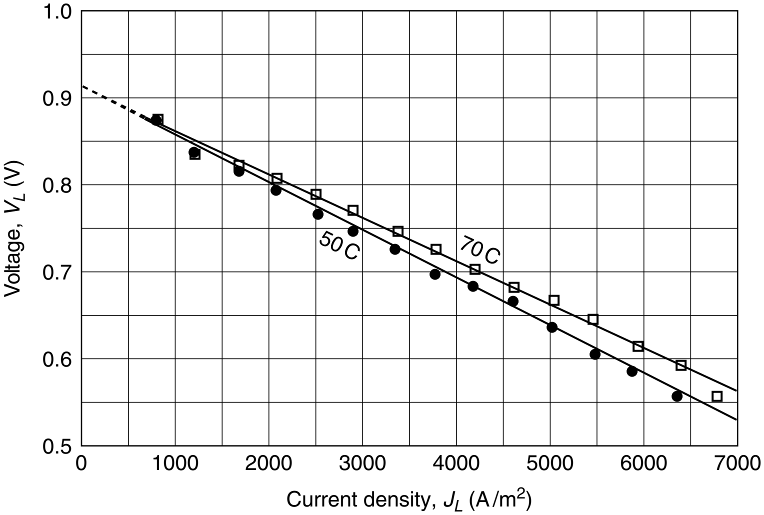


Figure 9.29 Characteristics of a given Ballard SPFC cell.

Since the active area of this particular cell is $A = 0.0232$ m^2, the equations above can be written in terms of the load current, I_L where $I_L = J_L \times A$,

$$\text{at } 50\ C, \quad V_L = 0.912 - 2.34 \times 10^{-3} I_L,$$
$$\text{and at } 70\ C, \quad V_L = 0.913 - 2.12 \times 10^{-3} I_L.$$

The open-circuit voltages are

$$\text{at } 50\ C, 0.912\ V \quad \text{or} \quad 77.4\% \text{ of } V_{rev}, \text{ which is } 1.178 \text{ V},$$
$$\text{and at } 70\ C, 0.913\ V \quad \text{or} \quad 78.0\% \text{ of } V_{rev}, \text{ which is } 1.171 \text{ V}.$$

The open-circuit voltage is only slightly influenced by the temperature. As explained before, V_{rev} became a bit smaller with the increase in temperature, while V_{oc} actually became marginally larger owing to improved kinetics. The internal resistance was affected by the temperature in a more substantial way. It fell from 2.34 mΩ to 2.12 mΩ (more than 9%) with the 20 K increase in temperature.

9.8.3.3 More Complete Empirical Characteristics of Fuel Cells

We will use a somewhat unorthodox procedure to fit an equation to the observed characteristics of a fuel cell. We start by recognizing that a cell must have an internal resistance, R_{int}. However, as we saw, this is not sufficient to accurately describe the cell performance. There must be an additional voltage drop, which we will call the **activation voltage**, V_{act}. We can imagine a V_{act} that has exactly the correct dependence on I_L needed to reproduce the observed values of V_L.

Our task is to find an analytical expression for such a V_{act} as a function of I_L. This should be possible, given a sufficiently large number of good experimental values of V_L–I_L pairs.

We will use real data from an experimental KHO fuel cell. These data were published in a book titled *Fuel Cells* by Will Mitchell, Jr. (1963). We scaled a graph that appeared on page 153 of the book. Although such scaling introduced additional noise in the data, it still permitted the carrying out of the necessary computer experimentation. The device was designated "New Cell" and refers to an old (1960) alkaline (KOH), high-pressure hydrogen–oxygen fuel cell described (in the Mitchell book) by Adams et al. It operated at 200 C and, to keep the electrolyte from boiling away, had to be pressurized to 42 atmospheres. Owing to its high operating temperature, this experimental cell is reasonably efficient, as witnessed by the large open-circuit voltage.

We rescaled the published data, obtaining a tabulation of V_L as a function of I_L. The values are plotted in Figure 9.30.

In the plot, we used an arbitrary (yet plausible) value for R_{int}. If this were the only loss mechanism, then the V-I characteristic would simply be the straight line marked "No activation voltage" in the figure. Owing, however, to the existence of an activation voltage, the true load voltage is then given by

$$V_L = V_{oc} - R_{int} I_L - V_{act}. \tag{9.96}$$

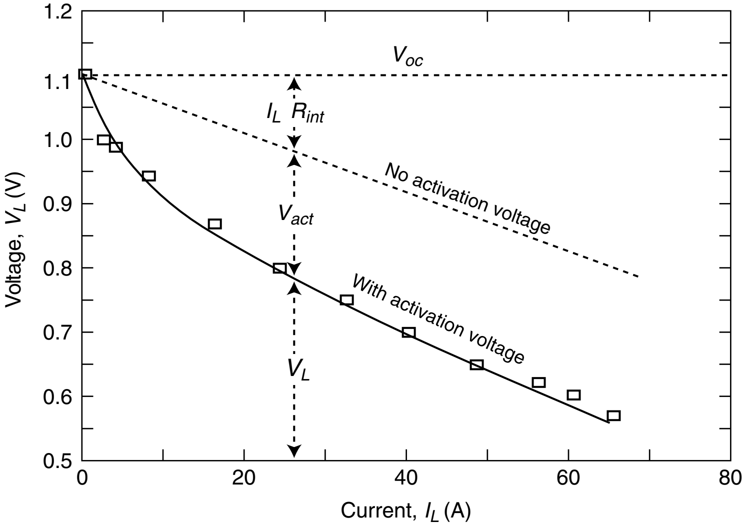


Figure 9.30 *V-I* characteristics of a high-pressure hydrogen–oxygen KOH fuel cell of 1960.

Solving for V_{act},

$$V_{act} = V_{oc} - R_{int} I_L - V_L. \tag{9.97}$$

V_{oc}, I_L, and V_L are experimentally observed values. R_{int} is chosen arbitrarily. Using these data, we are able to tabulate the activation voltage versus the natural log of the load current. This was done in Figure 9.31 for different choices of R_{int}. Of course, at this stage, we do not know which R_{int} is the correct choice. We do notice, however, that for a particular value of this resistance ($R_{int} = 0.0046$ ohm, in this example), the V_{act} vs. $\ln I_L$ plot is a straight line, whereas, for larger or smaller values of R_{int}, the plot is curved. A straight line means that there is a logarithmic dependence between V_{act} and I_L. Later we will show that, indeed, theory predicts such a simple dependence. Thus, one concludes that R_{int} has, in this case, a value around 4.6 milliohms and that V_{act} is given by

$$V_{act} = V_1 + V_2 \ln I_L, \tag{9.98}$$

where the parameters V_1 and V_2 are obtained from the tabulated values of V_{act} versus $\ln(I_L)$. This empirical expression is sometimes called the **Tafel equation**.

From the observed data in the example being discussed, it is possible to make a linear regression between V_{act} and $\ln(I_L)$ and, thus, to determine the values of the parameters V_1 and V_2 in Equation 9.98, which, for our example, becomes

$$V_{act} = 0.0277 + 0.0521 \ln I_L. \tag{9.99}$$

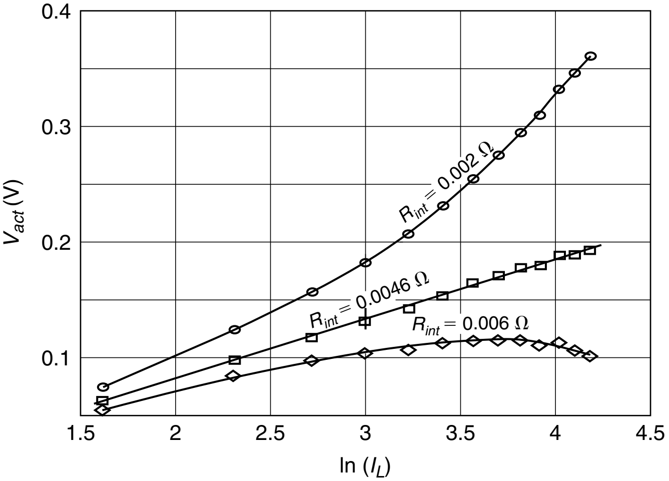


Figure 9.31 Only a specific value of R_{int} will yield a linear relationship between V_{act} and $\ln(I_L)$.

In obtaining the regression, we were careful to stay away from load currents too close to zero because this would cause $\ln(I_L)$ to blow up.

Equation 9.99 can also be written as

$$V_{act} = V_2 \ln \frac{I_L}{I_0} = 0.0521 \ln \frac{I_L}{0.588} \tag{9.100}$$

because

$$I_0 = \exp\left(-\frac{V_1}{V_2}\right) = \exp\left(-\frac{0.0277}{0.0521}\right) = 0.588. \tag{9.101}$$

Since the observed open-circuit voltage of the fuel cell in our example was 1.111 V, the V_L-$I_L$characteristics are

$$V_L = 1.111 - 0.0046 I_L - 0.0521 \ln \frac{I_L}{0.588}. \tag{9.102}$$

Equation 9.102 fits the measured data well, with one major exception. At low currents, the predicted V_L exceeds the observed value. At 0.1 A, the predicted value is 1.20 V, which is larger than the observed V_{oc}, and when $I = 0$, the equation predicts a value of $V_L = \infty$, which is patently absurd. As noted previously, our regression for calculating V_{act} is not valid when small load currents are considered. Can we modify Equation 9.99 or 9.100 so as to extend our model into the low-current range?

Equation 9.99 can be inverted,

$$I_L = I_0 \exp\left(\frac{V_{act}}{V_2}\right). \tag{9.103}$$

This expression reminds us of Boltzmann's law (see Chapter 2) which states: "the probability of finding molecules in a given spatial arrangement varies exponentially with the negative of the potential energy of the arrangement, divided by kT."

The potential energy of the electron in the presence of a voltage, V_{act}, is qV_{act}. Equation 9.103 can then be written as

$$I_L = I_0 \ \exp\left(\alpha \frac{qV_{act}}{kT}\right). \tag{9.104}$$

Here, we introduced the arbitrary factor, α, to adjust the magnitude of the argument of the exponential. Clearly, if Equation 9.104 is the same as Equation 9.103, then

$$\frac{V_{act}}{V_2} = \alpha \frac{qV_{act}}{kT}, \tag{9.105}$$

that is,

$$\alpha = \frac{kT}{qV_2}, \tag{9.106}$$

and for the present example in which $T = 473$ K,

$$\alpha = \frac{kT}{qV_2} = 0.783. \tag{9.107}$$

After all these manipulations, Equation 9.103 is simply a mathematical representation of most of the experimentally observed relationship between V_{act} and I_L. However, the equation fails badly in representing the obvious condition: that V_{act} must be zero when $I_L = 0$. On the other hand, Equation 9.108 will also fit the data, provided β is sufficiently large (because then the second term is essentially zero unless V_{act} is quite small). However, this same equation also fits the condition $V_{act} = 0$ for $I = 0$.

$$I_L = I_0 \exp\left(\alpha \frac{qV_{act}}{kT}\right) - I_0 \ \exp\left(-\beta \frac{qV_{act}}{kT}\right). \tag{9.108}$$

We are going to show that there are theoretical reasons for assuming that $\alpha + \beta = 1$, or $\beta = 1 - \alpha$. If so, Equation 9.108 becomes

$$I_L = I_0 \ \exp\left(\alpha \frac{qV_{act}}{kT}\right) - I_0 \ \exp\left((\alpha - 1) \frac{qV_{act}}{kT}\right), \tag{9.109}$$

and, for the current example,

$$\begin{aligned} I_L &= 0.588 \exp\left(0.783 \times 24.5 V_{act}\right) - 0.588 \ \exp\left(-0.217 \times 24.5 V_{act}\right) \\ &= 0.588 \left[\exp\left(19.2 V_{act}\right) - \exp\left(-5.32 V_{act}\right)\right], \end{aligned} \tag{9.110}$$

where 24.5 is the value of q/kT when the temperature is 473 K.

Equation 9.109 is known as the **Butler–Volmer equation**, which we will later derive from theoretical considerations.

The first term of the Butler–Volmer equation is, of course, Equation 9.104. The second term is equal to the first when $V_{act} = 0$ forcing $I_L = 0$ under such conditions. As V_{act} grows, the second term quickly decreases in magnitude, becoming negligible for even small values of the activation voltage so that Equations 9.104 and 9.109 then yield the same numerical result. Thus, Equation 9.109 can be made to represent with acceptable accuracy the relationship between I_L and V_{act}. If it were possible to invert the equation to express V_{act} as a function of I_L, then we would be able to write an analytical expression yielding V_L for a given I_L. Unfortunately, this is not possible. One way to tabulate the characteristics of a cell given R_{int}, I_0, and V_2 is to use V_{act} as an independent variable and to calculate I_L from the Butler–Volmer equation, and then to use Equation 9.96 to find V_L.

9.8.4 Open-circuit Voltage

Most of the losses in a fuel cell are current dependent—the internal resistance causes an output voltage drop proportional to the load current; chemical kinetics cause a voltage drop approximately proportional to the logarithm of this same current. There is, however, a loss mechanism that is current independent, one that occurs even when no current is drawn from the cell: the open-circuit voltage is invariably lower than the calculated reversible voltage.

The cause of this voltage drop is mostly the unwanted side-reactions that occur in the cathode of the cell owing to the very oxidizing conditions existing at that electrode. As an example, the protons that move through the electrolyte from anode to cathode tend to combine with the oxygen in the cathode, forming hydrogen peroxide,

$$2\mathrm{H}^+ + 2\mathrm{e}^- + \mathrm{O}_2 \rightarrow \mathrm{H}_2\mathrm{O}_2. \tag{9.111}$$

Another side-reaction oxidizes platinum,

$$\mathrm{Pt} + \mathrm{H}_2\mathrm{O} \rightarrow \mathrm{PtO} + 2\mathrm{H}^+ + 2\mathrm{e}^-. \tag{9.112}$$

When there is some fuel crossover, as always happens in direct methanol fuel cells or in other SPFC in which the electrolyte is too thin, then some fuel is oxidized at the cathode, generating a voltage that bucks the normal output of the cell and reduces the output voltage.

9.8.5 Reaction Kinetics

9.8.5.1 Reaction Rates

The electrons that constitute the load current of a fuel cell are generated by the oxidation reaction at the anode of the cell. The current that can be delivered is determined by the rate of electron generation, that is, by the chemical reaction rate. It is, therefore, of interest to investigate what influences such a rate.

Take, for instance, the reaction of nitric oxide and chlorine, producing nitrosyl chloride,

$$2\mathrm{NO} + \mathrm{Cl}_2 \rightarrow 2\mathrm{NOCl}, \tag{9.113}$$

which proceeds directly without intermediate products. It is an **elementary reaction**, one that cannot be broken down into smaller steps. In such reactions, the reaction rate is proportional to the concentration of the reactants. Thus, if the reaction

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{P} \tag{9.114}$$

is elementary, then the reaction rate, r,

$$r \equiv \frac{d\mathrm{P}}{dt} = k[\mathrm{A}][\mathrm{B}], \tag{9.115}$$

where k is the **reaction constant**, a quantity whose dimensions depend on the stoichiometry. The square brackets, [A], indicate concentration of A.

For the reaction,

$$2\ \mathrm{A} + 3\ \mathrm{B} \rightarrow \mathrm{P}, \tag{9.116}$$

$$r = k[\mathrm{A}][\mathrm{A}][\mathrm{B}][\mathrm{B}][\mathrm{B}] = k[\mathrm{A}]^2[\mathrm{B}]^3. \tag{9.117}$$

In general, for

$$a\ \mathrm{A} + b\ \mathrm{B} \rightarrow \mathrm{P}, \tag{9.118}$$

$$r = k[\mathrm{A}]^a[\mathrm{B}]^b. \tag{9.119}$$

On the other hand, the reaction of hydrogen and bromine, yielding hydrogen bromide,

$$\mathrm{H_2} + \mathrm{Br_2} \rightarrow 2\mathrm{HBr}, \tag{9.120}$$

has a reaction rate that is *not* $r = k[\mathrm{H}]^2[\mathrm{Br}]^2$ because it proceeds in a surprising five steps, starting with the dissociation of bromine, $\mathrm{Br_2}$ into 2 Br. It is, therefore, not an elementary reaction.

For all reactions, elementary or not, the exponents of the reaction rate formula are called **the reaction order** with respect to a given reactant. The **overall reaction order** is the sum of the individual reaction order for each reactant. Thus, the reaction of Equation 9.116 is second order in A, third order in B, and fifth order overall. For elementary reactions, the order is equal to the stoichiometric coefficient of the reaction formula. This is the **mass action law** proposed by Guldberg and Waage in 1864.

If the reaction is not elementary, that is, if it can be separated into a sequence of different elementary steps, then, although the reaction rate is still given by Equation 9.119, the exponents are not the stoichiometric coefficients. As a matter of fact, there is no way one can relate such coefficients to the reaction order. The reaction order must be determined experimentally for each particular reaction.

We have now to explain why reactions do not occur instantaneously and, incidentally, why the reaction constant is so strongly dependent on temperature.

9.8.5.2 Activation Energy

Consider an inert container filled with a mixture of hydrogen and oxygen at 300 K and in the correct stoichiometric proportions to form water. One would expect that the two gases would promptly react forming water because of the large thermodynamic driving force—the water vapor that would result from such a reaction has a free energy 228 MJ/kmole *lower* than that of the gas mixture: a system tends toward the configuration of the lowest possible free energy. It turns out, somewhat surprisingly, that actually nothing happens; the system turns out to be kinetically stable, the reaction does not occur or, in other words, it occurs at a negligible rate. This stability means that the gas mixture must be in a state of minimum free energy just as water is in a different, and much lower, minimum, as suggested in Figure 9.32. Separating these two minima, there must be an **activation barrier**, ΔG_{act_f}, that keeps the reaction from proceeding.

Indeed, the hydrogen molecule cannot react directly with an oxygen molecule. The oxygen molecule must first dissociate,

$$O_2 \to O + O. \tag{9.121}$$

This dissociation requires 493 MJ/kmole or 5.12 eV per O_2 molecule, the energy coming from thermal agitation collisions. Thus, only oxygen molecules that collide with more than 5.12 eV can react. The gas molecules have a Maxwellian energy distribution, and at 3000 K, about 48% of all

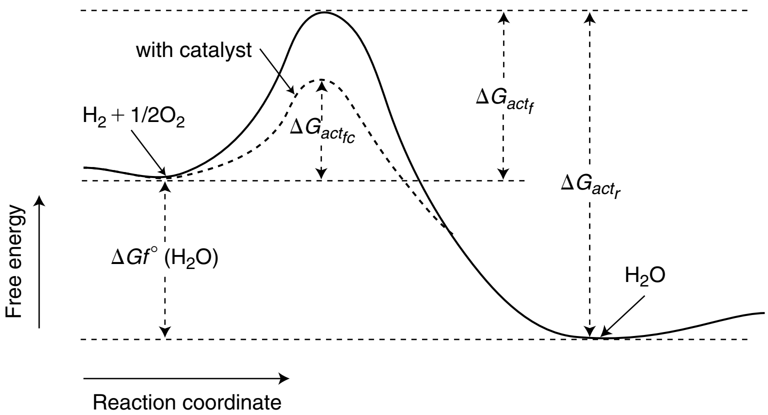


Figure 9.32 At 300 K, the hydrogen/oxygen mixture is kinetically stable because it occupies a local free energy minimum, while the water vapor occupies another, much lower, minimum. An **activation barrier**, ΔG_{act_f}, impedes the progress of the $H_2 + \frac{1}{2}O_2 \to H_2O$ reaction.

molecules have energy above this limit and plenty of atomic oxygen needed for the reaction would become available. However, the percentage falls to below 3% at 300 K, and the reaction would proceed much more slowly.[†]

Of course, the situation is much more complicated than the mere creation of atomic oxygen; other intermediate products have also to be generated for the reaction to occur. The need to create such intermediate compounds (**activated complexes**) sets the actual activation barrier.

The reaction rate, r, must be proportional to the number, N, of molecules that have more than the energy required to create the most energetic intermediate compound, that is, more than the activation barrier, qV_{act_f}. The subscript, f, indicates that we are discussing the barrier for the forward reaction from hydrogen plus oxygen to water. A reverse activation barrier, qV_{act_r}, controls the rate of the reverse reaction that decomposes water into its elements. Since the tail end of the Maxwellian distribution is, approximately, exponential,

$$\frac{\partial N}{\partial W} \propto \exp - \frac{W}{kT}, \tag{9.122}$$

and the number of molecules with energy larger than a given value, W_0 (shaded area in Figure 9.33), is approximately

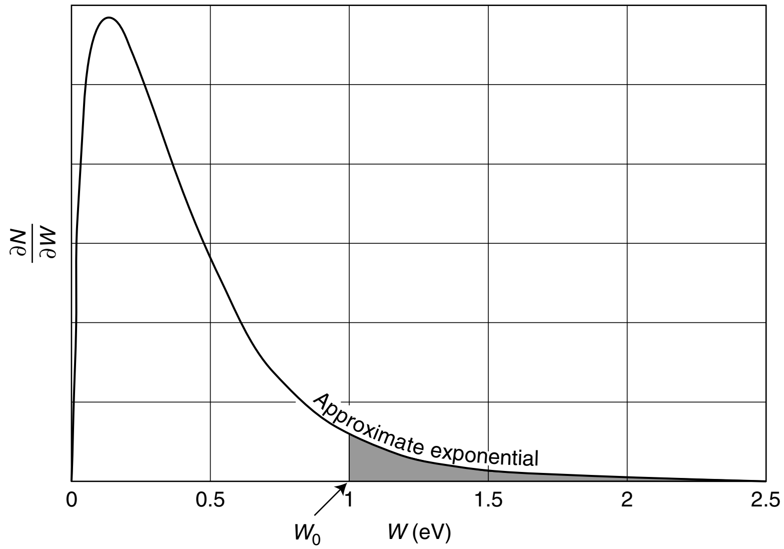


Figure 9.33 The tail end of a Maxwellian distribution is roughly exponential. Consequently, the number of molecules with more than a given energy, W_0, (1 eV in the drawing) is proportional to exp W_0/kT.

[†] The percentages were calculated using Equation 2.89 of Chapter 2.

$$N \propto \int_{W_0}^{\infty} \exp -\frac{W}{kT} dW = \exp -\frac{W_0}{kT} \rightarrow \exp -\frac{qV_{act}}{kT}. \qquad (9.123)$$

The mathematical relationship between reaction rate and temperature, called the **Arrhenius equation**,

$$r \propto \exp -\frac{qV_{act}}{kT} \rightarrow A \exp -\frac{qV_{act}}{kT} = A \exp -\frac{W_{act}}{RT}, \qquad (9.124)$$

was proposed by the Swedish physicist, Svante August Arrhenius, based on empirical data. It is actually an oversimplification but serves well to give a general feeling of what is going on. The factor, A, variously known as **frequency factor, pre-exponential factor**, or **steric factor**, incorporates the concentration effect of the previous subsubsection and is approximately constant over a small temperature range. W_{act} is the **activation energy**.

9.8.5.3 Catalysis

Returning to our "experiment," if we introduce into the container (still at 300 K) a small amount of finely divided platinum powder or, better yet, some platinum sponge,[†] we will find that the reaction will be enormously accelerated—the sponge will actually ignite the hydrogen/oxygen mixture. Clearly, the presence of platinum has reduced the activation barrier. This action is called **catalysis**, and it works by opening a reaction path involving less energetic activated complexes. This is suggested by the dotted line in Figure 9.32.

Some reactions are relatively simple, as is the oxidation of hydrogen in the anode of the fuel cell. Formally,

$$\mathrm{H_2 \rightarrow 2H^+ + 2e^-}. \qquad (9.125)$$

One way for the reaction to proceed is to first dissociate H_2 into 2 H, at a cost of 432 MJ/kmole of atomic hydrogen or 4.48 eV per H atom. However, a reaction path that involves a much smaller energy barrier is facilitated by the presence of platinum as a catalyst. It involves the adsorption of the H_2 molecule by the metal with a subsequent separation of the molecule into two atoms (M stands for metal—platinum in this example—and M...H represents an adsorbed hydrogen atom),

$$\mathrm{H_2} + M \rightarrow \mathrm{M \ldots H_2 \rightarrow 2(M \ldots H)}. \qquad (9.126)$$

The **chemisorbed** hydrogen (M...H) then ionizes, leaving an electron in the metal and making a proton available to the electrolyte.

[†] Platinum sponge is a highly porous form of metallic platinum, which exhibits a high surface-to-volume ratio. Catalysis, the phenomenon discussed here, is a surface action.

A substantially more complicated reaction path occurs at the cathode where oxygen is reduced,

$$O_2 + 4H^+ + 2e^- \rightarrow 2H_2O. \tag{9.127}$$

The oxygen reduction catalysis is further complicated by the repeated potential cycling of the fuel cell in automotive applications during stop and go driving. Such cycling leads to a dissolution of the platinum in the cathode. However, Zhang et al. (2007) have shown that nanosized gold clusters deposited on carbon-suspended platinum inhibit such dissolution without interfering with the catalysis.

9.8.6 The Butler–Volmer Equation

We derived the Butler–Volmer equation entirely from empirical data. We will now attempt to reach the same result based on electrochemical considerations.

9.8.6.1 Exchange Currents

When two dissimilar materials, at uniform temperature, are placed in contact with one another, a contact potential develops. The most familiar case (at least for the electrical engineer) is the potential that appears across a *p*-*n* junction.[†] In a single semiconductor crystal consisting of an *n*-region and a *p*-region, free electrons, more abundant in the *n*-side, diffuse toward the *p*-side, whereas holes from the *p*-side migrate to the *n*-side.

Were these particles uncharged, the diffusion process would stop only when the concentration became uniform across the crystal. This does not occur because a compensating drift current causes carriers to flow back in a direction opposite to that of the diffusion current. The drift current is driven by a **contact potential** created as follows: Being charged, the migrating electrons not only transport negative charges to the *p*-side but also leave uncompensated positively charged donors in the *n*-side. The holes also contribute to the accumulation of positive charges in the *n*-side and the uncompensated acceptors, to the accumulation of negative charges in the *p*-side. Thus, the *n*-side becomes positive and the *p*-side negative.

In an open-circuited junction in equilibrium, there is no net current. This is the result of the precise cancellation of the diffusion current by the drift current. Although these **exchange currents** add to exactly zero[††] they are not individually zero; their magnitude can be surprisingly large. In silicon, under normal conditions, they may be of the order of 1 million A cm^{-2}. The above is a good example of the dynamic equilibria that occur

[†] See Section 14.6.

[††] There are minute statistical fluctuations that give rise to the radio noise discussed in Chapter 8.

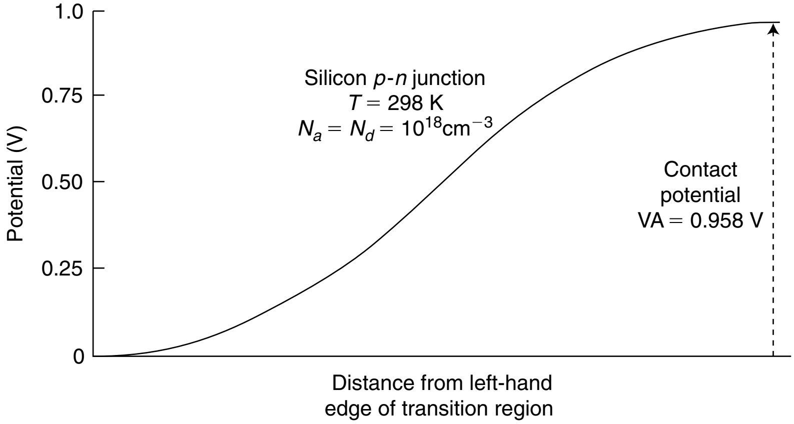


Figure 9.34 The potential varies with distance from the edge of a $p-n$ junction in a simple and predictable manner.

frequently in nature: a zero net effect is the result of the precise cancellation of two large opposing effects.

One peculiarity of the contact potential: it cannot be measured directly because around any closed loop the potentials cancel each other.[†]

The behavior of the electric potential across a *p-n* junction is easy to predict (see Figure 9.34). In the metal-electrolyte junction, the situation is more complicated because, at this interface, there is a change of carriers—in the metal electrode, the current is transported by electrons, while in the electrolyte, it is transported by ions, either positive or negative (**cations** or **anions**, respectively). Thus, the flow of current from electrode to electrolyte, or vice versa, always involves a chemical reaction.

Consider an inert metal electrode in simultaneous contact with an electrolyte and with adsorbed hydrogen atoms some of which are spontaneously ionized. We want to investigate the dependence of the current that flows through this system on the potential difference between the electrolyte and the electrode. Read the text box "Cell Potential" at the beginning of this chapter.

Measuring Electrode/Electrolyte Characteristics

To measure the electrolyte/electrode potential, one needs an electric (read electronic) contact with the electrolyte. Inevitably, this requires one more electrode in addition to the one whose behavior, we want to investigate. One can only measure the **cell potential**—the potential

(*Continues*)

[†] For more details, see Section 14.7 on photodiodes.

(*Continued*)

across a pair of electrodes—not the **half-cell potential**—the potential between an electrode and an electrolyte.

Electrochemists use a **reference electrode** whose potential (relative to the electrolyte), over a useful current range, is independent of the actual current. This is called a **nonpolarizable electrode**. Thus, although the absolute **working electrode** potential is unknown, the dependence of the current on potential can be observed.

The internationally accepted standard reference electrode is the **standard hydrogen electrode** (SHE), which, unfortunately, is difficult to use. More practical reference electrodes are commonly employed, such as the **silver-silver chloride electrode** or the **calomel**[†] electrode.

[†] Calomel is the common name for mercurous chloride or mercury (I) chloride, Cl-Hg-Hg-Cl. The word means "beautiful black" (*kalos melas*) in Greek.

The electrons stay in the metal, while the H^+ go into solution. As a consequence, the solution becomes more positive than the metal (potential, V, in Figure 9.35, which is akin to the contact potential discussed previously), causing some of the dissolved ions to be attracted back to the negatively charged electrode.[†] Just as in the *p*-*n* case, two exchange currents are established:

1. A forward current, i_f, carried by the ions that leave the metal and diffuse into the electrolyte under the influence of the gradient of ion concentration near the metal.
2. A reverse current, i_r, carried by the ions from the electrolyte that drift back to the metal under the influence of the electric field.

The total current is the sum of the two currents above,

$$i = i_f + i_r. \tag{9.128}$$

When there is no external electric connection to the electrodes, the two currents must, under steady-state conditions (indicated by the subscript, 0), be equal and opposite so that their sum is zero,

$$i_0 = i_{f_0} + i_{r_0} = 0. \tag{9.129}$$

[†] Observe that we have what essentially amounts to a capacitor: a negatively charged electrode separated from a positively charged electrolyte by a very thin solvent layer (**Helmholtz layer**, ≈ 0.3 nm).

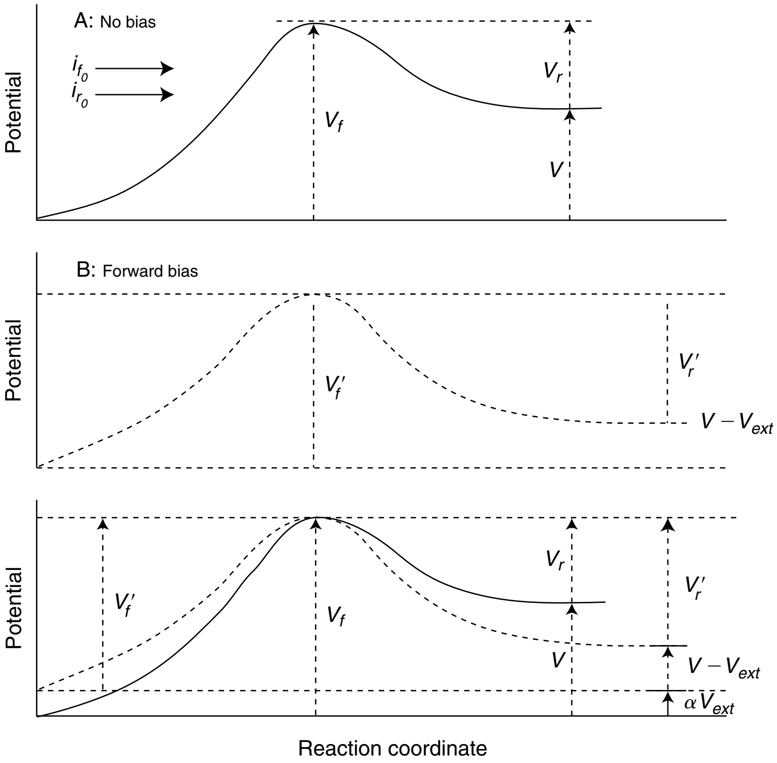


Figure 9.35 Potential versus reaction coordinate at a metal–electrolyte interface. A—Unbiased, B—Biased, C—Biased and unbiased superposed.

Notice the sign convention: $i_f > 0$, while $i_r < 0$.

The potential versus reaction coordinate plot, for the equilibrium condition, is suggested in Figure 9.35A. The forward current has to overcome the activation barrier, qV_f; hence, it must be of the form,

$$i_{f_0} = I_f \exp\left(-\frac{qV_f}{kT}\right), \tag{9.130}$$

while the reverse current has to overcome a smaller barrier, qV_r, because the electrolyte is at a higher potential, V (the contact potential, as it were), with respect to the electrode. Thus,

$$i_{r_0} = I_r \exp\left(-\frac{qV_r}{kT}\right). \tag{9.131}$$

Here, again, $I_r < 0$.

In absence of catalysis, V_f is large, and the exchange currents are small. Catalysts, by reducing V_f, cause a marked increase in these exchange currents, which are thus a good indicator of the kinetics of the reaction.

Now, assume that a voltage is applied to the system so that the potential between the electrode and the solution is reduced by an amount, V_{ext}. In other words, V_{ext} forward-biases the metal-electrolyte junction (see Figure 9.35B).

The application of this voltage will alter both the forward and the reverse activation barrier, but not necessarily by the same amount (this depends on the symmetry of the barrier) and not in the same direction: the forward barrier, the one that opposes the diffusion of ions from electrode to electrolyte is decreased, while the reverse barrier, the one that opposes the drift from electrolyte to electrode, is enhanced. The forward current, i_f, is increased, and the reverse current, i_r, is decreased so that they no longer cancel one another and a net external current will flow. Say that the forward bias is reduced by a fraction, α, of the applied external voltage, V_{ext}. The new forward barrier is now

$$V_f' = V_f - \alpha V_{ext}. \tag{9.132}$$

α is the **transfer coefficient** and necessarily must be larger than zero and smaller than 1.

This leaves a potential $(1 - \alpha)V_{ext}$ to alter the reverse barrier that becomes

$$V_r' = V_r + (1 - \alpha)V_{ext}. \tag{9.133}$$

If you are confused by signs, you may want to puzzle out the situation by studying Figure 9.35C in which we superposed the biased and unbiased potential curves.

The two exchange currents currents are now

$$i_f = I_f \exp\left[-\frac{q(V_f - \alpha\ V_{ext})}{kT}\right] = i_{f_0} \exp\left[\alpha \frac{q\ V_{ext}}{kT}\right], \tag{9.134}$$

$$i_r = I_r \exp\left[-q\frac{V_r + (1-\alpha)V_{ext}}{kT}\right] = i_{r_0} \exp\left[-(1-\alpha)\frac{q\ V_{ext}}{kT}\right]. \tag{9.135}$$

The total current that circulates in the external circuit is

$$\begin{aligned} i = i_f + i_r &= i_{f_0} \exp\left[\alpha \frac{q\ V_{ext}}{kT}\right] + i_{r_0} \exp\left[-(1-\alpha)\frac{q\ V_{ext}}{kT}\right] \\ &= i_{f_0} \exp\left[\alpha \frac{q\ V_{ext}}{kT}\right] - i_{f_0} \exp\left[-(1-\alpha)\frac{q\ V_{ext}}{kT}\right]. \end{aligned} \tag{9.136}$$

The expression in Equation 9.136 is the **Butler–Volmer equation** and is the same as that in Equation 9.109, derived empirically.

Although Equation 9.136 was derived here for the case of a half-cell reaction, it has exactly the same form as for the full-cell reaction case.

9.8.7 Transport Losses

So far, we have examined three important loss mechanisms in fuel cells: the open-circuit voltage drop, the internal resistance, and the activation energy losses. An additional loss mechanism occurs when a fuel cell is driven too hard—the transport losses.

At high currents, the reaction at the catalyst layers may become fast enough to deplete reactant concentration (and unduly raise product concentration). Reactants have to be delivered to the active part of the cell through tortuous paths; they have to be spread out evenly over the electrode surface by means of flow plates (through narrow flow channels) and then, most of the time, have to ooze through labyrinthine pores in the electrode. In general, the flow in the flow channels is governed by pressure differentials—it is a convective flow—while the motion through the porous electrodes is governed by concentration gradients. It is a diffusive flow described by

$$j = -qD\frac{dn}{dx}, \tag{9.137}$$

where j is the current density (proportional to the reactant flux), D is the diffusion constant, n is the concentration, and x is distance.

Assuming a constant concentration gradient,

$$j = -qD\frac{\Delta n}{\Delta x} = -qD\frac{n_X - n_0}{X}. \tag{9.138}$$

where X is the electrode thickness.

For a given n_0, maximum current density is reached when $n_X = 0$,

$$j_{\max} = qD\frac{n_0}{X} \tag{9.139}$$

from which

$$n_X = n_0 - \frac{X}{qD}j. \tag{9.140}$$

The voltage depends on concentration according to Nernst's equation (Equation 9.70):

$$\begin{aligned} V = V_0 + \frac{RT}{nF}\ln\left(\frac{n_X}{n_0}\right) &= V_0 + \frac{RT}{nF}\ln\left(\frac{n_0 - \frac{X}{qD}j}{n_0}\right) \\ &= V_0 - \frac{RT}{nF}\ln\left(\frac{j_{max}}{j_{max} - j}\right). \end{aligned} \tag{9.141}$$

$$\Delta V_{transp} \equiv V_0 - V = \frac{RT}{nF}\ln\left(\frac{j_{max}}{j_{max} - j}\right) = \frac{RT}{nF}\ln\left(\frac{i_{max}}{i_{max} - i_L}\right), \tag{9.142}$$

because the active cell areas cancel out. ΔV_{transp} is the voltage loss owing to the concentration drop caused by reactant diffusion through the electrode walls.

Although the above expression predicts reasonably well the *shape* of the ΔV_{transp} versus i_L dependence, it underestimates the actual value of ΔV_{transp}. In part, this is because we did not include the effect of reduced concentration on the value of i_{r_0} in the Butler–Volmer equation, that is, the effect of lower concentration on the kinetics of the reactions. We also did not consider the effects of pressure drops in the two flow plates.

In a modern fuel cell, a representative value of j_{max} might be some 10,000 A/m^2. This is a very high current density.

Figure 9.36 shows the calculated V-I characteristics of a hypothetical fuel cell, which is severely limited by transportation losses. The data used in the figure were internal resistance, $R_{int} = 0.0046\ \Omega$, exchange current, $i_0 = 0.5$ A, transfer coefficient, $\alpha = 0.783$, transport limiting current, $i_{max} = 110$ A, and operation temperature, $T = 363$ K. We multiplied the transport losses calculated from Equation 9.142 by three to (quite arbitrarily) compensate for the underestimation mentioned in one of the preceding paragraphs.

Figure 9.37, adapted from the paper by Ralph and Hogarth (2002), indicates the relative magnitude of the different losses discussed in this section. The most obvious observation is that the losses associated with the reduction of oxygen in the cathode are much larger than those associated

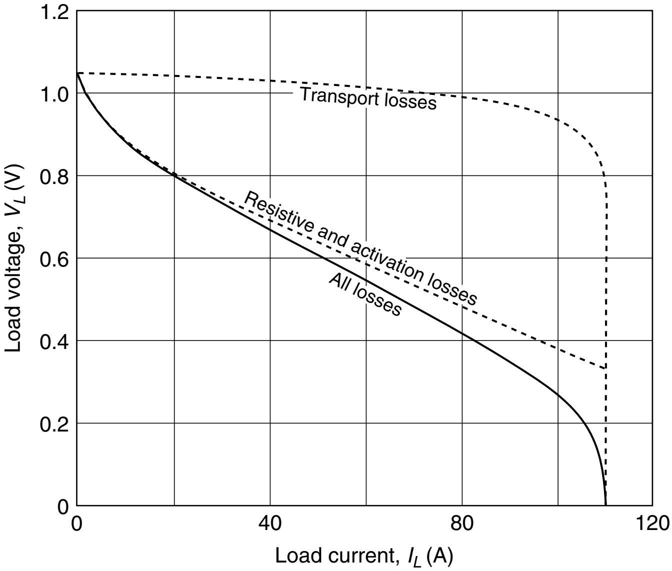


Figure 9.36 Transport losses tend to be insignificant at low currents and to increase abruptly as the load current approaches i_{max}.

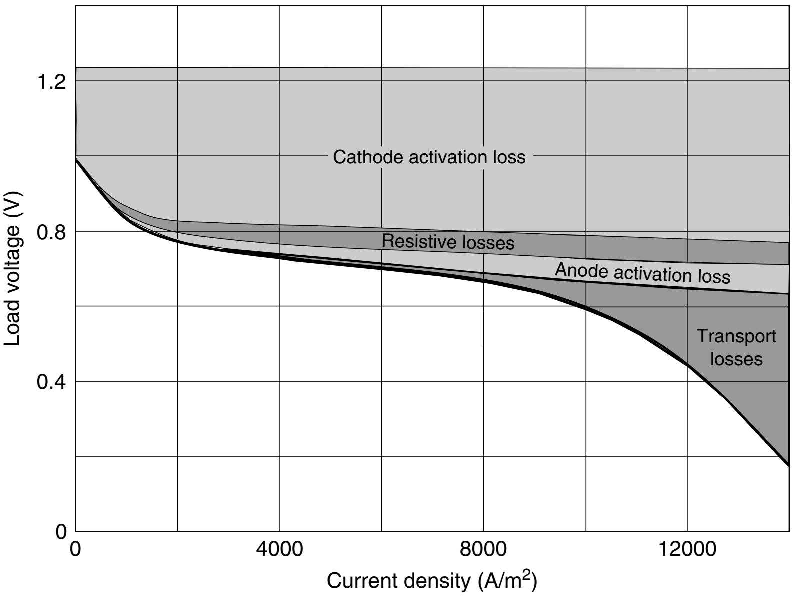


Figure 9.37 Losses in a typical SPFC are dominated by the adverse reaction rates for the reduction of oxygen in the cathode.

with the oxidation of hydrogen in the anode. Indeed, the voltage drop due to cathode kinetics exceeds 400 mV (at a plausible operating current), while the corresponding drop at the anode is less than 100 mV. The difference in reaction rates is more dramatically indicated by the low exchange currents at the cathode (10^{-8} to 10^{-6} A/m^2) compared with some 10 A/m^2 at the anode, when operating with pure platinum.

9.8.8 Heat Dissipation by Fuel Cells

Departures from reversibility constitute losses that appear as heat. Under thermal equilibrium, the rate of heat rejection is equal to the difference between the total power available from the reaction and the electric power delivered to the load. The power available from the reaction is $|\Delta\overline{h}|\dot{N}$, where $\Delta\overline{h}$ is the enthalpy change owing to the reaction per kilomole of product and $\dot{N}$ is the rate at which the product is being generated. If P_{heat} is the heat rejected and P_L is the electric power in the load,

$$P_{heat} = |\Delta\overline{h}|\dot{N} - P_L. \tag{9.143}$$

In practical cells, there are several mechanisms for heat generation:

1. The thermodynamic heat power, $P_{thermodynamic} = T|\Delta\overline{s}|\dot{N}$. This heat is rejected (rarely, absorbed) even by reversible cells.

2. The heat resulting from the difference between the reversible voltage, V_{rev}, and the open-circuit voltage, V_{oc}. It is $P_{oc} = (V_{rev} - V_{oc})I$ watts.
3. The heat dissipated in the internal resistance, R_{int}, of the cell. This amounts to $P_{Joule} = I^2 R_{int}$ watts.
4. The heat owing to other departures, V_{extra}, of the cell voltage from the simple $V_L = V_{oc} - IR_{int}$ behavior. These departures may be due to the activation voltage drop, V_{act}, or to the voltage drop resulting from electrolyte depletion, as mentioned. This amounts to $P_{extra} = V_{extra} I$.
5. The heat of condensation, $P_{cond} = |\Delta \overline{h}_{cond}|\dot{N}$, of the product water. If $\Delta \overline{h}$ of the reaction (Equation 9.143) is the value for formation of water vapor and the product water is removed as vapor, or if $\Delta \overline{h}$ is for the formation of liquid water, and the water is removed as liquid, then P_{con} must be taken as zero.

Example 4

Consider the Ballard fuel cell of Figure 9.29. What is the maximum power that can be transferred to a load, and what heat is generated? What is the efficiency of the cell? Use the V-I characteristics for 70 C, but, to simplify the problem, assume that the operating conditions are at RTP. The product water is removed from the cell in vapor form.

The V-J characteristic of the cells is

$$V_L = 0.913 - 49.3 \times 10^{-6} J. \tag{9.144}$$

Consequently, the power output is

$$P_L = V_L J = 0.913J - 49.3 \times 10^{-6} J^2 \text{ W m}^{-2}. \tag{9.145}$$

This is the power the cell delivers to a load per square meter of active electrode surface. The cell delivers maximum power when

$$\frac{dP}{dJ} = 0.913 - 98.6 \times 10^{-6} J = 0, \tag{9.146}$$

or

$$J = 9260 \text{ A/m}^2. \tag{9.147}$$

At this current, the cell would deliver 4230 W/m^2.

With 100% current efficiency, the rate of water synthesis is

$$\begin{aligned} \dot{N} = \frac{J}{q n_e N_0} &= \frac{9260}{1.60 \times 10^{-19} \times 2 \times 6.02 \times 10^{26}} \\ &= 48 \times 10^{-6} \text{ kilomoles } (\text{H}_2\text{O})\text{s}^{-1}\text{m}^{-2}. \end{aligned} \tag{9.148}$$

(*Continues*)

(*Continued*)

Hence, the energy input to the cell is

$$P_{in} = \Delta\overline{h}\dot{N} = 242 \times 10^6 \times 48 \times 10^{-6} = 11{,}600\,\mathrm{W/m}^2. \tag{9.149}$$

Of these, 4230 $\mathrm{W/m}^2$ appear as electric energy in the load, so that 11,600 – 4230 = 7370 $\mathrm{W/m}^2$ of heat are generated.

Notice that the Joule losses inside the cell amount to

$$P_{Joule} = R_{int}J^2 = 49.3 \times 10^{-6} \times 9260^2 = 4230\ \mathrm{W/m}^2. \tag{9.150}$$

This is, of course, equal to the power delivered to the load because, for maximum power transfer, the load resistance must equal the internal resistance of the generator.

The thermodynamic heat is

$$P_{therm} = T|\Delta\overline{s}|\dot{N} = 298 \times 44.4 \times 10^3 \times 48 \times 10^{-6} = 635\ \mathrm{W/m}^2. \tag{9.151}$$

The losses owing to V_{oc} being different from V_{rev} amount to

$$P_{oc} = (V_{rev}V_{oc})J = (1.185 - 0.913) \times 9260 = 2519\ \mathrm{W/m}^2. \tag{9.152}$$

Necessarily,

$$P_{in} = P_{therm} + P_{oc} + P_{Joule} + P_L. \tag{9.153}$$

The efficiency of this cell is

$$\eta = \frac{P_L}{P_{in}} = \frac{4230}{11{,}600} = 0.365. \tag{9.154}$$

In Example 4, the current delivered by the cell (9260 $\mathrm{A/m}^2$) exceeds the maximum current (7000 $\mathrm{A/m}^2$) in the Ballard data. This probably means that the cell cannot deliver all this power—that is, it cannot dissipate the 7400 $\mathrm{W/m}^2$ of heat it generates.

The efficiency of the cell, if operated at lower output levels, will be substantially larger than that in the extreme example above. In fact, at $J = 7000$, the efficiency would be about 45%.

9.8.8.1 Heat Removal from Fuel Cells

The operating temperature of a fuel cell depends on the type of cell. In any cell, if the temperature is to remain unchanged, an amount of heat power,

P_{remov}, equal to the generated heat power, P_{heat} must be removed from the device.

This heat removal can be accomplished passively or by using special heat exchange schemes. Heat can, for example, be removed by using a rate of air flow in excess of that which is required to satisfy the oxygen demand of the cell. This same stream of air may also be useful in removal of the reaction water.

Appendix: Batteries

A9.1 Introduction

Although Michael Faraday did demonstrate a mechanical electric generator in 1821 and, around 1832, several prototypical mechanical electric generators were built, one can date the beginning of widespread use of electric power to the work of Thomas Edison (1880) and Werner von Siemens (1881). Thus, when in 1859, Gaston Planté invented his lead–acid battery, there were no ready sources of electricity to be stored. Essentially, the Plante battery was a device designed to store the electricity generated by other batteries, a fact that did not promise a brilliant future for the invention. Nevertheless, the lead–acid battery has survived some 150 years and can be considered the most successful electricity storage device ever built. But now the tide is changing, and the growing need of more appropriate batteries has reinvigorated research into new chemistries. The need for higher gravimetric energy densities has turned our attention from one of the densest metals—lead—to the lightest—lithium. In between, we have sampled the periodic table of elements, trying mercury, cadmium, nickel, sulfur, sodium, and numerous other elements.

At present, lead–acid batteries have matured to the point where it is difficult to visualize any major improvement in their particular technology, while lithium is in the midst of rapid development.

A9.2 Capacity

In the operation of most vehicles, it is useful to know, at any given time, how far the vehicle can be driven before exhausting its energy supply. This is particularly true of the automobile. It would be ideal if an onboard instrument were capable of indicating with reasonable accuracy how many more kilometers the car could go before having to be refueled or recharged. This is clearly impossible because there is no way such an instrument can predict how the car will be driven—what speed, under what metereological conditions, under what road conditions, and whether the road is flat or is uphill or downhill. In internal combustion cars, we have come to accept a simpler instrument that measures the amount of energy left in storage, leaving the act of translating this energy into remaining range to the capacity of our brains to make such an estimate. In a liquid-fueled car, measuring stored energy is simply a matter of measuring fuel volumes (in general, fuel levels).

In electric, battery-driven, cars, the situation is much more complicated since one cannot determine the energy that can be extracted from a battery because this energy depends on the manner in which the battery is used. As a matter of fact, the state of charge (SOC) of a battery is almost invariably a measure of the amount of charge available, not of the amount of energy.

The amount of charge transferred by a charging device to a battery can be determined simply by having a record of the charging current and time,

$$C_{charge} = \int I dt. \tag{A.155}$$

The unit commonly used is **ampere-hours**, not the coulomb. C_{charge} is not the amount of charge taken up by the battery and, certainly, not the charge, C_{load}, that the battery delivers to the load. The ratio, C_{load}/C_{charge}, is variously called **charge acceptance** or **coulombic efficiency**, and is a function of the type and state of the battery, of the temperature and of the charging rate. Notice that we are talking about the ratio of *charges*, not of the ratio of energies, a quite diferent quantity. At its simplest, a battery can be modeled as a voltage source, V_{oc}, in series with an internal resistance, R_{int}; hence, the energy must include the I^2R_{int} battery looses, both in charge and discharge, and thus, is strongly dependent on the charging current, I.

The **capacity**[†], C_{load}, or simply C, is a somewhat fuzzy measure of how much charge a battery can deliver to a load. It is an imprecise number because it depends on temperature, age of the battery, state of charge, and rate of discharge. Formally, for a constant rate of discharge, I,

$$C = tI \quad \text{Ah}. \tag{A.156}$$

It has been observed that two identical, fully charged batteries, under the same circumstances, will deliver different charges to a load, depending on the selected discharge current. In other words, C is not constant, and the value of C for a fully charged battery is not an adequate description of the characteristic of the battery unless it is accompanied by an additional information: the **rated time of discharge** (assuming the discharge occurs under constant current).

[†] Capacity is measured in ampere-hours and has nothing to do with capacitance, measured in farads.

Peukert[†] showed, in 1897, that what is constant is

$$tI^n = constant, \tag{A.157}$$

where n, the **Peukert number**, is always somewhat larger than 1. For lead–acid batteries, a representative number is 1.2.

Peukert's law can be written as

$$tI \times I^{n-1} = \Lambda, \tag{A.158}$$

where Λ is a constant.

Consider an automotive battery with a Peukert number of 1.2, rated at 200 Ah when discharged at a uniform current for 20 hours that is, at a 10 A current. What is the capacity, C, of this battery if the discharge proceeds at 20 A?

From Equation A.158, we can calculate Λ, for $t = 20$ hr, $I = 10$ A, and $n = 1.2$,

$$\Lambda = 20 \times 10 \times 10^{1.2-1} = 317. \tag{A.159}$$

At $I = 20$ A,

$$C = \Lambda I^{1-n} = 317 \times 29^{1-1.2} = 174 \quad \text{Ah}. \tag{A.160}$$

We can easily find how long the above discharge will take simply by dividing its adjusted capacity by the discharge current:

$$t = \frac{C}{I} = \frac{174}{20} = 8.7 \quad \text{hr}. \tag{A.161}$$

Google has become a most valuable research tool, but by its very nature, it has also become a perpetuator of misconceptions. The vast majority of Google citations discussing Peukert's law, blandly state that the capacity of a battery is given by

$$C = tI^n, \tag{A.162}$$

where C is in ampere-hours, I in amperes, t in hours, and n is dimensionless.

Applying this equation to our previous example, we calculate,

$$t = \frac{C}{I^n} = \frac{200}{20^{1.2}} = 5.5 \quad \text{hr}. \tag{A.163}$$

(*Continues*)

[†] Possible pronunciation: *Poikert*, where the *e* sounds like the *ai* in *air*.

(*Continued*)

This differs substantially from the 8.7 hours we calculated before. The reason for this discrepancy is, of course, that Equation A.162, which states that the charge in ampere-hours is equal to the number of hours times the number of amperes raised to the exponent, n, is dimensionally incorrect.

A9.3 Ragone Plot

Consider a lead–acid battery with the following specifications:

$V_{oc} = 11.89 + 0.0074 \times SOC,$ where SOC is the state of charge in percent.
Capacity, $C = 80$Ah at a steady 4 A discharge rate (20-h discharge);
Internal resistance, $R_{int} = 0.0375\ \Omega$.
Peukert number $= 1.3$.
Mass, $m = 31\,\text{kg}$.
Volume, $\nu\ = 0.01\,\text{m}^3 (300 \times 170 \times 200\,\text{mm})$.

The total energy, W, the battery delivers to the load when operating for 20 hours at the rated current (4 A), is approximately

$$VI\Delta T = (V_{oc} - R_{int}I)I\Delta T, \tag{A.164}$$

which assumes that V_{oc} does not change as the battery is discharged (not quite true!). Taking $V_{oc} = 12.5$ V, we have

$$W = (12.5 - 0.0375 \times 4) \times 4 \times 80 \approx 4000 \quad \text{Wh}. \tag{A.165}$$

This confers to the battery a gravimetric energy density of $4000/31 = 129$ Wh/kg and a volumetric energy density of $4000/0.01 = 400{,}000$ Wh/m^3

What is the peak power that the fully charged battery can deliver to a load?

When fully charged ($SOC = 100$), the load voltage of the battery is

$$V = 11.89 + 0.0074 \times 100 - 0.0375I = 12.63 - 0.0375I, \tag{A.166}$$

the power delivered to the load is

$$P_L = (12.63 - 0.0375I)I = 12.63I - 0.0375I^2, \tag{A.167}$$

and the maximum power it can deliver can be found from

$$\frac{P_L}{dI} = 12.63 - 0.075I = 0, \tag{A.168}$$

from which, $I_{max} = 72.2 \quad \text{A}$. The maximum power is

$$P_{L_{max}} = 12.63 \times 72.2 - 0.0375 \times 72.2^2 = 738 \text{ W}. \tag{A.169}$$

The gravimetric power density of the battery is 738/31 $\approx$ 24 W/kg, and the volumetric power density is $738/0.01 = 73.8\,\text{kW/m}^3$.

Now let us examine a different type of electric energy storage system—a super capacitor. Maxwell produces a number of such units, among which is the BCAP 3000, which has the following specifications:

$V_{max} = 2.7\,\text{V}$.
Capacitance, $C = 3000\,\text{F}$.
Internal resistance (dc), $R_{int} = 0.00037\,\Omega$.
Volume, $\nu \; = 0.000475\,\text{m}^3$.
Mass, $m = 0.55\,\text{kg}$.

The energy stored in a capacitor is

$$W = \frac{1}{2}CV^2. \tag{A.170}$$

However, for practical reasons, the useful discharge must be stopped when the voltage drops to, say, $\frac{1}{2}V_{max}$. Thus the available energy is

$$\begin{aligned} W_{avail} &= \frac{1}{2}CV_{max}^2 - \frac{1}{2}C\left(\frac{V_{max}}{2}\right)^2 \\ &= \frac{3}{8}CV_{max}^2 = \frac{3}{8} \times 3000 \times 2.7^2 = 8{,}200 \quad \text{J} = 2.3 \text{ Wh}. \end{aligned} \tag{A.171}$$

The gravimetric energy density is 2.3/0.55 = 4.2 Wh/kg.

The maximum power is transferred to a load that matches the internal resistance of the capacitor,

$$P_{max} = \frac{2.7^2}{4 \times 0.000375} = 4{,}900 \text{ W}, \tag{A.172}$$

or a gravimetric power density of 8.8 kW/kg.

The results for the two different storage systems have been collected in the following table.

Storage system	**Energy density** **Wh/kg**	**Power density** **W/kg**
Battery	129	24
Capacitor	4.2	8,800

It is immediately obvious that these two storage systems behave quite differently. If you need a lot of energy delivered gradually (driving your car on a long highway, for instance), take the battery; if you need a powerful burst of energy lasting a short time (passing a driver on the highway), take the capacitor.

Let us include fuel cells in our tabulation. Ballard produces an automotive fuel cell, Mark1100, with the following characteristics:

Maximum net power: 110 kW continuous.
Maximum current: 440 A.
DC Voltage: 250 V.
Fuel: Gaseous hydrogen, commercial grade.
Oxidant: Air.
Temperature (maximum): 90 C.
Freeze start capability: -20 C.
Fuel pressure (nominal): 1.2 to 1.7 atmos (gauge).
Length $\times$ width $\times$ height: $804 \times 485 \times 210$ mm.
Weight: 120 kg.
Volume: 82 liters.

To estimate the gravimetric energy density, we have to decide how much fuel we are going to consider as part of the system. If, for instance, we decide we need to use the fuel cell for five continuous hours at full power (550 kWh), then, assuming 50% efficiency, we will need about 4 GJ of hydrogen or 28 kg. If the storage system is either a hydride system or a compressed gas system, some 6 kg of hydrogen can be carried in 100 kg of container plus gas. Thus, the 28 kg of hydrogen will add a mass of $100 \times 28/6 = 467$ kg to the system. This plus the 120 kg of fuel cell will amount to 567 kg and results in a gravimetric energy density of $550{,}000/587 = 937$ Wh/kg.

But if we want more fuel, then the energy density improves asymptotically to $550{,}000/467 = 1180$ Wh/kg.

The gravimetric power density for this cell when an infinite amount of fuel is supplied as above, that is, when we count only the mass of the fuel storage system, is $110{,}000/467 = 235$ W/kg. The table below tabulates the energy density and the power density of the three storage systems being compared

Storage system	**Energy density Wh/kg**	**Power density W/kg**
Battery	129	24
Capacitor	4.2	8800
Fuel cell	1180	235

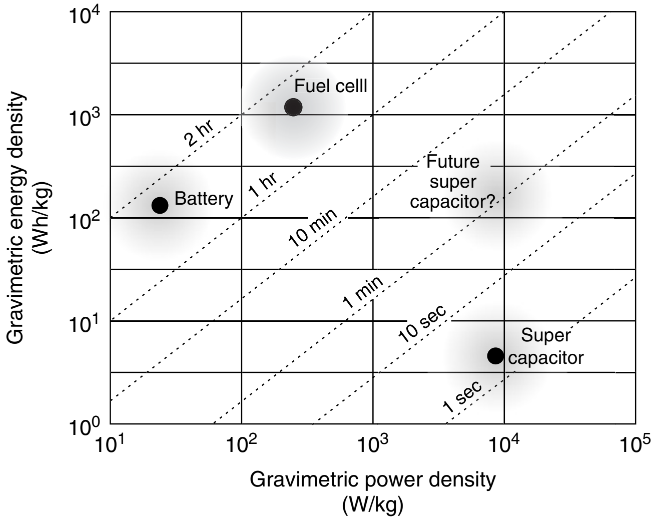


Figure 9.38 A Ragone plot reveals at a glance the relative characteristics of different energy storage systems. Ideally, one should try for the northeast corner of the chart.

To compare the performances of different energy storage system, let us plot their gravimetric energy density versus their gravimetric power density using logarithmic scales in what is called a **Ragone plot**, as we did in Figure 9.38. The ratio of the energy density (Wh/kg) to the power density (W/kg) is the number of hours the storage source can deliver full power (assuming a Peukert number of 1). On the plot, we have indicated, in dotted lines, constant discharge time isopleths. Efforts are being made to develop supercapacitors with capacitances some three orders of magnitude larger than those now available. This is not a physical impossibility and would solve an enormous problem in the electric energy storage technology.

A9.4 The Chemistry of Some Batteries

A9.4.1 Primary Batteries

There is no better way to start the description of how some electrochemical batteries are put together than to flash back to the year 1800, when the Italian scientist, Alessandro Volta (1745–1827), made one of the most significant breakthroughs in modern technology. Prior to Volta, electricity was generated only by electrostatic machinery—high voltages were created, but the currents were extremely fleeting, to the point that their effects could not be properly investigated. What Volta gave the world was a device capable of delivering sustained currents, thereby opening a whole new chapter of physics and technology.

(1800) Volta's Cell

Volta's cell consisted of a zinc disk separated from a silver (or copper) disk by a sheet of paper soaked in salt. The voltage generated was quite low (especially in view of the insensitive measuring instruments available at the time). This led to the stacking of many cells in series forming a "pile" or battery—actually, the first bipolar configuration ever built.

A "Volta" cell can be made by dipping a zinc and a copper electrode into a dilute (say, 10%) sulfuric acid solution. The zinc will oxidize:

$$\mathrm{Zn} \rightarrow \mathrm{Zn}^{++} + 2\,\mathrm{e}^{-}, \tag{A.173}$$

providing electrons for the external current; the zinc electrode becomes the anode of the cell. Zinc ions are soluble in water. The sulfuric acid, being a strong acid,[†] will mostly dissociate into ions:

$$\mathrm{H_2SO_4} \rightarrow 2\mathrm{H}^{+} + \mathrm{SO_4^{--}}. \tag{A.174}$$

The zinc ions combine with the sulfate ions, forming zinc sulfate. The protons, in the form of hydronium, $\mathrm{H^+(H_2O)}_x$, migrate through the electrolyte to the copper where they are reduced to hydrogen (by combining with the electrons arriving via the external circuit) and evolve as gas bubbles. The overall reaction is

$$\mathrm{Zn} + \mathrm{H_2} \rightarrow \mathrm{ZnSO_4} + \mathrm{H_2}. \tag{A.175}$$

By generating a high-energy chemical such as hydrogen and simply discarding it, the Volta cell cannot be a very efficient way of converting chemical energy into an electric one.

Two major difficulties further reduce the practicality of the Volta cell:

1. Soon after the current starts flowing, the copper electrode is covered with adhering small hydrogen bubbles that constitute an insulating layer severely limiting the current delivered. This is called **polarization**.
2. Even in the absence of current, the zinc is usually, corroded by the sulphuric acid, greatly reducing the battery's shelf life.

[†] The strength of an acid is a measure of the degree of its dissociation when in aqueous solution. Hydrochloric acid dissociates completely into $\mathrm{H^+}$ and $\mathrm{Cl^-}$; it is a very strong acid. Sulfuric acid is weaker, but is still a strong acid. Surprisingly, hydrofluoric acid, in spite of its corrosiveness, is a weak acid: when in water solution at room temperature, the concentration of $\mathrm{H^+}$ is less than 3% of the concentration of neutral HF molecules.

(1836) Daniell Cell

One way to avoid polarization is to stay away from reactions that cause hydrogen to evolve. John Frederic Daniell, a British chemist (1790–1845), did just that when he invented his eponymous cell.

Daniell cells were popular in the nineteenth century as a source of electricity, especially in telegraph systems. These cells consisted of a container divided into two compartments by a membrane permeable to ions. In one compartment, a zinc electrode was dipped in a zinc sulfate solution, and, in the other, a copper electrode was immersed in a copper sulfate solution.

The zinc anode oxidizes (i.e., loses electrons)

$$\mathrm{Zn} \rightarrow \mathrm{Zn}^{++} + 2\mathrm{e}^{-}. \tag{A.176}$$

The zinc is eroded, going into the solution in the form of ions.

At the other electrode (the cathode), the copper is reduced (i.e., the copper ions in the sulfate accept electrons) and the resulting metallic copper plates out onto the copper electrode:

$$\mathrm{Cu}^{++} + 2\mathrm{e}^{-} \rightarrow \mathrm{Cu}. \tag{A.177}$$

The cell will deliver current until it runs out of either zinc or sulfate, whichever is less. See Problem 9.2.

(1865) Leclanché Cell

Another way to avoid polarization is to interpose between the electrolyte and the cathode a substance capable of absorbing the hydrogen produced. The most popular avatar of this solution is the cell invented by the French electrical engineer, Georges Leclanché (1839–1882), that evolved into the very successful zinc–carbon cell, which until recently was the most common of all small batteries.

Zinc–carbon Cell

The zinc–carbon cell, just as all the previous described cells, used zinc as an anode and manganese dioxide (MnO_2) as a depolarizer. The cathode was a graphite rod surrounded by a bag full of a mixture (about 50/50) of manganese dioxide and powdered carbon. The carbon was added to increase the conductivity of the depolarizing region. The electrolyte is a mixture of zinc chloride, $ZnCl_2$, and ammonium chloride, NH_4Cl. One way to look at the complicated reactions in this type of cell is

Anode:

$$\mathrm{Zn} \rightarrow \mathrm{Zn}^{++} + 2\mathrm{e}^{-}. \tag{A.178}$$

The ammonium chloride in its aqueous solution dissociates

$$NH_4Cl \rightarrow NH_4^+ + Cl^-, \tag{A.179}$$

followed by

$$NH_4^+ \rightarrow NH_3 + H^+. \tag{A.180}$$

The ammonia combines with the zinc ion forming a zinc complex ion,

$$NH_3 + Zn^{++} \rightarrow Zn(NH_3)_2^{++}. \tag{A.181}$$

The cathode reaction gets rid of the hydrogen:

$$2MnO_2 + 2H^+ + 2e^- \rightarrow Mn_2O_3 + H_2O, \tag{A.182}$$

which all leads to the overall reaction:

$$Zn + 2MnO_2 + 2NH_4^+ \rightarrow Mn_2O_3 + Zn(NH_3)_2^{++}. \tag{A.183}$$

Alkaline Batteries

An improvement in maximum current and in the shelf life of the zinc–carbon battery, together with a better energy density, was achieved by the introduction of alkaline batteries, which, instead of using the zinc chloride/ammonium chloride electrolyte, use an alkaline, KOH, solution for this purpose. In addition, instead of a solid zinc anode, these cells used a powdered form of the metal, which has a much higher surface-to-volume ratio, thus permitting larger peak currents.

The anode reaction oxidizes the zinc-producing electrons that will constitute the useful current delivered by the battery,

$$Zn + 2OH^- \rightarrow ZnO + H_2O + 2e^-. \tag{A.184}$$

The cathode reaction reduces the manganese(IV) oxide, MnO_2, to manganese(III) oxide, Mn_2O_3,

$$2MnO_2 + H_2O + 2e^- \rightarrow Mn_2O_3 + 2OH^-. \tag{A.185}$$

This leads to an overall reaction,

$$Zn + 2MnO_3 \rightarrow ZnO + Mn_2O_3. \tag{A.186}$$

Zinc

From the preceding summary description of the various types of cells, it can be seen that zinc is the preferred anode material. Zinc has one major shortcoming, however: it is corrodible even when not delivering any current.

True, perfectly pure zinc is consumed only when a current is drawn, but the presence of impurities causes corrosion of the electrode even when the cell is inactive (the impurities form numerous microscopic electrochemical cells within the mass of the metal). To insure a long shelf life, the zinc is alloyed with mercury (**amalgamated**).

A9.4.2 Secondary Batteries

(1859) Lead–Acid

This extraordinary, first ever rechargeable battery, invented by the French physicist, Gaston Planté (1834–1889), has withstood the competition of numerous other batteries and is still in large-scale production after more than 150 years of use. Only the modern lithium ion cells seem to pose serious competition to Planté's invention.

Lead–acid batteries have survived mainly because they are a low-cost, high power-to-mass device, capable of delivering large surge currents, albeit for a short time owing to their unfavorable energy-to-mass ratio. (See A.3, Ragone Plots in Figure 9.38.)

The cell should be called a lead–lead oxide cell: the anode is metallic lead and the cathode is PbO_2. The electrolyte is a strong sulfuric acid solution that dissociates

$$H_2SO_4 \rightarrow 2H^+ + SO_4^{--}. \qquad (A.187)$$

The anode reaction (during discharge) is

$$Pb + SO_4^{--} \rightarrow PbSO_4 + 2e^-. \qquad (A.188)$$

The cathode reaction (during discharge) is

$$PbO_2 + SO_4^{--} + 4H^+ \rightarrow PbSO_4 + 2H_2O - 2e^-, \qquad (A.189)$$

so that the overall reaction during the discharge is

$$Pb + PbO_2 + 2H_2SO_4 \rightarrow 2PbSO_4 + 2H_2O. \qquad (A.190)$$

It can be seen that water is produced during the discharge, diluting the electrolyte. The degree of charge can therefore be determined by measuring the density of the electrolyte.

The original Planté battery was made by immersing fairly thick lead plates into a sulfuric acid solution. At this stage, the battery is entirely symmetrical—the two electrodes are identical. If, now, having selected one electrode at random as the anode, you drive a current through the system, you will find that, after a while, you have accumulated a certain amount of charge. Repeating the process a number of times gradually increases the capacity of the cell because the electrodes become more and more spongy, offering a much bigger surface area to the reaction. From the standpoint of mass production, this is an expensive way to make an accumulator. In 1881, Camille Fauré introduced a modification of the Planté cell that made it suitable for large-scale production. Fauré replaced the solid lead cathode by a latticework of lead filled with lead oxide, thus "preforming" this electrode.

The integrity of the electrode is somewhat compromised because little pieces of lead oxide are easily detached and fall to the bottom of the cell where they accumulate as a layer of "mud," eventually shorting out the device. Modern cells leave an ample volume at the bottom of the cell in which this mud accumulates without causing a short-circuit.

(1898) Nickel–Iron

In the early competition between steam, internal combustion, and electric automobiles, the electric suffered (as do modern electric cars) from a limited range. These cars had a large fraction of their weight in lead–acid batteries. Thomas Edison, (1847–1931), using his usual empirical approach, tested many materials in search of elements for a battery that would outperform the ones currently being used. He finally settled on an iron–nickel oxide combination which had a moderate success, being quite robust in terms of resistance to overcharge and to prolonged times of idleness.

The discharge reactions are

Anode:

$$\mathrm{Fe} + 2\mathrm{OH}^- \rightarrow \mathrm{Fe(OH)}_2 + 2\mathrm{e}^-, \qquad \text{(A.191)}$$

Cathode

$$2\mathrm{NiO(OH)} + 2\mathrm{H_2O} + 2\mathrm{e}^- \rightarrow 2\mathrm{Ni(OH)}_2 + 2\mathrm{OH}^-, \qquad \text{(A.192)}$$

Overall

$$2\mathrm{NiO(OH)} + \mathrm{Fe} + 2\mathrm{H_2O} \rightarrow 2\mathrm{Ni(OH)}_2 + \mathrm{Fe(OH)}_2. \qquad \text{(A.193)}$$

(1899) Nickel–Cadmium

In Sweden, Waldmar Jungner in 1899, modified Edison's battery by replacing the iron in the anode by cadmium. The resulting improved performance led to a configuration that was still being sold a few years ago.

NiCd batteries were particularly susceptible to "memory effect" or "voltage depression," a phenomenon causing the voltage of the cell to drop faster than expected during the initial phase of discharge, leading to an apparent reduction of capacity. This is a reversible phenomenon that can be corrected by a deep discharge followed by a normal recharging.

The discharge reactions are as follows:

Anode:

$$\mathrm{Cd} + 2\mathrm{OH}^- \rightarrow \mathrm{Cd(OH)}_2 + 2\mathrm{e}^-, \qquad (A.194)$$

Cathode:

$$2\mathrm{NiO(OH)} + 2\mathrm{H_2O} + 2\mathrm{e}^- \rightarrow 2\mathrm{Ni(OH)}_2 + 2\mathrm{OH}^- \qquad (A.195)$$

Overall:

$$2\mathrm{NiO(OH)} + \mathrm{Cd} + 2\mathrm{H_2O} \rightarrow 2\mathrm{Ni(OH)}_2 + \mathrm{Cd(OH)}_2. \qquad (A.196)$$

(1986) Nickel–Metal Hydride Battery (NiMH)

The Lithuanian born American engineer, Stanford Ovshinsky (1923–), made a significant alteration to the nickel–cadmium battery by using hydrogen in place of cadmium in the anode.

A metal alloy anode is immersed in a 30% (by weight) solution of KOH in water. This strong alkali dissociates almost completely into K^+ and OH^- ions. Hydrogen supplied to this electrode will combine with the hydroxyl ion forming water and liberating an electron, which becomes available to circulate through an external load constituting the useful output of the cell:

$$\mathrm{H} + \mathrm{OH}^- \rightarrow \mathrm{H_2O} + \mathrm{e}^-. \qquad (A.197)$$

The cathode is made of nickel oxyhydroxide (NiO[OH]) which, upon receiving an electron, is reduced to nickel hydroxide, liberating a hydroxyl ion:

$$\mathrm{NiO(OH)} + \mathrm{H_2O} + \mathrm{e}^- \rightarrow \mathrm{Ni(OH)}_2 + \mathrm{OH}^-. \qquad (A.198)$$

The resulting overall reaction is

$$\mathrm{NiO(OH)} + \mathrm{H} \rightarrow \mathrm{Ni(OH)_2}. \qquad (A.199)$$

Notice that Equations A.191, A.194, and A.197 are the same except for the different anode materials used (Fe→Cd→H). The same happens with Equations A.192, A.195, and A.198, and Equations A.193, A.196, and A.199. If the stoichiometry of Equations A.197 and A.198 seems different, it is because hydrogen is monovalent, whereas iron and cadmium are divalent.

One interesting consequence of using hydrogen as the anode material is that the anode reaction during discharge produces water. This compensates the water consumption by the cathode. Hence, in the NiMH battery the water level never changes, in contrast with the NiFe and NiCd batteries in which, during discharge, water is consumed at both electrodes, causing the volume of the cell to change and putting a strain on the container.

The OH^- concentration remains constant during the discharge because this ion is produced at the anode and consumed at the cathode.

The materials consumed—hydrogen and nickel oxyhydroxide—are regenerated during the charge when an external power supply forces electrons into the metal electrode. This causes the water to be electrolyzed into H^+ and OH^-. At the other electrode, the nickel hydroxide is oxidized into nickel oxyhydroxide-consuming hydroxyl ions.

Use of an hydrogen anode is attractive, but how can it be realized? In particular, how can one store the hydrogen generated during the charge phase? The answer is to use in the anode a metal alloy that can, by forming an easily reversible hydride, absorb and store the gas so none is evolved during charge, but becomes available on demand during discharge. This **hydride hydrogen storage** is discussed in some detail in Chapter 11.

Material requirements of the NiMH battery are complicated. The energy storage capacity of the battery depends on the amount of hydrogen that can be absorbed by the metal alloy electrode. It is necessary that the hydride formation be easily reversible (see Chapter 11), and this is determined by the enthalpy of formation of the hydride that must fall in the 25 to 50 MJ/kmole range. If the enthalpy of formation is too small, hydrogen will fail to react with the alloy and gas will evolve. If too large, the electrode will be oxidized. Surface properties of the metal alloy are critical in determining the catalytic activity, electric conductivity, and the porosity and area of the surface.

Nominal voltage delivered by each element of a NiMH battery is 1.2 V, some 60% that of a lead–acid cell.

A few years ago, the nickel–metal hydride battery had a clear edge on its competitors and was the technology of choice for powering camcorders, cellular phones, laptop computers, and other small electronic devices. This period of dominance was brief—the more advanced lithium ion technology has now cornered the market. The same is happening, at a more leisurely rate, in the field of automotive traction batteries. However, at present, the most popular hybrid automobile on the market—the Toyota Prius—still uses NiMH batteries.

Lithium Ion

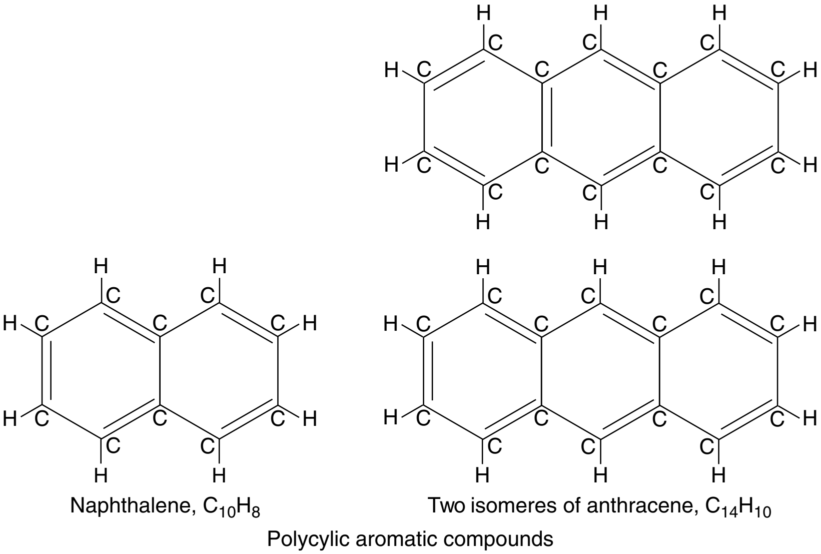
To help in understanding the operation of modern lithium–ion batteries, it is useful do become acquainted with the concept of intercalation compounds. Since graphite is a commonly used host of such compounds, we will say a few words about this material.

Graphene, Graphite, and Intercalation Compounds

In Chapter 13 we discuss the concept of aromaticity, typified by benzene. We show that the benzene ring with its six carbons is perfectly hexagonal; that is, all carbon-carbon bonds are of identical length. However, from the valence point of view, three of the six bonds should be double (hence shorter than the three single bonds). This would lead to a distorted hexagon. This discrepancy can be explained (read ahead in Chapter 13) by assuming that some of the bonding electrons are **delocalized**, a situation that characterizes **aromaticity**. There are many aromatic compounds, some containing one or more benzene-like rings. (see Figure 9.39).

In the figure we show two **planar-fused polyclycic aromatic** hydrocarbons. Chemists, for simplicity, omit in the diagram the hydrogen and the corresponding bonds and do not label the atoms with their corresponding symbols, H and C, as we have done.

Each of the rings has delocalized electrons, that is, is aromatic. However, the degree of aromaticity is variable. If all rings were a complete benzene ring, there would be some carbons with five bonds. To keep the number of bonds in each carbon at the correct value of four, the aromaticity of some rings has to be reduced. In many cases, this


Naphthalene, $C_{10}H_8$

Two isomeres of anthracene, $C_{14}H_{10}$

Polycylic aromatic compounds

Figure 9.39 Two planar fused polycyclic aromatic compounds: naphthalene and anthracene. Two isomeres of the anthracene are shown.

(*Continues*)

(*Continued*)

creates different isomers. See anthracene in the figure for which two different isomers are displayed. The one on top has the fully aromatic (benzene) ring in the middle of the molecule; the one on the bottom has this ring on the end of the molecule.

In the case of linear molecules as displayed, the hydrogen-to-carbon ration is

$$\frac{\mathrm{H}}{\mathrm{C}} = \frac{3+n-1}{3+2(n-1)}, \qquad \text{(A.200)}$$

where n is the number of rings. It can be seen that for a single ring (benezene), H/C= 1, and for an infinite number of rings, H/C = 0.5. However, if the molecule contains more than one row, some of the internal rings do not bind to hydrogens and the H/C ratio becomes much smaller. We can imagine a very large multirow planar molecule as depicted in Figure 9.40. In the limit, when $n \to \infty$, the H/C ratio tends toward zero. We have a planar polyclyclic aromatic hydrocarbon that, in reality has no hydrogen. It constitutes an essentially two-dimensional molecule and is called **graphene**.

Graphene is extremely strong in the directions along its plane. It is about one millionth of the thickness of a common page of paper! It has extraordinary properties and promises important technological applications. A graphene sheet rolled up into a cylinder forms a **carbon nanotube**, and if rolled up into a ball, it forms a **buckyball**. Here, we are interested in it because **graphite** is a stack of graphene sheets very weakly bound to the next solely by Van der Waals forces. It is

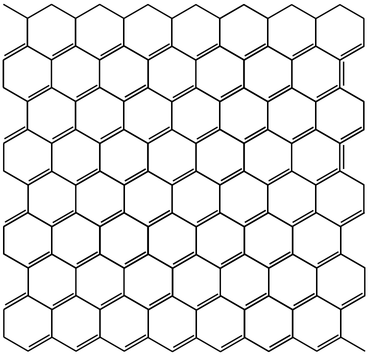


Figure 9.40 Graphene can be considered as the limiting case (when the number of rings tends toward infinity) of a general planar polycyclic aromatic hydrocarbon.

(*Continues*)

(*Continued*)

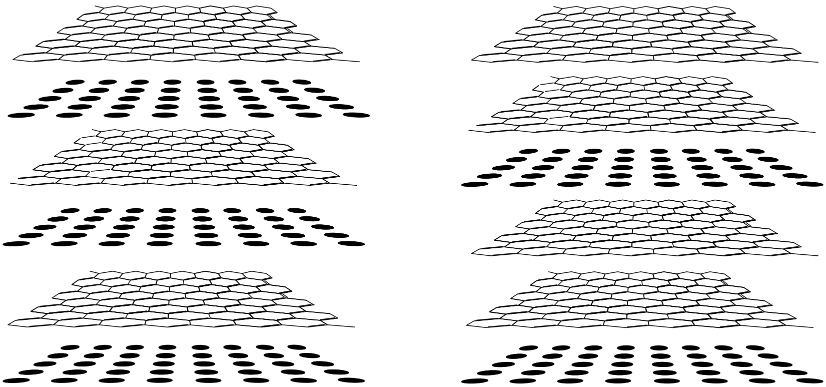


Figure 9.41 In stage 1 intercalation, graphene and guest layers alternate, (left), while in stage 2, there are 2 graphene layers for each guest layer (right), and in stage 3, there are 3, and so on.

very anisotropic and has other interesting properties, among which is its ability to form **intercalation compounds**.

Intercalation compounds are

> Compounds resulting from reversible inclusion, without covalent bonding, of one kind of molecule in a solid matrix of another compound, which has a laminar structure. The host compound, a solid, may be macromolecular, crystalline or amorphous.
>
> From the IUPAC Compendium of Chemical Terminology 2nd Edition (1997).

Graphite is a perfect host for intercalations, which can occur in different **stages**. (see Figure 9.41).

Notice that the guest molecules form flat sheets that often impose only moderate change in the structure of the host. The host acts as a container in which the guest can be (reversibly) stored. This possibility is exploited in the construction of lithium–ion cells.

For a modern battery, lithium is an obvious choice. It is the lightest of all metals (it has about half the density of water), and it has the highest reduction potential of all elements: −3.05 V (see the box "Cell Potential" in Section 9.2). If it were possible to pair lithium against fluorine, the cell would generate 5.92 V. This could lead to a cell with a very large gravimetric energy concentration, far surpassing that of any present-day cells.

The 1970s saw the commercialization of primary cells using metallic lithium anodes. However, all attempts to produce a rechargeable cell with this kind of anode failed, sometimes catastrophically, owing to the reactivity of the metal with electrolytes. A solution to this difficulty was found by using nonaqueous electrolytes and by replacing the lithium metal by graphite into which lithium ions could be inserted (see the preceding box on intercalation compounds).

Intercalation anodes had to be paired with appropriate cathodes, and it was found that Li_xCoO_2, Li_xNiO_2, or Li_xMnO_2 were acceptable and generated high voltages, in the neighborhood of 4 V. Although the realizable gravimetric charge concentration of such batteries was about the same as that of the NiCd ones, the high-output voltage gave the lithium cells a better than two-to-one advantage in energy concentration.

Figure 9.42 illustrates the fact that in any rechargeable battery, the essential action is the shuttling of ions back and forth though the electrolyte. Lithium-ion batteries are no exception. During charge, lithium ions are produced in the cathode by oxidation of a lithium-rich molecule. The ions then move through the electrolyte and are stashed in the anode into which they are inserted via an intercalation reaction. During discharge, the ions are extracted from the anode and driven to the cathode, which is then reduced.

The electrolyte is, typically, a lithium salt dissolved in an organic solvent. This would normally cause the solvent decomposition at the anode. However, during the initial charge, the lithium salts are deposited over this electrode, plating it with a solid layer of the otherwise disolved salt. This layer acts as a good conductor of lithium ions and prevents further decomposition of the solvent. Graphite, owing to its property as a good intercalation host, is the usual anode of current cells.

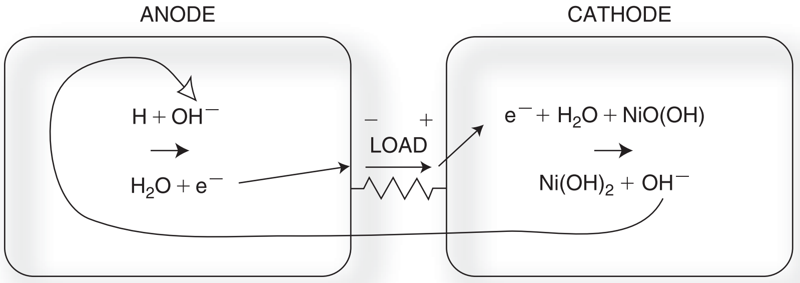


Figure 9.42 In a NiMH battery, during discharge, the hydroxyl ions move through the electrolyte from anode to cathode, while during the charge, they move in the opposite directions. In all batteries, ions shuttle back and forth between the electrodes.

A number of cathodes have been tried, each leading to a given average cell voltage . A common cathode material is lithium cobalt oxide, $LiCoO_2$, which, combined with a graphite anode, produces a cell with an average 3.7 V output. The gravimetric capacity of this cathode is 140 mAh/g.[†]

With the graphite/lithium cobalt oxide combination, the discharge reaction is

Anode:

$$x\mathrm{Li}^{+} + x\mathrm{e}^{-} + 6\mathrm{C} \rightarrow \mathrm{Li}_{x}\mathrm{C}_{6}. \tag{A.201}$$

Cathode:

$$\mathrm{LiCoO}_2 \rightarrow \mathrm{Li}_{1-x}\mathrm{CoO}_2 + x\mathrm{Li}^{+} + x\mathrm{e}^{-}. \tag{A.202}$$

Current lithium–ion batteries present a marked advantage over their predecessors. They have good gravimetric energy density, low self-discharge, and no memory problems. Yet, among other deficiencies, they have limited life and adverse aging (their performance decays with time whether in use or on the shelf).

The *Hindenburg* accident in 1937 gave a (undeserved) bad name to hydrogen. The explosion that occurred in a number of Li–ion batteries has done the same for these cells. One must remember that in a battery, the idea is to cram as much energy as possible in a very small volume. Li–ion batteries do this well and, consequently, are dangerous. The *Hindenburg* had hydrogen at, presumably, atmospheric pressure; this amounts to a bit more than 10 MJ/m^3. A small Li–ion battery in my camera delivers 740 mAh at 3.7 V and has the miniscule volume of 6 cm^3. This is more than 1 GJ/m^3, two orders of magnitude more than the energy density in the *Hindenburg.* Li–ion batteries deserve to be treated with respect!

In any battery, a short-circuit will lead to a rapid temperature rise. In the case of the sealed Li–ion cell, this will cause the internal pressure to rise to the bursting point or, if a vent is provided, to a venting of vaporized electrolyte. This vapor, being flammable, can be ignited, and the cell will form a small torch that can set fire to the other cells in the battery leading to spectacular explosions.

To in-sure safety, battery chargers must monitor the charging current and voltage and also the cell temperature. Since a deep discharge will damage the battery, there is both an upper and a lower limit to the cell voltage. Charging must be done according to strict and exact procedures. Typically,

[†] Notice the awkward units used in present-day literature. It would be much better if the gravimetric capacity were given in coulombs per kilogram (It would be C/kg). 1 mAH/g = 3600 C/kg.

at the beginning of charge, a constant current of, say, 1 C or slightly less is applied until the voltage reaches something like 4.2 V. The charger then switches to a constant voltage regimen, which stops when the charging current falls to 0.07 C or a bit more. The exact values are determined by the battery manufacturer and must be adhered to within 1%. The charger also monitors battery temperature (each cell is equipped with a thermistor) and the duration of the charge. Excess temperature or excess time will abort charging. Many batteries have four terminals—two are the usual "plus" and "minus" connections, one leads to the thermistor, and the fourth is connected to identification resistors that tell the charger what the rated capacity of the battery is.

To avoid damaging deep discharges, "smart" Li–ion batteries have built-in voltage monitoring circuits. In this case, the batteries suffer a modest self discharge when not in use.

One must recognize that the Li–ion battery is still in an early state of development and that major improvements appear to be possible. It is interesting that it is exactly in the one characteristic in which the batteries excel that major improvements can be expected in the near future. There are reasons to believe that major increases in energy density (possibly accompanied by much extended life) can be hoped for.

One of the limitations to energy density is the discharge capacity of graphite anodes currently used in most commercially available cells, which is, theoretically, 372 mAh/g. Silicon, on the other hand, has the highest known theoretical charge capacity—4200 mAh/g, more than 10 times that of graphite. However, silicon, as such, cannot be used because full insertion of lithium ions causes a 400% increase of volume and consequent pulverization of the anode, causing a dramatic battery capacity fading with each charge–discharge cycle. This is true even if small silicon particles of micrometer size are used. After cycling, the particles break down in size, and a large interface gap between adjacent grains results in unacceptable resistance to current flow.

Candace K. Chan et al. (2007) of Stanford University have come up with a very promising solution to the problem. They have constructed anodes made of silicon nanowire grown directly on stainless steel current collectors. These essentially unidimensional wires are thin enough to be immune to the strain caused by repeated insertion and extraction of lithium ions. The contact between wires and collector is robust, and the carrier transport along these wires is unaffected by cycling. Apparently, the technology for growing a forest of such wires rooted on the stainless steel current collector is amenable to simple mass production.

The capacity measured on first charging was 4277 mAh/g (the theoretical value, within experimental limits). However, this fell to a stable 3500 mAh/g in the next 20 cycles when the charge rate was C/5 (one-fifth of the rate capacity). Even at C/1, the capacity held at over 2000 mAh/g, five times better than the theoretical value for graphite.

Clearly, fixing the anode problem is not enough. Improvements can also be expected in the cathode and the electrolyte. The use of polymer electrolytes has already allowed considerable improvement on cells currently in production.

References

Adams, A. M., F. T. Bacon, and R. G. H. Watson, The high pressure hydrogen-oxygen cell, Chapter 4 of *Fuel Cells*, ed. Will Mitchell, Jr. Academic Press, **1963**.

Bard, A. J. and L. R. Faulkner, *Electrochemical Methods*, John Wiley, **2001**.

Barsoukov, E. and J. R. Macdonald, *Impedance Spectroscopy*, John Wiley, 2nd ed., **2005**.

Bessette, N. F., and J. F. Pierre, Status of Siemens Westinghouse tubular solid oxide fuel cell technology and development program, *Fuel Cells—Powering the 21st Century*, October 30–November 2, **2000**.

Boysen, Dane A., T. Uda, C. R. I. Chisholm, and S. M. Haile, High-performance solid acid fuel cells through humidity stabilization, Science **303**, p. 68, January 2, **2004**.

Chan, Candace K., H. Peng, G. Liu, K. McIlwrath, X. F. Zhang, R. A. Huggins, and Y. Cui, High-performance lithium battery anodes using silicon nanowires, *Nature*, Advanced on line publication **(2007)**.

Chu, Deryn, R. Jiang, and C. Walker, Methanol tolerant catalyst for direct methanol fuel cell applications, *Fuel Cells—Powering the 21st Century*, October 30–November 2, **2000**.

Dodelet, J. P., M. C. Denis, P. Gouérec, D. Guay, and R. Scholz, CO tolerant anode catalysts for fuel cells made by high energy ball-milling, *Fuel Cells—Powering the 21st Century, Fuel Cell Seminar*, October 30–November 2, **2000**.

Dohl, H., S. von Adrian, J. Divisek, B. Höhlein, and J. Meusinger, Development and process analysis of direct methanol fuel cell systems, *Fuel Cells—Powering the 21st Century*, October 30–November 2, **2000**.

Fuller, Timothy A., Larry J. Chaney, Dr. Tom L. Wolf, Jim Kesseli, James Nash and Joseph J. Hartvigson, A novel fuel cell/microturbine combined-cycle system, 2000. Fuel Cell Seminar, Portland, Oregon, October–November 2000.

Forbes, C. A., and J. F. Pierre, The solid fuel-cell future, *IEEE Spectrum*, October p. 40, **1993**.

Ghosh, D., M. E. Pastula, R. Boersma, D. Prediger, M. Perry, A. Horvath, and J. Devitt, Development of low temperature SOSF systems for remote power and home cogen applications, p. 511, *Fuel Cells—Powering the 21st Century, Fuel Cell Seminar*, October 30–November 2, **2000**.

Goldstein, Rocky, *Proton Conductors for Solid Electrolyte Fuel Cells*, Exploratory Research Letter, Electric Power Research Institute (EPRI), Palo Alto, CA.

Haberman, William L., and James E. A. John, *Engineering Thermodynamics with Heat Transfer*, Allyn and Bacon, **1989**.

Hibino, Takashi, A. Hashimoto, T. Inoue, J. Tokuno, S. Yoshida, and M. Sano, A low-operating-temperature solid oxide fuel cell in hydrocarbon-air mixtures, *Science* **288**, p. 2031, June 16, **2000**.

Koppel, Tom, *Powering the Future (The Ballard Fuel Cell and the Race to Change the World)*, John Wiley, **1999**.

Lal, Amit, and James Blanchard, The daintiest dynamos, *IEEE Spectrum*, September **2004**.

Mitchell, Will, Jr., *Fuel Cells*, Academic Press, **1963**.

Narayanan, S. T., T. I. Valdez, and F. Clara, Design and development of miniature direct methanol fuel cell sources for cellular phone applications. *Fuel Cells—Powering the 21st Century*, October 30–November 2, **2000**.

O'Hayre, R., S-W Cha, W. Colella, and F. B. Prinz, *Fuel Cell Fundamentals*, John Wiley, **2006**.

Ovshinsky, S. R., M. A Fetcenko, and J. Ross, A nickel metal hydride battery for electrical vehicles, *Science* **260**, p. 176, April 9, **1993**.

Pham, A. Q., B. Chung, J. Haslam, J. DiCarlo, and R. S. Glass, Solid oxide fuel cell development at Lawrence Livermore National Laboratory, *Fuel Cells—Powering the 21st Century,* p. 787, October 30–November 2, **2000**.

Ralph, T. R., and M. P. Hogarth, Catalysis for low temperature fuel cells, Part I: The cathode challenges, *Platinum Metals Rev.* 46(1), p. 3, **2002**.

Ralph, T. R., and M. P. Hogarth, Catalysis for low temperature fuel cells, Part II: The anode challenges, *Platinum Metals Rev.* 46(3), p. 117, **2002**.

Ralph, T. R., and M. P. Hogarth, Catalysis for low temperature fuel cells, Part III: Challenges for the direct methanol fuel cell, *Platinum Metals Rev.* 46(4), p. 146, **2002**.

Reddington, E., et al., Combinatorial electrochemistry: A highly parallel, optical screening method for discovery of better electrocatalysts, *Science*, 280, p. 1735, June 12, **1998**.

Rice, C., S. Ha, R. I. Masel, and A. Wieckowski, Catalysts for direct formic acid fuel cells, *J. Power Sources* 115 (2), pp. 229–235, April **2003**.

Service, R. F., The fast way to a better fuel cell, *Science*, 280, p. 1690, June 12, **1998**.

Watkins, David S., Research, development and demonstration of solid polymer fuel cell systems, *Fuel cell systems*, Leo J. M. J. Blomen and Michael N. Nugerwa, eds., Plenum Press, **1993**.

Yoon, S. P., S. W. Nam, T.-H. Lim, I.-H. Oh, H. Y. Ha, and S.-A. Hong, Enhancement of cathode performance by sol-gel coating of yttria-stabilized zirconia, *Fuel Cells—Powering the 21st Century, Fuel Cell Organizing Comm.*, pp. 611–614, October 30–November 2, **2000**.

Zhang, J., K. Sasaki, E. Sutter, and R. R. Adzic, Stabilization of platinum oxygen-reduction electrocatalysts using gold clusters, *Science* **315**, pp. 220–222, **2007.**

Further Reading

For those who want substantially more details about the workings of fuel cells, we recommend the book, *Fuel cell fundamentals* by R. O'Hayre et al.

Electrochemistry is described extensively in *Electrochemical methods.*

Diagnostic methods used to characterize fuel cells are examined in *Impedance spectroscopy.*

European Fuel Cell R&D Review, September **1994**, Argonne National Lab.

Fuel Cells, A Handbook (Revision 3), January **1994**, USDE, Office of Fossil Energy.

Halpert, Gerald, Sekharipuram R. Narayanan, Thomas Valdez, William Chun, Harvey Frank, Andrew Kindler, and Subbarao Surampudi (Jet Propulsion Laboratory), and Jack Kosek, Cecelia Cropley, and Anthony LaConti (Giner Inc.), Progress with the direct methanol liquid-feed fuel cell system, *IECEC*, **1997**.

Prater, Keith B., Polymer electrolyte fuel cells: A review of recent developments, *J. Power Sources* **51**, p. 129, **1994**.

Reynolds, W. C., Thermodynamic properties in SI, Department of Mechanical Engineering, Stanford University, **1979.**

Technology Development Goals for Automotive Fuel Cell Power Systems, July **1995**, Argonne National Laboratory.

Williams, Robert H., The clean machine, *Technology Review*, April **1994**.

PROBLEMS

9.1 Every substance is endowed with a certain amount of energy and a certain amount of entropy. Although entropy is well defined, energy has no absolute value; only changes in energy can be measured. For this reason (entirely by convention), the enthalpy of formation of elements in their natural state is taken as zero.

Consider aluminum and oxygen. In their natural states, their standard enthalpy of formation (i.e., the energy of formation at RTP) is, as we said, zero. Every kilogram of aluminum has (at RTP) an entropy of 1.05 kJ/K, whereas every kilogram of oxygen has an entropy of 6.41 kJ/K.

Aluminum burns fiercely, forming an oxide (Al_2O_3) and releasing energy. The standard enthalpy of formation of the oxide is -1.67 GJ/kmole. The entropy of the oxide is 51.0 kJ/K per kilomole.

According to the second law of thermodynamics, the entropy of a closed system suffering any transformation cannot diminish. It can, at best, remain unchanged as the result of the transformation, or else it must increase. If you add up the entropies of the aluminum and of the oxygen, you will discover that the sum exceeds the entropy of the oxide formed. This means that when aluminum combines with oxygen, only part of the chemical energy can be transformed into electricity, while the rest must appear as the heat associated with the missing entropy. That part that can (ideally) be converted into electricity is called the **free energy**.

Calculate the free energy of the aluminum/oxygen reaction.

9.2 Daniell cells used to be popular in the last century as a source of electricity, especially in telegraph systems. These cells consisted of a container divided into two compartments by a membrane permeable to ions. In one compartment, a zinc electrode was dipped in a zinc sulfate solution and, in the other, a copper electrode in a copper sulfate solution.

The zinc oxidizes (i.e., loses electrons):

$$\mathrm{Zn} \rightarrow \mathrm{Zn}^{++} + 2\mathrm{e}^{-}.$$

The zinc is eroded, going into the solution in the form of ions.

At the other electrode, the copper is reduced (i.e., the copper ions in the sulfate accept electrons, and the copper from the sulfate plates out onto the copper electrode):

$$\mathrm{Cu}^{++} + 2\mathrm{e}^{-} \rightarrow \mathrm{Cu}.$$

The cell will deliver current until it runs out of either zinc or sulfate, whichever is less.

Assume the cell has 95 g of zinc and 450 ml of a 0.1 M $CuSO_4$ solution. M stands for molarity: moles of solute per liter of solution. How long can this cell deliver a current of 2 A to a load?

9.3 A fuel cell has the following reactions:

$$\text{ANODE:} \qquad C + 2O^{--} \rightarrow CO_2 + 4e^-, \tag{1}$$

$$\text{CATHODE:} \qquad 4e^- + O_2 \rightarrow 2O^{--}. \tag{2}$$

Changes of enthalpy and free energy (RTP), per kmole of CO_2, are:

$$\Delta\overline{h}_f^\circ = -393.5 \text{ MJ},$$
$$\Delta\overline{g}_f^\circ = -394.5 \text{ MJ}.$$

What is the overall reaction? What is the ideal emf (i.e., V_{rev})? What is the difference in entropy between reactants and products?

Assume that the internal resistance of the cell is 1 milliohm. Otherwise, the cell behaves as an ideal voltage source. How much carbon is needed to deliver 1 MWh of electricity to the load in minimum possible time? What is the load resistance under such conditions?

9.4 The enthalpies and the free energies of formation (at RTP) of each species of interest in this problem are:

	$\Delta\overline{h}_f^\circ$	$\Delta\overline{g}_f^\circ$
	(MJ/kmole)	
CH_3OH (g)	−201.2	−161.9
CH_3OH (l)	−238.6	−166.2
O_2 (g)	0	0
CO_2 (g)	−393.5	−394.4
H_2O (g)	−241.8	−228.6
H_2O (l)	−285.9	−237.2

Owing to the reaction, the changes in enthalpy and in free energy are:

Methanol	**Water**	$\Delta\overline{h}^\circ$	$\Delta\overline{g}^\circ$
		(MJ/kmole)	**(MJ/kmole)**
liquid	gas	−638.5	−685.3
gas	gas	−676.5	−689.6
liquid	liquid	−726.5	−702.4
gas	liquid	−764.5	−706.7

Data from Dohle et al. (2000).

Consider methanol, a fuel that has been proposed for both internal combustion (IC) engines and fuel cells. Methanol can be derived from fossil fuels and also from biomass. Being liquid at RTP conditions, it is a convenient fuel for automobiles. It has reasonable reactivity in fuel cells.

In IC engines, methanol is first evaporated and then burned. The engine exhausts water vapor. In fuel cells, the methanol reacts in liquid form, but the product water is in vapor form.

1. What heat do you get by burning 1 kg of methanol in an IC engine?
2. How much electric energy will an ideal fuel cell (using methanol and air) produce per kg of fuel?
3. How much heat does the cell reject?
4. A practical Otto cycle engine has an efficiency of, say, 20%, while a practical methanol fuel cell may have an efficiency of 60% (this is the efficiency of the practical cell compared with that of the ideal cell). If a methanol fueled IC car has a highway performance of 10 km per liter, what is the performance of the fuel cell car assuming that all the other characteristics of the car are identical?
5. If you drive 2000 km per month and a gallon of methanol costs $1.20, how much do you save in fuel per year when you use the fuel cell version compared with the IC version? Can you think of other savings besides that in fuel?
6. You get a 10-year loan with yearly repayments of principal plus interest of 18% of the initial amount borrowed. By how much can the initial cost of the fuel-cell car exceed that of the IC car for you to break even? Assume that after 10 years the car is totally depreciated.
7. What is the open-circuit voltage of an ideal methanol fuel cell at RTP? To answer this question, you need to make an intelligent guess about the number of electrons freed per molecule of methanol.

In the above questions, assume 100% current efficiency and 100% efficiency of the electric motor.

9.5 You want to build a hydrogen manometer based on the dependence of the output voltage on the pressure of the reactants. Take an H_2/O_2 fuel cell at 298 K. Assume that it produces water vapor and and acts as an ideal cell. The oxygen pressure is maintained at a constant 0.1 MPa while the hydrogen pressure, p_{H_2}, is the quantity to be measured.

1. What is the output voltage when p_{H_2} is 0.1 MPa?
2. What is the output voltage when p_{H_2} is 1 MPa?
3. Develop an expression showing the rate of change of voltage with p_{H_2}. What is this rate of change when p_{H_2} is 0.1 MPa?

4. The output voltage of the cell is sensitive to temperature. Assume that a ±10% uncertainty in pressure measurement can be tolerated (when the pressure is around 1 MPa). In other words, assume that when a voltage corresponding to 1 MPa and 298 K is read, the actual pressure is 0.9 MPa because the temperature of the gases is no longer 298 K. What is the maximum tolerable temperature variation?

9.6 A certain gas, at 10^5 Pa, has a specific heat given by the expression

$$c_p = a + bT + cT^2$$

for 298 K < T < 2000 K.

$a = 27.7\ \text{kJ}\,\text{K}^{-1}\,\text{kmole}^{-1}$,
$b = 0.8 \times 10^{-3}\ \text{kJ}\,\text{K}^{-2}\,\text{kmole}^{-1}$,
$c = 10^{-6}\ \text{kJ}\,\text{K}^{-3}\,\text{kmole}^{-1}$.

At 298 K, the enthalpy of the gas is 0 and its entropy is 130.0 kJ K^{-1} kmole^{-1}. What are H, G, and S of the gas (per kilomole) at $T = 1000$ K and $p = 10^5$ Pa? Please calculate with four significant figures.

9.7 A fuel cell has the reactions:

$$\text{ANODE:} \qquad 2\text{A}_2 = 4\text{A}^{++} + 8\text{e}^-.$$

$$\text{CATHODE:} \quad 4\text{A}^{++} + \text{B}_2 + 8\text{e}^- = 2\text{A}_2\text{B}.$$

All data are at RTP.

The overall reaction, 2 A_2 + B_2 = 2 A_2B, releases 300 MJ per kmole of A_2B in a calorimeter. The entropies of the different substances are:

A_2:	200 kJ K^{-1} kmole^{-1},
B_2:	400 kJ K^{-1} kmole^{-1},
A_2B:	150 kJ K^{-1} kmole^{-1}.

A_2 and B_2 are gases, whereas A_2B is liquid.

What is the voltage of an ideal fuel cell using this reaction at RTP?

Estimate the voltage at standard pressure and 50 C.

How much heat does the ideal fuel cell produce per kilomole of A_2B, at RTP? What is the voltage of the cell if the gases are delivered to it at 100 MPa? The operating temperature is 25 C. If the internal resistance of the cell (operating at RTP) is 0.001 Ω, what is the maximum power the cell can deliver to a load? What is the fuel consumption rate of the cell under these circumstances? What is the efficiency of the cell?

9.8 Owing to its ceramic electrolyte, a fuel cell can operate at 827 C. Pure oxygen is used as oxidizer. The gases are fed to the cell at a pressure of 1 atmosphere. Use the following data:

	$\Delta\overline{h}_f^\circ$ (MJ per kmol)	$\Delta\overline{g}_f^\circ$ (MJ per kmol)	γ	$\overline{s}^\circ$ (kJ K^{-1} per kmol)
CO(g)	−110.54	−137.28	1.363	197.5
CO_2(g)	−393.51	−394.38	1.207	213.7
O_2(g)	0	0	1.341	205.0

The values of γ are those appropriate for the 25 C to 827 C interval.

We want to examine the influence of temperature on the performance of an ideal fuel cell.

1. Calculate the reversible voltage and the efficiency of the above (ideal) cell at both 25 C and 827 C.
2. As expected (if you did the calculations correctly), you will have found that both V_{rev} and η are larger at the lower temperature. Yet, the cell is operated at 827 C where its ideal performance is not as good. Why? Explain in some detail.

9.9 A fuel cell was tested in the laboratory and yielded the following:

Open-circuit voltage	0.600 V
Internal resistance	0.01 Ω
Voltage (I = 1 A)	0.490 V
Voltage (I = 10 A)	0.331 V

Thermodynamic data indicate that V_{rev} of a fuel cell is 0.952 V and that the enthalpy change of the reaction is 1.26 times the reaction free energy. Two electrons circulate in the load per molecule of product.

1. What power does the cell deliver to a load when the current is 5 A?
2. What is the heat power dissipated internally when $I_L = 5$ A?

Assume the Tafel equation is valid. The internal resistance given above is the slope of the straight-line portion of the v-i characteristic of the cell.

9.10 A hydrogen–oxygen fuel cell, operating at RTP, has the following v-i characteristics:

$$V_L = 0.8 - 0.0001\, I_L.$$

Assume 100% current efficiency.

1. What is the hydrogen consumption rate (in mg/s) when the cell delivers 1 kW to a load?
2. What is the heat generated by the cell? Liquid water is produced.

9.11 A fuel cell is prismatic in shape and measures $d \times 2d \times 2d$ (where $d = 33$ cm). It is fed with H_2 and O_2, which are admitted at 25 C and 1 atmosphere. Product water is exhausted at 1 atmosphere and 110 C.

The inside of the cell is an uniform temperature of at 110 C. The outside is maintained at 50 C by immersing it totally in running water admitted at 20 C and exhausted at 45 C. Liquid water has a heat capacity of 4 MJ per K per m^3. Assume that the temperature gradient across the walls is uniform. The walls are made of 10-mm-thick stainless steel with a heat conductivity of 70 W per m per K. The only energy input to the cell comes from the fuel gases admitted. Heat is removed by both the coolant water and the product water that is exhausted from the cell at a rate of $\dot{N}$ kmoles/s.

The load voltage is $V_L = 0.9 - R_{int}I$ volts. $R_{int} = 10^{-7}$ Ω.

1. What heat energy is removed every second by the coolant water?
2. What is the flow rate of the coolant water?
3. Express the heat removal rate by the product water in terms of $\dot{N}$.
4. What is the input power in terms of $\dot{N}$?

   What is the power delivered to the load in terms of $\dot{N}$?
5. Write an equation for thermal equilibrium of the cell using your results from above.
6. What is the value of I that satisfies the above equation?

To simplify the solution, assume that the fuel cell reaction proceeds at RTP and liquid water is produced at 25 C. This water is then heated up so that the exhaust is at the 110 C prescribed.

9.12 A fuel cell has a cooling system that allows accurate measurement of the heat dissipated and precisely controls the operating temperature, which is kept, under all circumstances, at 298 K.

When a current of 500 A is generated, the cooling system removes 350 W of heat, while when the current is raised to 2000 A, the amount of heat removed is 2000 W.

Estimate the heat removed when the current is 1000 A. The input gases are at RTP, and liquid water is created in the process.

Assume a linear dependence of the load voltage on the load current.

9.13 An ideal fuel cell operates at 1000 K. Two reactant gases (not necessarily H_2 and O_2) are fed in at 1 atmosphere. The reversible voltage is 1.00 V. What will the voltage be if gas A is fed in at 100 atmospheres and gas B at 200 atmospheres, both still at 1000 K? The reaction is

$$2A + 3B \rightarrow A_2B_3.$$

The product is liquid (in spite of the high temperature). A total of 6 electrons circulate in the load for each molecule of product.[†]

9.14 A fuel cell employs the following reaction:

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{AB}.$$

At STP, the relevant thermodynamic data are as follows:

	$\Delta\overline{h}_f^\circ$ [MJ/kmole]	$\overline{s}$ [kJ/(K kmole)]
A(g)	0	100
B(g)	0	150
AB(g)	−200	200

What is the reversible voltage of the above fuel cell? For each molecule of AB, two electrons circulate in the load.

9.15 A fuel cell battery is to be used in a satellite. It must deliver 2 kW at 24 V for one week. The mass of the cell must be the smallest possible.

The fuel cell manufacturer has a design with the following characteristics:

Open-circuit voltage: 1.10 V
Internal resistivity: 92×10^{-6} ohm m^2
Cell mass: 15 kg per m^2 of active electrode area

There is a linear relationship between V_L and I_L.

How many cells must be connected in series?
What is the total mass of all fuel cells in the battery?

9.16 The open-circuit voltage of a hydrogen–oxygen fuel cell operating at RTP is 0.96 V, and its internal resistance is 1.2 mΩ. The activation voltage drop is given by

$$V_{act} = 0.031 + 0.015 \ln I,$$

where I is in amperes.

From thermodynamic data, the reversible voltage, V_{rev}, is known to be 1.20 V.

Two hundred of the above cells are connected in series, forming a battery that feeds a resistive load rated at 2.5 kW at 100 V.

What is the actual load voltage?
How much heat does the battery generate internally?

[†] The very large difference in pressure between the two reactant gases would require a strong diaphragm or electrolyte. This suggests a ceramic electrolyte and hence the high operating temperature. Clearly, the realizability of this fuel cell is highly doubtful.

9.17 A fuel cell battery is fed by H_2 and O_2 (both at 300 K) and produces water vapor that promptly condenses inside the cell.

ΔT is the difference, in kelvins, between the temperature of the active area of the cell and 300 K, which is supposed to be the temperature of the cooling liquid and of the environment.

Two main mechanisms remove heat from the cells:

1. Some heat is conducted away at a rate of 40 W per m^2 of active electrode surface for each kelvin of ΔT. This, of course, implies some cooling system whose exact nature is irrelevant as far as this problem is concerned.
2. To simplify the solution, assume that water vapor is synthesized in the cell at the temperature of the incoming gases (300 K) and then immediately condenses at this temperature and heats up by an amount ΔT to reach the operating temperature of the cell. The water is then removed, carrying with it a certain amount of heat and thus cooling the cell. If the temperature of the product water exceeds 100 C, assume that the cell is pressurized so that water does not boil. However, assume that all the reactions actually occur at RTP, that is, use thermodynamic data for RTP.

   The V-J characteristic of the cell is

   $$V = 1.05 - 95.8 \times 10^{-6} J,$$

   where J is the current density in A/m^2.

The current efficiency is 100%.

Although the cell will operate at conditions that differ from RTP, use RTP thermodynamic data to simplify the problem.

1. The battery is not going to be operated at full power because it probably will exceed the maximum allowable temperature. Nevertheless, calculate what the equilibrium temperature would be if full power operation were attempted.
2. In fact, the battery will operate at a much lower power. It must deliver 20 kW to a load at 12 V. It consist of several cells connected in series. The mass of each cell is 15 kg for each m^2 of active electrode area.

The battery must deliver this power for a week. The total mass (fuel plus battery) must be minimized. Ignore the mass of the fuel tanks.

How many cells must be employed?
What is the total mass?
How many kg of H_2 and how many of O_2 are needed?
What is the operating temperature of the cell?

9.18 Fill in the answers as follows:

If output voltage rises, mark "R"; if it falls, mark "F"; if there is no effect, mark "N."

	Ideal Fuel Cell	**Practical Fuel Cell**
Higher temperature		
Higher reactant pressure		
Higher product pressure		

9.19 A fuel cell, generating water vapor, has a straight-line V-I characteristic:

$$V_L = V_0 - R_{int} I$$

Both V_0 and R_{int} are temperature dependent and are given by the following expressions, over the temperature range of interest.

$$V_0 = \beta_0 (1 + \alpha_V T) V_{rev},$$
$$R_{int} = (1 + \alpha_R T) R_{int_0}.$$

The coefficients are:

1. $\beta_0 = 0.677$.
2. $\alpha_V = 443.5 \times 10^{-6}$ per K,
3. $R_{int_0} = 0.00145\ \Omega$.
4. $\alpha_R = -1.867 \times 10^{-3}$ per K.

What are the efficiencies of the fuel cell at 298 K and at 500 K when feeding a 1 milliohm load?

9.20 A fuel cell battery is to be used aboard the space station. The bus voltage (the voltage the battery has to deliver under full load) is 24 V when delivering 30 kW. Since the space craft has a hydrogen and oxygen supply, the battery will use these gases, which are delivered to it at 1 atmosphere and 298 K. A manufacturer-submitted sample cell was tested in the laboratory with the following results:

When no current is drawn from the cell, its voltage is 1.085 V. When delivering 2000 A, the voltage is 0.752 V. A straight-line relationship was found to exist between V and I. The cell masses 75 kg, and, when taken apart, it was found that the active electrode area is $1.5\,\mathrm{m}^2$. It is clear that if the battery is to deliver 30 kW under 24 V, it must generate a current of 1250 A. Since all the cells are in series, this is also the current through each cell.

The sample cell operates with a current density of $1250/1.5 = 833.3\,\mathrm{A/m}^2$. If the manufacturer constructs a cell, in all aspects identical to the sample, except with a different active electrode areas, S, the new cell must still deliver the 1250 A but under different current density and, consequently, under different cell load voltage. Since the battery load voltage must still be 24 V, the battery will contain a different number, N, of cells.

Assume that the mass of the new cell is proportional to the active area of the electrodes. The total mass (mass, M_B, of the battery plus mass, M_F, of the fuel, H_2 and 0_2) is to be minimized for a 30-day-long mission during which the battery delivers a steady 30 kW at 24 V. Ignore the mass of the fuel tanks. The current efficiency is 100%.

Calculate this minimum total mass. How many cells are needed in series?

In the cell above, assume that water is synthesized as vapor at the temperature of the incoming gases (298 K) and promptly condenses into a liquid and then heats up to T_{op}, the operating temperature of the device. The product water is continuously removed from the cell at this latter temperature. In addition, a cooling system also removes heat. It does this at a rate of 6 W per degree of temperature difference ($T_{op} - 298\,\mathrm{K}$) for each square meter of active electrode surface.

What is T_{op} when the battery delivers 30 kW as in the first part of this problem?

9.21 A hydrogen–oxygen fuel cell has the following characteristics when both reactants are supplied at the pressure of 1 atm:

$$V_{oc} = 0.75 + 0.0005T \;\; \mathrm{V},$$
$$R_{int} = 0.007 - 0.000015T \;\; \Omega.$$

Estimate, roughly, the power this fuel cell delivers to a 5-milliohm load.

In any fuel cell, heat may be removed by

1. circulation of a coolant.
2. excess reactants that leave the cell at a temperature higher than the input temperature.
3. products that leave the cell at a temperature higher than that at which they were synthesized.

To simplify this problem, assume that the contribution of mechanism a is always 30 times that of mechanism c and that of mechanism b is negligible. Assume that the reactants are fed in at 298.2 K and that the product water is created as a vapor at this temperature and then heated to T_{op} by the heat rejected by the cell.

Pure hydrogen and pure oxygen are supplied at 1 atm.

1. What is the temperature of the cell when temperature equilibrium has been reached?
2. What is the load current and the power delivered to the load under the above conditions?

9.22

R_L (Ohms)	I (Amps)
0.05	14.98
0.10	8.23
0.15	5.71
0.20	4.37
0.25	3.54

The reactions in a fuel cell are

$$\begin{aligned}\text{Anode}: &\quad A \rightarrow A^{++} + 2e^{-}\\ \text{Cathode}: &\quad A^{++} + 2e^{-} + B \rightarrow AB.\end{aligned}$$

The gases, A and B, are fictitious (and so are their properties). The atomic mass of A is 16 daltons and that of B is 18 daltons. Both A and B behave, over the temperature range of interest, as if they had 5 degrees of freedom, while the product, AB, as if it had 7.

The fuel cell was tested in a laboratory by observing the current delivered as the load resistance was altered.

The results are displayed in the table.

1. What are the open-circuit voltage and the internal resistance of the cell?
2. Careful calorimetric observations show that when the fuel cell is delivering 10.0 A to a load, the heat dissipated internally is 3.40 W. From this information, determine the ΔH of the reaction.
3. Notice that P_{heat} depends on $(V_{rev} - V_{oc})I$. This would suggest that it is possible to determine V_{rec} from the knowledge of P_{heat}. Demonstrate that it is not possible to do so, that is, that for a fixed V_{oc} and R_{int}, P_{heat} is not sensitive to the value of V_{rev}.
4. In order to estimate the ΔG of the reaction, an external voltage was applied to the fuel cell so as to cause it to act as an electrolyzer. When this external voltage was 1.271 V, the electrolyzer produced A at a rate of 2.985 g/hour. Making plausible assumptions, estimate the ΔG of the reaction.

9.23 A hydrogen/oxygen fuel cell has the V-J characteristic (at RTP):

$$V_L = 0.98 - 10^{-3}J.$$

The active area of its electrodes is 0.444 m^2. The water is exhausted from the cell in gaseous form.

1. What is the rate of heat production when the cell is connected to a load of

   1.1. Open circuit?

   1.2. Short circuit?

   1.3. A load that maximizes the power output?

2. What are the efficiencies of the cell under the three conditions above?
3. What is the efficiency of the cell if it delivers half of its maximum power? Use the more efficient solution.
4. Assume that V_{oc} is a constant fraction of V_{rev}. Thus, under all circumstances $V_{oc} = (0.98/1.185) \times V_{rev} = 0.827V_{rev}$.

   What is the V-I characteristic of the cell when fed air at 1 atmos and 25 C instead of oxygen?
5. To simplify this problem assume that the only way to remove heat from the cell is via the exhaust stream, which consists of water vapor and excess input gases. The input gases (hydrogen and air) are at 298 K. Assume that the water is produced at 298 K and then heated by the fuel cell to the exhaust temperature, T.

   What is the value of T when the cell, fed by the minimum amount of air that satisfies the oxygen requirement of the device, produces the electric output of Item 3 (half its maximum power). Although the oxidizer is air, not pure oxygen, use, for simplicity, the V-I characteristic for pure oxygen as given in Item 1.
6. If you made no mistake, you have found that the temperature calculated in the preceding item is too high. Much more vigorous cooling will be necessary. This can be accomplished by injecting much more air than is required by the stoichiometry. Assume that the temperature increase should not exceed 80 K. How much must the flow of air be compared with that required in Item 5?

9.24 The EV1 was an exceedingly well-designed automobile, having excellent aerodynamics and, all over, low losses. With an energy supply of 14 kWh, it had range of over 100 km. Its 100-kW motor allowed very good acceleration, making it a “sexy” machine. The problem was that, no matter how good a battery it used, it took a long time to recharge it. If instead of a battery, it had used fuel cells, then *refueling* would take only minutes versus hours for *recharging*.

Imagine that you want to replace the NiMH batteries by a fuel cell battery, which, of course, must supply 100 kW of power. The *V*-*I* characteristic of the available hydrogen/oxygen fuel cell operating at RTP is

$$V_l = 1.1 - 550 \times 10^{-6} I.$$

The maximum internal heat dissipation capability is 300 W. Product water exits the cell in vapor form.

The fuels cells deliver energy to a power conditioning unit (inverter) that changes dc input into ac power. The efficiency of this unit can be taken as 100%.

1. What is the input voltage of the power conditioning unit; in other words, what is the voltage that the fuel cell battery (at 100 kW) must deliver assuming the the smallest possible number of individual cells are used.
2. The 100 kW are needed only for acceleration. For cruising at 110 km/h, only 20 kW are required.[†] How many kg of hydrogen are needed to provide a range of 800 km to the car (using 20 kW)?
3. If the hydrogen is stored at 500 atmospheres, how much volume does it occupy at 298 K?

9.25 A single-chamber low-operating-temperature solid oxide fuel cell somewhat similar to the one described by Hibino et al. (2000) when operated at a current density of 6000 A/m^2, delivers a load voltage that depends on the thickness of the electrolyte in the manner indicated in the following table.

Electrolyte thickness (mm)	**Load Voltage (V)**
0.15	0.616
0.35	0.328
0.50	0.112

The cell has essentially straight *V*-*I* characteristics. Its specific resistance, $\Re$ (see Subsection 9.8.3.1), consists of two components, $\Re_1 + \Re_2$, where $\Re_1$ is the resistance of the electrolyte and $\Re_2$ represents all other resistances of the cell. The open-circuit voltage is 0.892 V.

1. If it were possible to use a vanishingly thin electrolyte, what maximum power would the cell be able to deliver?

[†] Just a wild guess!

2. What would be the corresponding load resistance if the cell has an effective electrode area of 10 by 10 cm?
3. Compare the power output of the cell with that for the cell with 0.15-mm-thick electrolyte.

9.26 Solid oxide fuel cells manufactured by Siemens Westinghouse have a very pronounced curvature in their V-J characteristics. One class of cells using "ribbed" units behaves according to

$$V_L = 0.781 - 1.607 \times 10^{-6} J - 6.607 \times 10^{-9} J^2,$$

where J is the current density in A/m^2 and V_L is the load voltage in V.

1. What is the open-circuit voltage of the cell?
2. What is the voltage of the cell when delivering maximum power?

9.27 The V-I characteristics of a given H_2/O_2 fuel cell (measured with incredible precision) are tabulated as shown. See plot. The measurements were made at RTP. Water leaves the cell as a gas.

1. Calculate the efficiency of the cell when 10 A are being delivered.
2. Calculate the rate of heat generated by the cell when $I_L = 10$ A.
3. Visually, the characteristics appear as a straight line for sufficiently large current. This suggests that, in the relatively large current region, one can use the equation

$$V_L = V_{oc} - R_{app} I,$$

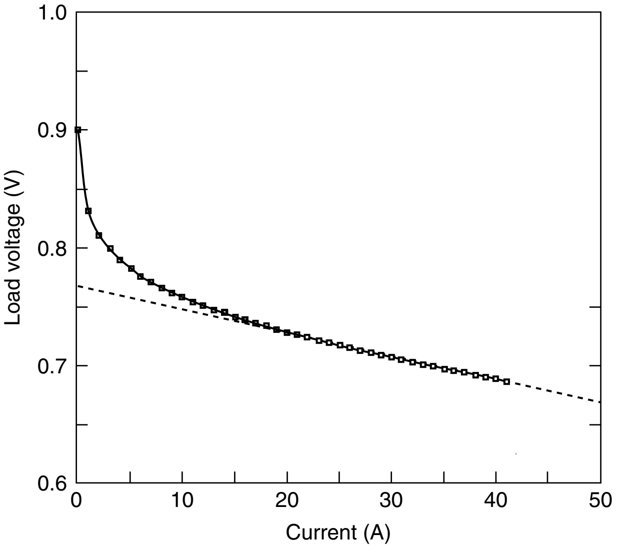
where R_{app} is the apparent internal resistance of the cell as inferred from the straight line. Estimate the value of R_{app} using

3.1. the region $30 \le I \le 41$ A.

3.2. the region $10 \le I \le 41$ A.

4. For each of the values of R_{app} above, determine the magnitude of the various sources of heat (Joule effect, etc.) when the cell delivers 10 A to the load. Clearly, because you used a straight-line approximation, the activation voltage does not contribute to the heat calculation.
5. Now, determine accurately the value of the internal resistance, R_{int}; that is, include the activation voltage in the V-I characteristics.
6. Write equations describing the manner in which the load voltage depends on the load current. Check the values obtained against the tabulated data. Do this for $I_L = 40$ A and for $I_L = 0.5$ A.
7. Explain why your equation overestimates V_L at small currents.

Load current (A)	Load voltage (V)	Load current (A)	Load voltage (V)
0.00	0.90000		
0.5	0.84919		
1.00	0.83113	21.00	0.72622
2.00	0.81144	22.00	0.72390
3.00	0.79922	23.00	0.72163
4.00	0.79018	24.00	0.71942
5.00	0.78291	25.00	0.71725
6.00	0.77677	26.00	0.71514
7.00	0.77142	27.00	0.71306
8.00	0.76664	28.00	0.71103
9.00	0.76230	29.00	0.70902
10.00	0.75831	30.00	0.70706
11.00	0.75461	31.00	0.70512
12.00	0.75114	32.00	0.70322
13.00	0.74786	33.00	0.70134
14.00	0.74476	34.00	0.69949
15.00	0.74180	35.00	0.69766
16.00	0.73896	36.00	0.69586
17.00	0.73624	37.00	0.69408
18.00	0.73361	38.00	0.69232
19.00	0.73107	39.00	0.69058
20.00	0.72861	40.00	0.68886
		41.00	0.68715



9.28 Although low-voltage automotive batteries have been standardized at 12 V, no such standards have been agreed on for automotive traction batteries. Some hybrid cars use 275 V motors and 275 V batteries (some use 550 V motors powered by 275 V batteries.)

Consider a fuel cell battery rated 100 kW at 275 V. It uses pure hydrogen and pure oxygen, both at 1 atmosphere pressure. The battery, consisting of 350 cells, operates at 390 K. To simplify the problem, assume 298 K thermodynamics. Assume a linear V vs. I characteristic for the fuel cells.

1. What is the hydrogen consumption (in kg of H_2 per hour) when the battery delivers 100 kW?
2. The retarding force on a car can be represented by a power series in U (the velocity of the car):

   $$F = a_0 + a_1 U + a_2 U^2. \tag{1}$$

   $a_1 U$ represents mostly the force associated with the deformation of the tires. $a_2 U^2$ is the aerodynamic retarding force and is

   $$a_2 = \frac{1}{2}\rho C_D A U^2, \tag{2}$$

   where $a_0 = 0$, $\rho = 1.29$ kg/m^3 is the air density, $C_D = 0.2$ is the drag coefficient, and $A = 2$ m^2, is the frontal area of the vehicle.
3. When delivering 50 kW, the battery voltage is 295 V. When cruising at a constant, moderate speed of 80 km/hr the car, uses only 15 kW. What is the range of the car under such conditions if the hydrogen tank can store 4 kg of the gas? This assumes flat, horizontal roads and no wind.
4. How slow must the car drive to do 1000 km on 4 kg of H_2?

9.29 To test a fuel cell in a laboratory, an ac voltage generator (peak-to-peak voltage $v_{pp} = 0.001$ V) was connected in series with the load, and an ac ammeter (peak-to-peak current i_{pp}) was used to measure the load current fluctuations caused by the varying V_L. The frequency used was low enough to cause any reactive component in the measurement to be negligible. The following measurements were obtained:

I_L (A)	V_L (V)	i_{pp} (A)
5.34	0.956	0.186
10.67	0.936	0.366

It was observed that there was a 180° phase relationship between v_{pp} and i_{pp}—that is, that increasing the voltage actually reduced the current.

Calculate the true internal resistance of the cell.

9.30 Hydrogen–oxygen fuel cell. Although the temperature of the cell will vary throughout its operation, use thermodynamic data for RTP so as not to complicate the computation.

Each cell is 3 mm thick and has a total area of 10 by 10 cm.

The density of each cell is equal to twice the density of water, and the specific heat capacity of the cell is 10% of the specific heat of water. This means that it takes 24 J of heat to raise the cell temperature by 1 kelvin.

Under all circumstances, the product water is removed in vapor form.

The highest allowable operating temperature of the cell is 450 K.

Although heat is removed from the cell by several different mechanisms, the net effect is that the rate of heat removal is proportional to $T - 300$: In fact, the heat removal rate, $\dot{Q}_{rem} = 0.3(T - 300)\,\mathrm{W}$.

Laboratory tests reveal that when the load current is 2 A, the load voltage is 0.950 V, and when the load current is 20 A, the load voltage falls to 0.850 V.

1. Write an equation relating V_L to I_L, assuming a linear relationship between these variables.
2. What is the maximum power the cell can deliver?
3. Show that this maximum power cannot be delivered continuously because it would cause the cell temperature to exceed the maximum allowable operating temperature.
4. What is the maximum power that the cell can deliver continuously to a load?
5. Although the cell cannot deliver maximum power continuously, it can do so for a short time if it starts out cold—that is, if its initial temperature is 300 K. It will generate more heat than it can shed, and its temperature will rise. How long can the cell (initially at 300 K) deliver maximum power to the load without exceeding 450 K?

9.31 An ideal hydrogen/air fuel cell operates at 298 K. A mixture of hydrogen and water vapor is fed in at 3 atmospheres pressure. Moist air is also fed in at this same pressure. In both the fuel and the oxidant streams, the partial pressure of the water vapor is 0.5 atmosphere. What is the voltage the ideal cell delivers to a 1-ohm load? Please calculate the voltage to a millivolt precision.

9.32 A 24-cell hydrogen/air fuel cell battery operating at RTP consists of cells having the characteristics,

$$V_L = 1.05 - 0.001 I_L. \tag{3}$$

Water is exhausted from the battery as a vapor.

The heat removal system can, at most, remove 5278 W of heat. What is the maximum power the battery can deliver (under steady-state conditions) to a load?

9.33 A domestic fuel cell system in a rural area is to be fed by butane. This gas is to be steam reformed, and the resulting carbon monoxide is to be shifted to hydrogen. Assuming no losses, how many kg of hydrogen can be extracted from each kg of butane?

9.34 A hydrogen/oxygen fuel cell operating at 298 K is fed (on the anode side) a mixture of hydrogen and water vapor—for each kg of hydrogen there are 1.8 kg of water vapor. The total pressure of the mixture is 3 atmospheres.

On the cathode side, it is fed moist air also at 3 atmospheres—for each kg of air there are 0.125 kg of water. Assume that air consists of 20% oxygen and 80% nitrogen, by volume.

The product is liquid water (remember to use a ΔGf° of -237.2 MJ/kmole, not -228.6).

The characteristics of the fuel cell are as follows.

- The open-circuit voltage, V_{oc}, is 139 mV lower than the reversible voltage (V_{rev}) owing to unavoidable unwanted side-reactions, mostly at the cathode.
- The internal resistance, R_{int}, of the cell is 5.1 mΩ.
- The exchange currents add up to $I_0 = 0.60$ A.
- The transfer coefficient, α, is 0.7.

Transport losses are negligible in the region of operation.

a. What is the reversible voltage of this cell? Please show four significant figures.

b. The fuel cell is connected to a 10-mΩ load. How much heat is generated internally by the cell?

c. The fuel cell input gases contain water vapor (steam), and the fuel cell produces liquid water. Assume all of the water vapor contained in the input gases is lost. Does the fuel cell produce enough water to moisten the input gas streams when delivering the current of part b?

9.35 The V-I characteristics of a fuel cell can be represented with reasonable accuracy, by the equation,

$$V_L = V_{oc} - R_{int} I_L - V_2 \ln \frac{I_L}{I_0}. \tag{4}$$

Let $V_{oc} = 1.05$ V and $R_{int} = 0.005$ ohm. When a 0.16-ohm load resitor is used, the current is 5.416 A, and when the load resistance is increased to 0.17 ohm, the current falls to 5.125 A.

What is the value of the constant, I_0 in the formula?

Do not use numerical or trial-and-error solutions.

9.36 A lab test of a hydrogen/oxygen fuel cell consisted of adjusting the load resistance, R_L, until a preselected load current, I_L, was precisely achieved. This yielded the following results:

R_L (Ω)	I_L (A)
1.0643	1.000
0.1944	5.000
0.0129	50.00

a. What power does the fuel cell deliver to the load when $I_L = 20$ A?

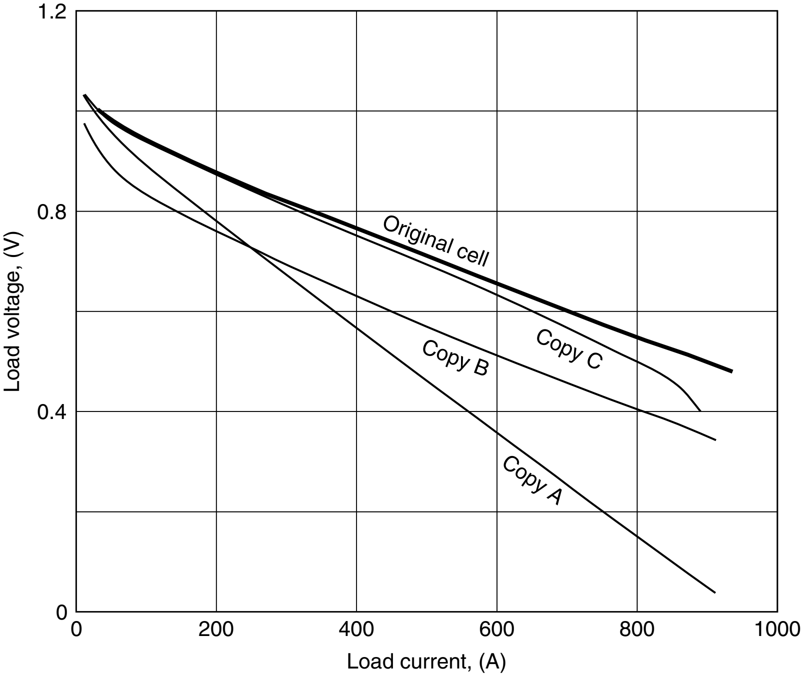
b. What is the activation voltage when $I_L = 20$ A?

c. By using the data for 1 A and 50 A, estimate a value for R_{app}, and using this value estimate the power delivered to the load when $I_L = 20$ A.

The open-circuit voltage of the cell (at 90 C which is the operating temperature in this problem) is 1.10 V.

9.37 *All answers must be short and terse. No dissertations, please.*

A fuel cell manufacturer has accurate data on a rival manufacturer's solid-polymer, hydrogen/oxygen fuel cell known for its excellent performance. The data are plotted in the figure and are labeled Original cell.



The development department of the first manufacturer did its best to build a cell identical to the original but had limited success, as demonstrated by the V-I plot labeled Copy A. All copies operate at the same temperature and pressures as the original. They all were fed pure, uncontaminated, hydrogen and clean air.

1. Identify the one error made in Copy A. What tipped you off? What would you do to rectify this error?
2. Do the same for Copy B.
3. Do the same for Copy C.
4. In addition to the three copies of the original fuel cell whose characteristics appear in the figure, there is a fourth copy (Copy D) that has a plot almost exactly parallel to that of the original bur is, roughly, 0.1 lower in voltage. For this one also, answer the questions as before.

9.38 Let us compare the performances of two different fuel cell cars, one equipped with a small fuel cell (SFC) and one with a large fuel cell (LFC). The SFC car has slightly more power than the average power needed for the trip, while the LFC car has much more power than the average needed and is capable of fast accelerations and can climb steep hills at a pretty good clip—it is much more "sexy." In this particular test, the cars drive on a flat, horizontal road with no external wind. They drive at the same uniform speed, and the fuel cell (in both cases) delivers a steady 10 kW to the wheels. Either fuel cell is actually a battery of 250 elementary, hydrogen/air, fuel cells connected in series. In order to simplify calculations, model the elementary fuel cell as a voltage generator (open-circuit voltage, $V_{oc} = 1.0$ V), in series with an internal resistance, R_{int}.

In the case of SFC, $R_{int} = 0.0055$ ohm, in the case of LFC, $R_{int} = 0.0011$ ohm. In either case, water vapor is produced by the reaction.

Assume that the trip lasts exactly five hours and that the cars travel at a constant speed of 90 km/h. Calculate (for both SFC and LFC):

1. How many kg of hydrogen will be consumed?
2. How much reserve power, P_{acc}, is available for surges? That is, what is the difference between the maximum power the battery can deliver and the steady power needed for driving the car at the specified speed?
3. Select either the SFC or the LFC car as the one you would recommend for production. Defend your selection by pointing out the advantages and disadvantages of the two types.

9.39 A given electric car experiences a retarding force given by

$$F = 80 + 5v + 0.25v^2,$$

where v is the velocity of the car.

Assume a straight, horizontal, windless highway.

The car carries 20 lead–acid batteries connected in series, each having a 300-Ah capacity when discharged at a steady 15-A current. The batteries have a Peukert number of 1.2. Assume that the open-circuit voltage of each battery is 12.0 V independently of the state of charge, SOC. Each battery has an internal resistance of 40 milliohms.

1. What is the range of the car when the electric motor (100% efficient) is running at a current of 15 A?
2. Assume that you want to cruise at 85 km/hr. What is the car's range?

Chapter 10
Hydrogen Production

10.1 Generalities

In mid-2008, if you looked up “Hydrogen use” in Google (English), you would find about 1,550,000 entries. Better than any other statistics, this attests to the enormous interest in this gas. Alone for the production of ammonia, there was a 46-fold increase in hydrogen utilization between 1946 and 2003, when over 19 million tons were produced. The impending massive introduction of fuel cells into the economy will cause the demand for hydrogen to rise much more rapidly in the near future. Techniques for both bulk production and local generation (especially in vehicles) will be perfected by established industries and by a host of start-ups, all trying to profit from the new market. An interesting study of hydrogen production methods can be found in the paper by Brinkman (see References).

It is important to start with the clear understanding that, though extremely abundant, hydrogen, unlike fossil fuels, is not a source of energy. Much of the existing hydrogen is in the form of water—hydrogen ash—and considerable energy is required to extract the desired element. Hydrogen is, at best, an excellent vector of energy. It holds great promise as

1. Fuel for land and sea vehicles, especially when used in high-efficiency fuel cells.
2. Fuel for large air- and spacecraft owing to its high energy-to-weight ratio when in cryogenic form.
3. Industrial and domestic fuel for generation of heat and electricity.
4. A means for transporting large quantities of energy over long distances.

The advantages of hydrogen include:

1. Low pollution
   Hydrogen burns cleanly, producing only water. It is true that, depending on the flame temperature when burned in air, small amounts of nitrogen oxides may also be generated. Pollution, however, may be associated with some hydrogen production processes.
2. Controllability
   At ambient temperatures, hydrogen reacts extremely slowly with oxygen. Catalysts permit adjusting the reaction speed over a large range from very low-temperature flames to intense ones.

3. Safety

   Hydrogen's reputation as a dangerous gas stems mostly from the spectacular 1937 explosion of the *Hindenburg* in Lakehurst, New Jersey, when 36 people were killed. Yet, a good case can be made that the explosion actually proved how *safe* the gas is. Indeed, the *Hindenburg* carried 200,000 cubic meters of hydrogen, equivalent to 2.5×10^{12} joules of energy. An energetically equivalent amount of gasoline would correspond to over 80 cubic meters, which could form a pool of fiery liquid covering the area of some 15 football fields.[†]

   3.1 Being the lightest of all gases, it quickly rises and disperses, while liquid fuels form pools that spread the fire.
   3.2 The smallness of the hydrogen molecule causes this gas to leak easily through tiny cracks and holes, making it difficult to accumulate in explosive concentrations.
   3.3 Owing to its low density, a given volume of hydrogen contains little energy and thus represents a much smaller hazard than natural gas or gasoline (the vapor of the latter contains 20 times the energy of H_2 on a same-volume basis).
   3.4 At 1 atmosphere, the auto-ignition temperature for hydrogen is about 580 C, whereas that for gasoline is as low as 260 C. The likelihood of accidentally starting a fire is much higher with gasoline.
   3.5 Hydrogen/air mixtures with less than 4.1% fuel (in volume) will not catch fire while the flammability limit for gasoline is 1%.
   3.6 A pure hydrogen flame radiates little energy, allowing firemen to get much closer to the site of a fire.
   3.7 Hydrogen is totally nontoxic and can be inhaled in high concentration (of course, it can asphyxiate you and can also cause you to explode if hydrogen-filled lungs are accidentally ignited).

Accumulation of hydrogen in high points of equipment or buildings can be prevented by installing catalysts that cause the (relatively) slow oxidation of the gas and its conversion to water. Odorants such as mercaptans can be added to the hydrogen to alert people to any escaping gas.

The numerous processes for the production of hydrogen include:

1. chemical,
2. electrolytic,
3. thermolytic,
4. photolytic, and
5. biological.

[†] This assumes a perfectly flat cement surface. If the spill is over a flat, compacted earth surface, the area will be something like 20% of that on a nonabsorbing surface.

Hydrogen production can fall into one of several categories, including the following:

1. Production of hydrogen in massive amounts at stationary plants as, for instance, in the production of ammonia.
2. Production of hydrogen in small amounts by compact on-board plants for use in fuel cell vehicles. This last application is only now being developed and promises to become of significant economic interest.
3. Production of hydrogen in modest amounts for the food industry and for other small consumers. Frequently, electrolytic processes are employed because they yield purer gas.
4. Production of hydrogen for use in compact residential or local electricity (and hot water) generation.

10.2 Chemical Production of Hydrogen

10.2.1 Historical

On September 19, 1783, in the presence of Louis XVI, a hot-air balloon built by the Montgolfier brothers rose up carrying a duck, a rooster, and a sheep. This marked the first recorded aeronautical accident: the sheep kicked the rooster, breaking its wing. Otherwise, the experiment was a success and encouraged humans to try an air trip by themselves. On November 21 of the same year, Jean-François Pilâtre de Rozier, a French chemist, and his countryman, François Laurent Marquis d'Arlandes, stayed aloft for 25 minutes and rose to 150 m altitude. It is noteworthy that no sheep was taken along and that the Montgolfiers, prudently, also stayed on the ground.

Hydrogen balloons were introduced surprisingly early—on December 1, 1783, only 10 days after the first manned hot-air flight, Jacques-Alexandre-César Charles and one of the Robert brothers[†] made their ascent. The test was a little too successful; after a short trip, Robert disembarked to salute the onlookers, causing the lightened balloon to rise quickly to 2700 m carrying the flustered Charles, who, eventually, opened a valve dumping some hydrogen and returned safely to the ground. Charles is the physicist celebrated in Charles's law (1787), which states that the volume of a fixed amount of gas at constant pressure is proportional to the temperature. This is, of course, a part of the ideal gas law.

The more primitive technology of hot-air balloons has survived and has grown greatly in popularity, whereas no one uses hydrogen for this purpose any more.

[†] Marie-Noël Robert. Strangely, he and his brother, Anne-Jean, had feminine first names.

Early on, hydrogen was produced by passing steam over red-hot iron filings. The iron combines with oxygen in the water, liberating hydrogen:[†]

$$3\mathrm{Fe} + 4\mathrm{H_2O} \rightleftharpoons \mathrm{Fe_3O_4} + 4\mathrm{H_2}. \tag{10.1}$$

The gas was then washed by bubbling it through water.

The iron oxide formed is the ferroso-ferric oxide in which iron appears in two different states of oxidation—"ferrous" (iron II) usually forming pale green salts, and "ferric" (iron III) usually forming yellow-orange-brownish salts. The mineral magnetite is naturally occurring Fe_3O_4, a compound that is also used in ferrites employed in some electronic devices.

After 1850, hydrogen for balloons was frequently produced from the reaction of iron with sulfuric acid. The high price of the sulfuric acid led to costly hydrogen.

10.2.2 Metal–Water Hydrogen Production

Small amounts of hydrogen are produced even now by making aluminum chips react with caustic soda (NaOH). This is sometimes the source of the gas used in meteorological balloons:

$$2\mathrm{Al} + 2\mathrm{KOH} + 6\mathrm{H_2O} \rightarrow 2\mathrm{KAl(OH)_4} + 3\mathrm{H_2}. \tag{10.2}$$

Jerry Woodall of Purdue University proposes the production of hydrogen by decomposition of water using aluminum's affinity for oxygen. The basic reaction is

$$2\mathrm{Al} + 3\mathrm{H_2O} \rightarrow 3\mathrm{H_2} + \mathrm{Al_2O_3}. \tag{10.3}$$

This reaction is strongly exothermic—it causes an enthalpy change of 812 MJ per kmole of alumina produced, and there is a strong thermodynamic driving force (the ΔG change due to the reaction is -864 MJ/kmole). Nevertheless, aluminum appears perfectly stable in water as attested by its common use in cooking utensils. The reason is the passivation that results from the very quick formation of a tough layer of oxide that isolates the metal from contact with the oxygen in the air. The oxide—alumina—is very hard (it is the abrasive, corundum[††]) and quite refractory (it melts at about 2330 K). It adheres strongly to the aluminum surface.

Aluminum is combustible, burning with an intense flame. Fortunately, because of the passivation, it is hard to ignite. Aluminum aircraft are usually safe from fire, but, if the metal is ignited, they will burn spectacularly.

[†] The iron/water reaction is now being considered as a way of storing hydrogen. See Chapter 11.

[††] Corundum is usually clear but can acquire beautiful colors when contaminated by impurities: when deep red, it is called *ruby*, and when of other colors, it is called *sapphire*.

Aluminum can be dissolved in certain substances. It makes sense that when dissolved, the aluminum, having no fixed surface that can be passivated, should react with water-producing hydrogen. Indeed, this is what happens when aluminum is dissolved in gallium—one of the few metals (together with cesium, rubidium, francium, and mercury) that are liquid at room temperature.[†] An alloy containing, by weight, 2.5% aluminum and 97.5% gallium remains liquid down to below 50 C. Pouring water on this liquid causes an abundant evolution of hydrogen, consuming aluminum but leaving the gallium untouched.

Increasing the percentage of aluminum to 80% produces (a metastable) alloy that is quite solid at room temperature but, surprisingly, still reacts vigorously with water. This alloy, patented by Jerry Woodall, is more attractive for hydrogen production than the gallium-rich liquid not only because it is an easily handled solid but also because only a modest amount of gallium is involved. To be sure, when all the aluminum is used up, much of the gallium can be recovered. Nevertheless, the less gallium, the better the economics. Both the alumina and the aluminum-depleted gallium alloy are recycled. This requires a system for collecting these leftover materials. The alumina is then electrolytically reduced back to aluminum for reuse.

One should ask, what are the energetic efficiency, the specific gravimetric hydrogen concentration, and the overall cost? The last question depends on numerous assumptions and usually leads to optimistic conclusions during the system development phase and, often, to later disappointments.

Energetically, as we saw, the aluminum/water reaction releases 812 MJ of heat (which is wasted) and 858 MJ in the form of 3 kmoles of hydrogen, all per kilomole of alumina produced. Thus only half of the chemical energy of the aluminum/water reaction is transformed into hydrogen. To regenerate the aluminum from the alumina, we must use an electrolytic process that has a theoretical requirement of 23 MJ per kg of aluminum. Although the efficiency of the process improved throughout the twentieth century, it seems improbable that it will ever much exceed the present-day value of about 50%. This means that some 50 MJ of electricity are needed to regenerate 1 kg of aluminum, which can yield 16 MJ of hydrogen. Thus, considering only the energy to recover the aluminum, the energy efficiency of producing hydrogen by the proposed method is 32%.

Gravimetrically, we can see from Equation 10.3 that 1 kmole of aluminum (27 kg) yields 1.5 kilomoles of hydrogen (3 kg). This would by itself be an acceptable mass ratio of 5.5%. Observe that we did not include the mass of the water needed for the reaction because this water can be supplied from the very exhaust of the fuel cell that is using the produced hydrogen. About 6% is what one can currently expect of the very best compressed hydrogen or metal hydride systems. What our computation

[†] Provided you live in the tropics: the melting point of gallium is 29.7 C.

does not take into account is the mass of all the equipment needed for the aluminum/water reaction.

10.2.3 Large-scale Hydrogen Production

The bulk of the hydrogen produced in the world is made from fossil fuels. Oil, naphtha, and natural gas are still the main materials used. Owing to their growing scarcity, some effort is being made to use the more abundant coal, although the high sulfur content of many coals has led to serious ecological concerns.

Hydrocarbons and alcohols, among other substances, can yield hydrogen when submitted to **partial oxidation, steam reforming**, or **thermal decomposition**. These processes lead to a mixture of CO and H_2 called **syngas**.

When any of the above reactions is used in a fuel processor to feed fuel cells with pure hydrogen, the efficiency, η, can be defined as

$$\eta \equiv \frac{\text{Lower heat value of the hydrogen delivered to the fuel cell}}{\text{Higher heat value of feed stock} + \text{higher heat value of fuel used for heat}}.$$

10.2.3.1 Partial Oxidation

Partial oxidation can be carried out noncatalytically (POX) or catalytically (**ATR** or **autothermal reaction**).

Partial oxidation is preferred when the raw material is a heavier fraction of petroleum, while steam reforming is more convenient for lighter ones. However, small fuel processors for automotive use based on partial oxidation of methanol are being seriously considered.

In the partial oxidation process, air is used as oxidant, and this results in nitrogen being mixed with the hydrogen produced, reducing the partial pressure of the hydrogen and, consequently, lowering the fuel cell output.

Partial oxidation is accomplished by reacting a fuel with a restricted amount of oxygen:

$$C_nH_m + \frac{n}{2}O_2 \rightarrow nCO + \frac{m}{2}H_2. \tag{10.4}$$

Thus, for the case of methane,

$$CH_4 + \frac{1}{2}O_2 \rightarrow CO + 2H_2. \tag{10.5}$$

These reactions take advantage of oxygen having a greater affinity for carbon than for hydrogen.

10.2.3.2 Steam Reforming

In steam reforming, the fuel reacts with water that adds its hydrogen to that from the fuel and does not introduce any nitrogen into the reformate.

This contrasts with the partial oxidation process. Steam reforming of a generalized hydrocarbon proceeds according to

$$C_nH_m + nH_2O \rightarrow nCO + \left(\frac{m+2n}{2}\right) H_2. \tag{10.6}$$

This reaction is also known as the **carbon-steam** reaction.

As an example, consider carbon itself (let $m = 0$ to cancel out the hydrogen in the hydrocarbon). Note that in this case, all the hydrogen comes from the water, and the fuel contributes only energy:

$$C + H_2O \rightarrow CO + H_2. \tag{10.7}$$

Consider also methane,

$$CH_4 + H_2O \rightarrow CO + 3H_2. \tag{10.8}$$

10.2.3.3 Thermal Decomposition

Thermal decomposition of alcohols can be exemplified by the methanol and ethanol reactions indicated below:

$$CH_3OH \rightarrow CO + 2H_2 \tag{10.9}$$

and

$$C_2H_5OH \rightarrow CO + H_2 + CH_4. \tag{10.10}$$

All the hydrogen comes from the fuel used.

10.2.3.4 Syngas

Syngas, the mixture of CO and H_2 that results from all the reactions discussed so far, can be used directly as fuel. It can even be directly used in molten carbonate and ceramic fuel cells, but, owing to the presence of the carbon monoxide, it is totally incompatible with low-temperature fuel cells such as SPFCs.

Syngas has been used as domestic and industrial fuel, but its low energy per unit volume makes it unattractive if it has to be pumped to a distant consumer. For such application, the gas can be **enriched** by transforming it into methane (see Equation 10.16). This is the basis of many coal gasification processes. Observe that the preceding syngas is dangerously poisonous owing to the carbon monoxide it contains.

An important use of syngas is as a feedstock for the production of an amazing number of chemicals. Many of these have an H/C ratio substantially larger than that of syngas. For this reason, and for its use in low temperature fuel cells, a hydrogen enriching step may be needed. This is known as a **shift** reaction.

10.2.3.5 Shift Reaction

The shift reaction promotes the combination of carbon monoxide with water. The result is carbon dioxide and more hydrogen:

$$CO + H_2O \rightarrow CO_2 + H_2. \tag{10.11}$$

By using the shift reaction, it is possible to adjust the H/C ratio of syngas over a wide range of values. For fuel cells, the shift reaction is used to (nearly) eliminate all the CO.

As an example, consider the production of hydrogen from natural gas (methane):

$$CH_4 + H_2O \rightarrow CO + 3H_2 \tag{10.12}$$

followed by

$$CO + H_2O \rightarrow CO_2 + H_2 \tag{10.13}$$

with the overall result:

$$CH_4 + 2H_2O \rightarrow CO_2 + 4H_2. \tag{10.14}$$

Notice that the heat of combustion of methane into liquid water is 890 MJ/kmole, while the combustion of 4 kilomoles of hydrogen, again into liquid water, yields $4 \times 286 = 1144$ MJ. Thus, in the above reaction, the products have more energy than the reactants—the reaction is endothermic. The extra energy comes from the heat necessary for the reaction to proceed. In fixed installations, this heat usually comes from the combustion of hydrocarbons:

$$C_nH_m + \left(n + \frac{m}{4}\right) O_2 \rightarrow nCO_2 + \frac{1}{2}mH_2O. \tag{10.15}$$

In the more compact automotive and residential uses, the heat may conveniently come from the combustion of part of the hydrogen in the reformate. See the example further on.

10.2.3.6 Methanation

The transformation of syngas into methane, part of the process of transforming any fossil fuel into the (usually) more valuable “natural gas,” is called **methanation**. Besides being of great industrial importance, methanation is of interest to us in this text because it provides a technique for eliminating most of the CO impurity from the stream of hydrogen produced from carbon-bearing fuels. The methanation reaction is

$$CO + 3H_2 \rightarrow H_2O + CH_4. \tag{10.16}$$

This is the reverse of steam reforming of methane shown in Equation 10.8. Incidentally, carbon dioxide can also be transformed into methane:

$$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O. \tag{10.17}$$

10.2.3.7 Methanol

In addition to being a valuable fuel and chemical, methanol is an important intermediate in the production of many other chemicals. This is particularly true because it is the only substance that can be produced singly and with good efficiency from syngas. Efficient production of other substances leads to mixtures that can be difficult to separate.

Methanol may become the fuel of choice for fuel cell cars. It can be produced from syngas:

$$CO + 2H_2 \rightarrow CH_3OH. \tag{10.18}$$

This reaction, discovered in 1902 by Paul Sabatier and Jean Baptiste Senderens, is the base of the **Fischer–Tropsch** process that attained such fame in Germany during World War II, generating liquid fuels from coal.

The plant that produces methanol bears a strong resemblance to an ammonia plant. The difference lies in the type of syngas and in the catalysts used in the reactor.

It should be pointed out that the end product of syngas-based products is controlled by adjusting temperatures and pressures and by the choice of catalysts.

Methanol can also be produced directly from biomass such as wood.

There are some problems with the widespread use of methanol as an automotive fuel. The main problem is its toxicity. This is one reason to prefer formic acid (see the corresponding subsection), which is far less toxic. The other is its low volumetric energy concentration, which is the lead motivation for the development of higher alcohols and other liquids as automotive fuels. See Chapter 13.

10.2.3.8 Syncrude

Syncrude is the term that describes the liquid products resulting from coal liquefaction. Liquefaction is a more efficient process of converting coal than gasification. It requires little water and can use all sorts of coal, including bituminous coals that tend to cake when submitted to gasification. There are essentially four syncrude processes:

1. The Fischer–Tropsch process similar to that used for production of methanol. Here, however, selectivity is not desired: instead of pure methanol, the process yields a complex mixture of hydrocarbons.
2. **Pyrolysis**, the destructive distillation of coal in the absence of air results in gases, liquids, and solids (char). Coal is flash heated

because prolonged heating will cause the liquid fraction to crack forming gases.

3. Direct hydrogenation of coal.
4. Solvent extraction of liquids. The solvents used are produced in a preliminary direct hydrogenation step.

10.2.4 Hydrogen Purification

Hydrogen derived from electrolysis (see later) comes close to being acceptably pure when leaving the electrolyzer. On the other hand, when hydrogen is derived from fossil fuels, it is accompanied by many impurities including massive amounts of CO_2, objectionable traces of CO and, in some processes, large amounts of nitrogen. In addition, the feed stock itself may contain undesirable components, such as sulfur, which must be removed prior to processing.

10.2.4.1 Desulfurization

If the feedstock is in gaseous form, sulfur can be removed by spraying it with a calcium-based (limestone, for instance) slurry. The SO_2 in the gas reacts with the slurry, producing, sulfites or sulfates, which are then removed.

Catalysts containing molybdenum disulfide with low concentrations of cobalt or nickel can be used to convert sulfur-bearing molecules in heavy crude into H_2S gas.

A number of other desulfurization processes exist.

10.2.4.2 CO_2 Removal

Syngas as well as biogas (see Chapter 13) contains large percentages of carbon dioxide, which, at best, acts as a dilutant. In the presence of water, a destructive acid is formed that can damage equipment and pipelines. Removal of CO_2 is a technique central to sequestering schemes aimed at reducing CO_2 emissions into the atmosphere.

The removal of the CO_2 can be accomplished by a number of processes including

1. Chemical methods that use, for example, calcium hydroxide to absorb the carbon dioxide, forming calcium carbonate. In a later step, the carbonate is regenerated to hydroxide.
2. Physical processes called **temperature swing adsorption (TSA)**, which take advantage of the solubility variation of CO_2 with temperature. The solvents may be water, methanol, or one of the three ethanol amines (mono- or MEA, di- or DEA, and tri- or TEA).
3. Currently, the most popular technique is **pressure swing adsorption (PSA)**, which employs the ability of certain substances such as some zeolites to selectively adsorb carbon dioxide (or some other substances) at high pressure and then desorb them when the

pressure is lowered. In an effort to design better adsorbers, organic molecules are being investigated. See Atwood, Barbour, and Jerga (2004).

4. Partial removal of CO_2 can be achieved by the use of special membranes (cellulose acetate, for example) that display a higher permeability to carbon dioxide than for other molecules. CO_2 is a relatively large molecule; hence the selectivity cannot be based on pore size. In other words, the membranes do not act as filters—they are nonporous. Carbon dioxide dissolves into the membrane, diffuses through it, and emerges on the other side. Some of the useful gas is lost in the process: 85% can be recovered if 3% CO_2 is tolerated in the exhaust, and 90–92% is recovered if 8% CO_2 in the exhaust is acceptable.

10.2.4.3 CO Removal and Hydrogen Extraction

Hydrogen extraction (with removal of most of the CO) can be achieved by means of metallic membranes that allow the passage of H_2 but not of other gases. Again, although known as "filters," they are nonporous and depend on the dissociation of H_2 into H, which then forms a hydride (see Chapter 11) that diffuses rapidly through the sheet and reconstitutes the molecular hydrogen on the other side. This is a two-step mechanism: the dissociation and the hydridization and diffusion. Tantalum allows the second step to proceed efficiently but is a poor catalyst for the dissociation. Palladium performs both steps very well but is expensive. One possible solution is the use of tantalum with a very thin palladium plating. Since hydrogen tends to embrittle the palladium, the palladium is alloyed with gold, silver, or copper (typically 60% Pa and 40% Cu).

Sizable flow rates through the membrane without excessive pressure differentials demand small thicknesses. This economizes palladium but can result in very fragile sheets that may have imperfections in the form of minute pinholes. Unwanted gases seeping through such pinholes destroy the selectivity of the "filter." Deposition of very thin but very uniform layers of palladium on top of a high-porosity substrate may solve the problem. The degree of undesirable porosity of the membrane can be inferred from a measurement of a helium flux through it. Helium, unlike hydrogen, can only cross the membrane by passing through the pinholes.

Data from IdaTech (see Figure 10.1) show the relationship between the thickness of the membrane and the flux rate of hydrogen—that is, the number of kilowatts of hydrogen that flow each second through each square meter of membrane (kg/s being translated into watts using the lower heat value of hydrogen). The data were measured with the membrane at 400 C and a pressure differential of 6.6 atmos. The difference in permeability of smooth compared to etched membranes is obvious. Etched membranes are rough and thus present a larger surface area than smooth ones, thereby improving the surface reaction with hydrogen. This effect is more noticeable

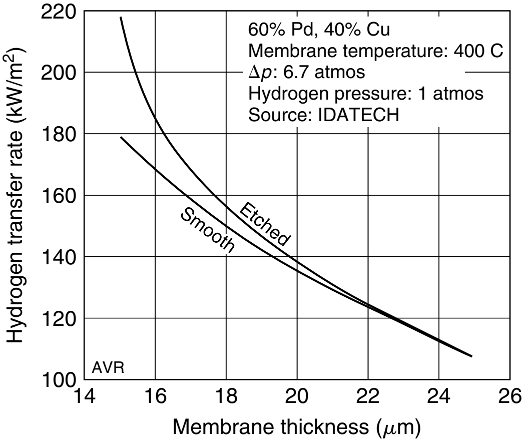


Figure 10.1 Rate of hydrogen transport through a palladium–alloy membrane. Thin etched membranes transport hydrogen better than the smooth ones, owing to the increased surface area.

with thin membranes because, in the thicker ones, the rate of permeation is controlled mostly by the gas diffusion through the bulk of the material.

Palladium is an expensive metal. A 1-kW fuel cell operating at 60% efficiency requires a hydrogen input of around 1700 W. A 17-μm thick membrane will produce, at a Δp of 6.7 atmospheres, hydrogen at a rate of about 170 kW/m^2; hence a total effective membrane area of about 0.01 m^2 or 100 cm^2 is required. If the total area of the palladium sheet is 1.5 times larger than this and if the average thickness of the sheet is 20 μm, then the total amount of metal is 0.3 cm^3. Palladium constitutes 60% of the alloy or 0.18 cm^3. This corresponds to about 2.2 grams. Owing to the use of palladium in automobile catalytic converters, the price of the metal has risen considerably in the last decade. At the current $25/gram, the cost of palladium in a 1-kW fuel cell would be $55, not a negligible amount.

If it proves possible to reduce the palladium thickness to, say, 1 or 2 μm, the cost of the material would fall substantially. Also, the designer could opt for a lower Δp across the membrane.

10.2.4.4 Hydrogen Production Plants

Large-scale hydrogen production starting from fossil fuels (frequently, for the production of ammonia) is a mature technology. A typical setup is shown in Figure 10.2. The first step in the process is frequently the desulfurization of the feedstock because sulfur tends to poison the catalysts required in some of the subsequent steps. Next, syngas is produced by steam reforming. This is, as pointed out previously, an endothermic reaction that requires heat input.

The shift reaction that eliminates most of the CO releases substantially less heat than that needed to drive the reforming (or decomposition)

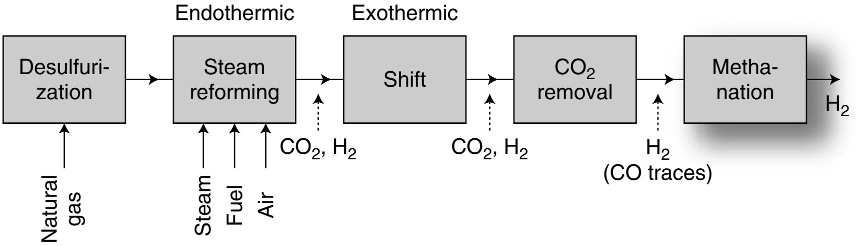


Figure 10.2 Typical hydrogen production sequence.

reaction. The shift reaction, being exothermic, profits from operation at low temperatures owing to equilibrium considerations. However, this influences unfavorably the kinetics of the reaction. Consequently, attention has been given to the development of good catalysts. Early catalysts based on nickel, cobalt, or iron oxide required temperatures above 700 K. Modern copper-based catalysts allow operation at temperatures as low as 520 K. After the shift reaction, the gas contains large amounts of carbon dioxide mixed with hydrogen. This requires a CO_2 removal step. The final step in the hydrogen production sequence is the elimination of residual CO, which otherwise would alter the catalyst in the ammonia synthesis process.

10.2.5 Compact Fuel Processors

Compact fuel processors for use in automobiles or for residential applications are vigorously being developed at the moment. Much of the work has been concentrated on adapting the classical industrial hydrogen production techniques to the much miniaturized requirements of the in situ hydrogen generators for automotive uses.

Before discussing an example of a miniature steam reforming plant, we will present an example of calculations required to specify the device. For this, we will need some thermodynamic data, which, for convenience, we summarize in Table 10.1.

IdaTech (of Bend, OR) has developed a series of processors that work on the lines of the above example.† They can handle different fuels such as methanol, methane, and others. Refer to Figure 10.3. An equimolar pressurized methanol/water mixture, or equivalent feedstock, is forced into the **reforming region** having first been heated and vaporized in a coil-type heat exchanger exposed to the high temperature of the **combustion region**. In the reforming region, the methanol is converted to H_2 and CO. The CO is water gas shifted into CO_2 and more hydrogen. However, considerable amounts of impurities are left. To separate the hydrogen from these impurities, part of the gas is allowed to pass through a palladium filter

†U.S. Patents 5,861,132, 5,997,594, and 6,152,995.

Table 10.1 Thermodynamic Data

Substance	Formula	Heat of combustion	Heat of combustion	$\overline{h}^{*}_{\text{vap}}$	$\overline{h}_f^{\circ}$	$\overline{g}_f^{\circ}$	$\overline{s}^{\circ}$
		(Higher)	(Lower)	kJ/	MJ/	MJ/	kJ K^{-1}/
		MJ/kmol	MJ/kmol	kmol	kmol	kmol	kmol
Carbon	C	−393.52			0	0	5.74
C dioxide	CO_2				−393.52	−394.36	213.80
C monoxide	CO	−282.99			−110.53	−137.15	197.65
Ethanol (g)	C_2H_5OH	−1409.30	−1277	42.34	−235.31	−168.57	282.59
Ethanol (l)	C_2H_5OH				−277.69	−174.89	160.70
Hydrogen	H				218.00	203.29	114.72
Hydrogen	H_2	−285.84	−241.80		0	0	130.68
Hydroxyl	OH				39.46	34.28	183.7
Methane	CH_4	−890.36	−802.16		−74.85	−50.79	186.16
Methanol (g)	CH_3OH	−764.54	−676.34	37.9	−200.67	−162.00	239.70
Methanol (l)	CH_3OH				−238.66	−166.36	126.80
Oxygen	O				249.19	231.77	161.06
Oxygen	O_2				0	0	205.04
Water (g)	H_2O			44.1	−241.82	−228.59	188.83
Water (l)	H_2O				−285.83	−237.18	69.92

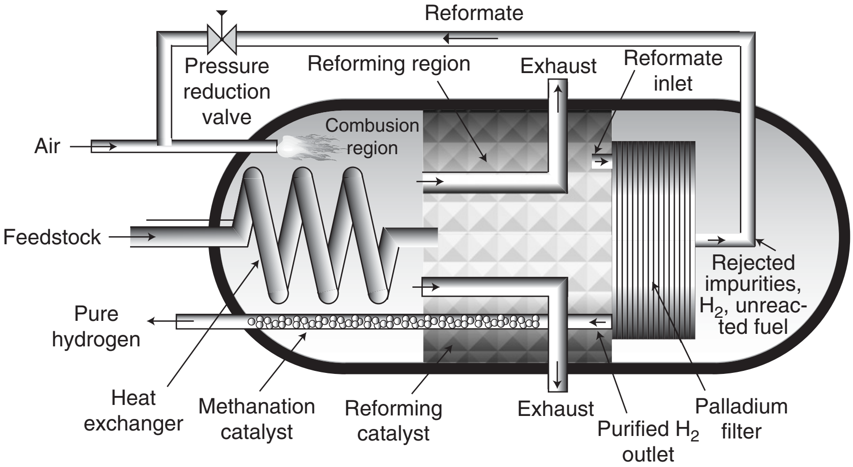


Figure 10.3 One configuration of the IdaTech fuel processor.

Example

Consider a methanol-to-hydrogen converter appropriate for vehicular or for residential use. Such a converter or **fuel processor** might employ the direct methanol decomposition reaction,

$$CH_3OH \rightarrow CO + 2H_2. \qquad (10.19)$$

(Continues)

(*Continued*)

If the methanol were evaporated and burned, it would yield (see Table 10.1) 676.34 MJ per kilomole of water vapor being produced. If the alcohol is first decomposed, then the resulting 2 kilomoles of hydrogen would yield 2 × 241.82 = 483.64 MJ, while the carbon monoxide would yield an additional 282.99 MJ, for a total of 766.63 MJ.

In other words, the products of the decomposition yield, upon combustion, more energy than the original fuel—the decomposition is endothermic and requires a heat input of 766.63 − 676.34 = 90.29 MJ per kilomole of methanol.

It is important to get rid of much of the CO and to extract additional H_2. This can be accomplished by employing a shift reaction,

$$CO + H_2O \rightarrow CO_2 + H_2. \tag{10.20}$$

When burned, CO releases 282.99 MJ/kmol, while the product hydrogen releases 241.8. The reaction is thus exothermic, providing 282.99 − 241.8 = 41.19 MJ of heat per kilomole of CO.

The overall reaction is

$$CH_3OH + H_2O \rightarrow CO_2 + 3H_2, \tag{10.21}$$

which, of course, is endothermic requiring 90.29 − 41.19 = 49.00 MJ per kilomole of methanol.

For each kilomole of methanol, 1 kmole of water is required. The system works by taking in the correct alcohol/water mixture, evaporating it, and heating it to the reaction temperature, which may be between 200 and 600 Celsius. The energy budget is

$$\Delta H_{_{TOTAL}} = \Delta H_{_{VAP}} + \Delta H_{_{C_P}} + \Delta H_{_{REACT}} + \Delta H_{_{LOSS}}. \tag{10.22}$$

$\Delta H_{_{VAP}}$ is the amount of energy required to vaporize the fuel/water mixture. For the example under consideration,

$$\begin{aligned}\Delta H_{_{VAP}} &= \Delta H_{VAP_{METHANOL}} + \Delta H_{VAP_{WATER}}\\ &= (37.9 \times 10^6 + 44.1 \times 10^6)\\ &= 82.0 \times 10^6 \ \text{ J/kmole of methanol.}\end{aligned} \tag{10.23}$$

$\Delta H_{_{C_P}}$ is the energy required to raise the temperature of the vaporized fuel/water to the operating temperature. Since the specific heat of most substances varies considerably with temperature, and since the exact operating temperature in this example is unspecified, we will

(*Continues*)

(Continued)

use representative values of c_p: 37 kJ K^{-1} for water and 39 kJ K^{-1} for methanol.

$$\begin{aligned} H_{c_p} &= c_{p_{METHANOL}}\Delta T + c_{p_{WATER}}\Delta T \\ &= 37\times 10^3\Delta T + 39\times 10^3\Delta T = 76\times 10^3\Delta T. \end{aligned} \tag{10.24}$$

If, somewhat arbitrarily, we take $\Delta T = 250$ K—that is, if we chose a reaction temperature of $298+250 = 548$ K or 275 C—then the energy to vaporize the water/fuel mixture is

$$\Delta H_{c_p} = 19\times 10^6 \text{ J/kmole of methanol.} \tag{10.25}$$

ΔH_{REACT} was calculated at the beginning of this section. It amounts to 49 MJ per kilomole of methanol.

Finally, ΔH_{LOSS} can be minimized by good insulation. We will consider it to be negligible in this example. $\Delta H_{TOTAL} = 82+19+49 =$ 150 MJ/kmole of methanol. This energy can be obtained by burning $150/242 = 0.62$ kilomole of hydrogen. Hence, of the 3 kmoles of hydrogen produced, about 20% must be diverted to the burner and about 80% is available at the output. The fuel processor converts 676 MJ of methanol into 575 MJ of hydrogen, a ratio of 85%.

If energy were not expended in vaporization and heating the fuel/water mixture and if there were no losses, then the heat required would be solely that needed to drive the decomposition: 49 MJ/kilomole of methanol. This requires 0.2 kmole of hydrogen, and the useful hydrogen output would be 2.8 kmoles or 676 MJ. In other words, the fuel processor would have, as expected, 100% efficiency.

The output gas stream is at a temperature much higher than that of the input fuel/water mixture. It appears that some of the output heat could be used to preheat the input and thus increase the overall efficiency of the conversion.

that is selectively permeable to H_2. A small part of the hydrogen produced in the reforming region (mixed with the rest of reformate leftover gases), instead of going through the palladium membrane, is forced out into the combustion region where it burns, combining with air introduced into the apparatus by means of a blower.[†] A spark igniter, not shown in the

[†] When operating with a fuel cell, the air for the fuel processor may be supplied by the fuel cell cathode exhaust stream.

figure, starts the combustion. This hydrogen (part of the nonpurified reformate) provides the energy necessary to drive the reforming reaction. No external source of heat is needed, except during start-up. At start-up, a small electric heater raises the temperature of the reforming region to initiate the reaction. Start up times as short as three minutes have been demonstrated.

Even after being filtered through the palladium, a small amount of CO and CO_2 is still present in the hydrogen. Before leaving the equipment, the hydrogen flows through an outlet tube inside of which an appropriate catalyst causes these impurities to be transformed into methane, a gas that is innocuous as far as the catalysts used in fuel cells are concerned. High purity hydrogen, containing less than 1 ppm of CO and less than 5 ppm of CO_2, is produced by this fuel processor.

In some fuel processor configurations, the membrane is formed into a tube, while in others it is left in planar shape. The latter solution, adopted by IdaTech, leads to a very compact filter: the membrane consists of a stack of elements interconnected by manifolds so that all elementary filters are in parallel. The stack, see, for example, Figure 10.3, has one reformate inlet, one outlet for the purified hydrogen, and one outlet for the rejected impurities plus some hydrogen and some unreacted fuel.

Preparation of the palladium–alloy membrane by IdaTech starts with a sheet of the material, typically 50 micrometers thick, which is held in place by a rectangular frame. A center region of the sheet is etched, while the borders are left untouched and retain the original thickness. The etch area is defined by a piece of filter paper attached to the surface. The etchant, generally aqua regia,[†] saturates the paper, which distributes the chemical evenly over the selected area.

10.2.5.1 Formic Acid

The fuel processors described in the previous subsubsection operate at relatively high temperatures (>200 C). There are, however, clear advantages in working at room temperature. This has been done by Loges et al. (2008) using the catalytic decomposition of formic acid.

Formic acid, HCOOH, is an inexpensive liquid of much lower toxicity than methanol, but able to carry, by weight, only about one third the hydrogen carried by the alcohol. Formally, it decomposes as:

$$\mathrm{HCOOH} \rightarrow \mathrm{H_2} + \mathrm{CO_2}. \tag{10.26}$$

[†] Aqua regia ("royal water") is a mixture of 25% nitric acid and 75% hydrochloric acid. It has the property of dissolving gold, although neither acid will do this by itself.

Actually, to make the reaction proceed, formic acid is combined (via an addition reaction) with an amine such as triethylamine, abbreviated NEt_3. This results in a formate or a formic acid adduct.[†] A commercial ruthenium catalyst, tris(triphenylphosphine)ruthenium(II) chloride, $RuCl_2(PPh)_3$ is used. Here "Ph" is the symbol for "phenyl." When the hydrogen evolves, the amine is recovered.

According to the investigators who developed the process, the gas is pure enough to be fed to a SPFC after passing it through a simple charcoal filter to eliminate traces of the amine. This implies that no CO is generated. Typically, the formic acid–amine ratio is 5:2.

10.3 Electrolytic Hydrogen

10.3.1 Introduction

Production of hydrogen by electrolysis is a relatively old art that has found industrial application in the food industry and in other activities that need only a moderate amount of the gas. Hydrogen produced by electrolysis has the advantage of being easily purified, whereas that produced from fossil fuels tends to contain several hard to remove contaminants.

To a small extent, electrolytic hydrogen has been used for the synthesis of ammonia, but the low cost of oil prevalent up to 1972 and again in the 1980s made this technology unattractive. However, the growing cost and uncertainty in the supply of fossil fuels may popularize electrolytic ammonia in countries where there is abundant hydroelectric power.

By far the most important electrolytic ammonia plant is one operated by Norsk Hydro in Glenfjord (Norway). This company had a long tradition of using electricity for the production of nitrogen fertilizers even before the invention of the Harber–Bosch ammonia process. Early in this century, Norsk Hydro used an electric arc operating in air to cause atmospheric nitrogen and oxygen to combine forming nitrogen oxides, to be converted into nitrates in an additional step. This is, of course, a simple but inefficient process of "fixing" atmospheric nitrogen.

In the 1920s, the plant was modified to produce ammonia by direct combination of electrolytic hydrogen with nitrogen extracted from the air. The Glenfjord plant also produced heavy water. At its peak, the plant used 380 MW of electricity and produced 1300 tons of ammonia and 85 kg of D_2O per day. At $1/g, the heavy water sales amounted to over $30 million per year. The current price of heavy water is about $0.10/g and, for this reason, the Glenfjord operation became less profitable and the production was reduced to 600 ton/day of ammonia.

[†] For more details on the chemistry involved, read Chapter 11.

Another large electrolytic ammonia plant was operated by Cominco in Trail, B.C. (Canada). Production was about 200 ton/day. The operation was discontinued when, after the oil embargo, it became more profitable to export hydroelectric energy to the United States and switch to the abundant (in Canada) natural gas for ammonia synthesis.

Originally, the Trail plant used four different unipolar electrolyzers designed, respectively, by Fauser, Pechkranz, Knowles, and Stuart, each manufacturer accounting for 25% of the total capacity. The plant was later rebuilt, and all electrolyzers were replaced by a Cominco design known as Trail-Cells.

An ammonia plant with a capacity of roughly 400 ton/day was erected in Aswan (Egypt) using, originally, Demag electrolyzers, which were replaced by Brown Bovery equipment. A plant with 60 ton/day, capacity using Lurgi electrolyzers was installed in Cuzco (Peru).

Ammonia plants are not the only consumers of electrolytic hydrogen. In fact, the vast majority of such plants use hydrogen from fossil fuels rather than from electrolysis. Electrolytic hydrogen is preferred by food and pharmaceutical industries owing to the ease with which high purity gas can be obtained—99.999% purity is not uncommon.

For certain applications, an electrolyzer/fuel cell combination constitutes an excellent way to store energy.

One important future application of electrolyzers is in **hydrogen gas stations** for refueling fuel cell vehicles.

10.3.2 Electrolyzer Configurations

10.3.2.1 Liquid Electrolyte Electrolyzers

Until relatively recently, almost all water electrolyzers used liquid electrolytes. Owing to its low conductivity, pure water cannot be used as an electrolyte: it is necessary to add a substance that increases the conductivity. Although acids satisfy this requirement (and are used in some electrolyzers), they are frequently avoided because of corrosion problems. Alkaline electrolytes are better; they almost invariably use potassium hydroxide in concentrations of 25% to 30% (about 6 M). Potassium is preferred to sodium owing to its higher conductivity. Care must be taken to avoid too much contact between air and the solution because the CO_2 in the air can combine with the hydroxide producing carbonate.

Smaller units tend to be of the **tank** type (unipolar), whereas larger units use the **filter press** (bipolar) configuration.

Tank electrolyzers consist of single cells in individual tanks in which many plates can be connected in parallel.

Bipolar electrolyzers are arranged so that all but the two end electrodes act, simultaneously, as anode for one cell and cathode for the next. The cells are automatically connected in series. In such a configuration, it is convenient to have many small cells in series instead of a few large cells as in the tank case. Consequently, bipolar electrolyzers operate with higher voltages than unipolar ones, a circumstance that reduces the cost of the power supply because, for the same power delivered, high-voltage devices cost less than high-current ones. In addition, higher voltages correspond to lower currents and demand substantially cheaper connecting bars.

To prevent the mixing of the gases generated, the cell is divided into two compartments separated by a diaphragm that, though impermeable to gases, allows free passage to ions. Asbestos is used as a diaphragm in aqueous electrolytes. It is inexpensive but is attacked by the electrolyte when the temperature exceeds 200 C.

KOH electrolyzers become more efficient when operated at higher gas pressure because this causes the gas bubbles (whose presence reduces the effective area of the electrode to which they stick) to become smaller. Pressure is generally raised by throttling the gas outlets. Lurgi produces a KOH electrolyzer that operates at 3 MPa (30 atmospheres).

KOH cells produce wet hydrogen containing fine droplets of potassium hydroxide. Steps must be taken to remove the potassium hydroxide and to dry the gas. This type of cell requires electrolyte concentration and level controls.

Electrodes must be chosen to fulfill the following requirements:

1. Corrosion resistance
2. Catalytic action
3. Large surface

In KOH electrolytes, the cathode (H_2 electrode) can be made of iron because catalysis problems are minor and iron is cheap and stable in alkaline solutions. The anode is usually nickel—in general, nickel-plated iron. The plating is made spongy to increase the effective surface. KOH cells present the disadvantage of being very bulky.

10.3.2.2 Solid-Polymer Electrolyte Electrolyzers

Solid-polymer electrolyzers are dramatically more compact than KOH electrolyzers and do not require electrolyte controls. They can operate at much higher current densities than liquid electrolyte devices. They have the disadvantage of requiring de-ionized water and of producing wet gases. Ion-exchange membranes are close to ideal electrolytes. Their advantages include the following.

1. The electrolyte can be made extremely thin, leading not only to great compactness but also to reduced series resistance owing to the short path between electrodes. Electrolyte thicknesses as small as 0.1 mm are used.
2. No diaphragm is needed: the ion-exchange membrane allows the motion of ions but not of gases.
3. The electrolyte cannot move. It has constant composition, and no electrolyte concentration controls are needed.

4. There are no corrosives either in the cells or in the gases produced.
5. Notwithstanding its thinness, the membrane can be strong enough to allow large pressure differentials between the H_2 and the O_2 sides. Differentials as large as 3 MPa (30 atmospheres) are permitted in some cells.
6. Large current densities are possible.
7. Extremely long life (20 years ?) seems possible without maintenance.

10.3.2.3 Ceramic Electrolyte Electrolyzers

In Chapter 9, we discussed ceramic fuel cells using the anion conductor yttria-stabilized zirconia as electrolyte. Cation-conducting ceramics have been proposed by Garzon et al. (2002) for use in electrolyzers. They work at temperatures between 450 and 800 C, use steam as feedstock, but produce pure, dry hydrogen. High temperature improves the kinetics of the reaction. $SrCe_{0.95}Yb_{0.05}O_{2.975}$ is an example of the material used as electrolyte.

10.3.3 Efficiency of Electrolyzers

Electrolyzers are the dual of fuel cells—much of their theory can be learned from studying Chapter 9 of this book. Before we can compare the performance of different electrolyzers, we have to say a few words about the efficiency of such devices.

We saw that the efficiency of an ideal fuel cell is the ratio of ΔG, the free energy change of the reaction, to ΔH, the enthalpy change. For a given ΔH, an amount, ΔG, of electricity is generated, and an amount, $\Delta H - \Delta G$, of heat has to be *released*.

An ideal fuel cell, being perfectly reversible, can operate as an electrolyzer: when ΔG units of energy are supplied, gases capable of releasing ΔH units of energy on recombining are formed and a quantity of heat $\Delta H - \Delta G$ is *absorbed* by the electrolyzer from the environment. This means that the ideal water electrolyzer acts as a heat pump. Thus, the efficiency of an ideal electrolyzer is

$$\eta = \frac{\Delta H}{\Delta G}, \tag{10.27}$$

which is larger than unity.

Consider a water electrolyzer operating at RTP. The change in enthalpy for the reaction is 285.9 MJ/kmole, while the free energy change is 237.2 MJ/kmole (all per kilomole of water).† Thus the ideal efficiency of

†In Chapter 9, we used 228.6 MJ/kmole as the free energy change, while here, in Chapter 10, we are using a larger value—237.2 MJ/kmole. The reason is that fuel cells synthesize water vapor (even though it may immediately condense into a liquid), while electrolyzers usually dissociate liquid water. However, when working with steam electrolyzers, the value to use is that of Chapter 9.

a water electrolyzer, with reactant and products at RTP, is

$$\eta = \frac{285.9}{237.2} = 1.205. \tag{10.28}$$

If we divide top and bottom of Equation 10.28 by qn_eN_0, we obtain quantities with the dimension of voltage, and we can express the efficiency as a voltage ratio,

$$\eta = \frac{V_H}{V}, \tag{10.29}$$

where V_H is the hypothetical voltage that an ideal fuel cell would generate if it could convert all the enthalpy change of the reaction into electricity. At RTP,

$$\eta = \frac{1.484}{1.231} = 1.205 \tag{10.30}$$

and the efficiency of any electrolyzer at RTP is

$$\eta = \frac{1.484}{V}, \tag{10.31}$$

where V is the required operating voltage.

This simple relation between required voltage and efficiency leads manufacturers to specify their equipment in terms of operating voltage or else in terms of **overvoltage** (the difference between the operating voltage and 1.484 V). If the operating voltage is less than 1.484 V, the electrolyzer absorbs heat and tends to work below ambient temperature. If the operating voltage is higher than 1.484 V, the electrolyzer must shed heat and tends to operate at above ambient temperature.

The amount of heat exchanged with the environment is

$$\dot{Q} = (V - 1.484)I. \tag{10.32}$$

If $\dot{Q} < 0$, the electrolyzer operates endothermically.

Clearly, the operating voltage of an electrolyzer depends on the current forced through the device. In a simple model, the device can be represented by a voltage generator of magnitude V_{oc} in series with an internal resistance accounting for the losses. The V-I characteristic is then a straight line as indicated in Figure 10.4.

The cell in the figure is assumed to have a characteristic given by

$$V = 1.40 + 0.001\ I. \tag{10.33}$$

In this case, $V_{oc} = 1.40$ V is the voltage obtained by extrapolation as the current forced into the electrolyzer tends toward zero. The efficiency was calculated from Equation 10.31 and the heat exchange from Equation 10.32.

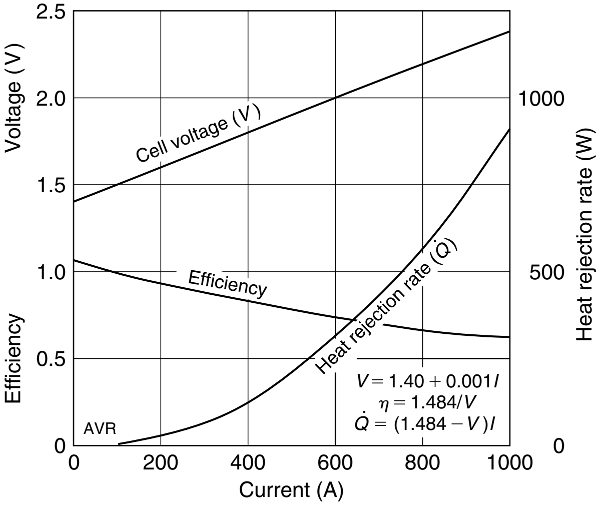


Figure 10.4 Voltage, efficiency, and heat rejection as a function of current in a typical electrolyzer.

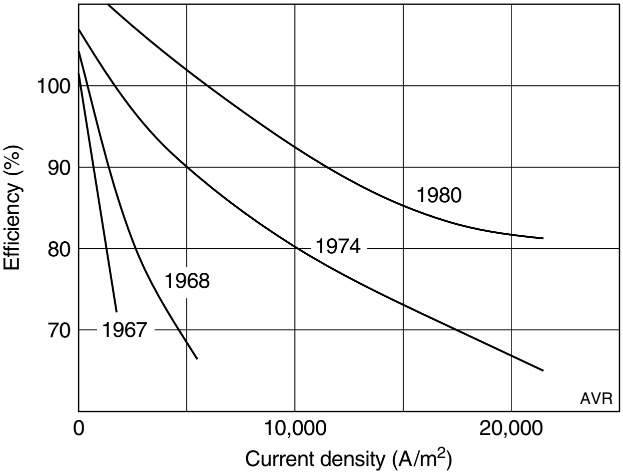


Figure 10.5 Efficiency versus current density of various GE (SPE) cells showing progress in the development.

Owing to the scale used, one cannot see from the figure that the values of $\dot{Q}$ are negative when the current is less than 80 A. Typical existing commercial electrolyzers (KOH) operate with voltages around 2 V, or 74% efficiency.

Solid-polymer electrolyzers (SPEs) perform substantially better than KOH ones. Figure 10.5 shows how the efficiency of some GE SPE cells depends on current density. The lines represent different stages of development. The rapid progress made between 1967 and 1974 was mainly due to improved catalysts, especially the anodic one (oxygen electrode). Notice that the 1974 model will operate endothermically at low current densities.

The recommended operating point of any electrolyzer is determined by economics. At low current densities, the efficiency is high—that is, a large mass of hydrogen is produced for a given amount of energy used. However, operation at low current densities results in little hydrogen produced per dollar invested in the electrolyzer. On the other hand, at high-current densities, although plenty of gas is produced per unit investment cost, little is produced per unit energy used.

There is an optimum current density that results in a minimum cost for the hydrogen produced. This depends on the cost of electricity, the interest rate, and so on. See Problem 10.1.

A manufacturer's claim that its unit operates at, say, 80% efficiency is not sufficient information. Even primitive electrolyzers will operate at these efficiency levels if the current density is kept low enough. Good electrolyzers will operate economically at high-current densities.

A typical, good KOH electrolyzer should operate at some $4\,\mathrm{kA/m^2}$, while solid-polymer devices will operate at the same efficiency with $20\,\mathrm{kA/m^2}$.

The hydrogen production rate, $\dot{N}$, is strictly proportional to the current, I, forced through the electrolyzer.

$$\dot{N} = \frac{I}{2qN_0}\ \text{kmoles/sec.} \tag{10.34}$$

The rate does not depend on the efficiency. However, the voltage necessary to drive a given current through the device is inversely proportional to the efficiency.

10.3.4 Concentration-Differential Electrolyzers

In a normal electrolyzer, the minimum amount of electric energy required is equal to the free energy change, ΔG, necessary to separate water into its constituent molecules. In concentration-differential electrolyzers, part of this minimum energy can come from nonelectric sources.

Consider, for instance, the device shown in Figure 10.6. It is a common ion-exchange membrane electrolyzer, except that the anode is bathed in a concentrated KOH solution while the cathode is in distilled water. As indicated in the figure, the KOH dissociates into K^+ and OH^-. The membrane, permeable to positive ions but not to negative ions or to electrons, allows the migration of K^+ to the cathode where it reacts with water, regenerates KOH, and forms H^+ ions. The hydroxyl ion at the anode has its negative charge removed by the external power supply and decays into water, oxygen, and electrons. The oxygen is one of the outputs of the system. The electrons, pumped by the power supply to the cathode, recombine with the protons to form hydrogen—the other output of the system.

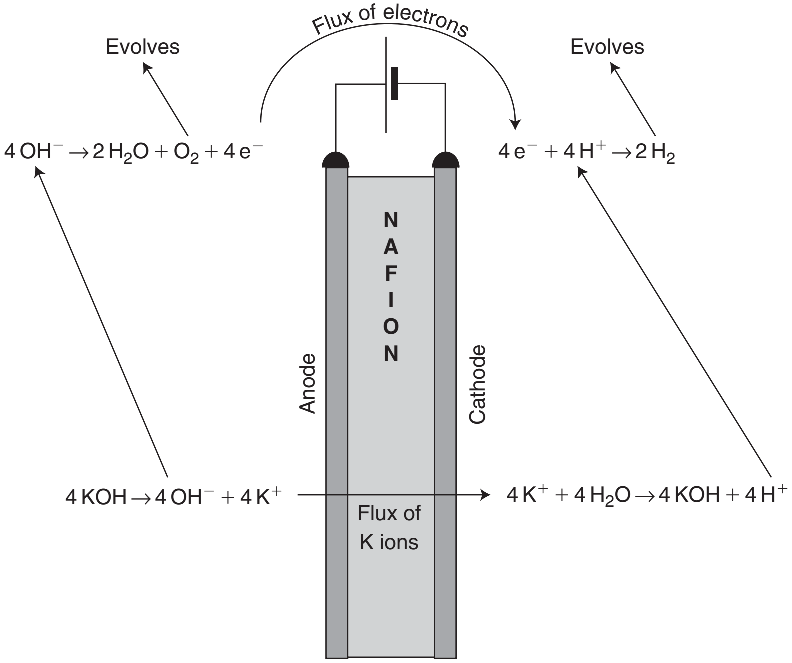

Flux of electrons
Evolves
Evolves
$4\,OH^- \rightarrow 2\,H_2O + O_2 + 4\,e^-$
$4\,e^- + 4\,H^+ \rightarrow 2\,H_2$
NAFION
Anode
Cathode
$4\,KOH \rightarrow 4\,OH^- + 4\,K^+$
$4\,K^+ + 4\,H_2O \rightarrow 4\,KOH + 4\,H^+$
Flux of
K ions

Figure 10.6 Reactions in a concentration-differential electrolyzer.

Owing to the migration of the K^+ ions, the anode becomes more negative than the cathode, so that the electrolyzer acts as a battery in series with the external power supply. To accomplish the electrolysis, the external source must provide only a voltage equal to that required by normal electrolyzers *minus* the internal "built-in" electrolyzer voltage.

As we are going to see, the built-in voltage is

$$V = 2.3\frac{kT}{q}(pH_A - pH_C), \tag{10.35}$$

where pH_A and pH_C are, respectively, the anode and cathode pH.

With concentrated KOH in the anode side and distilled water in the cathode side, we have $pH_A = 14$ and $pH_C = 7$, which leads to a built-in voltage of 0.42 V.

Thus, an ideal concentration-differential electrolyzer, at RTP, requires a voltage of $1.23 - 0.42 = 0.81$ volts and would have an efficiency of $1.48/0.81 = 1.83$. Even smaller voltages are required if, instead of distilled water, an acid solution is used to bathe the cathode.

Clearly, as the operation of the cell progresses, the KOH solution in the anode is depleted and that at the cathode is increased. To operate on a continuous basis, it is necessary to maintain the concentration disequilibrium. This can be done by washing the cathode continuously with distilled water and reconcentrating the weak KOH solution that results. This reconcentrated solution is then fed to the anode.

The energy necessary to perform the regeneration is low-grade thermal energy of far less utility than the electric energy saved.

What is the voltage developed by a concentration cell? In Chapter 7, we derived an expression for the open-circuit voltage across a concentration cell:

$$V = \frac{kT}{q} \ln \frac{N_A}{N_C}. \tag{10.36}$$

Since $\ln N = 2.3 \log N$ and, by definition, $-\log N \equiv pH$,

$$|V| = 2.3\frac{kT}{q}(pH_A - pH_C). \tag{10.37}$$

10.3.5 Electrolytic Hydrogen Compression

For many applications such as ammonia production or delivery of hydrogen to fuel cell cars, hydrogen must be available at high pressure. Thus, a hydrogen plant must frequently include a compressor.

An electrolyzer can produce gases at pressures above that of the environment provided its outlets are throttled, as is the case of the Lurgi, Teledyne, and GE equipment. This, however, limits the maximum pressure to values well below those necessary in some applications.

Alternatively, the pressure of the environment itself can be raised by housing the electrolyzer in a pressure vessel. For economic reasons, one needs compact equipment; otherwise the cost of the pressure vessel becomes prohibitive.

From Table 10.2, it can be seen that only solid-polymer electrolyzers (SPEs)—and perhaps some of the proposed positive ion ceramic electrolyzers—lend themselves to such an arrangement. The specific volume of the solid-polymer electrolyzer is almost two orders of magnitude smaller than that of pressurized KOH machines.

The energy cost of introducing liquid water into the pressure vessel is minimal owing to the negligible volume of the liquid. There is, however,

Table 10.2 Specific Volume of Various Electrolyzers

Manufacturer	**Specific volume ($\mathbf{m^3/MW}$)**
Norsk-Hydro	45
Lurgi or Teledyne	20
General Electric (SPE)†	0.3

†See Figure 10.7.

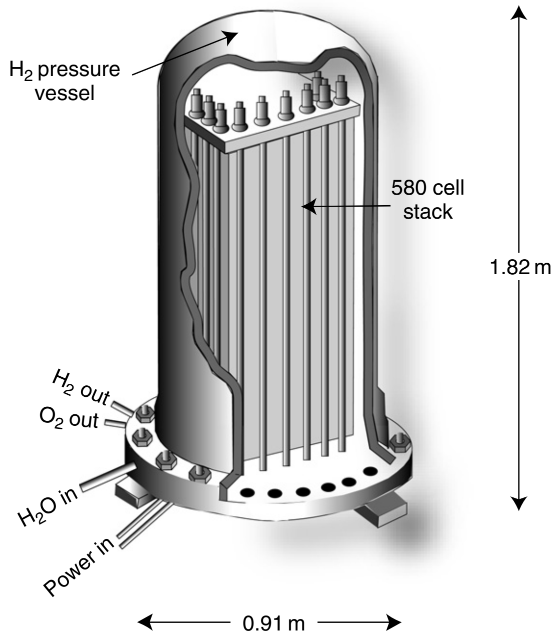


Figure 10.7 Owing to their compactness, SPE electrolyzers can be stacked in a pressure vessel to deliver gases at high pressure. Proposed GE 5 MW unit.

an energy cost in compressing “electrolytically” the product gases. Theoretically, per kilomole of hydrogen produced, an electrolyzer delivering oxygen at pressure p_{ox} and hydrogen at pressure p_h requires an excess energy, $RT_0\ \ln(p_{ox}^{1/2} p_h p_o^{-3/2})$ compared with one delivering the same mass of gases at a pressure p_0, provided all gases involved are at the temperature T_0. This is, of course, the energy required to compress the gases isothermally.

In practice, the compression energy is somewhat higher owing to the decrease in the efficiency of the ion-exchange membrane electrolyzer with increasing pressure (as a result of the partial permeability of the membrane).

The great advantage of electrolytic compression is the simplicity and economy in maintenance since there are no moving parts in the system.

10.4 Thermolytic Hydrogen

10.4.1 Direct Dissociation of Water

It is well known that water vapor, at high temperatures, will dissociate into hydrogen and oxygen. Hydrogen could, in principle, be separated out by using an appropriate device such as the palladium filter discussed in the subsection on CO removal. Although, at a first glance, this scheme may seem attractive, its implementation is difficult.

Consider an experiment that consists of introducing 1 kilomole of water vapor (and nothing else) into a cylindrical container equipped with a piston weighed down so as to maintain a constant pressure of 1 atmosphere. The water is then heated to 3000 K. This pressure and this temperature have been chosen arbitrarily to serve as an example.

If all the molecules in the container remained as H_2O, one could look up their free energy by consulting an appropriate table. However, at least some of the water vapor will dissociate into its components, hydrogen and oxygen, according to

$$H_2O \longleftrightarrow H_2 + \frac{1}{2}O_2, \tag{10.38}$$

and the mixture, containing, H_2O, H_2, and O_2, will have a different free energy.

If the dissociation were complete, the gases in the container would consist of 1 kilomole of hydrogen and 0.5 kilomole of oxygen. The free energy of this gas (still at 1 atmosphere and 3000 K) will be found to be larger than that of the pure H_2O vapor. Notice that the 1 kilomole of H_2O has been transformed into 1 kilomole of H_2 plus 0.5 kilomole of O_2—the total number of kilomoles is now, $N_{TOTAL} = 1.5$. Thus, the partial pressure of the hydrogen would be 1/1.5 atmospheres, and that of the oxygen must be 0.5/1.5 atmospheres.

At any realistic temperature, water will only partially dissociate. Let F be the fraction of dissociated water. If so, the number of kilomoles of undissociated water (from the 1 kmole of water introduced into the container) is $(1 - F)$. The number of kilomoles of H_2 is F, and that of O_2 is $F/2$. The mixture consists of

$$(1 - F)H_2O + FH_2 + \frac{1}{2}FO_2. \tag{10.39}$$

The total number of kilomoles in the gas mixture is

$$N_{TOTAL} = (1 - F) + F + \frac{1}{2}F = 1 + \frac{1}{2}F. \tag{10.40}$$

The partial pressure of the three species in a mixture whose total pressure is p atmospheres is

$$p_{H_2O} = \frac{1 - F}{1 + F/2}p = \frac{1 - F}{1 + F/2}, \tag{10.41}$$

$$p_{H_2} = \frac{F}{1 + F/2}p = \frac{F}{1 + F/2}, \tag{10.42}$$

$$p_{O_2} = \frac{F/2}{1 + F/2}p = \frac{F/2}{1 + F/2}. \tag{10.43}$$

We took $p = 1$ atmosphere according to the conditions we selected for this example.

The free energy of a species varies with pressure as

$$\overline{g}_i = \overline{g}_i^* + RT \ln p_i, \tag{10.44}$$

where $\overline{g}_i^*$ is the free energy of species *i per kilomole at 1 atmosphere pressure*. See sub-subsection 9.7.4.4 “Free Energy Dependence on Pressure” in Chapter 9.

The free energy of the gases in the mixture being considered is

$$\begin{aligned} G_{mix} = {} & (1-F)\left[\overline{g}^*_{\mathrm{H_2O}} + RT \ln\left(\frac{1-F}{1+F/2}\right)\right] \\ & + F\left[\overline{g}^*_{\mathrm{H_2}} + RT \ln\left(\frac{F}{1+F/2}\right)\right] + F/2\left[\overline{g}^*_{\mathrm{O_2}} + RT \ln\left(\frac{F/2}{1+F/2}\right)\right] \\ = {} & (1-F)\overline{g}^*_{\mathrm{H_2O}} + F\overline{g}^*_{\mathrm{H_2}} + \frac{1}{2}F\overline{g}^*_{\mathrm{O_2}} + RT\left[(1-F)\ln\left(\frac{1-F}{1+F/2}\right)\right. \\ & \left. + F \ln\left(\frac{F}{1+F/2}\right) + F/2 \ln\left(\frac{F/2}{1+F/2}\right)\right]. \end{aligned} \tag{10.45}$$

Equation 10.45 is plotted in Figure 10.8, which shows that the free energy of the water–hydrogen–oxygen mixture at 3000 K and 1 atmosphere has a minimum when 14.8% of the water is dissociated. At this point, the reverse reaction $H_2 + \frac{1}{2}O_2 \rightarrow H_2O$ occurs at exactly the same rate as the forward reaction, $H_2O \rightarrow H_2 + \frac{1}{2}O_2$: equilibrium is established.

To find the point of equilibrium, we must seek the value of F that minimizes G_{mix}:

$$\begin{aligned} \frac{dG_{mix}}{dF} = {} & -\overline{g}^*_{\mathrm{H_2O}} + \overline{g}^*_{\mathrm{H_2}} + \frac{1}{2}\overline{g}^*_{\mathrm{O_2}} \\ & + RT\left[(1-F)\frac{-3}{2(1-F)(1+F/2)} + \frac{1}{1+F/2} + \frac{1}{2}\frac{1}{1+F/2}\right. \\ & \left. - \ln\left(\frac{1-F}{1+F/2}\right) + \ln\left(\frac{F}{1+F/2}\right) + \frac{1}{2}\ln\left(\frac{F/2}{1+F/2}\right)\right] \\ = {} & -\overline{g}^*_{\mathrm{H_2O}} + \overline{g}^*_{\mathrm{H_2}} + \frac{1}{2}\overline{g}^*_{\mathrm{O_2}} \\ & + RT\left[- \ln\left(\frac{1-F}{1+F/2}\right) + \ln\left(\frac{F}{1+F/2}\right) + \frac{1}{2}\ln\left(\frac{F/2}{1+F/2}\right)\right] = 0 \end{aligned} \tag{10.46}$$

because

$$(1-F)\frac{-3}{2(1-F)(1+F/2)} + \frac{1}{1+F/2} + \frac{1}{2}\frac{1}{1+F/2} = 0. \tag{10.47}$$

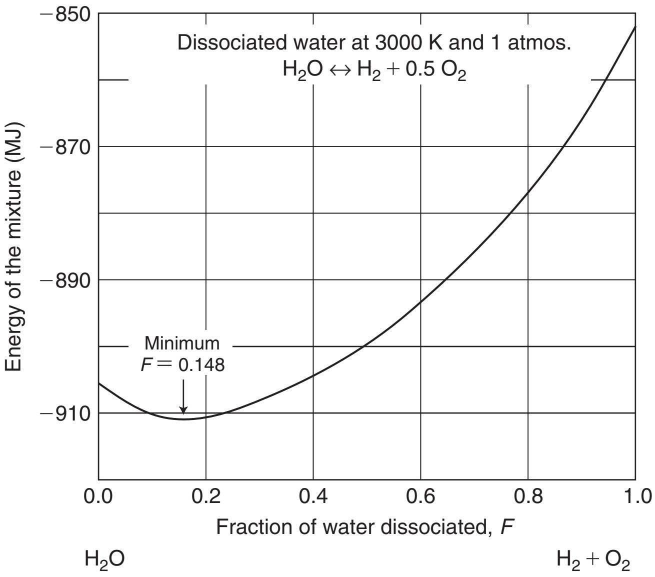


Figure 10.8 Free energy of a water–hydrogen–oxygen mixture as a function of the fraction, F, of dissociation.

The arguments of the natural logarithms are the different gas pressures. Hence,

$$\frac{dG_{mix}}{dF} = -\overline{g}^*_{\mathrm{H_2O}} + \overline{g}^*_{\mathrm{H_2}} + \frac{1}{2}\overline{g}^*_{\mathrm{O_2}} + RT\Big[-\ln p_{\mathrm{H_2O}} + \ln p_{\mathrm{H_2}} + \frac{1}{2}\ln p_{\mathrm{O_2}}\Big] = 0, \tag{10.48}$$

$$\frac{dG_{mix}}{dF} = -\overline{g}^*_{\mathrm{H_2O}} + \overline{g}^*_{\mathrm{H_2}} + \frac{1}{2}\overline{g}^*_{\mathrm{O_2}} + RT\ln\left(\frac{p_{\mathrm{H_2}}p_{\mathrm{O_2}}^{\frac{1}{2}}}{p_{\mathrm{H_2O}}}\right) = 0. \tag{10.49}$$

Define an **equilibrium constant**, K_p,

$$K_p \equiv \frac{p_{\mathrm{H_2}}p_{\mathrm{O_2}}^{\frac{1}{2}}}{p_{\mathrm{H_2O}}}, \tag{10.50}$$

then

$$K_p = \exp\left(\frac{\overline{g}^*_{\mathrm{H_2O}} - \overline{g}^*_{\mathrm{H_2}} - \frac{1}{2}\overline{g}^*_{\mathrm{O_2}}}{RT}\right). \tag{10.51}$$

K_p depends on the temperature and on the *exact way the chemical equation was written.* The K_p for the $\mathrm{H_2O}\rightleftharpoons\mathrm{H_2} + \frac{1}{2}\mathrm{O_2}$ reaction is not the same as that for the $2\mathrm{H_2O}\rightleftharpoons 2\mathrm{H_2}+\mathrm{O_2}$ reaction. *However, K_p does not depend on pressure.* Indeed, referring to Equation 10.51, we see that the different $\overline{g}_i^*$ are values at 1 atmosphere and do not change if the actual pressure is changed. Furthermore, if an inert gas, say, argon, is mixed with the water vapor, it does not alter the value of K_p because its $\overline{g}^*$ is zero.†

The relationship between the equilibrium constant, K_p, and the fraction, F, of water that has been dissociated can be determined by expressing the partial pressures as a function of F as was done in Equations 10.41, 10.42, and 10.43, which are for the particular case in which the total pressure of the mixture is 1 atmosphere. For the more general case, when the mixture is at some arbitrary pressure, p, the partial pressures become

$$p_{H_2O} = \frac{1-F}{1+F/2}p, \tag{10.52}$$

$$p_{H_2} = \frac{F}{1+F/2}p, \tag{10.53}$$

$$p_{O_2} = \frac{F/2}{1+F/2}p. \tag{10.54}$$

Inserting these partial pressures into Equation 10.50, one obtains,

$$K_p = \frac{F^{\frac{3}{2}}}{\sqrt{2-3F+F^3}}p^{1/2}. \tag{10.55}$$

Since, as was pointed out, K_p is pressure independent, F must depend on pressure. For instance, we saw that for $T = 3000$ K and $p = 1$ atmosphere, 14.8% of the water dissociates ($F = 0.148$). However, if the pressure is increased to 100 atmospheres, the amount of dissociation falls to 3.4%. This is, of course, predicted by Le Chatelier's rules. In the reaction

$$H_2O \rightleftharpoons H_2 + \frac{1}{2}O_2,$$

an increase in pressure drives the equilibrium to the left because there are more kilomoles in the right-hand side than in the left-hand side.

It can be seen from the information developed above that the direct thermal dissociation of water would require impractically high temperatures. Figure 10.9 shows that at the temperature at which palladium melts (1825 K), only an insignificant fraction of the water is dissociated when the pressure is 1 atmosphere. This means that the partial pressure of the produced hydrogen is much too small to be useful.

Increasing the pressure of the water does not help the situation because the degree of dissociation falls dramatically (Figure 10.10).

†Nitrogen would behave differently because, at the high temperatures employed, it would react with oxygen, forming nitrogen oxides.

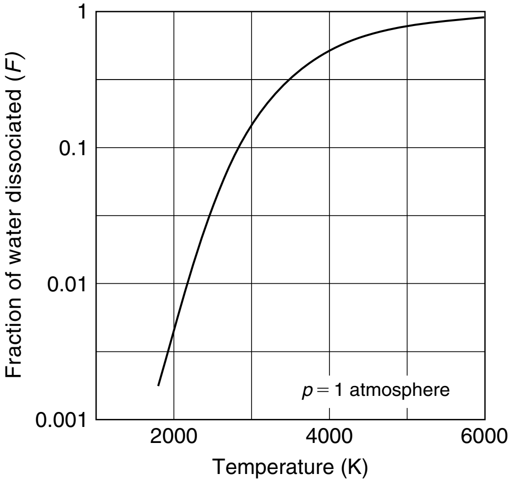


Figure 10.9 Water dissociation as a function of temperature. Pressure: 1 atmosphere.

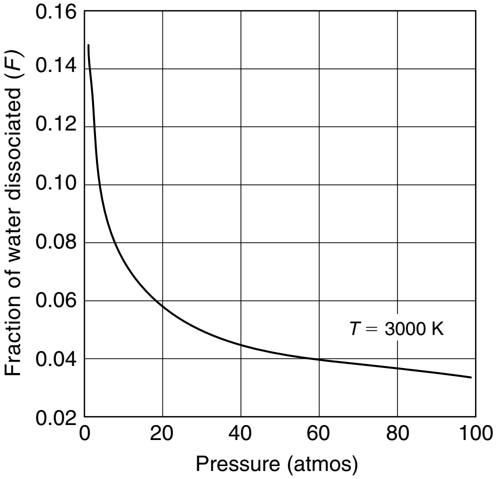


Figure 10.10 Water dissociation as a function of pressure. Temperature: 3000 K.

The definition of equilibrium constant can be extended to more complicated reactions. Thus, for instance, for the reaction

$$\mu_A A\ \mu_B B \rightleftharpoons \mu_C C\ \mu_D D, \tag{10.56}$$

the equilibrium constant is

$$K_p = \frac{p_A^{\mu_A} p_B^{\mu_B}}{p_C^{\mu_C} p_D^{\mu_D}}. \tag{10.57}$$

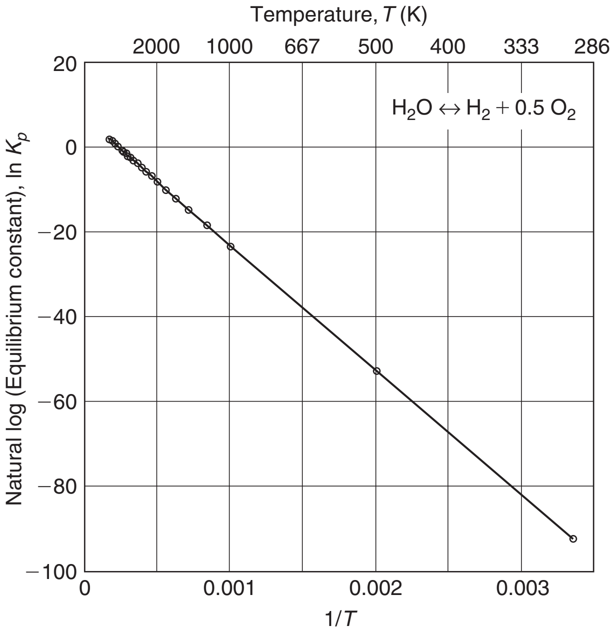


Figure 10.11 Natural logarithm of the equilibrium constant for the dissociation of water versus inverse temperature.

The equilibrium constant can then be related to the equilibrium composition by

$$K_p = \frac{N_A^{\mu_A} N_B^{\mu_B}}{N_C^{\mu_C} N_D^{\mu_D}} \left(\frac{p}{N_{TOTAL}}\right)^{\Delta\mu}, \tag{10.58}$$

where $\Delta\mu \equiv \mu_C + \mu_D - \mu_A - \mu_B$, and the N_i are for equilibrium composition, not for the theoretical reaction (Equation 10.56).

The value of the equilibrium constant, K_p, for different reactions can be found in thermodynamic tables. Figure 10.11 plots K_p versus $1/T$ for the $H_2O \leftrightarrow H_2 + \frac{1}{2}O_2$ reaction.

A linear regression of the data displayed in Figure 10.11 yields

$$K_p = 847.3 \exp -\frac{246 \times 10^6}{RT}. \tag{10.59}$$

−246 MJ/kmole is (approximately) the energy of formation of water. This is another example of the Boltzmann equation.

10.4.2 Chemical Dissociation of Water

To circumvent the difficulties encountered with the direct dissociation of water, several chemical reactions have been proposed. In all of them, the intermediate products are regenerated so that, at least theoretically, there is no consumption of materials other than water itself. The temperatures required for these reactions must be sufficiently low to permit practical implementation of the process. In particular, it is desirable that the temperatures not exceed 1100 K to make the process compatible with nuclear fission reactors.

One proposed reaction chain is indicated in Equations 10.60 to 10.64. The highest temperature step operates at 730 C. Overall efficiency is 50%. The major disadvantage is the use of corrosive hydrobromic acid.

$$Hg + 2HBr \rightarrow HgBr_2 + H_2 \tag{10.60}$$

$$HgBr_2 + Ca(OH)_2 \rightarrow CaBr_2 + HgO + H_2O \tag{10.61}$$

$$HgO \rightarrow Hg + \frac{1}{2}O_2 \tag{10.62}$$

$$CaBr_2 + 2H_2O \rightarrow Ca(OH)_2 + 2\ HBr \tag{10.63}$$

The overall reaction is:

$$H_2O \rightarrow H_2 + \frac{1}{2}O_2. \tag{10.64}$$

Another, more complicated, chain of reactions for the thermolytic production of hydrogen see Figure 10.12 is:

1. Reaction 1

$$Cr_2O_3(s) + 4Ba(OH)_2(\ell) \overset{600\,C}{\longrightarrow} 2Ba_2CrO_4(s) + 3H_2O(g) + H_2\,(g) \tag{10.65}$$

2. Reaction 2

$$2BaCrO_4(s) + Ba(OH)_2(\ell) \overset{850\,C}{\longrightarrow} Ba_3(CrO_4)_2(s) + H_2O(g) + \frac{1}{2}O_2\,(g) \tag{10.66}$$

   Notice that, since hydrogen evolves from Reaction 1 and oxygen from Reaction 2 which occur in different parts of the equipment, it is easy to separate these gases. Input reactants are Cr_2O_3 and $BaCrO_4$. Products (besides H_2 and O_2) are Ba_2CrO_4 and $Ba_3(CrO_4)_2$. To recover the reactants, the last two chromates are made to react with water at low temperature:

3. Reaction 3

$$2Ba_2CrO_4(s) + Ba_3(CrO_4)_2 + 5H_2O \rightarrow Cr_2O_3(s) + 2BaCrO_4(s) + 5Ba(OH)_2(d) \tag{10.67}$$

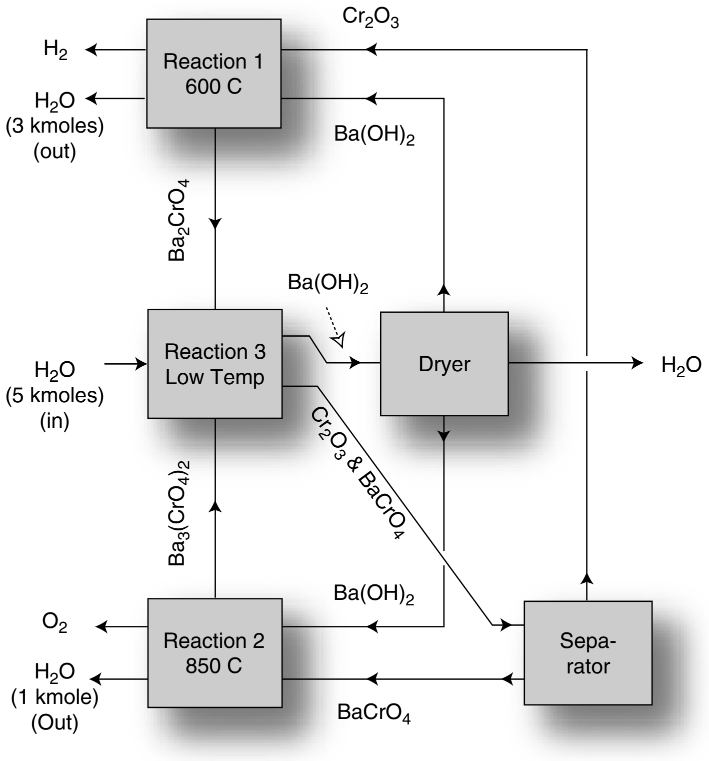


Figure 10.12 Block diagram of the barium chromate cycle.

As pointed out in Chapter 1, the heavy-metal nuclear reactors currently under consideration as a means of reviving the popularity of fission reactors may be used as a source of heat to drive thermochemical hydrogen production processes. The high-temperature, low-pressure mode of operation of these reactors would facilitate the design of the necessary chemical machinery.

10.5 Photolytic Hydrogen

10.5.1 Generalities

The technology for using solar light energy to produce hydrogen is well established. One certainly can produce this gas through entirely nonpolluting processes by using photovoltaic converters whose output drives electrolyzers. The main effort here is to develop processes that can accomplish this transformation more economically.

Water can be decomposed (and synthesized) according to

$$2H_2O \rightleftharpoons 2H_2 + O_2. \tag{10.68}$$

Though simple looking, the reaction is the result of a chain of events that leads through the formation of several intermediate substances and can follow different paths depending on the circumstances and the catalysts.

The structural formula of water, H–O–H, suggests that the first step in the dissociation must be the breaking of the H–O bond so that H and OH (or the corresponding ions) are formed. Next, the OH is dissociated into O and another H, then these atoms coalesce into diatomic molecules. For the case in which no ionized species take part in the reactions, the different energies involved are given in Table 10.3.

Figure 10.13 indicates schematically the progress of the reaction. The exact sequence of events may vary.

Although there is only a 2.51 eV energy difference between a molecule of water and its elements, to dissociate it directly, it is necessary to overcome a 5.15 eV barrier. Thus, photons or phonons with less than 5.15 eV cannot initiate the direct water dissociation reaction.

For direct lysis the input energy must be 5.15 + 4.40 = 9.55 eV per molecule of water. The result is a hydrogen/oxygen mixture with 2.51 eV more energy than the initial water. The efficiency of this "fuel" production is 2.51/9.55 = 0.263. The rest of the input energy will appear as heat as the result of the **back reaction** of the **activated intermediate** species.

Table 10.3 Dissociation Energies

	eV/ molecule	**MJ/ kmole**
$H_2O \rightarrow H + OH$	5.15	496.2
$OH \rightarrow H + O$	4.40	423.9
$H + H \rightarrow H_2$	−4.48	−431.7
$O + O \rightarrow O_2$	−5.12	−493.3
$H_2O \rightarrow H_2 + \frac{1}{2}O_2$	2.51	241.8

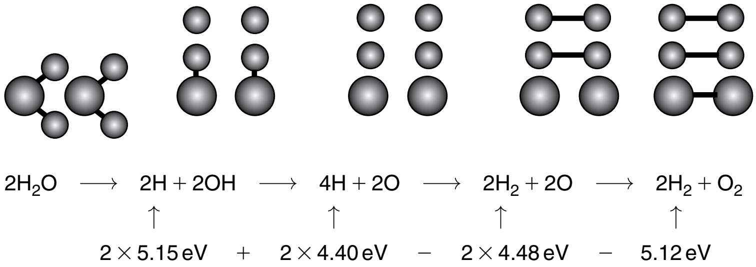

$2H_2O \longrightarrow 2H + 2OH \longrightarrow 4H + 2O \longrightarrow 2H_2 + 2O \longrightarrow 2H_2 + O_2$
↑ ↑ ↑ ↑
2 × 5.15 eV + 2 × 4.40 eV − 2 × 4.48 eV − 5.12 eV

Figure 10.13 One possible path in the decomposition of water.

10.5.2 Solar Photolysis

Figure 10.14 shows the cumulative energy distribution in the solar spectrum. It can be seen that about 22% of the energy is associated with photons having more than 2.51 eV, the energy necessary to dissociate water. However, as was discussed in the preceding subsection, without catalysts, photons with more than 5.15 eV are necessary to initiate the reaction. Thus, practical photolytic hydrogen production depends on the development of appropriate catalysts.

A catalyst, X, may, for instance, be oxidized in the presence of water under the influence of light:

$$X + H_2O \rightarrow X^+ + H + OH^- \tag{10.69}$$

$$X^+ + \frac{1}{2}H_2O \rightarrow X + H^+ + \frac{1}{4}O_2 \tag{10.70}$$

$$H^+ + OH^- \rightarrow H_2O \tag{10.71}$$

The threshold energy for this reaction is 3.8 eV, somewhat less than the 5.15 eV necessary in the absence of catalysts. The reaction would use only 3% of the solar energy and, of that, a great deal would be lost in the $H^+ + OH^- \rightarrow H_2O$ back reaction.

Actually, for a ground-based system, the available energy is considerably less than 3% because the higher frequencies of the solar spectrum are selectively absorbed by the atmosphere.

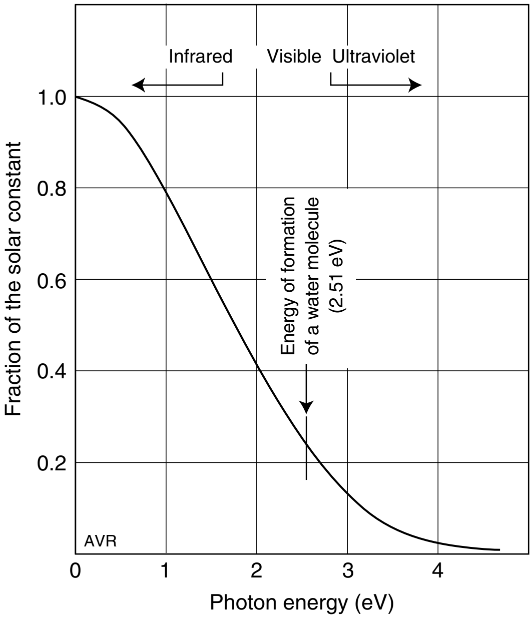


Figure 10.14 Cumulative energy distribution in the solar spectrum.

A somewhat more favorable reaction is one in which the catalyst, Y, is reduced:

$$Y + H_2O \rightarrow Y^- + H^+ + OH \tag{10.72}$$

$$Y^- + H_2O \rightarrow Y + OH^- + \frac{1}{2}H_2 \tag{10.73}$$

$$OH^- + H^+ \rightarrow H_2O \tag{10.74}$$

$$\frac{1}{2}OH + \frac{1}{2}OH \rightarrow \frac{1}{2}H_2O + \frac{1}{4}O_2 \tag{10.75}$$

A third reaction avoids intermediate species that will back-react.

$$Z + H_2O \rightarrow ZO + H_2 \tag{10.76}$$

$$ZO \rightarrow Z + \frac{1}{2}O \tag{10.77}$$

The threshold for this reaction is 2.9 eV, and it could use 13% of the solar energy. However, its realizability is uncertain.

The above energy considerations paint a dubious picture of the possibilities of using direct sunlight for the production of hydrogen. But, as usual, there are many other aspects to be considered.

Direct photolysis of water runs into a fundamental difficulty: water is transparent to most of the sunlight. Only frequencies above some 1600 THz are absorbed, and this represents less than 0.01% of the solar radiation in space, and much less of the solar radiation that reaches the surface of the earth (see Chapter 12). Nevertheless, these analyses have not discouraged investigators from trying to find economical ways of harvesting solar energy, transforming it into chemical energy. If a plant leaf can do it, so should we.

Only by 1972 was it possible to demonstrate the dissociation of water by direct absorption of light without the help of an externally applied bias. Fujishima and Honda used an n-TiO_2 single crystal and a platinum counter-electrode in the arrangement depicted in Figure 10.15. Titanium dioxide[†] was used because of its stability in liquid environments. However, it is a white pigment used in common paints. This means that it reflects all of the incident visible light, absorbing only ultraviolet of frequency larger than 1580 THz (190 nm). This is due to its large band gap (3 eV). Thus, titania uses only a small part of the energy available in the solar spectrum, one of the causes of the low efficiency of the Fujishima and Honda cell (<1%).

Titania, or any other material, dipped in an electrolyte will develop a contact potential. A transition layer in which the potential changes from that of the semiconductor to that of the electrolyte is formed on the surface. This layer is thin (a few tenths of nanometers), so the electric fields are intense. In the case of n-type titania, the direction of the field is such

[†] Titanium dioxide—titania—occurs in the form of rutile, anatase, and brookite.

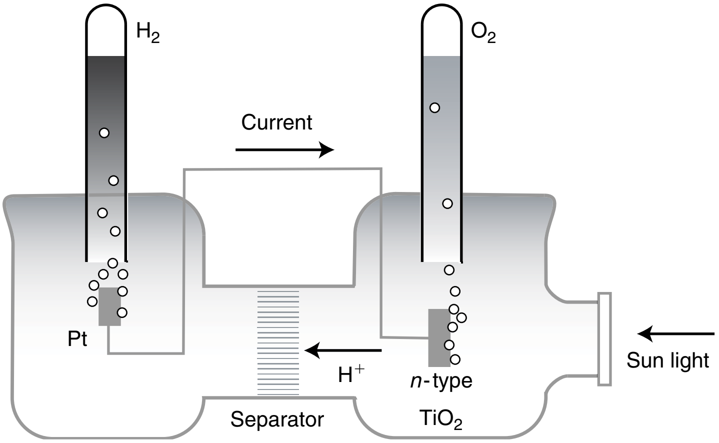


Figure 10.15 Sunlight was used to produce hydrogen and oxygen in a 1972 experiment by Fujishima and Honda.

that the electron created by the incident photon is driven deeper into the semiconductor while the corresponding hole is injected into the electrolyte. A charge separation occurs even though there is no p-n junction. The hole injected into the (aqueous) electrolyte oxidizes the water,

$$2p^+ + H_2O \rightarrow \frac{1}{2}O_2 + 2H^+, \tag{10.78}$$

causing oxygen to evolve at the photoelectrode. Meanwhile, the proton from the above reaction migrates through the electrolyte to the platinum electrode where it combines with the electron (that came from the TiO_2 via the external connection) and is reduced to H and eventually associates to form H_2, evolving at the metal electrode,

$$2e^- + 2H^+ \rightarrow H_2. \tag{10.79}$$

These results are encouraging but, owing to the low efficiency, are hardly practical.

One obvious cause for the low efficiency is, as pointed out earlier, the excessive band gap of titania. By replacing this material by valence semiconductors such as GaAs, much higher efficiencies have been realized. Licht, Wang, and Mukerji (2000) have built cells with AlGaAs/Si in a bipolar configuration (band gaps of 1.6 eV and 1.1 eV, respectively) that split water with 18.3% efficiency. Unfortunately, cells using this class of semiconductors have a limited life owing to severe corrosion problems.

Back to titania. Doping will, of course, alter the band gap of a semiconductor. Kahn et al. (2002) report that incorporating carbon will reduce the band gap of titania from the original 3 eV to 2.32 eV and correspondingly increase the water splitting efficiency to some 8.5%. If you want to know why reducing the band gap increases the efficiency, read the beginning of Chapter 14.

Another way to increase the efficiency of photochemical cells is to use **dye sensitization**. For more details, go to Chapter 14.

10.6 Photobiologic Hydrogen Production

The majority of living organisms must respire, that is, must consume oxygen and release carbon dioxide to fuel their anabolism. This is true of plants. However, plants, under the influence of light, also perform photosynthesis, the opposite of respiration: they fix atmospheric carbon dioxide and release oxygen. Thus, when exposed to light, plants tend to be net oxygen producers and, in the dark, net oxygen consumers.[†]

In principle, some plants could generate the oxygen needed for their respiration by extracting it from water and releasing hydrogen. This occurs, for instance, with certain algae. Since, in darkness, the plant is almost dormant, the amount of hydrogen released is small. An enzyme—hydrogenase—promotes such release. Hydrogenase is inhibited by the presence of oxygen and, thus, does not work when photosynthesis is active.

Sulfur deprivation reversibly inactivates photosynthesis, allowing the production of hydrogen in larger amounts when the plant is exposed to light.

Melis et al. (2000) have demonstrated a photobiological hydrogen production system using the alga *Chlamydomonas reinhardtii.* The system proceeds in two stages:

1. Algae are grown normally and build up their store of carbon compounds.
2. Sulfur is withdrawn from the system, and the algae, still exposed to light, release hydrogen, consuming some of the accumulated carbon.

After Stage 2 has depleted much of the carbon, Stage 1 is reinitiated and the plant is again "refattened."

At present, it appears that substantially more research is required not only to perfect the parameters of the system but also to develop more efficient strains or algae.

The overall sunlight-to-hydrogen gas efficiency will be modest because photosynthesis is less than 8% efficient (see Chapter 13) and light is needed not only for photosynthesis but also for the hydrogen release. Nevertheless, what really counts from the practical point of view is the cost of the gas produced and not the efficiency of the system. It remains to be seen if the proposed Melis system will be economically attractive.

[†] Some plants will shut off photosynthesis when exposed to too much light, and will, then, only respire. See "Photorespiration" in the end of Section 13.4 of Chapter 13.

References

Atwood, Jerry L., Leonard J. Barbour, and Agoston Jerga, A new type of material for the recovery of hydrogen from gas mixtures, *Angewandte Chemie* **43**, pp. 2948–2950, **2004.**

Brinkman, Greg, *Economics and environmental effects of hydrogen production methods* http://www.puaf.umd.edu/faculty/papers/fetter/students/Brinkman.pdf.

Garzon, Fernando H., R. Mukundan, and Eric L. Brosha, Enabling science for advanced ceramic membrane electrolyzers, Proc. 2002 U. S. DOE Hydrogen Program Review, NREL/CP-610-32405, **2002**.

Khan, Shahed U. M., Mofareh Al-Shahry, and William B. Ingler, Jr., Efficient photochemical water splitting by a chemically modified n-TiO_2, *Science* **297**, pp. 2243–2245, **2002.**

Khan, Shahed U. M., Mofareh Al-Shahry, and William B. Ingler, Jr., Response to Comments on "Efficient photochemical water splitting by a chemically modified n-TiO_2," *Science* **301**, p. 1673, September 19, **2003.**

Licht, S., B. Wang, and S. Mukerji, Efficient solar water splitting, exemplified by RuO_2-catalyzed AlGaAs/Si photoeletrolysis, *J. Phys. Chem. B*, **104**, pp. 8920–8924, **2000**.

Loges, Björn, Albert Boddien, Henrik Junge, and Matthias Beller, Controlled generation of hydrogen from formic acid amine adducts at room temperature and application in H2/O2 fuel cells, *Angew. Chem. Int. Ed.* **47**, pp. 3062–3965, **2008**.

Melis, Anastasios, Liping Zhang, Marc Forestier, Maria L. Ghirardi, and Michael Seibert, Sustained photobiological hydrogen gas production upon reversible inactivation of oxygen evolution in green the alga *Chlamydomonas reinhardtii.*, *Plant Physiol.* **122**, pp. 127–136, January **2000**.

Pham, Ai-Quoc, High efficiency steam electrolyzer, Proc. of the 2000 US DOE Hydrogen Program Review, NREL/CP-610-32405.

PROBLEMS

10.1 Let C_D be the price of an electrolyzer plant expressed in dollars per kW of hydrogen produced with a given current density, J_0. The number of output kW is the heat power generated if the hydrogen were burned. There is no compelling reason to operate the electrolyzer at the nominal (J_0) current density. It can be operated at any other current density, J, within the permissible range. The efficiency of the electrolyzer depends linearly on the current density:

$$\eta = \mathrm{a} + \mathrm{b}J.$$

The cost of the hydrogen produced, C_H, is

$$C_H = C_{INV} + C_{OP} + C_E. \qquad \$/(\text{kg of H}_2)$$

where

C_{INV} = part of the cost attributable to the investment in the plant,
C_{OP} = maintenance and operation cost (assume this to be zero), and
C_E = cost of the electricity, a function of c_e, the price of electricity.

Let Θ be the **plant factor**—that is, the fraction of the total time during which the plant is actually operating, and R be the yearly cost of the capital, expressed as a fraction of the borrowed capital.

Develop an expression for the value of J that minimizes the cost of the hydrogen. It must be a function of C_D, J_0, R, Θ, c_e, a, and b.

What is this optimum value of J for the case in which

$$C_D = 100\ \$/\text{kW},$$
$$J_0 = 10{,}000\ \text{A/m}^2,$$
$$R = 0.2\ \text{per year},$$
$$\Theta = 1,$$
$$c_e = 10\ \$/\text{MWh},$$
$$a = 0.74,\ \text{and}$$
$$b = -6 \times 10^{-6}\ \text{m}^2/\text{A}?$$

10.2 An electrolytic cell having 100% current efficiency, operates at a voltage of 1.9 V when the current is 20 kA. Its operating temperature is 86 C. The cell happens to be completely heat insulated: heat can only be removed by the flow of reactants and of products and by water flowing through a cooling system built into the cell. Both feed and cooling water enter the system at 25 C. The cooling water leaves at 80 C. The gases leave at 85 C. Assume that the enthalpy change owing to the reaction is independent of temperature (285.9 MJ per kilomole of water).

1. What is the hydrogen production rate in kg/hour?
2. What is the feed water consumption rate?
3. What is the flow rate of the cooling water?
4. A consultant was called in, and he improved catalysis in the anode reducing the operating voltage (at 20 kA) to 1.475 V. The cooling water can now be shut off. What is the new operating temperature?

10.3 Consider an ideal water electrolyzer installed at sea level and an ideal hydrogen/oxygen fuel cell installed 1000 m higher up.

The oxygen and the hydrogen produced electrolytically rise through a pipe and feed the fuel cell. Water is produced by the fuel cell and flows down another pipe turning a turbine that drives a electric generator. The turbine/generator combination has 100% efficiency.

Here is a simple argument:

Since both electrolyzer and fuel cell are ideal (reversible), the electric energy generated by the fuel cell should be exactly the energy needed by the electrolyzer. But there is some energy, W, generated by the flowing product water and this constitutes a net “profit.”

Clearly,

1. we have finally invented a perpetual motion machine, or
2. we are extracting energy from some nonobvious source, or
3. there is a flaw in the argument.

If you believe (a), you are in trouble. It must be either (b) or (c). If it is (b), describe the hidden source and demonstrate that it delivers precisely W units of energy. If (c), explain what is wrong with the argument and demonstrate that W plus the energy delivered by the fuel cell is exactly equal to the energy required by the electrolyzer. Assume that the temperature of the electrolyzer is equal to that of the fuel cell.

10.4 Hydrogen, to be produced electrolytically, is needed at 40 MPa and 25 C. The engineers who are designing the plant must consider two possibilities:

1. Produce hydrogen at 0.1 MPa and then compress it mechanically to 40 MPa.
2. Pressurize the electrolyzer to 37 MPa and remove the hydrogen at 40 MPa. This assumes that the internal pressure difference between the oxygen and the hydrogen side is 3 MPa and that the

oxygen is produced at the environmental pressure of the electrolyzer.

When operating at 0.1 MPa, the electrolyzer efficiency is 85%. However, when pressurized, its efficiency falls to 80%.

The mechanical compressor has 65% efficiency.

How much energy is needed to produce 1 ton of hydrogen in each case? Calculate the energy required to supply the feed water (only the energy necessary to force the water into the pressurized environment).

The high-pressure oxygen can be expanded to 0.1 MPa through a turbine. How much energy can you recover if you have a 100% efficient turbo-generator? What is the temperature of the exhaust oxygen assuming an adiabatic (completely insulated) turbo-generator? Assume the oxygen does not condense.

10.5 An electrolyzer consists of 100 cells, each with 1 m^2 of effective area. The efficiency of each cell is given by

$$\eta = 1.205 + \mathrm{b}J,$$

where J is the current density in $\mathrm{A/m}^2$.

When $J = 1000\,\mathrm{A/m}^2$, the voltage required by each cell is 1.310 V.

The electrolyzer is installed in a cubical room (3-m edge) whose walls, floor, and ceiling conduct heat at a rate of 50 $\mathrm{W/m}^2$ per kelvin of temperature difference. No other source or sink of heat is in the room. Outside temperature is 30 C.

When a current of 1000 A is forced through the cells,

1. How many kg of H_2 are produced per day?
2. How many kg of O_2 are produced per day?
3. How many kg of H_2O are consumed per day?
4. What is the equilibrium temperature of the room?
5. What current causes the room to reach the lowest possible temperature?

10.6 A water electrolyzer operating at RTP requires 1.83 V to produce 1 metric ton of hydrogen per day when operating continuously.

1. What current does it draw?
2. How many m^3 of water does it use per day?
3. How many MJ of heat does it either reject or absorb per day (state which)?

10.7 We want to estimate the performance of a system for the production of ammonia in geographically dispersed plants. Input energy is to be electrical.

The agricultural plot to be served by each ammonia plant is to be 20 by 20 km in effective cultivated area.

Nitrogen fertilization is to be intensive: 40 kg of ammonia per hectare per year (1 ha = 10,000 m^2).

The electrolyzers will operate at a current density that results in 85% overall practical efficiency (at RTP). Current efficiency is 100%.

1. Assuming that the plants operate 24 hours per day, how many tons (1000 kg) must each plant produce per day?
2. How much power does each electrolyzer demand?
3. What is the voltage of each electrolyzer cell?
4. What is the current through each electrolyzer cell?
5. Assume that instead of delivering the gases at 1 atmosphere, the electrolyzer is required to deliver them at 400 atmospheres. Clearly, additional electric power will be needed. Assume that this additional power is exactly the power for isothermal compression of the gases. How much is this additional power?
6. In the traditional ammonia process, 80% of the cost of production is the cost of energy. The price of ammonia in the international market is \$200/ton. Thus, \$40/ton cover all the nonenergy-related costs. How much can you afford to pay for 1 kWh and still make a 10% profit? Remember that the \$40/ton remains unchanged.

10.8 A hydrogen producer has a battery of 100 cells connected in series. A *v*-*i* test shows that a current of 35.6 A must be driven into any of the cells to achieve a 1.482 V potential drop across it. As the current is decreased, the voltage also decreases and, as the current approaches 0, the voltage approaches 1.376 V.

A production rate of 1 liter (0.001 m^3) of H_2 per second at RTP is required. What voltage must be applied to the battery?

Model each cell as an ideal electrolyzer in series with an opposing voltage generator and a resistance of constant value (independent of the current).

10.9 Consider an ideal electrolyzer in series with a 0.01 Ω resistance. The gas outlets are connected to small closed vessels of equal volume so that when the device operates, gas pressure builds up. Initially, both the oxygen pressure and the hydrogen pressure are 1 atmosphere. The electrolyzer operates at 298 K.

A 1.333 V constant-voltage power supply is connected to the electrolyzer, in series with the 0.01 Ω resistor. The current is monitored.

1. What is the initial current?
2. With time, the pressures in the gas containers rise. Assuming these containers do not rupture, what is the maximum hydrogen pressure that the system will reach?

10.10 A spherical balloon is filled with hydrogen until the internal pressure equals that of the surrounding air. The bag is made of a material that can expand or contract but contributes negligibly to the pressure of the gas. The temperature of the hydrogen is equal to that of the air. The balloon, when empty, weighs 32 kg. When the air pressure is exactly one atmosphere and its temperature is 0 C, the balloon has a diameter of 10 m.

1. What is its net lifting force? How does the lift force depend on the external pressure and temperature?
2. To fill the balloon, a water electrolyzing plant is employed. The plant consists of 100 cells, each operated at 2000 A and 1.92 V. How long must this plant operate to produce the required amount of hydrogen?
3. How much feed water is used per second (in liters per second)?
4. What is the rate of heat production by the plant?

10.11 A battery of 12 water electrolyzers connected in series (so that the same current flows through each unit) operates in a room maintained at 298 K. The product gases are withdrawn under a pressure of 1 atmos. A voltage of 17.784 V is applied and the resultant current is 1200 A.

1. Calculate the production rate of hydrogen in kg/day.
2. Calculate the water consumption in liters/day.
3. Calculate the operating temperature of the electrolyzers.

10.12 How much power does an ideal electrolyzer operating at 298.2 K require to produce 1 kg of hydrogen per hour at a pressure of 400 atmospheres? The oxygen is produced at 370 atmospheres.

10.13 You have been hired to run a hydrogen production plant. The fixed cost (amortization of the equipment and the buildings, salaries, taxes, etc.) amounts to $c_F = \$2000$/day regardless of how much hydrogen is produced.

The electrolyzer consists of N cells connected in series, each having the characteristic,

$$V = V_0 + R_{int} I. \tag{10.80}$$

The utility will provide electric energy at a price, c (in \$/kWh), that can vary from day to day, depending on availability.

Each day you must adjust the hydrogen production rate, $H_{prod.rate}$ (kg/day) so as to minimize the cost of the gas. To this end, you must develop a formula that tells you the optimal production rate as a function of c.

Calculate the hydrogen production rate that leads to the cheapest gas for the case in which there are 250 series connected cells, each one having the characteristic

$$V = 1.420 + 20 \times 10^{-6} I. \tag{10.81}$$

Do this for an electric energy cost of 2 cents per kWh.

10.14 Consider the electrolyzer of Problem 10.13, operating at 20,000 A.

Calculate:

1. The total voltage that must be applied.
2. The hydrogen production rate in kg (H_2)/day.
3. The rate of water consumption in m^3/day.
4. The heat power rejected.

10.15 A hydrogen production system using direct dissociation of water is to operate at 1500 K. To have adequate hydrogen flux through a palladium filter used to separate the gas from oxygen and water vapor, it is necessary to have a hydrogen pressure differential of 5 atmospheres across the membrane. Assume that the pure hydrogen side is at 1 atmosphere. What pressure must be used on the water vapor side? Repeat for $T = 3000$ K.

10.16 An electrolyzer is made of a MEA that has a specific resistance, $\Re = 65\ \mu\Omega\ \text{m}^2$. Each cell is to produce 100 grams of hydrogen per hour. The current efficiency is 100%.

1. What is the current through each cell?
2. The V-I characteristic is

$$V = V_{rev} + V_{offset} + RI, \tag{10.82}$$

where V_{offset} is a constant offset voltage of 0.1 V.

The price of electricity is \$0.05/kWh. The price of the electrolyzer is proportional to the active area of the cell and is \$10,000 per square meter of active area. The cost of money is 12% per year. The only costs to be considered are those of investment and of electric energy.

Temperature of operation is 298 K. The plant operates continuously throughout the year. In assembling the electrolyzer, you can choose the active area of the cell. What current density results in the most economical hydrogen production?

10.17 It is difficult to design containers that can operate at the extreme high temperature required to thermally dissociate water. Let us assume, optimistically, that temperatures of 2800 K can be handled. A 1-m^3 canister contains 100 g of liquid water when cold. It is then heated to 2800 K, and part of the water will dissociate into hydrogen.

1. How many grams of hydrogen are formed?
2. The amount of free hydrogen is going to be small. What would happen if instead of 100 g of water, you had used 10 kg?

10.18 Ammonia is perhaps the most important of fertilizers. It provides nitrogen essential to the growth of plants. In 2000, worldwide production of ammonia exceeded 120 million tons per year.

All ammonia is produced by the Harber–Bosch process:

$$3H_2 + N_2 \leftrightarrow 2NH_3. \tag{10.83}$$

Enthalpy of formation: $\Delta \overline{h}_{f^\circ_{NH_3}} = -46.19\,\mathrm{MJ/kmole}$.

1. Is the equilibrium toward the ammonia side favored by raising or by lowering the pressure in the reactor?
2. Is the equilibrium toward the ammonia side favored by raising or by lowering the temperature of the gases? Clearly, it is not possible to take the favorable conditions of pressure and temperature to an extreme. Give one good reason for limiting pressure and temperature extremes.

10.19 Nitric oxide, NO, is a gas of great importance in the biology of mammals where it plays a central role in the operation of cells. It is also a serious pollutant because it readily combines with oxygen to form the toxic gas, NO_2. The nitrogen in the air can react with the oxygen according to

$$\frac{1}{2}N_2 + \frac{1}{2}(O)_2 \leftrightarrow NO. \tag{10.84}$$

This causes the formation of the undesired NO which, reacting with atmospheric oxygen and water vapor, is converted to nitric acid and causes acid rain to fall. NO is also a source of photochemical smog and a destroyer of the ozone layer.

Here are some pertinent thermodynamic data:

Enthalpies of formation:

	298.15 K	6000 K	
NO	90.37	298.87	MJ/kmole

Entropies:

	298.15 K	6000 K	
N_2	95.7	146.3	kJ K^{-1}kmole^{-1}
O_2	205.0	313.3	kJ K^{-1}kmole^{-1}
NO	210.6	369.4	kJ K^{-1}kmole^{-1}

Equilibrium constant, $K_p = 4.522 \exp - \dfrac{90.58 \times 10^6}{RT}$.

1. Does the reaction that generates NO from N_2 and O_2 proceed spontaneously at room temperature? If you say it does, then you have to explain how come there is so little NO in the air. If you say it does not, then how come there is a problem with NO pollution?

2. What happens when the gases are at 6000 K? If the reaction converts nitrogen and oxygen into nitric oxide, what fraction of the gas mixture, at equilibrium, consists of NO?
3. You will have found from Item 2 that at 6000 K (easily reached during a lightning stroke) substantial amounts of NO are produced. But the air quickly cools down, and as we saw from Item 1, at low temperature the equilibrium is toward complete dissociation of NO. So, how come NO lingers around constituting a serious pollutant?

10.20 Ammonia, NH_3, is produced by direct combination of nitrogen extracted from air with hydrogen obtained from fossil fuels, from water or from both. A large industrial ammonia plant can produce more than 1000 tons of the substance, per day.

In some countries, it may be better to have a distributed ammonia production system to avoid the difficulties of transporting it over long distances. Assume we want to produce only 3 tons of ammonia in each of a number of small plants using electrolytic hydrogen. The plants will work continuously, 24 hours per day.

a. How many kilograms of hydrogen will have to be produced each day?

b. Consider an electrolyzer cell represented by a simple voltage source, $V_0 = 1.35\,\text{V}$, in series with an internal resistance, $R_{int} = 0.0005\ \Omega$. Operating conditions are at RTP.

When operating at 80% efficiency, how many cells (connected in series) have to be used to produce hydrogen at the rate required by Item a?

c. How much heat must each cell reject when operating under the conditions of Item b?

d. If the current driving the electrolyzer is reduced sufficiently, the device will act as a heat pump. Calculate what current maximizes the amount of heat drawn in by each cell. Calculate the amount of heat drawn in under such conditions.

10.21 An ideal solid-polymer electrolyzer operating at 60 C produces H_2 and O_2 at 1 atmosphere pressure. It draws a current of 300 A and produces hydrogen at a rate of 1.557×10^{-6} kmoles/sec. What is the production rate (still using 300 A) if the operating temperature is raised to 90 C and all pressures are doubled?

10.22 What is the partial pressure of O_2 when 9 g of pure water are heated to 3000 K in a 100 liter container?

10.23 The current through a 2-cell electrolyzer is 200 A. The total voltage is 4 V. What is the hydrogen production rate? What is the heat power generated?

Chapter 11
Hydrogen Storage

As one approaches the city, one sees a landscape unmarred by unsightly towers and dangling transmission lines. Suburbs and center are devoid of wire-carrying poles, and the electric system is totally immune to storm-felled trees and branches. The only carbon dioxide produced has been exhaled by men, animals, and, at night, by plants. Cooking fires and heating systems release only water vapor, as do cars and buses. No ozone, no carbon monoxide, no nitrogen oxides can trace their birth to devices created by humans.

Each factory, each office, each house derives its energy from ultra-pure hydrogen either generated in situ or brought in by reliable underground pipelines. During the day, gleaming roofs collect sunlight and generate more electricity than can be used. The excess is sent to electrolyzers and stored locally as high-pressure hydrogen to be used at night or as fuel for cooking and heating. High-precision, inexpensive oscillators keep the fuel cell generating electricity at the exact prescribed frequency, its phase updated periodically by low-frequency radio signals.

When needed, additional hydrogen, from nonpolluting sources, is imported through existing pipes. The energy is totally ecologically friendly. Utopic? Yes, but entirely realizable with current technology, yet impossible with current economics!

To make this picture a reality, great advances must be made in bringing down the cost of numerous processes, not the least of which is the technique for storing hydrogen, the topic of this chapter.

The qualities of hydrogen have been extolled by enthusiasts. Indeed, once it has been produced (a process that can occur without pollution), you have the most ecologically friendly fuel possible. It suffers, however, from a major drawback—its extremely low density. Per cubic meter, hydrogen has but one-third of the combustion energy of methane. Nevertheless, a given pipeline has about the same power-carrying capacity with the two gases: the lower specific energy of hydrogen is almost exactly compensated by its lower viscosity. Thus, the bulk distribution of hydrogen should not present insurmountable problems. A more serious problem arises when hydrogen has to be supplied to a moving vehicle. One must either generate it in the vehicle itself or find a convenient way to store it aboard. In this chapter, we will examine the storage alternative.

Hydrogen can be stored as an element, or it can be extracted, as needed, from some hydrogen-rich substance using an onboard extraction process:

1. Processes that alter the state or the phase of hydrogen **(hydrogen-only systems)**:
   - 1.1 Compression of the gas (see Subsection 11.1), or a combination of compression and refrigeration.
   - 1.2 Liquefaction of the element (see Subsection 11.2). Owing to its low critical temperature, hydrogen must be cooled to some 20 K to remain liquid in nonpressurized vessels.
2. Processes that associate hydrogen to other substances:
   - 2.1 Adsorption of the gas on some appropriate substrate such as activated carbon.
   - 2.2 Chemical combination of hydrogen so as to create a hydrogen-rich compound. Such compounds can be:
     - 2.2.1 Compounds in which H_2 is tightly bound, requiring a relatively complex chemical process for the recovery of the gas. There are, for instance, substances like methanol (discussed in Chapter 10), ethanol, ammonia, and water itself that can be thought of as "carriers" of hydrogen.
     - 2.2.2 Compounds that can be reversibly transformed into another substance with a higher (or lower) hydrogen content.
     - 2.2.3 Metal hydrides that are metal–hydrogen compounds that can release and absorb hydrogen reversibly by a simple change of temperature.

A number of characteristics have to be considered when evaluating hydrogen storage systems. They include

1. **Gravimetric concentration (GC).** This is the ratio of the mass of the stored hydrogen to the overall mass of the (loaded) storage and retrieval system. The dimensions are kg per kg—that is, it is a dimensionless parameter.
2. **Volumetric concentration (VC).** This is the ratio of the mass of the stored hydrogen to the total volume of the storage and retrieval systems. The dimensions are kg per m^3 [ML^{-3}].
3. **Turnaround efficiency.** This can be the ratio of the retrieved hydrogen to the amount of input hydrogen, or the ratio of the energy retrieved to the input energy.
4. **Dormancy.** This is the ability of the system to retain its hydrogen over a long period of time.

In addition, there are considerations of cost (capital, maintenance, operating, and replacement costs), safety, ease of utilization, and so on.

11.1 Compressed Gas

For compressed gas containers, the main quantity of interest is the gravimetric concentration—that is, the ratio of the mass of the maximum amount of gas that can be stored to the mass, M_{cont}, of the container, where the maximum amount of gas corresponds to gas at just under the burst pressure, p_{burst}, of the vessel. This ratio is proportional to the **performance factor** (PF) of the container:

$$PF \equiv \frac{p_{burst}V}{M_{cont}}. \tag{11.1}$$

In SI, the performance factor has the dimensions of joules per kilogram.

For a given material and a given technology for building a compressed gas container, the mass of the container is proportional to the pressure so that the ratio of stored gas mass to the mass of the container is independent of the storage pressure. Hence, the only way to improve the PF is to use better materials and better technology in the construction.

Small quantities of hydrogen, as used in chemical laboratories, can be conveniently stored in simple steel pressure cylinders, usually at 150 atmospheres. For fuel cell vehicles (FCVs), compressed hydrogen may be a practical way to carry the necessary fuel. It is certainly the simplest storage system, and it requires no special equipment to retrieve the gas. What is needed are containers with a good PF.

Canisters of modern design store the hydrogen in a plastic container that is impermeable to the gas but that, by itself, is unable to resist any significant pressure. This resistance is provided by a lining of carbon fiber-epoxy composite, a layer protected by an outer shell resistant to considerable mechanical damage. Such canisters can operate with pressures of up to 700 atmospheres but must pass a burst test of 1650 atmospheres. In addition, they are tested by cycling the pressure 500,000 times. Note that even if one plans to use these canisters for 20 years, refueling them every week, only 1000 cycles are used.

The failure mode of these canisters is relatively benign—they do not explode in shrapnel but fail by delamination of the wrapping. They are designed to leak before bursting.

The gravimetric energy concentration target established by the Department of Energy (DOE) was 3.8% for the year 2005 (3.8 kg of hydrogen in a total system mass of 100 kg). This is still disappointingly low, but the DOE hopes for 5.0% in 2010 and 7.6% by 2020. The current refueling rate is 2 kg of H_2/minute (energetically, 280 MJ/min). To estimate if this is acceptable, consider a gasoline car that has, typically, a 50-liter tank

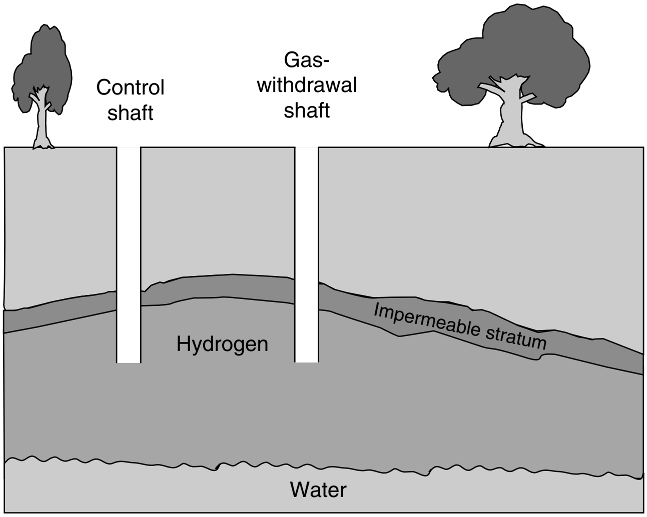


Figure 11.1 Hydrogen storage in an aquifer.

corresponding to a stored energy of 1800 MJ. A hydrogen fuel cell car has, at least, twice the efficiency of an internal combustion one; hence it needs only 900 MJ to have the same range. The refueling time would be slightly over 3 minutes.

For very large-scale storage, it may be possible to keep hydrogen in underground formations, such as porous rocks, old mines, caves, aquifers, and exhausted natural gas deposits. At present, there is little experience with underground storage of hydrogen. However, results of the helium storage in Amarillo, Texas, suggest that there will be little difficulty with this technology. Figure 11.1 depicts an arrangement for keeping large amounts of gas in an aquifer, provided an impermeable layer of rocks forms an adequate roof over the structure. In Amarillo, 8.5×10^8 m^3 of helium are stored without problems. It should be noted that helium has leakage characteristics similar to those of hydrogen. At STP, 8.5×10^8 m^3 of hydrogen correspond to 10,000 TJ of stored energy.

To gain an idea of how much this storage capacity is, one can compare it with that of one of the world's largest pumped storage facilities[†] in Ludington, Michigan. This facility has a capacity of 54 TJ, or nearly 200 times less than that of a hydrogen-filled Amarillo reservoir.

Another storage arrangement for hydrogen would be the very pipelines used for transporting the gas. A typical trunk pipeline for natural gas can be over 1000 km long. It may have a 1.2-m diameter and operate at 6 MPa

[†] In pumped storage plants, normal hydroelectric machinery can be reversed so as to pump water back into the reservoir during times when there is surplus electric energy in the system.

(60 atmospheres). The hydrogen stored in such a pipeline would correspond to an energy of 1000 TJ, nearly 20 times the capacity of the Ludington reservoir.

11.2 Cryogenic Hydrogen

Although hydrogen was first liquefied in 1898, it was only recently, through the efforts of NASA, that the technology for production and storage of large quantities of the liquid was developed.

The largest storage unit in existence is one at Cape Canaveral, with a capacity of $3375\,\mathrm{m}^3$. Since the density of liquid hydrogen is $71\,\mathrm{kg\,m}^{-3}$, the facility can accumulate 240,000 kg of liquid hydrogen, or 34 TJ, just a little less than the capacity of the Ludington reservoir we have been using for comparison.

There are two different species of hydrogen molecules: para- and ortho-hydrogen. In the first, the spin in the two atoms that constitute the molecule are in opposite directions, whereas in the second, the spins are in the same direction.

In the liquid state, para-hydrogen ($p\text{-}H_2$) has lower enthalpy than the ortho form ($o\text{-}H_2$). At the boiling point of H_2 (20.4 K at 0.1 MPa), the difference is $1.406\,\mathrm{MJ\,kmole}^{-1}$.

In hydrogen, gaseous or liquid, the reaction

$$p\text{-}H_2 \leftrightarrow o\text{-}H_2 \tag{11.2}$$

goes on continuously, and an equilibrium concentration of each species is established. The equilibrium at STP corresponds to 25% $p\text{-}H_2$ and 75% $o\text{-}H_2$, whereas at 20.4 K, the equilibrium shifts to 99.79% $p\text{-}H_2$. Owing to the slow kinetics of the reaction, freshly cooled hydrogen tends to have an excess of the ortho variety, and its transformation into the para variety results in the release of heat, causing the liquid to boil even though no external heat is supplied.

Freshly condensed hydrogen, even if kept in a perfectly adiabatic container, will lose 1% of its mass during the first hour and 50% during the first week. To minimize such losses, $o\text{-}H_2$ is catalytically converted to $p\text{-}H_2$ during the liquefaction process. Levels of 95% $p\text{-}H_2$ are desirable.

Liquid hydrogen has been considered as a fuel for aircraft. Lockheed investigated the performance of a supersonic airplane designed to carry 234 passengers 7800 km at Mach 2.7. A kerosene-powered plane with such a capability would have a gross weight of 232 tons, of which 72 tons would be fuel. A hydrogen plane with equivalent performance would have a gross weight of only 169 tons, of which less than 22 tons would be fuel.

Hydrogen-driven commercial airplanes will probably not be seen in the near future. Present-day design efforts for transport planes in the Mach 3 range are based on jet fuel (kerosene) engines. However, the proposed space plane will probably need hydrogen as a fuel. It will be a hypersonic transport (perhaps Mach 8) capable of taking off from a conventional runway and achieving orbital flight.[†]

One of the problems associated with high speeds while inside the atmosphere is the high temperatures generated. The **stagnation temperature** of a body moving through a gas (the temperature reached by the gas at the point in which its flow speed relative to the body is zero) is given by

$$\frac{T}{T_{amb}} = 1 + \frac{\gamma - 1}{2} M^2. \tag{11.3}$$

See insert.

For air, $\gamma = 1.4$ and our equation becomes

$$\frac{T}{T_{amb}} = 1 + 0.2M^2. \tag{11.4}$$

For M = 2.5 and $\gamma = 1.4$, $T/T_{amb} = 3.25$ and for M = 25, it is 226. This means that, at the latter velocity, if the ambient temperature is 300 K, the stagnation temperature is 67,800 K.

Clearly, no material exists that can operate at such temperatures. This means that the heat developed at the leading edges of the fuselage, wings, and control surfaces must be efficiently removed. Part of this can be achieved by radiation and conduction and in part by refrigeration. The liquid hydrogen fuel can be used to cool critical regions of the plane prior to being conveyed to the engine in gaseous form.

To be liquefied, hydrogen must be of high purity. Most other gases will freeze during the process and will tend to clog the pipes. If oxygen ice is formed, explosions may result. Specifications generally call for less than 10 ppm of O_2.

Practical liquefaction machines require about 40 MJ to condense 1 kg of H_2. This energy cannot be recovered; the resulting storage turnaround efficiency is $143/(143 + 40) = 0.78$ or 78%.

The cost of a cryogenic plant does not scale linearly with the rate of production; it grows with $\dot{M}^{0.7}$, where $\dot{M}$ is the rate of production.

Stagnation Temperature

The sum of the enthalpy and the kinetic energy of a flowing gas is constant:

$$c_pT + \frac{1}{2}u^2 = c_pT_\infty + \frac{1}{2}u_\infty^2. \tag{11.5}$$

(*Continues*)

[†] A supersonic combustion ramjet (scramjet) engine seems to be essential for the success of hypersonic transport planes. On March 27, 2004, an experimental scramjet engine was tested in free flight. It operated for 10 seconds and accelerated to Mach 7 (8000 km/hr).

(*Continued*)

The preceding is, of course, per unit mass. Let T_∞ be the undisturbed—that is, ambient temperature and $u_\infty \equiv v$ be the undisturbed wind velocity—that is, the velocity of the object. At the **stagnation point**, the velocity is, by definition, zero. Hence,

$$c_pT = c_pT_{amb} + \frac{1}{2}v^2, \tag{11.6}$$

$$\frac{T}{T_{amb}} = 1 + \frac{1}{2}\frac{v^2}{c_pT_{amb}}. \tag{11.7}$$

The speed of sound is $c = \sqrt{\gamma R T_{amb}}$; therefore,

$$\frac{T}{T_{amb}} = 1 + \frac{\gamma R}{2c_p}\frac{v^2}{c^2} = 1 + \frac{1}{2}\gamma R M^2, \tag{11.8}$$

where $M \equiv v/c$ is the **Mach number**.
But

$$\gamma = \frac{c_p}{c_v} \tag{11.9}$$

and

$$c_p = c_v + R; \tag{11.10}$$

thus

$$c_p = \frac{\gamma R}{\gamma - 1} \tag{11.11}$$

$$\frac{T}{T_{amb}} = 1 + \frac{\gamma - 1}{2}M^2. \tag{11.12}$$

11.3 Storage of Hydrogen by Adsorption

Both hydrogen molecules and methane can be readily adsorbed on carbon. The gases are held in place by weak van der Waals forces so that the energy necessary to retrieve the fuel is small.

Carbon systems can be combined with other techniques: the gas can be pressurized and the temperature can be lowered. Typically, an adsorption system using activated carbon requires 20 to 40 atmosphere pressure and liquid nitrogen temperature to hold the hydrogen. These requirements severely limit the practical application of such systems. They achieve 5%

to 6% gravimetric concentration (GC), just about the same as that of good metal hydrides systems (discussed in Subsection 11.4.5).

Carbon nanotubes promise to uptake hydrogen much more effectively. Single wall carbon nanotubes have been reported as yielding gravimetric concentrations of up to 10% when operated at 120 K and 0.4 atmospheres. N. M. Rodriguez (Northeastern University) claims 40% GC in graphite nanotubes. Chambers, Parle, and Baker (1998). report 67% GC at room temperature and 120 atmospheres in graphite nanofibers with herringbone structure.

All these systems require low temperatures or high pressures. Chen et al. (1999) describe a system that operates at 1 atmosphere and at some 200 C to 400 C. It uses lithium doped carbon nanotubes and achieves a GC of 20%.

It should be noted at this point that if each carbon atom were to attach 1 hydrogen atom, then the GC would be only 8%.

11.4 Storage of Hydrogen in Chemical Compounds

11.4.1 Generalities

The main difficulty encountered in the storage of hydrogen is, as pointed out, the low density of the gas. It is possible to substantially increase the packing density by associating hydrogen with other substances. The storage and retrieval processes consist then of the synthesis of a hydrogen-rich compound followed, when the gas is needed, by its dissociation.

The requirements of a practical hydrogen storing compound include:

1. High storage capacity
   The density of liquid hydrogen is 71 kg m^{-3}. Many hydrogen-rich compounds have packing densities that exceed this value. As an example, consider three common hydrides listed in Table 11.1.
   To achieve high **volumetric storage capacity**, the hydride must have a high hydrogen-packing density. To achieve a high **gravimetric storage capacity**, the hydride must be of relatively low density.

Table 11.1 Hydrogen-Packing Densities in Different Compounds

Technical† name	**Common name**	**kg(H$_2$)m^{-3}**	**kg(H$_2$)kg^{-1}**
Oxygen di-hydride	water	111	0.111
Nitrogen tri-hydride	ammonia	113	0.147
Di-nitrogen tetra-hydride	hydrazine	126	0.125

† *With an apology to chemists, we have characterized the three substances as "hydrides," to emphasize their hydrogen-carrying properties.*

Table 11.2 Enthalpies of Formation of Some Hydrogen-Rich Compounds

Compound	**Energy of formation (per kg of H_2)(MJ)**
Water (ℓ)	143
Ammonia (g)	15.4
Hydrazine	−12

2. Low reaction energy
   Hydrogen as a fuel is almost always, used by combining it with oxygen and producing water. This releases 143 MJ per kilogram of liquid water. Clearly, the energy of formation of the hydride in which hydrogen is stored must be substantially lower than this value for the storage system to be useful. For instance, 15.4 MJ are necessary to dissociate enough ammonia to produce 1 kg of hydrogen—a theoretical turnaround efficiency of $143/(143 + 15.4) = 90\%$. This is acceptable. On the other hand, water cannot be used because it has a theoretical turnaround efficiency of 0%.

   In some cases, the energy required to promote the dissociation of the hydrogen-carrying compound can be low-grade heat energy. Thus the practical impact of a lower turnaround efficiency can be small.

   The enthalpies of formation of the three compounds in the preceding table are shown in Table 11.2.
3. Reversibility
   The reaction should be easily reversible; that is, it should be easy to shift the equilibrium to either the H_2 or the hydride side.
4. Kinetics
   The hydrogen fixing and releasing reactions must occur rapidly, at relatively low temperatures and without requiring expensive catalysts.
5. Separability
   It should be easy to separate the dissociation products. Ideally, the products should be gaseous hydrogen plus solid residues.
6. Low corrosiveness

11.4.2 Hydrogen Carriers

Clearly, one way to store and transport hydrogen is to synthesize a hydrogen-rich substance and then, as needed, generate hydrogen by a chemical “reforming” process like some of those described in Chapter 10. Of great interest for fuel cell vehicles (FCVs) is the use of methanol, a liquid fuel, which can be reformed with relative ease. Considerable effort is being devoted to the design of simple onboard reformers, but there are difficulties:

1. Efficiency
   Assume natural gas as the feedstock. If hydrogen is produced from this gas in a practical device, $\approx 90\%$ of the heat of combustion of the initial raw material appears in the H_2. If natural gas is converted to methanol, only 71% of the heat of combustion of the raw material is available from the alcohol produced. While hydrogen can be used directly by a fuel cell, methanol has to be reformed onboard, and only 77% of the heat of combustion of the alcohol becomes available as hydrogen for the cell. Thus, the relative efficiencies are 90% for hydrogen and only 77% of 71% or an overall 55% for methanol.
2. CO_2 emission
   Again, assume that both methanol and hydrogen are derived from natural gas. Fueling an FCV with methanol results in an overall emission of CO_2 1.5 times that resulting from using hydrogen.
3. Contaminants
   Fuel cells for automotive use will most likely be low-temperature SPFCs that are sensitive to the presence of impurities, such as CO, that degrade the catalyst. Such impurities will be present in the hydrogen whether it is produced directly from natural gas or through an intermediate methanol step. The latter requires reforming aboard the vehicle where, plausibly, good purification of the hydrogen will be more difficult and more expensive than at a central hydrogen generating plant.
4. Environmental danger
   If the use of methanol ever becomes as widespread as that of gasoline today, there will unavoidably be some spills, as happens occasionally with petroleum products. The consequences of methanol spills may, however, be much more serious than those of oil or gasoline. The last two fuels do not mix with water and float on its surface. Methanol mixes with water in any proportion, and a major spill may contaminate an aquifer. This will render the water of a given region undrinkable because methanol is quite poisonous, blinding or killing those who ingest a sufficient quantity of it.

11.4.3 Water Plus a Reducing Substance

In Chapter 10, it was mentioned that balloonists of the last century produced hydrogen by passing steam through a bed of iron filings. The iron oxidizes into rust and the water is reduced to hydrogen in a reversible reaction:

$$3Fe + 4H_2O \leftrightarrow Fe_3O_4 + 4H_2. \tag{11.13}$$

Driving this reaction toward the left constitutes a way of storing hydrogen. There are, however, substantial difficulties in this scheme.

Unfortunately, the reaction cannot be driven to completion either in the forward, or hydrogen-producing, direction or in the reverse direction.

The temperatures that must be used are quite high (above 1000 C).

Although iron oxidation is exothermic, the heat released is far less than that needed to boil the required water and superheat the resulting steam.

11.4.4 Formic Acid

Photosynthesis operates by hydrogenating carbon dioxide transforming it into glucose (see Chapter 13). A simpler reaction is the direct hydrogenation of CO_2, transforming it into formic acid. Formally,

$$CO_2 + H_2 \rightarrow HCOOH. \tag{11.14}$$

Hydrogenation is an example of an **addition reaction** in which a substance reacts with a double or triple bond, opening it up and incorporating itself into the product. Hydrogenation of fats or organic oils removes some or all multiple C=C or C≡C bonds, reducing the degree of saturation of the original fat or oil. Consult Chapter 13.

Since carbon dioxide, O=C=O, has two double bonds, it can be hydrogenated twice according to Equation 11.14.

The hydrogenation is accomplished in the presence of an amine and is mediated by a ruthenium, rhodium, or iridium catalyst (see Figure 11.2). The resulting formic acid can be made to release the hydrogen on demand by means of a room temperature decomposition (still in the presence of an amine). CO_2 is released and reused during the next storing phase; the system is CO_2 neutral. The production of hydrogen from formic acid, discussed in Chapter 10, was proposed by Loges et al. (2008).

11.4.5 Metal Hydrides[†]

A majority of the requirements listed in Subsection 11.4.2 can be met by a class of substances loosely called **metal hydrides**, or simply hydrides.

A number of elements form unstable hydrides (hydrides that can easily be reversed). Magnesium, iron, titanium, zirconium, yttrium, lanthanum, and palladium are examples. Hydrides of elements are called **binary**. **Ternary hydrides**—hydrides formed by a combination of hydrogen with a binary compound—are more promising. A typical example are the hydrides of TiFe. The addition of a third element, leading to **quaternary hydrides**, increases even further the degree of freedom in choosing the characteristics

[†] It is important to distinguish the noun "hydrate" from the noun "hydride." Hydrate describes a combination of an element or a radical with water, while hydride describes the combination with hydrogen. "To hydrate" is the action of creating a hydrate. The *Oxford English Dictionary* does not list the verb "to hydride," which we will use in this chapter to indicate the action of creating a hydride.

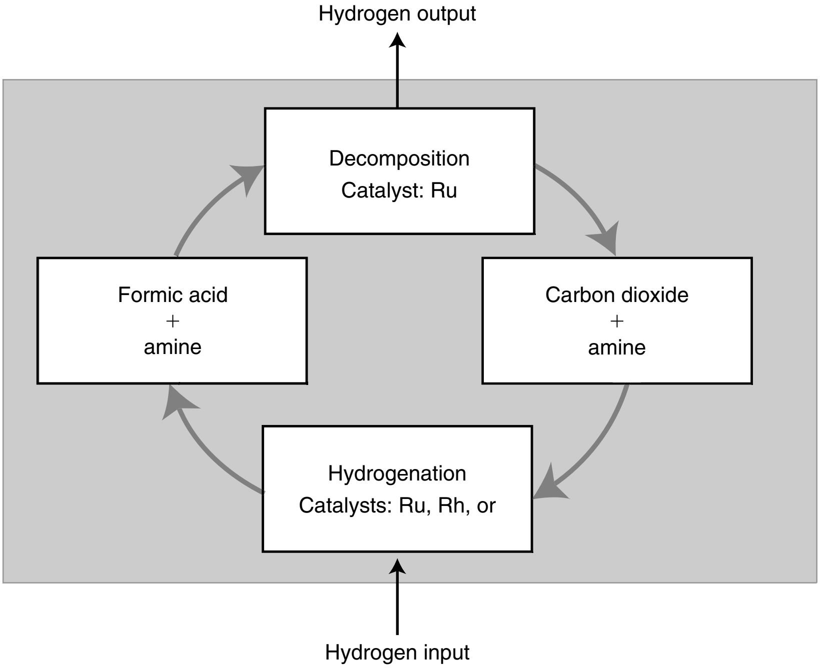


Figure 11.2 Formic acid hydrogen storage system.

of the system. As can be seen, a large number of combinations are possible. Research in this area has only scratched the surface.

The exchange of hydrogen (or any other gas) with a solid is called **sorption**. If hydrogen is being fixed, the reaction is called **absorption**; if hydrogen is being released, it is called **desorption**.

Hydride characteristics are best inspected by examining the pressure vs. hydrogen concentration isotherms of the material. Such data are measured by an apparatus similar to the one shown schematically in Figure 11.3.

A certain amount of **activated** (see further on) granules composed, for example, of elements A and B forming an alloy, AB, is placed in the sample holder. The exact number of kilomoles of the alloy, μ_{AB}, has previously been determined. Valves a and c are open and valve b is closed. The sample is degassed by heating it to a high temperature and extracting the released vapors by means of a vacuum pump.

Valves a and c are then closed, and b is opened filling the reservoir with a known volume of hydrogen at known pressure and temperature—that is, with a known amount of the gas. With b and c closed, a is opened and hydrogen is absorbed. This causes the temperature of the powder to rise because the absorption is exothermic. The system is then returned to the initially selected temperature, and the equilibrium pressure is observed.

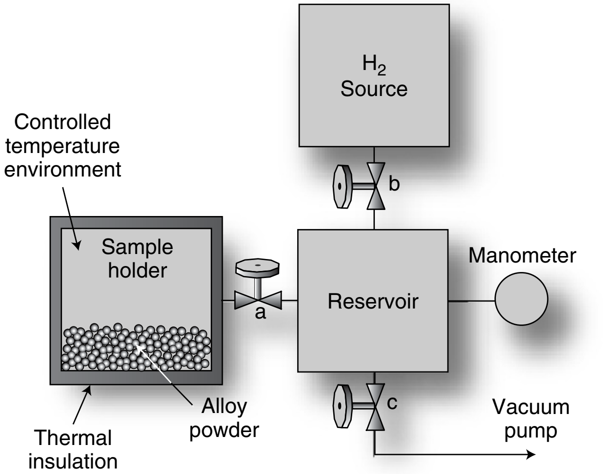


Figure 11.3 Apparatus for the determination of pressure versus concentration isotherms.

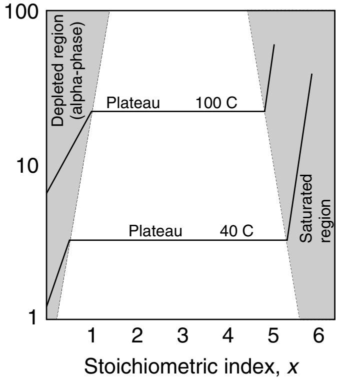


Figure 11.4 Idealized pressure-stoichiometry plot for hydrides.

In this manner, the amount, μ_H, of hydrogen taken up by the granules can be calculated. Either the **hydrogen/metal ratio**, H/M, or the stoichiometric index, x, of the formula ABH_x, is determined. For instance, $ABH_{0.4}$ has a H/M ratio of $0.4/(1+1) = 0.2$. Many pressure/composition diagrams in this chapter have ordinates scaled in both x and H/M.

The procedure is repeated with increasing amounts of hydrogen, and H/M (or x) is tabulated versus the equilibrium pressure, p. In an extremely idealized way, the resulting plot may look like the one shown in Figure 11.4.

The solid material in the system can be in one of several states:

1. It may be simply the initial metallic alloy (called the **alpha-phase**).
2. It may be a single hydride of the alloy (called the **beta-phase**).
3. It can, in some cases, be a double hydride of the alloy (called the **gamma-phase**).[†]

At low-hydrogen concentrations, the pressure depends strongly on x: hydrogen either occupies the interstices between the granules or is dissolved in the alloy without altering its chemical composition. The corresponding region on the pressure-stoichiometry diagram is the **depleted region**.

As the concentration increases, a **plateau** region is reached where pressure is nearly independent of concentration. This is a region of equilibrium between the alpha- and the beta-phase. The higher the hydrogen concentration, the larger the fraction of the alloy in the beta-phase and the smaller in the alpha-phase. In this region, the alloy is hydrided at discrete locations where its composition is ABH. This occurs because it is energetically more favorable to react chemically than to dissolve. An equilibrium between hydrided sites, ABH, and nonhydrided sites, AB, is established. The more hydrogen is introduced, the larger the ratio of the hybrided to the non hybrided sites. The exact proportion is indicated by the value of x in ABH_x. The plateau persists until all the alloy is hydrided—that is, until all the material has the composition ABH—it is all in the beta-phase. If the amount of hydrogen is increased further, the excess will dissolve in the beta-phase, and the pressure again rises rapidly. This is the **saturated region**. In some cases, a dihydride (gamma-phase) is formed, and a second plateau may appear.

Real pressure-stoichiometry curves depart from the ideal shape described above. The transition from the alpha- to the beta-phase is smooth, not abrupt. The plateau may exhibit a small slope, and a sorption hysteresis may become evident: the plateau pressure can be noticeably higher during absorption than during desorption. In some cases, a second plateau may be reached where a dihydride, ABH_2, is formed. This second plateau is called the **beta prime-phase**. See Figure 11.5.

Different plateau pressures correspond to different temperatures. The width of the plateau becomes narrower as the temperature increases. Above a given temperature, no plateau is formed, as indicated in Figure 11.6.

Figure 11.7 shows how the plateau pressure can be tailored to a given specification by a relatively small change in alloy composition. Three different alloys, Ni_5La, $Ni_5La_{0.8}Nd_{02}$, and $Ni_5La_{0.8}Er_{02}$, display, at 40 C, plateau pressures of about 3, 5, and 10 atmospheres, respectively.

[†] Double hydride does not mean that there are necessarily two hydrogen atoms in the hydride formula. For instance, the single hydride of the $LaNi_5$ system is $LaNi_5H_5$.

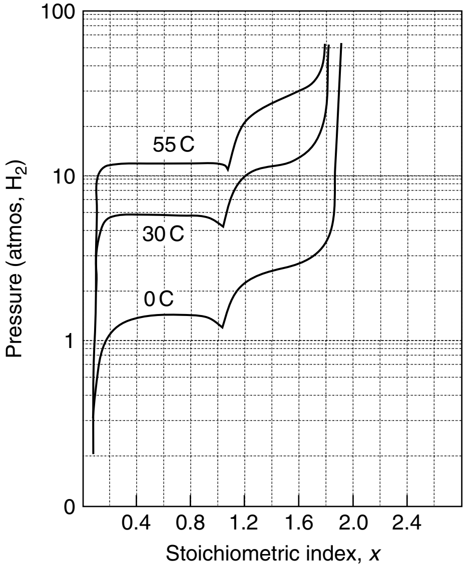


Figure 11.5 Pressure vs. stoichiometric index, x, for TiFe.

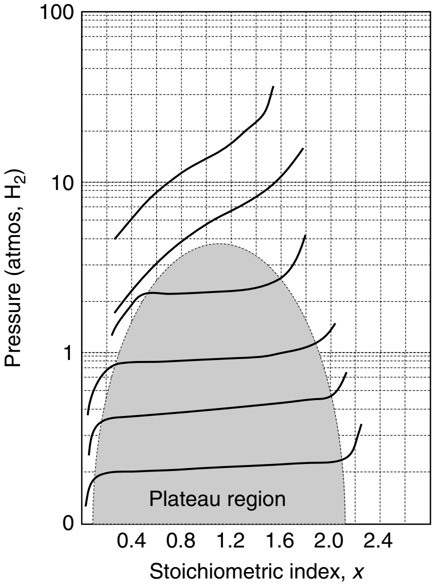


Figure 11.6 The plateau region becomes progressively narrower as the temperature increases.

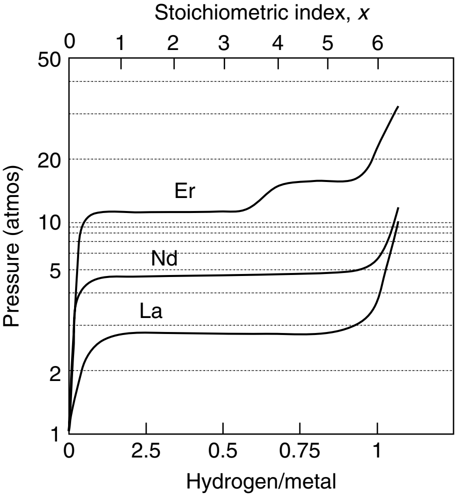


Figure 11.7 The plateau pressure can be tailored by relatively small composition changes of the alloy.

When a metallic compound is first prepared, it will usually not react well with hydrogen at room temperature, probably owing to an oxide layer on the surface.

Many alloys require activation, which consists of heating granules of the material to some 300 to 500 C in high vacuum to outgas them, and, afterward, expose them to high purity hydrogen.

Not surprisingly, the lattice constant of the hydride is not the same as that of the parent material—absorption of hydrogen is accompanied by an increase in volume.

In the $LaNi_5$ case, for instance, there is a 25% expansion. Since hydrides tend to be brittle, the expansion causes the material to break into a fine powder. The breaking up of particles helps the activation because it exposes new clean surfaces. TiFe needs a thorough activation, while $LaNi_5$ needs only a very simple one.

The tendency to expand during absorption and contract during desorption can lead to mechanical difficulties. During desorption, the contracting particles constitute a more compact powder that can accommodate itself into new positions in the container and, upon subsequent expansion, can provoke serious strains on its walls.

The theoretical volumetric concentration (VC) of hydrogen in hydrides can exceed that of liquid hydrogen (see Table 11.3). These data, however, overestimate the concentration and, consequently, overestimate the corresponding volumetric energy concentration because the bulk density of the hydride powder is much less than that of the hydride itself, owing to the imperfect packing of the individual granules.

Table 11.3 Hydrogen Concentration in Various Hydrides

	Mass of H_2 in Hydride %	**Mass of H_2 in Hydride (kg/m^3)**	**Energy density**	
			(MJ/kg)	**(GJ/m^3)**
$H_2(\ell)$	100	71	143	10.2
$H_2(g)$, STP	100	0.089	143	0.013
LaH_3	2.1	108	3.0	15.4
MgH_2	7.6	101	10.0	14.4
TiH_2	4.0	153	5.7	21.9
VH_2	3.8	95	3.0	13.6
ZrH_2	2.1	122	3.0	17.4
$LaNi_5H_5$	8.7	89	2.0	12.7
Mg_2NiH_4	3.6	81	4.5	11.6
$TiFeH_{1.95}$	1.85	101	2.6	14.4

For instance, the realizable hydrogen concentration in $LaNi_5H_5$ is about 45 kg of hydrogen per cubic meter, about half of the number that appears in the table. Nevertheless, on a per volume basis, hydrides can pack roughly the same energy as liquid hydrogen, whereas in a per mass basis, hydrides pack only a few percent of the liquid hydrogen energy. For applications, such as airplanes, in which weight considerations are of primary importance, hydrides do not loom as promising energy storage methods. For locomotives, barges, trucks, and buses, they may be useful. Hydrides may even become attractive for automobiles.

Compare Mg_2NiH_4 with gasoline: 100 kg of hydride store, at best, $3.6 \times 143 = 515$ MJ of fuel, whereas 100 kg of gasoline represent 4700 MJ, almost a 10-fold advantage. However, the usual efficiency of a gasoline engine is around 20%, while that of a hydrogen fuel cell is around 60%; this reduces the gasoline advantage by a factor of 3.

A hydride fuel cell car (of current technology) has a weight advantage over an electric car using advanced batteries.

One additional practical consideration is the rapidity with which a hydride storage system can be charged and discharged. This depends, of course, on the kinetics of the reaction. Magnesium, owing to its small density would be attractive were it not for its inability to absorb hydrogen under normal conditions. The kinetics of the reaction can be improved by the addition of nickel as a catalyst. A 5% nickel doping already helps noticeably, but a much larger amount of nickel is generally used, such as in the Mg_2Ni alloy.

If fast recharge rates are required, the total amount of absorbed hydrogen may be smaller than when a slow recharge is used. FeTi can be (relatively quickly) charged to $FeTiH_{1.6}$, while to achieve $FeTiH_{1.95}$, an overnight charge may be needed.

The capacity of the storage system is also determined by the maximum amount of hydrogen that can be dissolved in the alloy (alpha-phase). The

298 K plateau for FeTi ends when $FeTiH_{0.4}$ is reached. Any attempt to extract more hydrogen is accompanied by a great reduction in hydrogen pressure. Thus, for quick charge-discharge cycles, this particular alloy can only be operated between $FeTiH_{0.4}$ and $FeTiH_{1.6}$, realizing a turnaround capacity only 60% of the total theoretical capacity of the system.

11.4.5.1 Characteristics of Hydride Materials[†]

Some of the distinguishing characteristics of materials used in hydride systems are as follows.

Plateau Slope

Simple thermodynamic considerations, as discussed in the next section, predict horizontal plateaus. In practice, the plateau pressure rises somewhat with increasing hydrogen content.

The slope of the plateau pressure can be represented by the **slope parameter**, $d\ln(p_d)/d(H/M)$, where p_d is the plateau pressure of the desorption isotherm. In Figure 11.8, the dashed line through the 25 C desorption isotherm intercepts the $H/M = 0$ axis at 9.1 atmospheres and the $H/M = 1.2$ axis, at 14.8 atmospheres.

Hence

$$\frac{d\ln p_d}{d(H/M)} = \frac{\ln 14.8 - \ln 9.1}{1.2} = 0.41. \tag{11.15}$$

This is a reasonably large slope parameter. TiFe, for instance, tends to have a slope parameter of 0.00, whereas some (but not all) calcium-based alloys have parameters larger than 3. When an alloy solidifies (during the initial manufacture), the component elements sometimes tend to segregate—that is, to settle out. This seems to be the main reason for the appearance of a plateau slope, because, from thermodynamic grounds, a perfectly uniform alloy should exhibit no slope at all. Annealing of the material *prior* to crushing can reduce the plateau slope. Slope parameters and other characteristics are listed in Tables 11.4, 11.5, and 11.6.

Sorption Hysteresis

As mentioned previously, the plateau pressure during absorption is usually somewhat larger than that during desorption. In other words, there is a **sorption hysteresis** when the alloy is cycled. See, for instance, Figures 11.8, 11.9, 11.11, and 11.12.

Hysteresis is an irreversible process caused by the heat generated by the plastic deformation of the crystal as the lattice expands or contracts during the sorption cycle.

[†] All alloys using the HY-STOR trademark are produced by Ergenics, Inc, and much of the data used in this text are from a paper by Houston and Sandrock (1980). In the formulas, "M" stands for Mischmetall (German), a mixture of rare earths generally extracted from monazite sands. Their effect on the plateau pressure depends strongly on their cerium-to-lanthanum ratio.

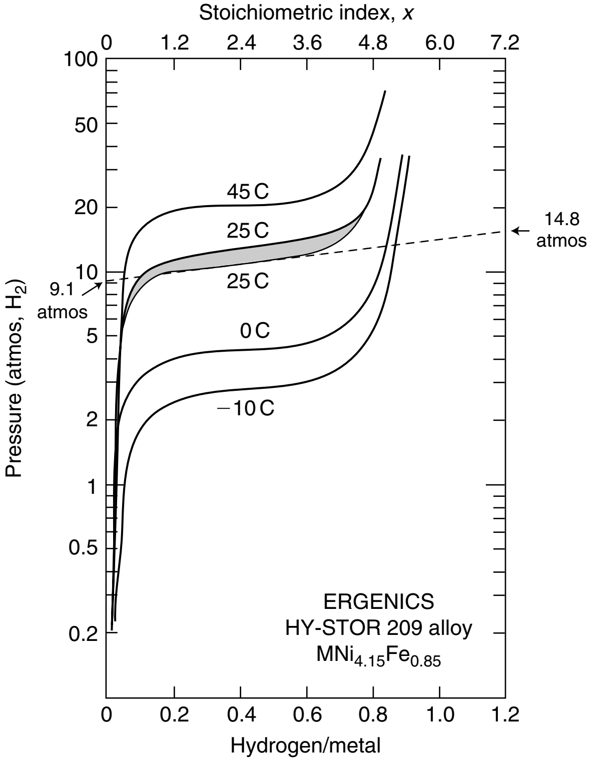


Figure 11.8 Measurement of the slope parameter in the pressure-composition isotherm.

The degree of hysteresis is indicated by the ratio of the absorption plateau pressure, p_a, to the desorption pressure, p_d, measured at a H/M value of 0.5 and, frequently, at 25 C.

It is commonly assumed that the degree of hysteresis is fairly temperature independent.

Usable Capacity

The usable capacity is defined, somewhat optimistically, as the change in the H/M ratio in a hydride when the pressure is lowered from 10 times the plateau pressure to 0.1 times this plateau pressure. This is a 100:1 pressure range. A more realistic definition would be based on a much smaller pressure range.

In Figure 11.10 (Alloy $Fe_{0.8}Ni_{0.2}Ti$), the plateau pressure of the 70 C isotherm is about 0.9 atmosphere. At 10 times this pressure, the H/M ratio is 0.65, whereas at 0.1 times the plateau pressure, $H/M = 0.02$. This leads to $\Delta(H/M) = 0.63$. In other words, the hydride releases 0.63 kilomole of H (0.63 kg) for each kilomole of metal.

The alloy consists of 0.80 kilomole of iron (massing $0.8 \times 55.8 = 44.6$ kg), 0.2 kilomole of nickel (massing $0.2 \times 58.7 = 11.7$ kg), and 1 kilomole

Table 11.4 Hydrogen Capacity and Heat Capacity, Selected Hydrides

Hydride	**HY-STOR alloy***	**Hydrogen capacity**		**Heat capacity**
		$\Delta(H/M)$	**wt.%**	**J/(kg K)**
FeTi	101	0.90^c	1.75	540
$Fe_{0.9}Mn_{0.1}Ti$	102	0.92^d	1.79	540
$Fe_{0.8}Ni_{0.2}Ti$	103	0.62^e	1.21	500
$CaNi_5$	201	0.71	1.39	540
$Ca_{0.7}M_{0.3}Ni_5$	202	0.96	1.60	500
$Ca_{0.2}M_{0.8}Ni_5$	203	0.74	1.08	440
MNi_5	204	1.01	1.41	420
$LaNi_5$	205	1.02	1.43	420
$LaNi_{4.7}Al_{0.3}$	207	0.95	1.36	420
$MNi_{4.5}Al_{0.5}$	208	0.83	1.20	420
$MNi_{4.15}Fe_{0.85}$	209	0.82	1.15	—
Mg_2Ni	301	1.31^f	3.84	750
Mg_2Cu	302	0.75^g	2.04	750

**HY-STOR is a trademark of Energics, Inc.*
c – 30 C; d – 40 C; e – 70 C; f – 25 C; g – 325 C.

Table 11.5 Thermodynamic Data Selected Hydrides

Hydride	**HY-STOR alloy***	**ΔH_f MJ kmole^{-1} of H_2**	**ΔS_f kJ K^{-1}kmole^{-1} of H_2**
MNi_5	204	−20.9	−96.8
$Ca_{0.2}M_{0.8}Ni_5$	203	−24.3	−108.7
$MNi_{4.15}Fe_{0.85}$	209	−25.1	−104.8
$Ca_{0.7}M_{0.3}Ni_5$	202	−26.8	−100.4
FeTi	101	−28.0	−106.1
$MNi_{4.5}Al_{0.5}$	208	−28.0	−104.8
$Fe_{0.9}Mn_{0.1}Ti$	102	−29.3	−107.0
$LaNi_5$	205	−31.0	−107.7
$CaNi_5$	201	−31.8	−101.2
$LaNi_{4.7}Al_{0.3}$	207	−33.9	−106.8
$Fe_{0.8}Ni_{0.2}Ti$	103	−41.0	−118.8
$ZrCr_2$		−46.0	−98.3
Mg_2Ni	301	−64.4	−122.3
Mg_2Cu	302	−72.8	−142.3
Mg		−77.4	−138.3

**HY-STOR is a trademark of Energics, Inc.*

Table 11.6 Deviations from the Ideal Selected Hydrides

Hydride	HY-STOR alloy[*]	Plateau slope[a] $\frac{d \ln p_d}{d(H/M)}$	Hysteresis parameter P_a/P_d
FeTi	101	0.00[c]	1.89
$Fe_{0.9}Mn_{0.1}Ti$	102	0.65[d]	1.85
$Fe_{0.8}Ni_{0.2}Ti$	103	0.36[e]	1.05
$CaNi_5$	201	0.19	1.17
$Ca_{0.7}M_{0.3}Ni_5$	202	3.27	1.11
$Ca_{0.2}M_{0.8}Ni_5$	203	0.98	1.48
MNi_5	204	0.54	5.2
$LaNi_5$	205	0.09	1.21
$LaNi_{4.7}Al_{0.3}$	207	0.48	1.05
$MNi_{4.5}Al_{0.5}$	208	0.36	1.12
$MNi_{4.15}Fe_{0.85}$	209	0.43	1.18
Mg_2Ni	301	0.02[f]	—
Mg_2Cu	302	0.17[g]	—

[*] *HY-STOR is a trademark of Energics, Inc.*
a – 25 C plateau; c – 30 C; d – 40 C; e – 70 C; f – 25 C; g – 325 C.

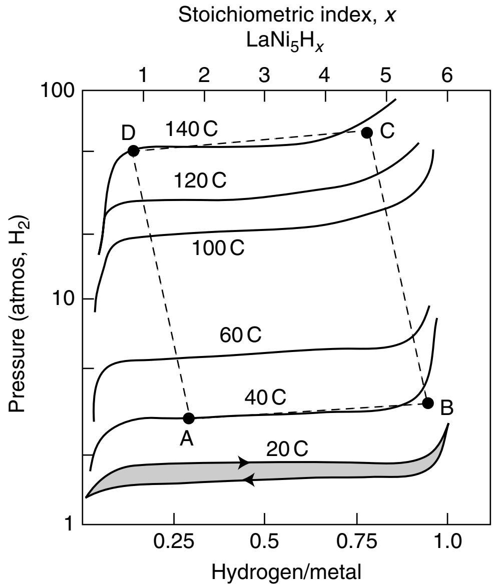


Figure 11.9 Hysteresis in the sorption isotherms of the $LaNi_5$.

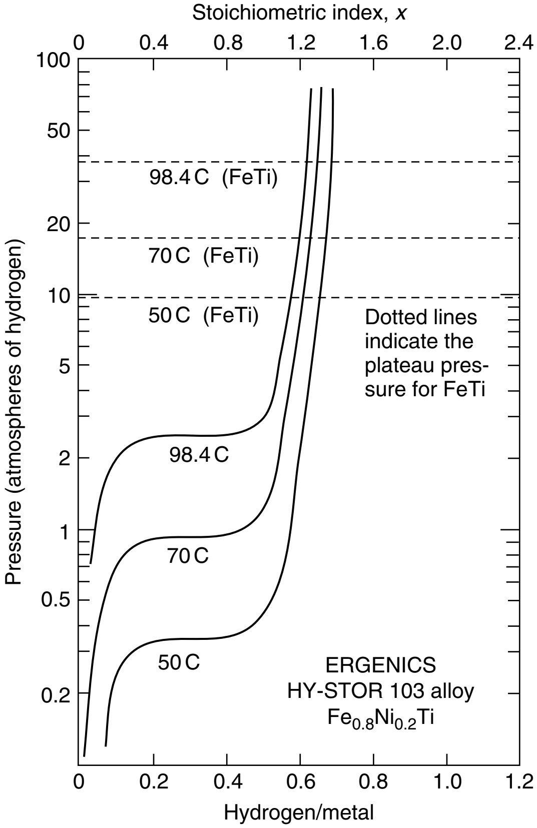


Figure 11.10 Addition of small amounts of Ni causes a substantial change in the characteristics of the FeTi alloy of Figure 11.5.

of titanium (massing 47.9 kg). This adds up to 2 kilomoles of metal massing $44.6 + 11.7 + 47.9 = 104.2$ kg, or 52.1 kg/kmole. Thus, the released hydrogen represents $0.62/52.1 = 0.0012$ of the mass of the hydride, or 1.2%.

Heat Capacity

Hydride systems are activated by temperature changes. It is important to know the heat capacity of the different alloys in order to properly design the systems. The values of heat capacities of a number of alloys are listed in Table 11.4.

Plateau Pressure Dependence on Temperature

The temperature dependence of the plateau pressure is a function of the thermodynamic properties of the material and is examined in more detail in the next section. By altering the composition of the alloy, it is possible to tailor the characteristics of the material to suit many different applications.

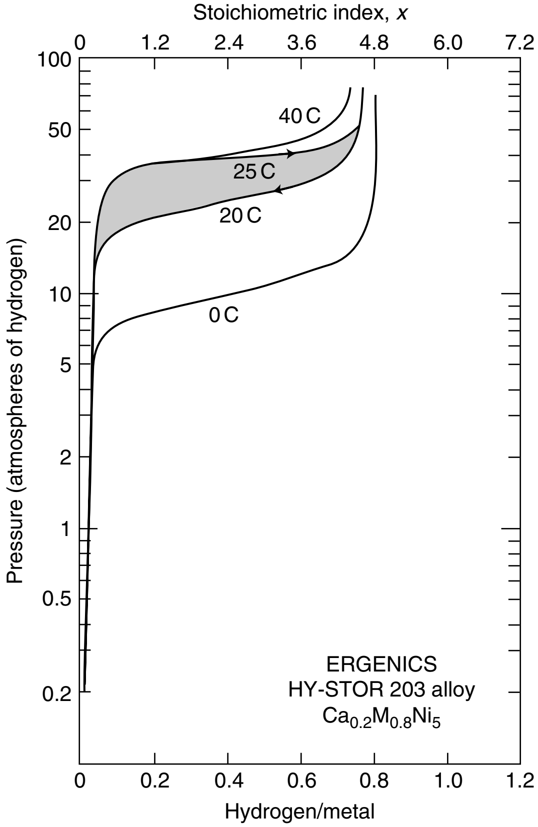


Figure 11.11 Characteristics of the $Ca_{0.2}M_{0.8}Ni_5$ system.

From Figure 11.5, it can be seen that the popular FeTi system has a plateau pressure above 10 atmospheres when the temperature is 55 C.

The substitution of Ni for some of the Fe dramatically reduces the plateau pressure (see Figure 11.10); 20% Ni is sufficient to lower the pressure (at 50 C) to 0.3 atmospheres. In addition, the nickel greatly reduces the hysteresis (although this is not shown in the figures) and facilitates the activation of the material. The $Ca_{0.2}M_{0.8}Ni_5$ system (see Figure 11.11) exhibits a plateau pressure of some 30 atmospheres at 25 C, whereas the calcium-free $MNi_{4.15}Fe_{0.85}$ system has a plateau pressure of only 10 atmospheres at the same temperature (Figure 11.8). Lower than atmospheric plateau pressure at room temperature can be achieved, for instance, by adding aluminum to the $LaNi_5$ system (Figure 11.12).

The reaction of hydrogen with metallic compounds involves, of course, enthalpy changes: absorptions are exothermic, and desorptions are endothermic. The enthalpy changes can be determined by calorimetric techniques. More often, they are determined from the pressure versus concentration isotherms.

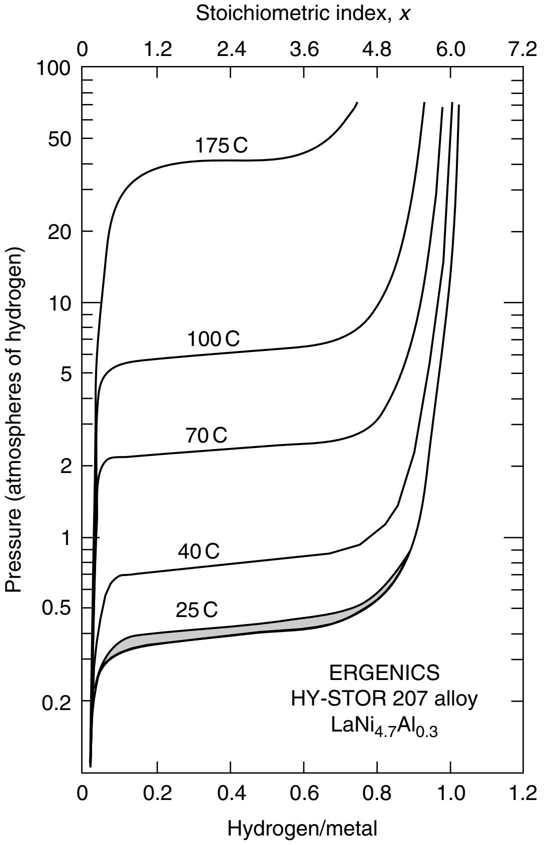


Figure 11.12 Characteristics of the $LaNi_{4.7}Al_{0.3}$ system.

11.4.5.2 Thermodynamics of Hydride Systems

By scaling the data of Figure 11.5 (using more isotherms than shown), one can obtain an empirical relationship between the plateau pressure, p (expressed in atmospheres), and the temperature, T:

$$\ln p = 12.7 - 3360\frac{1}{T}. \tag{11.16}$$

This relationship can be written in the form of Boltzmann's equation:

$$p = p_0 \exp\left(-\frac{W_a}{RT}\right) \tag{11.17}$$

or

$$p = 328 \times 10^3 \exp\left(-\frac{28 \times 10^6}{RT}\right). \tag{11.18}$$

$R = 8314\,\mathrm{JK^{-1}kmoles^{-1}}$ is the gas constant. The enthalpy change of hydriding, ΔH, in this case is equal to $3360 \times 8314 = 28\,\mathrm{MJ\,kmole^{-1}}$, a value close to the enthalpy change of the sorption reaction determined calorimetrically at STP.

Another way of interpreting the empirical relation of Equation 11.16 is to observe that the change in free energy in a gas compressed isothermally is

$$\Delta G = RT \ln r \ \mathrm{J/kmole}, \tag{11.19}$$

where $r = p/p_0$ is the compression ratio. If the pressures are expressed in atmospheres and the reference pressure, p_0, is one atmosphere, then

$$\Delta G = RT \ln p. \tag{11.20}$$

But

$$\Delta G = \Delta H - T\Delta S = RT \ln p. \tag{11.21}$$

$$\ln p = -\frac{\Delta S}{R} + \frac{\Delta H}{RT}. \tag{11.22}$$

Comparing Equations 11.16 and 11.22,

$$\Delta H = -3360 \times 8314 = -28\,\mathrm{MJ\,kmole^{-1}}, \tag{11.23}$$

$$\Delta S = -12.7 \times 8314 = -106\,\mathrm{kJ\ K^{-1}kmole^{-1}}. \tag{11.24}$$

The negative sign corresponds to the case of absorption: the enthalpy change is negative because absorptions are exothermic. The entropy change is negative because hydrogen is in a more ordered state in the hydride than in the gas. If the order in the hydride were perfect, the entropy change would be $-130\,\mathrm{kJ\,K^{-1}\ kmole^{-1}}$, the negative of the entropy of the gas at STP. It can be seen that hydrogen is nearly perfectly ordered in the hydride. Indeed, all metallic hydrides have roughly the same ΔS for the reaction, as can be seen from Table 11.5.

Equation 11.22 as well as the data in Table 11.5 would only be valid if ΔH and ΔS were temperature independent, which they are not, and if the plateau slope were zero (plateau pressure independent of stoichiometry) and in absence of hysteresis (i.e., if the plateau pressure for absorption were equal to that for desorption). Such conditions do not occur in practice and, consequently, the data in the table are to be used only for a first estimation of the hydride performance.

Deviations from the ideal behavior described in the preceding paragraph are illustrated in Table 11.6.

At 1 atmosphere, $p/p_0 = 1$, and $\ln p/p_0 = 0$. Consequently, from Equation 11.22

$$T = \Delta H/\Delta S. \tag{11.25}$$

But all hydrides have roughly the same ΔS of about $100\,\mathrm{kJ\,K^{-1}\,kmole^{-1}}$, so that

$$T \approx 10^{-5}\Delta H, \tag{11.26}$$

or, expressing the enthalpy change in MJ,

$$T \approx 10\Delta H. \tag{11.27}$$

This formula yields the temperature at which a given hydride, in its plateau region, is in equilibrium with 1 atmosphere of hydrogen. It is an indication of the stability of the material. Yttrium and cerium hydrides are the most stable of all. They have to be heated to some 1400 K for their plateau pressure to reach 1 atmosphere.

Plateau pressures can be tailored by suitably doping certain alloys. $LaNi_4Al$ has a plateau pressure of 0.002 atmosphere at 298 K, while at the same temperature, $GdNi_5$ has a plateau pressure of 150 atmospheres. It is also possible to "design" materials with fairly flat plateaus (useful in compressors and heat pumps). Mischmetall is inexpensive but results in markedly sloping plateaus.

11.5 Hydride Hydrogen Compressors

Metallic hydride systems lend themselves to the construction of hydrogen compressors that have no moving parts (other than some valves). Since such compressors also store a substantial amount of gas, they may be of particular interest in processes in which it is necessary to both compress and store the gas. This is, for instance, the case of electrolytic ammonia plants operating with intermittent electricity supplies or of "hydrogen gas stations."

A hydrogen compressor may consist of a hydride-containing vessel equipped with a heat exchanger through which either a heating or a cooling fluid circulates. See Figure 11.13. Two external valves permit the control of inflow and outflow of the gas. Heating up the saturated hydride causes the pressure to rise. Moderate changes in temperature result in substantial increases in pressure owing to the exponential relationship between these variables.

When the compressor delivers hydrogen, it needs additional heat to supply the desorption energy and maintain the pressure constant. Having exhausted the hydride, the system pumps cooling fluid into the heat exchanger so as to permit the recharging at a low pressure.

We will use an example to illustrate the operation of a hydrogen compressor. Consider the $LaNi_5$ system of Figure 11.9. The compressor works by cycling the temperature from 40 C to 140 C and back. Arbitrarily, we start at a moment when the alloy is saturated and at 40 C (Point B, of the

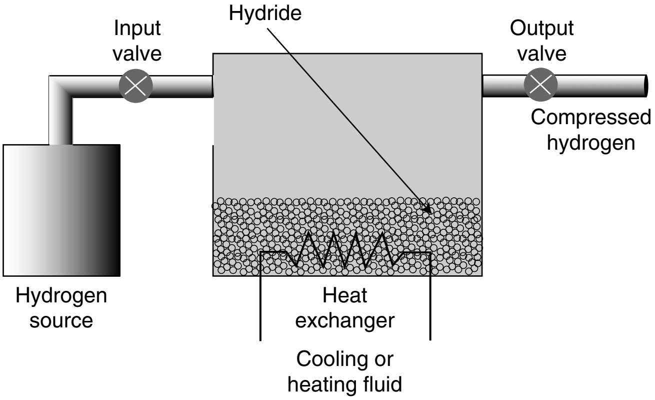


Figure 11.13 Schematic diagram of a hydrogen compressor.

graph), when the stoichiometric coefficient is 5.64; that is, the empirical formula of the alloy is $\text{LaNi}_5\text{H}_{5.64}$. The enthalpy change owing to hydrogen absorption is, in this case, $\Delta f = -30.2$ MJ K^{-1} kmole^{-1}.[†] The value of ΔS is -107.7 kJ K^{-1} kmole^{-1}; hence, the plateau pressure at 40 C (313 K) is approximately

$$P_{plateau} = \exp\left(-\frac{\Delta S}{R} + \frac{\Delta H}{RT}\right) = \exp\left(-\frac{-107.7}{8314} + \frac{-28 \times 10^6}{8314 \times 313}\right)$$
$$= \exp(1.35) = 3.8 \quad \text{atmos.} \tag{11.28}$$

The free hydrogen gas in the "dead space" is in equilibrium with the gas in the alloy; that is, it is at 3.8 atmospheres.

The cycle starts by heating the hydride to 140 C (point C). An amount of heat, $\Delta H_{B\rightarrow C}$ must be added. The plateau pressure rises to some 6.5 MPa (64 atmospheres), and the free hydrogen pressure, still in equilibrium with the alloy, also goes to 6.3 MPa—there is now a larger amount of free gas. Some hydrogen desorption occurs: heating the system causes a (small) reduction in the value of x, which, as scaled from the figure, is now 4.69 kmoles. The amount desorbed was 0.95 kmole of H or 0.48 kmole of H_2.

As high-pressure hydrogen is withdrawn from the compressor (by opening the output valve), the heat of desorption of the gas withdrawn from the vessel would cause the system to cool down. To keep the pressure constant, the temperature must not change. An external source must supply a compensating amount of heat. The pressure will remain constant until the low end of the plateau, point D, is reached. The total heat supplied for desorption is $\Delta H_{C\rightarrow D}$. At point D, $x = 0, 80$.

[†] Table 11.4 lists the reaction enthalpy as -31 MJ K^{-1} kmole^{-1}. Small discrepancies of this kind are to be expected because the alloy characteristics are quite sensitive to impurities, and so exact values measured for ΔH vary somewhat from sample to sample.

Withdrawing hydrogen beyond point D results in a drop of pressure even if the temperature is still held at 140 C. This triggers the control unit that closes the output valve and changes the flow through the heat exchanger from the heating fluid to the cooling fluid, bringing the temperature down to the initial 40 C. An amount of heat, $\Delta H_{D\rightarrow A}$, is removed from the system. The pressure falls to something like 0.38 MPa, and x rises to 1.75.

The input valve is now opened, admitting hydrogen from the source (the source pressure must be somewhat higher than 0.38 MPa). Hydrogen is absorbed by the alloy and would raise the temperature of the system. However, the coolant removes an amount of heat, $\Delta H_{A\rightarrow B}$ keeping the temperature constant until point B is reached, completing the cycle.

To answer the question of how efficient a hydride hydrogen compressor is, we must first agree on what is the useful output of such a machine. Hydrogen enters the compressor at a some (relatively) low pressure, p_{in}, and, presumably, at room temperature, T_{in}, and leaves the compressor at a much higher pressure, p_{out}, and at a somewhat higher temperature, T_{out}. So, the work done on the gas involves raising both its pressure and its temperature. However, almost invariably, there is no interest in warm hydrogen—it probably will cool down anyway before being used; what is of interest is the pressure increase, that is, the *isothermal* compression of the gas:

$$W = RT_A \ln \frac{p_{out}}{p_{in}} \text{ J/kmole.} \tag{11.29}$$

This is, again, a rough approximation because, if the hydrogen cools, its pressure will fall below the pressure, p_{out}.

The energy input is the sum of $\Delta H_{B\rightarrow C} + \Delta H_{C\rightarrow D}$. Notice that the heat rejected in the phases D→A→B is swept away by the coolant and cannot be recovered. As a consequence, the efficiency is

$$\eta = \frac{RT_A \ \ln(p_C/p_A)}{\Delta H_{B\rightarrow C} + \Delta H_{C\rightarrow D}}. \tag{11.30}$$

The heat input, $\Delta H_{C\rightarrow D}$, is, to first order, the heat of desorption of the gas which, in this example, is 30.2 MJ/kmole. In practice one must consider that the actual input to the heat exchanger must be somewhat higher because there is a necessary temperature difference between the exchanger and the alloy; otherwise, heat would not flow into the alloy. $\Delta H_{B\rightarrow C}$ has numerous components, as outlined in the accompanying example.

Example

What is the efficiency of a $LaNi_5$ system compressing hydrogen from 0.3 to 5 MPa. with a temperature change from 40 to 140 C. The work

(Continues)

(*Continued*)

done is

$$W = 8314 \times (273 + 40) \ln \frac{5}{0.3} = 7.3 \quad \mathrm{MJ\,kmole}^{-1}. \tag{11.31}$$

The enthalpy change for the desorption reaction is 30.2 MJ/kilomole. The (very optimistic) upper limit for the efficiency is found by neglecting $\Delta H_{B\to C}$. It is

$$\eta = 7.3/30.2 = 0.24 \quad \text{or} \quad 24\%. \tag{11.32}$$

The practical efficiency is considerably lower than the one estimated above. The factors that contribute to a reduced performance include:

1. Ignoring the heat, $\Delta H_{B\to C}$, necessary to raise the temperature of the system from 40 to 140 C. The components of $\Delta H_{B\to C}$ are

   1.1 The heat of desorption of some hydrogen. It is proportional to the heat of formation of the hydride and the fraction of the total hydrogen that was desorbed. This latter amount depends, of course, on the pressure step and on the relative magnitude of the "dead volume" inside the compressor vessel, and this, in turn, depends on the grain size and the packing of the alloy. Typically, the "dead volume" amounts to some 60% of the volume of the vessel. In Figure 11.9 the stoichiometric index varies as shown in Table 11.7.

   Notice that the difference between the values of the stoichiometric index, x, at points C and D is the same as that between points B and A because the amount of hydrogen desorbed must equal that absorbed.

   The number of kilomoles of hydrogen pumped per cycle per kilomole of alloy is $\frac{1}{2}(x_C - x_D)$, the factor $\frac{1}{2}$ corresponding to the diatomic nature of the hydrogen molecule. This amounts to 1.95 kilomoles.

Table 11.7 Stoichiometric Indices throughout the Compressor Cycle

Point	Stoichiometric index
A	1.75
B	5.64
C	4.69
D	0.80

(*Continues*)

(*Continued*)

The number of kilomoles of hydrogen desorbed when the temperature is raised is $\frac{1}{2}(x_B - x_C)$, or 0.48 kilomole. Thus, one must add $(0.48/1.95) \times 30.2 = 7.4$ MJ of heat to pump 1 kilomole of hydrogen. This brings the efficiency down to $7.3/(30.2 + 7.4)$ or 19.4%.

1.2 The heat capacity of the hydride whose temperature must also be raised from 40 to 140 C. The heat capacity of many metals and alloys is roughly 30 kJ/K per kilomole. That of the hydride is somewhat higher. For each kilomole of hydrogen pumped, about 0.5 kilomole of alloy are needed. The heat required to elevate its temperature is then $0.5 \times 30 \times (140 - 40) = 1.5$ MJ per kilomole of H_2. The efficiency is now $7.3/(30 + 7.4 + 1.5)$ or 18.8%.

1.3 The heat capacity of the vessel itself.

2. Ignoring the heat losses of the system to the environment. The equipment must be well insulated to reduce such losses. Increasing the size of the vessel—that is, the system capacity—increases the volume-to-surface ratio and reduces the relative heat losses. Larger compressors operate with proportionally smaller heat losses.
3. Failing to take into account the need to have the temperature of the heating fluid higher than that of the hydride. There must be a temperature difference across the walls of the heat exchanger. Thus, the temperature swing of the hydride is smaller than that of the heat exchanger fluid.
4. Disregarding hysteresis. As pointed out previously, the absorption plateau pressure tends to exceed by an amount, Δp, the desorption pressure. The two 20 C isotherms in Figure 11.9 illustrate this phenomenon. The area defined by the absorption–desorption contour represents wasted energy. Since Δp is nearly independent of pressure, the hysteresis is, proportionally, more important at lower temperatures where the pressures are smaller.

Laboratory-built $LaNi_5$ compressors exhibit efficiencies of some 6%, showing that the contribution of the second-order loss mechanisms is significant. Compared with mechanical hydrogen compressors (typical efficiency of 60%), hydride compressors operate with modest efficiencies. However, they use low-grade heat for driving energy, heat frequently available from some cooling process in another part of the plant.

$LaNi_5$ has been suggested as the working alloy owing to its large absorption capacity, its fast reaction rate, the levelness of its plateaus, and the favorable pressure ranges.

Three units of the type we have described can be combined so that while one is delivering the gas, the others are at different phases of the recharging process. This insures a steady delivery of hydrogen. Typically, an individual cycle will last 10 or more minutes.

11.6 Hydride Heat Pumps

A hydride absorbs heat when it releases hydrogen. This suggests its use as a heat pump.

Consider the thermally driven heat pump shown schematically in Figure 11.14. We want to pump an amount of heat, Q_C, from a cold source at a temperature, T_C, into an environment at a higher temperature, T_R.

Since heat will not flow spontaneously from a colder reservoir to a hotter one, the pumping requires the expenditure of a certain amount of additional (heat) energy, Q_H, from a hot source. The total amount of heat dumped into the warm environment is

$$Q_R = Q_C + Q_H. \tag{11.33}$$

The cold source may be the outside of a house in winter, while the hot source may be from collected sunlight, burning fuel, or an electric heater. Q_R may be used to warm the house.

In the reversible case

$$\frac{Q_R}{T_R} = \frac{Q_C}{T_C} + \frac{Q_H}{T_H}, \tag{11.34}$$

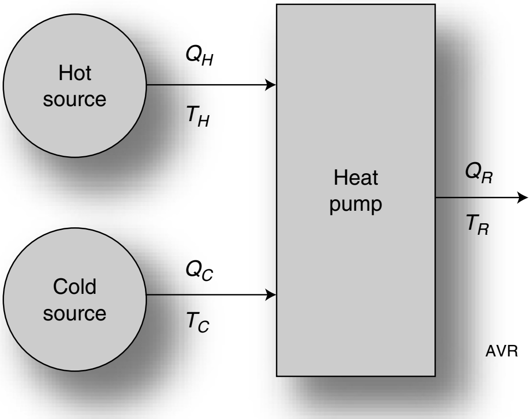


Figure 11.14 A thermally driven heat pump.

from which the Carnot efficiency of the process can be derived:

$$\eta_{CARNOT} = \frac{Q_R}{Q_H} = \frac{1 - T_C/T_H}{1 - T_C/T_R}. \tag{11.35}$$

With an outside temperature of 0 C (273 K), a house temperature of 25 C (298 K), and a hot source at 100 C (373 K), the Carnot efficiency of the heat pump is

$$\eta_{CARNOT} = \frac{1 - 273/373}{1 - 273/298} = 3.2. \tag{11.36}$$

To avoid talking about efficiencies larger than unity, the value in equation 11.36 is called the **coefficient of performance** (COP). A COP of 3.2 means that for each joule of input to the heat pump from the hot source, 3.2 joules of heat are delivered to the house.

A heat pump can be built by using two hydrides with matching properties. Consider two containers, one of which (Container B) is inside the house and at a temperature T_R—the room temperature. The type of alloy is such that at the system pressure of the moment, p_0, it is fully saturated with hydrogen. Outside the house, there is another container (Container A) interconnected with B so that both experience the same pressure, p_0. The material in A is chosen so that, although the temperature is lower (it is the outside temperature, T_C), the alloy is completely depleted.

As an example, let

Room temperature, T_R	300 K,
Outside temperature, T_C	270 K,
Hot source temperature, T_H	400 K,
Entropy change (absorption)(Alloy A), ΔS	-110 kJ K^{-1} per kilomole of H_2,
Entropy change (absorption) (Alloy B), ΔS	-110 kJ K^{-1} per kilomole of H_2,
Enthalpy change (absorption) (Alloy A), ΔH	-26 MJ per kilomole of H_2,
Enthalpy change (absorption) (Alloy B), ΔH	-30 MJ per kilomole of H_2.

By applying Equation 11.22, one calculates for this initial step the plateau pressures for Hydride A at 270 K as 5.2 atmospheres and for Hydride B at 300 K as 3.3 atmospheres.

Assume that enough hydrogen exists in the system to make its pressure $p_0 = 4$ atmospheres. This pressure is higher than the plateau pressure of hydride B. Thus, the hydride is saturated. However, 4 atmospheres is less than the plateau pressure of Hydride A, which consequently must be depleted (it is in its alpha-phase).

The first step in the operation consists of moving Container A inside (its temperature rises to T_R, and it continues depleted).

An amount of heat, ΔH_B, is then delivered by a heat source to Container B, raising its temperature to T_H and causing the desorption of the hydrogen. The pressure is high enough to force the absorption of hydrogen by the alloy in Container A, which sheds into the room ΔH_A units of heat for each kilomole of gas absorbed.

In our example, when the temperature of Alloy B is raised to 400 K, its plateau pressure is raised to about 67 atmospheres, while that of Alloy A (now at 300 K) is 16 atmospheres. Of course, the hydrogen pressure in the system does not reach the 67 atmospheres; as Alloy B is heated (taking in 30 MJ per kilomole from the hot source), it releases some gas, and the system pressure starts going up. When it reaches a value slightly above 16 atmospheres, the gas is absorbed by Alloy A and the pressure remains constant unless the alloy saturates. This absorption causes the shedding of 26 MJ of heat energy into the room for each kilomole absorbed.

The second step consists of removing the heat source from Container B and moving Container A outside again. The temperature in B falls to T_R and that in A, to T_C. But these are exactly the initial temperatures that led to B being saturated and A being depleted. The saturation of B causes this container to shed ΔH_B units of heat into the room, while the depletion of A causes it to absorb ΔH_A units of heat from the cold outside.

In this step, Alloy B has its plateau pressure lowered back to 3.3 atmospheres, while the system pressure is (initially) still 16 atmospheres. Hydrogen is absorbed, and the absorption heat of 30 MJ/kilomole is returned to the room.

The cycle has been completed. The room received the heat ΔH_A from Container A during Step 1 and ΔH_B from Container B in Step 2. This latter amount of heat came from the hot source, but ΔH_A came from the cold outside.

The coefficient of performance is

$$\mathrm{COP} = \frac{\Delta H_B + \Delta H_A}{\Delta H_B} = 1 + \frac{\Delta H_A}{\Delta H_B}. \tag{11.37}$$

The room received 26 MJ during Step 1 from Alloy A and 30 MJ during Step 2 from Alloy B, a total of 56 MJ all at a cost of 30 MJ from the hot source. The coefficient of performance is 56/30 = 1.86.

Clearly, there is no necessity of moving containers physically from the outside to the inside of the house. An equivalent effect can be achieved by circulating inside and outside air at the appropriate moments during the cycle.

It is obvious that a similar system can be used as a refrigerator.

References

Although numerous works have been published on hydrides and their applications, there is one, fairly early paper, that is widely regarded as the “Bible” of this topic. It is van Mal’s Ph.D. dissertation (notice the lat inized form of the Dutch name, a charming medieval practice still observed in Europe[†]):

Van Mal, Harmannus Hinderikus, Stability of ternary hydrides and some applications, Ph.D. dissertation, Technische Hogeschool Delft (Netherlands), May **1976**.

Amankwah, K.A.G., et al. Hydrogen storage on superactivated carbon at refrigeration temperatures, *International Journal of Hydrogen Energy* 14(9), pp. 437–447, **1989**.

Chambers, A. C. Park, and R.T.K. Baker, *J. Phys. Chem. B* **102**, p. 4253, **1998**.

Chen, P., X. Wu, J. Lin, and K.L. Tan, *Science* **285**, p. 91, July 2, **1999**.

Houston, E. L., and G. D. Sandrock, Engineering properties of metal hydrides, *J. Less-Common Metals* **74**, pp. 435–443, **1980**.

Loges, Björn, Albert Boddien, Henrik Junge, and Matthias Beller, Controlled generation of hydrogen from formic acid amine adducts at room temperature and application in H2/O2 fuel cells, *Angew. Chem. Int. Ed.* **47**, pp. 3062–3965, **2008**.

Williams, Robert H., The clean machine, *Technology Review*, April **1994**.

[†] Remember that the famous Mercator projection so frequently used in map making is due to Gerhard Kremer, the Flemish mathematician. Compare Kremer (Dutch), Krämer (German = shopkeeper), and Mercator (Latin). Other famous latinizations include Catharina Elisabeth Textor, Johann Wolfgang von Goethe's mother, who improved her name from Weber to Textor and Joachim Neumann who became Joachim Neander, after whom Neanderthal is named. Neander = néos + andro (anēr) = neu + mann = new + man.

PROBLEMS

11.1 An alloy, Mg_2Ni, interacts with hydrogen so that the following reversible reaction takes place:

$$Mg_2Ni + 2H_2 \leftrightarrow Mg_2NiH_4.$$

The thermodynamic data for the absorption reaction are (per kilomole of H_2):

$$\Delta H f^0 = -64.4\,\text{MJ},$$

$$\Delta S f^0 = -122.3\,\text{kJ/K}.$$

The relevant atomic masses (in daltons) are

H	1
Mg	24.3
Ni	58.7

In the following questions, all pressures are in the plateau region.

1. What is the dissociation pressure at 300 K?
2. At what temperature is the equilibrium pressure 1 MPa?
3. What are the proportions, by mass, of the elements in Mg_2Ni?
4. A vessel with an internal volume of 1000 cm^3 contains 1.56 kg of Mg_2NiH_4. The density of this material is 2600 kg/m^3. The saturated hydride is placed in the evacuated vessel. The temperature is raised to 400 C, but no hydrogen is allowed to escape. How many kg of hydrogen are desorbed? How many remain in the hydride?
5. How much heat must be added to desorb the rest of the hydrogen? Assume that desorption stops when the empirical formula of the material reaches the $Mg_2NiH_{0.4}$ composition. The hydrogen desorbed is removed from the vessel so that the pressure remains constant.
6. If a hydrogen compressor were built using the vessel above, and if the hydrogen were fed into the system at 10^5 Pa and 85 C, what would the theoretical efficiency of the compressor be? The output gas is at 85 C and 5 MPa. Neglect all losses.

11.2 Figure 11.8 shows the pressure-composition isotherms for the reaction of hydrogen with $LaNi_5$. Assume that, in the plateau region, the isotherms are horizontal—that is, that the pressure does not depend on the composition. Assume further that the pressure at 40 C is 0.3 MPa and at 140 C is 5 MPa.

One kg of $LaNi_5H_5$ is placed inside a pressure vessel and heated to 140 C. Hydrogen is slowly withdrawn until the hydride is left with the empirical formula $LaNi_5H_2$.

Molecular masses: La, 138.9; Ni, 58.7.

The ΔH for absorption for this alloy is -32.6 MJ per kilomole of H_2, and the ΔS is -107.7 kilojoules per kelvin per kilomole.

1. To maintain the temperature at 140 C during the desorption, must heat be supplied to or withdrawn from the hydride?
2. How many joules of heat must be exchanged with the hydride to perform the desorption?
3. What is the change in entropy of the system during desorption? Assume that ΔS is temperature independent. Does the entropy increase or decrease during the desorption?
4. How many kilograms of hydrogen were withdrawn?

11.3 Hydrogen is to be transported. Two solutions must be considered:

1. Liquefy it.
2. Convert it to ammonia. Later, when hydrogen is to be used, the ammonia can be catalytically cracked with 85% recovery of hydrogen; 46.2 MJ per kg of ammonia are required to dissociate the gas.

Purely from the energy point of view, which is the more economical solution?

11.4 Calorimeter measurements show that when a compound, AB, reacts with H_2 forming a hydride ABH, 18.7 MJ of heat are released for each kilomole of H_2 absorbed. The manufacturer of this hydride consults you about the advisability of shipping it (saturated with hydrogen) inside a normal gas pressure cylinder (which is rated at 200 atmospheres). During shipment the cylinder may be exposed to the sun. Without having any additional data on the hydride, what would you advise the manufacturer? Could you tell him the maximum temperature that the hydride can safely reach?

11.5

1. Estimate the enthalpy change for the hydrogen absorption reaction of an alloy that has a plateau pressure of 1 atmosphere when $T = 0$ C.
2. Estimate the enthalpy change for the hydrogen absorption reaction of an alloy that has a plateau pressure of 1 atmosphere when $T = 30$ C. Assume that Alloy A is in an environment that causes its temperature to be 0 C (perhaps the outside of a house) and that Alloy B is at 28 C (say, inside a room). The alloy containers are interconnected by means of a pipe so that the hydrogen pressure is the same in both. The amount of hydrogen in the system is such that, in any phase of the cycle, when one alloy is saturated, the other is depleted.

3. What is the system pressure?
4. Which alloy is depleted, and which contains most of the hydrogen?
5. Raise the temperature of Alloy A to 28 C.

What is the plateau pressure? What is the actual hydrogen pressure? Circulate hot water (from a hot source) through the heat exchanger of Alloy B and heat it up to 100 C (373 K). Maintain this temperature.

6. What is the pressure of the system, and what happens to the two alloys? Return the temperature of Alloy B to 28 C, and that of Alloy A to 0 C (by putting it in contact with the exterior).
7. What happens?
8. Tabulate all the sorption heat inputs and outputs of the system and the corresponding temperatures. Ignore all other heat inputs and outputs. For example, ignore the heat necessary to change the temperature of the system.
9. How much heat is delivered (per kmole of hydrogen) to the environment at 28 C in a complete cycle as described above?
10. Of this amount of heat, how much must come from the hot water source (and must be paid for)?
11. What is the coefficient of performance of this heat pump—that is, what is the ratio of the heat delivered to the environment at 28 C to the heat required from the hot water?

11.6 A recently developed binary compound is to be used as a hydrogen storage medium because it forms a reversible (ternary) monohydride.

The data of the system include:
Molecular mass of the compound (no hydrogen): 88 daltons.
Density of the compound (hydrided or depleted): 8900 kg m^{-3}.
Enthalpy of hydriding: -26.9 MJ per kilomole of H_2.
Change of entropy owing to absorption of hydrogen: -100 kJ K^{-1} per kilomole of H_2.
Type of isotherm: single plateau.
Heat capacity of the compound: 200 J K^{-1} kg^{-1}.
Heat capacity of the container and of the hydrogen gas: negligible.
Heat capacity of water: 4180 J K^{-1} kg^{-1}.
Density of hydrogen: 0.089 kg m^{-3}.

2.5 kg of the compound (activated, in a fine powder, at 0 C) are placed inside a container measuring internally 10 by 10 by 10 cm.

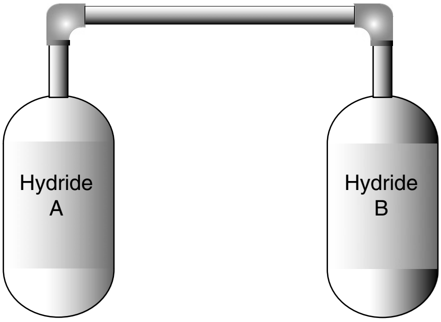
We want to adjust the system so that the alloy is saturated and the gas pressure is exactly the plateau pressure. To find this point, we will have to observe the behavior of the pressure in the alloy canister as hydrogen is added. The pressure initially rises; then as the plateau is reached, it stabilizes until saturation is achieved. We will first overshoot this point; then, having removed the hydrogen source, we will purge some of the gas, observing the pressure and stopping the purge just when the pressure falls to the previously observed plateau pressure. During this whole operation, the temperature is carefully maintained at 0 C.

A 5-atmosphere hydrogen source is used to fill the container. The gas is applied long enough to allow equilibrium to be established. Any heat absorbed or released by this operation is removed or added to the system so that at the end of the operation, the pressure is still 5 atmospheres and the temperature is still 0 C. Next the hydrogen source is removed. The pressure inside the container remains at 5 atmospheres. A valve is cracked open, and hydrogen leaks out while the pressure is monitored. As soon as the pressure stabilizes, the valve is closed.

1. How many grams of hydrogen were lost in the bleeding?
2. How much hydrogen remains in storage? Express the loss as a percentage of the stored gas.
3. What is the pressure of the stored gas?

The container is now immersed in a tank of water at 40 C. This tank contains 0.3 liter of water and is entirely adiabatic. Its walls have negligible heat storage capacity.

4. What is the temperature of the system (water tank plus alloy container) after equilibrium is reached?



11.7 An inventor proposes the following device to cool drinks at a picnic. It consists of two sturdy containers (something like small portable oxygen bottles) one of which (Container A) can be placed inside a styrofoam box in which there are 12 beer cans. The other (Container B) is outside the box and can be placed over a fire. A pipe interconnects the two containers (see the figure). The containers are filled with different alloys capable of absorbing hydrogen. Alloy A is TiFe. Alloy B has to be described.

When both containers are at the same temperature, 298 K, Alloy A is depleted and B is saturated.

1. For this to happen, what are the required characteristics of Alloy B? Establish a relationship between the thermodynamic parameters of the two alloys.
2. Which of the alloys from the following table can be used as Alloy B in the device under consideration?

Hydride	Δ H MJ/kmole	ΔS kJ/K per kmole
AB	−21.0	−96.5
CD	−26.1	−99.4
EF	−27.9	−106.8
GH	−32.1	−101.8
IJ	−32.6	−110.5
KL	−33.4	−98.3

To operate the system, Container A is placed in a bucket of water and kept at 25 C. This requires refreshing the water occasionally. Container B is placed over the picnic fire so that hydrogen is transferred to Container A whose hydride becomes saturated.

Next, Container A is placed inside the styrofoam box, in contact with the 355-ml beer cans which, for good thermal contact, are immersed in 4.5 liters of water.

Container B is now cooled to 298 K, returning the system to its original state.

The heat required to desorb the hydrogen from A will cool the 12 cans of beer from 25 C to 10 C.

Assume that beer behaves as water, at least as far its thermal capacity is concerned. The styrofoam box is essentially adiabatic. Assume that during the cycle the composition of the hydride in A varies from $TiFeH_{0.95}$ to $TiFeH_{0.4}$

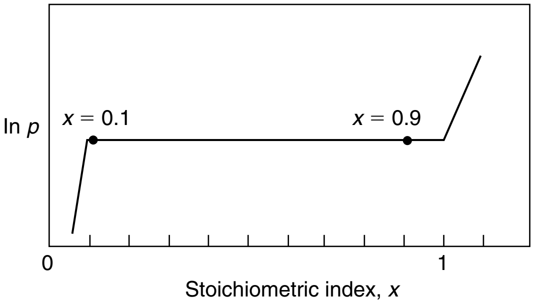
The atomic mass of Ti is 47.9 daltons and that of Fe is 55.8 daltons.

3. Estimate the minimum mass of TiFe required.

11.8 Two canisters are interconnected by a pipe. Canister A contains TiFe and is at 300 K, while Canister B contains $CaNi_5$ and is at 350 K. The system is filled with hydrogen at a pressure of 4 atmospheres.

In which canister is the bulk of the hydrogen? No guesses, please! Use the thermodynamic data of Table 11.4.

11.9 *WARNING: This problem contains units, such as grams, centimeters, and so on, that are not of the SI.*



Consider a vessel containing a hydrogen-absorbing alloy, AB.

Vessel:

Volume: 200 cm^3
Thermal insulation: adiabatic
Heat capacity: negligible

Alloy:

Formula mass: 120 daltons
Amount: 200 g
Heat capacity: 1700 J/K per kg of alloy (same for hydrided alloy).

Hydride:

Heat of formation of ABH (absorption): −30.0 MJ per kmole of H_2
Entropy change owing to absorption: −110 kJ/K per kmole of H_2
Density (of $ABH_{0.9}$): 1600 kg/m^3.

System:

Initial temperature: 300 K
Initial hydrogen pressure: equilibrium.

Hydrogen is forced into the vessel until the average alloy composition is $ABH_{0.9}$ What is the minimum pressure required to force all the needed hydrogen into the vessel? How much hydrogen is forced in?

11.10 A perfectly adiabatic (heat-insulated) vessel has an internal volume of 100 cm^3 and contains 240 g of an alloy powder, AB, that forms a monohydride, ABH. The thermodynamic data for absorption are:

$$\Delta H = -28 \text{ MJ per kilomole of } H_2$$
$$\Delta S = -100 \text{ kJ/K per kilomole of } H_2$$
$$\text{Heat capacity, } c_v = 400 \text{ J/K per kg of alloy}$$

Additional data include

$$\text{Formula mass of the alloy} = 150 \text{ daltons}$$
$$\text{Density of the alloy} = 8000\,\text{kg/m}^3$$
$$\text{Bulk density of the alloy powder} = 4000\,\text{kg/m}^3$$

The vessel has been charged with hydrogen so that the pressure is 10 atmospheres. The system is at 30 C.

1. Hydrogen is withdrawn. How many milligrams of the gas can be removed without causing a change in the temperature of the hydride? Remember that the container is adiabatic. The heat owing to the work that the withdrawn hydrogen may exert is exchanged with the gas outside the hydride container and does not influence the temperature of the container.
2. If more hydrogen is released, it will cause the cooling of the hydride. Assume that the vessel has no heat capacity. How much hydrogen is released if the pressure falls to 1 atmosphere?
3. What is the value of x in the empirical formula ABH_x after the above desorption?

11.11 The Pons and Fleishman cold fusion experiment employs an electrolytic cell consisting of a palladium negative electrode and a platinum positive electrode. The electrolyte is a concentrated solution of LiOH in D_2O. The palladium electrode is a cylindrical rod 10 cm long and 1.2 cm in diameter. Just prior to the experiment, the rod is completely degassed by heating it up in a vacuum.

When a 0.5. A current is forced through the cell, nothing unusual happens for a long time. To be sure, normal electrolysis occurs with D_2 evolving at the palladium and 0_2 at the oxygen electrodes. The D_2O used up is continually replenished.

In some rare instances, it is claimed that, after the electrolysis has proceeded for a long time, heat suddenly begins to be produced in substantial amounts—73 W, in this case. This heat production rate is sustained for 120 hours, after which the cell is disconnected.

If you look up the enthalpies of formation of all palladium compounds, you will find that the largest value is associated with the formation of palladium hydroxide: 706 MJ/kmole.

You will also find that the atomic mass of palladium is 106 daltons and that the density of the metal is 12 g cm^{-3}.

Can you prove that the energy generated is not chemical in nature?

To explain the delay, assume that D_2-D_2 fusion occurs at a rapid rate only if the deuterium is packed with sufficient density and that this will happen only when the palladium is completely saturated with deuterium and the formation of palladium di-deuteride begins. How long would you expect the cell to operate before it heats up?

11.12 A canister contains a mixture of two alloys (Alloy 1 and Alloy 2). A hydrogen source equipped with a valve is connected to this canister. A measured amount of the gas can be delivered to it.

Describe the behavior of p_{system} (the hydrogen pressure, in pascals, read by a manometer connected to the canister) versus the amount, μ_{H_2} of H_2 (in moles) introduced into the system. Sketch a rough p_{system} versus μ_{H_2} graph.

Estimate all the break points in the above sketch—that is, all the values of μ_{H_2} at which the curve changes abruptly its character.

Do the above for $T = 400\,\text{K}$, following the detailed instructions in Items 1 through 5.

Here are some data:

Internal volume of the canister: 1 liter.

The idealized p versus x characteristics of the alloys are as follows:

Region 1: The equilibrium pressure is proportional to the stoichiometric coefficient, x, for $0 < p < p_{plateau}$. When $x = x_{crit_\ell}$, p reaches $p_{plateau}$.
Region 2: The plateau pressure is perfectly constant until x reaches a critical value, x_{crit_u}.
Region 3: For $x > x_{crit_u}$, the pressure rises linearly with x with the same slope as that of Region 1.

Alloy 1: 0.8 kg of alloy AB having a density of $2750\,\text{kg/m}^3$, $\Delta S = -110\,\text{kJ K}^{-1}\text{kmole}^{-1}$, $\Delta H = -35\,\text{MJ kmole}^{-1}$.

At 400 K, $x_{crit_{\ell 1}}$ (the minimum value of x in the plateau region) is 0.3, and $x_{crit_{u1}}$ (the maximum value of x in the plateau region) is 3.55.

Alloy 2: 1.0 kg of alloy CD having a density of $2750\,\text{kg/m}^3$, $\Delta S = -90\,\text{kJ K}^{-1}\text{kmole}^{-1}$, $\Delta H = -28\,\text{MJ kmole}^{-1}$.

At 400 K, $x_{crit_{\ell 2}}$ (the minimum value of x in the plateau region) is 0.3, and $x_{crit_{u2}}$ (the maximum value of x in the plateau region) is 4.85.

The formula masses are

A. 48 daltons.
B. 59 daltons.
C. 139 daltons.
D. 300 daltons.

The density of the above alloys is (unrealistically) the same whether or not hydrided.

Define "gas-space" as the space inside the canister not occupied by the alloys.

1. Calculate the volume, $V_{gas-space}$ of the gas space.
2. What is the number of kilomoles of each alloy contained in the canister?
3. Tabulate the plateau pressures of the two alloys for 300 and 400 K.
4. Describe, in words, the manner in which the system pressure, p_{system}, varies as hydrogen is gradually introduced in the system. Use μ_{H_2} as the measure of the number of moles of H_2 introduced. Sketch a p_{system} vs μ_{H_2} graph.

5. Calculate the values of μ_{H_2} that mark the break points (points of abrupt change in the p_{system} behavior—that is, points where the pressure changes from growing to steady, or vice versa).

11.13 A container able to withstand high pressures has an internal capacity of 0.1 m^3. It contains 490 kg of an alloy, AB, used to store hydrogen.

The properties of this alloy are:

Atomic mass of A	60 daltons
Atomic mass of B	70 daltons
Density of AB	8900 kg/m^3
Heat capacity	1 kJ K^{-1}kg^{-1}
ΔH	−25 MJ/kmole of H_2
ΔS	−105 kJ K^{-1}/kmole of H_2
Depletion end of plateau	x = 0.01
Saturation end of plateau	x = 1

x is the stoichiometric coefficient of hydrogen in ABH_x. The values correspond to 300 K. The plateau pressure is essentially independent of x in the above interval. Above $x = 1$, the equilibrium pressure rises very rapidly with x so that x does not appreciably depend on the hydrogen pressure.

1. What volume inside the container can be occupied by gas?
2. How many kilomoles of alloy are inside the container?
3. What is the plateau pressure of the hydrogen in the alloy at 300 k?
4. Introduce 10 g of H_2. Give an upper bound for the pressure of the gas in the container.
5. Now introduce additional hydrogen so that the total amount introduced (Steps 4 and 5) is 100 g. The temperature is kept at 300 K. What is the pressure of the gas?
6. What is the stoichiometric index, x, of the hydride, ABH_x?
7. Finally, introduce sufficient hydrogen so that the total amounts to 4 kg. The temperature remains at 300 K. What is the pressure of the gas?
8. Assume that the container has negligible heat capacity. Withdraw adiabatically (i.e., without adding any heat to the system) 3 kg of hydrogen. What will be the pressure of the gas inside the container after the 3 kg of hydrogen has been removed?

11.14 Two hydrogen-storing alloys have the following properties:

Hydride	ΔH MJ kmole^{-1}	ΔS kJ K^{-1} kmole^{-1}
A	−28	−100
B	−20	?

What must the value of ΔS_B be to make the plateau pressure of the two hydrides be the same when $T = 400\,\text{K}$?

ΔS_B is, of course, the ΔS of Alloy B.

How do you rate your chances of finding an alloy with the properties of Alloy B? Explain.

11.15 TiFe is sold by Energics, Inc. under the label HY-STOR 101. Pertinent data are found in Tables 11.4 through 11.6. Treat this alloy as "ideal" (no hysteresis).

The atomic mass of iron is 55.8 daltons, and that of titanium is 47.9 daltons. The saturated alloy ($TiFeH_{0.95}$) is at 350 K and is in a perfectly adiabatic container with negligible heat capacity.

A valve is opened, and hydrogen is allowed to leak out until the pressure reaches 2 atmospheres. What is the composition of the hydride at the end of the experiment; that is, what is the value of x in $TiFeH_x$?

To simplify the problem, assume that there is no "gas space" in the container (patently impossible). Also, neglect the heat capacity of the hydrogen gas and any Joule-Thomson heating owing to the escaping gas.

11.16 A canister contains an alloy, AB, that forms a monohydride, ABH. The canister is perfectly heat-insulated—that is, adiabatic—and contains 0.01 kmole of the alloy and a free space of 600 ml, which initially is, totally empty (a vacuum).

The molecular mass of the alloy is 100 daltons, and its thermodynamic characteristics for absorption are $\Delta H = -25\,\text{MJ}$ and $\Delta S = -100\,\text{kJ/K}$, all per kilomole of H_2. For simplicity, make the unrealistic assumption that the plateau extends from $x = 1$ all the way to $x = 0$. x is the stoichiometric coefficient in ABH_x. Assume also that the plateau is perfectly horizontal and that there is no hysteresis. The heat capacity of the alloy is $500\,\text{J kg}^{-1}\text{K}^{-1}$. Again, for simplicity, assume that neither the canister itself nor the hydrogen gas has significant heat capacity. Finally, assume that the volume of the alloy is independent of x.

Introduce 0.002763 kmole of H_2 into the canister. Both canister and hydrogen are at 300 K. What will the gas pressure be inside the canister?

11.17 HELIOS is an electric airplane developed by AeroVironment to serve as a radio relay platform. It is supposed to climb to fairly high altitudes (some 30 km) and to orbit for a prolonged time (months) over a given population center fulfilling the role usually performed by satellites. Although its geographic coverage is much smaller, HELIOS promises to be substantially more economical.

The plane is propelled by 14 electric motors of 1.5 kW each. Power is derived from photovoltaic cells that cover much of the wing surface. In order to stay aloft for many days, the plane must store energy obtained during the day to provide power for nighttime operation. The solution to this problem is to use a water electrolyzer that converts the excess energy, provided during daytime hours by the photovoltaics, into hydrogen and oxygen. The gases are then stored and during darkness feed a fuel cell that provides the power required by the airplane.

Although the specifications of the HELIOS are not known, let us take a stab at providing the outline of a possible energy storage system. To that end, we will have to make a number of assumptions that may depart substantially from the real solution being created by AeroVironment.

1. Calculate the amount of hydrogen and of oxygen that must be stored. Assume that during takeoff and climbing to cruise altitude, the full 1.5 kW per motor is required, but for orbiting at altitude only half the above power is required. Assume also that the power needed for the operation of the plane (other than propulsion, but including the energy for the radio equipment) is 3 kW. Assume also that the longest period of darkness lasts 12 hours. The fuel cell has an efficiency of 80%.
2. The amount of fuel calculated in Item 1 must be stored. Assuming STP conditions, what is the volume required?
3. Clearly, the volumes calculated in Item 2 are too large to fit into HELIOS. The fuel cell busses being operated experimentally in Chicago have hydrogen tanks that operate at 500 atmospheres and that allow a gravimetric concentration of 6.7% when storing hydrogen. If such tanks were adopted for the HELIOS, what would be the mass of the fully charged fuel storage system?
4. Assume that the efficiency of a mechanical hydrogen compressor is 60% and that of an oxygen compressor is 80%. How much energy do these compressors require to compress the gases isothermally to their 500° atmosphere operating temperature. Assume for simplicity that the electrolyzer that produces these gases is pressurized to 5 atmospheres.
5. While orbiting, during daylight, what is the total energy that the photovoltaic collectors have to deliver? The electrolyzer is 80% efficient.

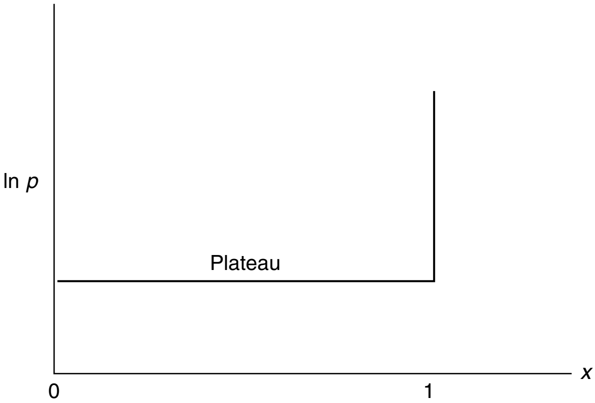
11.18 Two 100-liter canisters are interconnected by a pipe (with negligible internal volume). Canister A contains 37.2 kg of FeTi and Canister B, 37.8 kg of $Fe_{0.8}Ni_{0.2}Ti$.

Although the gas can freely move from one canister to the other, there is negligible heat transfer between them. Thus the gas can be at different temperatures in the two canisters. The gas always assumes the temperature of the alloy it is in contact with.

The pertinent data are summarized in the following the two boxes:

Element	Atomic mass (daltons)	Density (kg/m^{-3})
Ti	47.90	4540
Fe	55.85	7870
Ni	58.71	8900

Alloy	ΔH (MJ kmol^{-1})	ΔS (kJ^{-1} kmol^{-1})
FeTi	−28.0	−106.1
$\mathrm{Fe_{0.8}Ni_{0.2}Ti}$	−41.0	−118.8



To simplify the solution, assume that the $\ln p$ vs. x characteristics of the alloys consist of a perfectly horizontal plateau followed by a vertical line in the saturated region. In other words, the characteristics look as sketched in the accompanying figure. The alloys are depleted when $x = 0$ and are saturated when $x = 1$. Also, neglect all the hydrogen dissolved in the saturated alloy.

Initially, Canister A is at 300 K and Canister B, at 400 K. They have been carefully evacuated (the gas pressure is zero).

Assume that the density of the alloys is the average of that of the component elements.

Enough hydrogen is introduced into the system so that one of the alloys becomes saturated. This will cause the temperature of the alloys to change.

To simplify things, assume that the alloys and the canisters themselves have negligible heat capacity.

1. Does the temperature increase or decrease?
2. The temperature is now adjusted to the values of 300 K and 400 K, as before. How many kg of hydrogen had to be introduced into the system to make the gas pressure, at this stage, 10%

higher than the plateau pressure of the saturated alloy while leaving the other alloy depleted? Please be accurate to the gram.

3. Now raise the temperature of Alloy A to 400 K. Describe what happens:

   3.1 What is the final gas pressure?

   3.2 What is the stoichiometric value of H in each alloy?

   3.3 How many joules of heat had to be added?

   3.4 How many joules of heat had to be removed?

11.19 A canister with 1 m^3 capacity contains 3000 kg LaNi_5. Although initially the system was evacuated, an amount, μ_1, of hydrogen is introduced so that the pressure (when the canister and the alloy are at 298 K) is exactly 2 atmospheres. Assume that once the alloy is saturated, it cannot dissolve any more hydrogen; that is, assume that the p-x characteristic just after the beta-phase is vertical.

The density of lanthanum (atomic mass 138.90 daltons) is 6145 kg/m^3 and that of nickel (atomic mass 58.71 daltons) is 8902 kg/m^3. Assume that the density of the alloy is equal to that of a mixture of 1 kilomole of lanthanum with 5 kilomoles of nickel.

1. What is the value of μ_1?

2. 100 MJ of heat are introduced into the canister whose walls are adiabatic and have no heat capacity. Ignore also the heat capacity of the free hydrogen gas.

   Calculate the pressure of the free hydrogen gas.

11.20 Please use $R = 8314$.

Here is the setup:

A hydrogen source is connected to a canister containing an alloy, A, that can be fully hydrided to AH.

The hydrogen source is a high-pressure container with an internal capacity of $V_s = 1$ liter. It is charged with enough hydrogen to have the gas at $p_0 = 500$ atmospheres when the temperature is $T_0 = 300$ K. The container is in thermal contact with the hydride canister. In steady state, the container, the hydrogen, and the canister are all at the same temperature. The heat capacity of the container is 300 J/K.

The whole system—container and canister—is completely adiabatic: no heat is exchanged with the environment.

There is a pipe connecting the hydrogen source to the hydride canister. A valve controls the hydrogen flow. According to the Joule–Thomson law, the gas escaping from the source and flowing into the canister will warm up. However, in this problem, assume that there is no Joule–Thomson effect.

Initially, while the hydrogen delivery valve is still shut off, there is no gas pressure inside the canister, and canister, and contents are at 300 K.

Canister:

Volume, $V = 1$ liter.

Heat capacity, $c_{can} = 700\,\mathrm{J/K}$.

The alloy has the following characteristics:

Amount of hydride, $m = 5.4\,\mathrm{kg}$.

Density of hydride, $\delta = 9000\,\mathrm{kg/m^3}$.

Molecular mass of alloy, A, 100 daltons.

Heat of absorption: $-28\,\mathrm{MJ/kmoles}$.

Entropy change of absorption: $-110\,\mathrm{kJ\ K^{-1} kmole^{-1}}$.

Heat capacity, $c_{hyd} = 500\,\mathrm{J\ kg^{-1}\ K^{-1}}$.

The $\ln p$ versus x characteristics of the alloy are a perfectly horizontal plateau bound by vertical lines corresponding to the depleted and the saturated regions.

The valve is opened and hydrogen flows into the canister. If you wait long enough for the transients to settle down, what is the pressure of the gas in the hydrogen source? Is the hydride in the plateau region?

11.21 A hydrogen source consists of a 5-liter container with hydrogen at an initial pressure of 500 atmospheres. Throughout the experiment the hydrogen in this container remains at a constant temperature of 300 K.

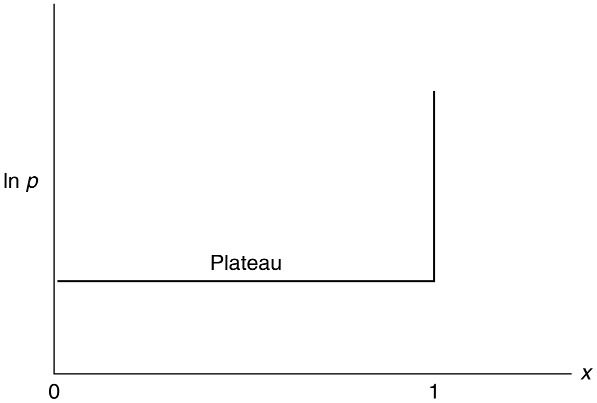
a. How many kilomoles, μ_0, of hydrogen are initially in the container?

A separate 2-liter canister ("alloy canister") contains 3.5 kg of a metallic alloy, AB, with the following properties:

Density	δ	3.5 kg
Heat of absorption	ΔH	−30 MJ/kmole (H2)
Entropy change (absorp.)	ΔS	−100 kJ/(kmole K)
Formula mass		100 daltons
Composition (fully hydrided)	ABH	

Assume an extremely simplified ln(p) versus x characteristic:

- Horizontal plateau.
- No depletion region.
- Vertical saturation region.



The following heat capacities are relevant:

Canister	100	J/K
Alloy	440	J/K per kg
Hydrogen (H_2)	20,800	J/K per kmole

b. If the alloy were fully hydrided, how many kilomoles of hydrogen would be absorbed by the 3.5 kg of the alloy in the canister?

c. The alloy does not completely fill the canister. A "dead space" is left, which is initially a vacuum but will later contain some hydrogen. What is the volume of this dead space?

Before the experiment is began, the alloy is completely degassed—hydrogen is neither in the dead space nor absorbed by the alloy. Both canister and alloy are at 300 K.

The experiment consists of opening the valve and waiting until the system settles into steady state.

As hydrogen flows from the source to the alloy canister, the pressure in the former drops (but the temperature stays put), while the pressure in the dead space rises (and so does the temperature).

d. When steady state is reached, there is a unique pressure in the dead space. The calculation of this pressure is laborious. Assume that it is 10 atmospheres. Calculate the number of kilomoles, μ_{ds} of H_2 in the dead space and the number of kilomoles, μ_{abs}, of H_2 absorbed by the alloy. Prove that the assumed pressure cannot correspond to a steady-state situation.

e. Calculate the correct steady-state pressure in the dead space. Check the correctness of any assumption you may have made.

11.22 The (very idealized) data of the TiFe alloy are:

$\Delta H = -28$ MJ kmoles^{-1}, for absorption.

$\Delta S = -106.1\,\text{KJ K}^{-1}\text{kmoles}^{-1}$, for absorption.

Specific heat capacity, $c = 540\,\text{J kg}^{-1}\text{K}^{-1}$.

Density $= 6200\,\text{kg/m}^3$.

The plateau pressure extends all the way from $x = 0$ to $x = 1$. A this latter point, the pressure rises independently of x.

In all, 100 kg of the above alloy are placed inside a 160-liter canister, which is then carefully pumped out so that the internal pressure is essentially zero. The container is perfectly heat insulated from the environment and (together with the alloy inside) is at 298.0 K.

A large, 1000-liter, separate container filled with hydrogen (pressure = 500.00 atmospheres) is kept at a constant 298.0 K throughout the whole experiment.

A pipe, equipped with a shutoff valve, interconnects the two containers. The valve is initially closed.

The experiment begins with the momentary opening of the valve, allowing some hydrogen to enter the alloy-containing canister. This process lasts long enough for the hydrogen source pressure to fall to 475.54 atmospheres, at which moment the valve is shut. Disregard any Joule–Thomson effect on the temperature of the gas entering the alloy container; that is, assume that the hydrogen enters this container at 298.0 K.

1. How much hydrogen was allowed into the alloy-containing canister?
2. What is the final pressure of the hydrogen in this alloy-containing canister after steady state has been reached?

Ignore the heat capacity of the H_2. The heat capacity of the canister is 20,000 J K^{-1}

11.22 *Solving this problem "exactly" is somewhat laborious. What we want is simply an estimate of the final temperature. You must make some simplifying assumptions. If you do this the hard way, you will lose some points (even if you get the right answer) because you will be wasting time. Reasonable simplifying assumptions will lead to estimates with less than 2% error, which is good enough for government work.*

A small metal canister with a volume of 2 liters contains 3 kg of $TiFeH_{0.95}$ (HY-STOR alloy 101) at 300 K. Alloy density is $6200\,\text{kg/m}^3$. The atomic mass of titanium is 47.90 daltons, and that of iron is 55.85 daltons. The canister has no heat capacity and is completely heat-insolated from the environment. Estimate the temperature of the gas in the cylinder after 2 g of H_2 are withdrawn.

11.23 An adiabatic canister having a volume of $0.05\,\text{m}^3$ contains 30 kg of alloy HY-STOR 205 at 300 K. However, there is no gas of any kind in this canister.

The heat capacity of the canister by itself is negligible.

Admit $\mu_{in} = 0.1$ kilomole of hydrogen into the canister. After a while, things will settle to a new equilibrium. What is the pressure of the gas?

The density of HY-STOR 205 (totally depleted) is 8400 kg/m^3.

11.24 The hydrogen distribution system of a given city delivers the gas at a pressure of 20 atmos. An automobile refueling station must increase the pressure to 400 atmospheres to fill the pressurized gas containers in the cars, which typically have a storage capacity of 6 kg of the gas.

Design a hydride hydrogen compressor for this application. The compressor should have the ability to deliver 6 kg of compressed hydrogen (or somewhat less) in one single "stroke," that is, one single compression cycle. To make sure that there is an adequate gas flow from the city pipeline to the input of the compressor, the plateau pressure of the hydride used, at 25 C, should be a bit lower than the 20-atmosphere pipe line pressure. Assume, however, that the intake plateau pressure is 20 atmos.

Normally, you should have complete freedom to select the alloy to be used; however, to avoid disparate solutions to this problem, we will impose some constraint in this choice. The alloy must have:

$\Delta S_{absorption} = -106.8$ kJ K^{-1}kmole^{-1}.

A beginning of the plateau (transition between depletion and plateau) when the stoichiometric coefficient, $x_{beg} = 0.15$.

An end of the plateau (beyond which the alloy saturates) when $x_{end} = 1.05 - 0.00033T$ where T is the temperature in kelvins.

Use idealized characteristics of an alloy to build a hydrogen compressor capable of raising the pressure from 0.5 atmosphere to 50 atmospheres. Idealized characteristics have horizontal plateaus. The alloy has a heat capacity of 540 J kg^{-1}K^{-1}.

The hydrogen compressor has the following phases:

a. Intake. Starts from point A of the characteristics and move to point B, at constant pressure.

b. Compression. Begins at point B and goes to point C. Temperature and pressure increase.

c. Exhaust. Begins at point C and goes to point D. Gas is delivered at constant pressure.

d. Reset. Goes from point D to point A completing the cycle. Both pressure and temperature are reduced.

During the operation of the compressor, you must always remain in the plateau region. You are not allowed to go into either the depletion or the saturation region. It makes sense to design the system so that the end of the exhaust phase (point D) is exactly at the beginning of the plateau (this allows the delivery of the largest amount of hydrogen). Since point A must have a stoichiometric index larger than that of point D, point A cannot be at the beginning of the plateau. Use a stoichiometric index at A, $x_a = 0.36$.

The alloy you are going to employ has the formula AB and forms an hydride ABH. Point A has an atomic mass of 48 and B, 56 daltons. The alloy has a density of 7000 kg/m^3. The internal, empty, volume of the compressor is V_{empty}. Owing to the granular nature of the alloy, only 60% of this volume is actually occupied by the alloy; the rest is "dead space," usually occupied by H_2 gas. The granules exactly fill V_{empty}; hence, the "dead space" is only the intergranular space.

1. What is the ΔH (absorption) that the alloy must have.
2. Determine to what temperature must the alloy be heated to achieve the desired compression.
3. Assume that in the intake phase (A→B) the compressor takes in 6 kg of hydrogen. Calculate how many kilograms of the alloy are required.
4. The above mass of alloy granules, exactly fills the internal volume of the compressor. This leaves, as dead space, the intergranular volume. Calculate the internal volume of the compressor and the volume of the dead space.
5. How much H_2 is desorbed while going from point B to point C?
6. Considering only the energy for desorption form point B to point C and that from point C to point D plus the energy necessary to heat up the alloy from the intake temperature to the exhaust temperature, estimate the efficiency of the compressor. The compressor itself has zero heat capacity; ignore it.

Chapter 12
Solar Radiation

12.1 The Nature of the Solar Radiation

The sun radiates in all regions of the spectrum, from radio waves to gamma rays. Our eyes are sensitive to less than one octave of this, from 400 to 750 THz (750 to 400 nm), a region known, for obvious reasons, as **visible**. Though narrow, it contains about 45% of all radiated energy. At the distance of one astronomical unit, the power density of the solar radiation is about 1360 W m^{-2}, a value called **solar constant**, which is not really constant. It varies a little throughout the year, being largest in January when the Earth is nearest the sun.

The expression **power density** is used to indicate the number of watts per square meter. This is also known as **energy flux**. We will use the expression **spectral power density** to indicate the power density per unit frequency interval or per unit wavelength interval.

Roughly, the distribution of energy over different spectral regions is

Infrared and below ($f < 400\,\text{THz}, \lambda > 750\,\text{nm}$)	46.3%
Visible ($400\,\text{THz} < f < 750\,\text{THz}, 400\,\text{nm} < \lambda < 750\,\text{nm}$)	44.6%
Ultraviolet and above ($f > 750\,\text{THz}, \lambda < 400\,\text{nm}$)	9.1%

A much more detailed description of the solar radiation is given in Table 12.1 which shows the fraction, G, of the solar constant associated with frequencies larger than a given value, f. These data, plotted in Figure 10.14 (see Chapter 10 on hydrogen production), correspond to the spectral power density distribution shown in Figure 12.1. For comparison, the spectral power density distribution of black body radiation (6000 K) is plotted for constant wavelength intervals and constant frequency intervals. Notice that these two distributions, though describing the very same radiation, peak at different points of the spectrum. To understand the reason for this apparent paradox, do Problem 12.5.

All of these observations refer to radiation outside Earth's atmosphere. The power density of solar radiation on the ground is smaller than that in space owing to atmospheric absorption. Radiation of frequencies above 1000 THz ($\lambda < 300\,\text{nm}$) is absorbed by the upper atmosphere, causing photochemical reactions, producing photoionization, and generally heating up the air. However, this part of the spectrum contains only 1.3% of the solar constant. The ozone layer near 25-km altitude absorbs much of it. Ozone is amazingly opaque to ultraviolet around 250 nm (1200 THz), even though the layer is very thin: if the atmosphere were at uniform sea-level density, it would be 8 km thick—the ozone layer would then measure 2 mm.

Table 12.1 Cumulative Values of Solar Power Density
Fraction of Total Power.
Data from F. S. Johnson.

f	***G***	***f***	***G***	***f***	***G***	***f***	***G***
(THz)		**(THz)**		**(THz)**		**(THz)**	
43	0.9986	176	0.9083	536	0.3180	779	0.0778
50	0.9974	188	0.8940	541	0.3120	789	0.0735
60	0.9951	200	0.8760	545	0.3050	800	0.0690
61	0.9948	214	0.8550	550	0.2980	811	0.0642
63	0.9945	231	0.8290	556	0.2900	822	0.0595
64	0.9941	250	0.7960	561	0.2830	833	0.0553
65	0.9938	273	0.7570	566	0.2760	845	0.0510
67	0.9933	300	0.7090	571	0.2690	857	0.0469
68	0.9929	316	0.6810	577	0.2630	870	0.0427
70	0.9923	333	0.6510	583	0.2560	882	0.0386
71	0.9918	353	0.6170	588	0.2490	896	0.0346
73	0.9913	375	0.5790	594	0.2420	909	0.0308
75	0.9905	400	0.5370	600	0.2350	923	0.0266
77	0.9899	405	0.5270	606	0.2280	938	0.0232
79	0.9891	411	0.5180	612	0.2200	952	0.0233
81	0.9883	417	0.5080	619	0.2130	968	0.0166
83	0.9874	423	0.4980	625	0.2060	984	0.0150
86	0.9863	429	0.4880	632	0.1980	1000	0.0130
88	0.9852	435	0.4780	638	0.1900	1017	0.0106
91	0.9839	441	0.4670	645	0.1820	1034	0.0085
94	0.9824	448	0.4560	652	0.1750	1053	0.0070
97	0.9808	455	0.4450	659	0.1670	1071	0.0059
100	0.9790	462	0.4330	667	0.1590	1091	0.0051
103	0.9772	469	0.4210	674	0.1510	1111	0.0042
107	0.9747	476	0.4090	682	0.1440	1132	0.0035
111	0.9721	484	0.3970	690	0.1370	1154	0.0029
115	0.9690	492	0.3840	698	0.1300	1176	0.0025
120	0.9657	500	0.3720	706	0.1240	1200	0.0021
125	0.9618	504	0.3650	714	0.1170	1224	0.0018
130	0.9571	508	0.3590	723	0.1100	1250	0.0016
136	0.9520	513	0.3520	732	0.1030	1277	0.0014
143	0.9458	517	0.3450	741	0.0970	1304	0.0011
150	0.9387	522	0.3390	750	0.0908	1333	0.0008
158	0.9302	526	0.3320	759	0.0860	1364	0.0006
167	0.9203	531	0.3250	769	0.0819		

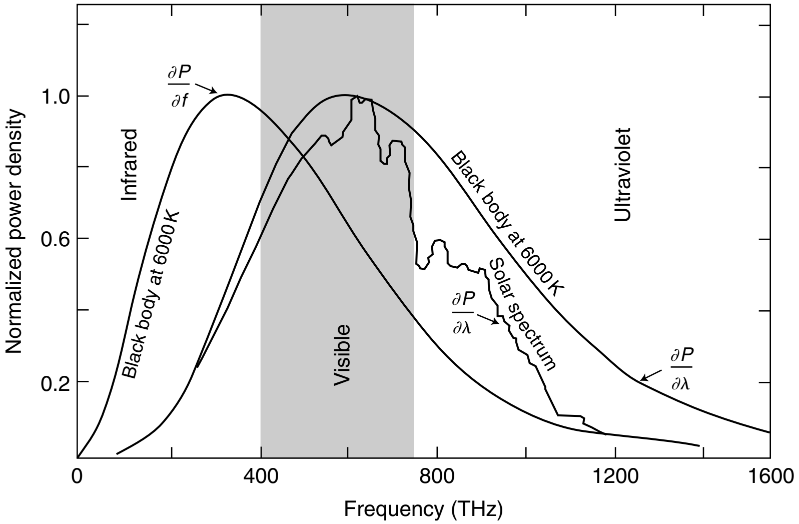


Figure 12.1 The solar power density spectrum compared with that of a black body.

Although solar radiation is generated by several different mechanisms, the bulk of it is of the black body type. The energy per unit volume per unit frequency interval inside a hollow isothermal black body is given by Planck's law:

$$\frac{dW}{df} = \frac{8\pi h}{c^3} \frac{f^3}{\exp(hf/kT) - 1} \quad \text{J m}^{-3}\text{Hz}^{-1}. \tag{12.1}$$

In the preceding expression, W is the energy concentration. The energy flux is equal to the energy concentration times the speed of light (just as a particle flux is equal to the particle concentration times the speed of the particle). Energy flux is, as we stated, the same as power density, P:

$$\frac{dP}{df} \propto \frac{f^3}{\exp(hf/kT) - 1} \quad \text{W m}^{-2}\text{Hz}^{-1}. \tag{12.2}$$

In terms of wavelength,

$$\frac{dP}{d\lambda} \propto \frac{\lambda^{-5}}{\exp(hc/kT\lambda) - 1} \quad \text{W m}^{-2} \text{ per m of wavelength interval.} \tag{12.3}$$

When one tries to match a black body spectrum to that of the sun, one has the choice of picking the black body temperature that best fits the *shape* of the solar spectrum (6000 K) or the temperature that, at one astronomical unit, would produce a power density of 1360 W m^{-2} (5800 K).

12.2 Insolation

12.2.1 Generalities

Insolation[†] is the power density of the solar radiation. In Section 12.1, we saw that the insolation on a surface that faces the sun and is just outside Earth's atmosphere is called the solar constant. It has a value of $1360\,\mathrm{W\,m^{-2}}$.

It is convenient to define a **surface solar constant**—that is, a value of insolation on a surface that, at sea level, faces the vertical sun on a clear day. This "constant" has the convenient value of about $1000\,\mathrm{W\,m^{-2}}$ or "one sun." At other than vertical, owing to the larger air mass through which the rays have to pass, the insolation is correspondingly smaller.

American meteorologists depart from the SI and define a new—and unnecessary—unit called a langley. It is one gram calorie per cm^2 per day. To convert langleys to $\mathrm{W\,m^{-2}}$, multiply the former by 0.4843.

The insolation depends on:

1. the orientation of the surface relative to the sun, and
2. the transparency of the atmosphere.

Caveat

In the following discussion, we will make a number of simplifications that introduce substantial errors in the results but still describe in general terms the way insolation varies throughout the year and during the day. Some of these errors can, as indicated further on, be easily corrected, and those that remain are of little consequence for the planners of solar energy collection systems.

The major sources of errors are:

1. We assume that the length of the period between successive sunrises is constant throughout the year. That is not so—see "The Equation of Time" in Appendix B to this chapter.
2. The time used in our formulas is the mean local time and differs from the civil time, which refers to the time measured at the center of each time zone.

   Corrections for effects a and b can easily be made by introducing the "Time Offset." (See Appendix A.)
3. Our formulas consider only the geometry of the situation. The presence of the atmosphere causes diffraction of the light so that the sun is visible even when it is somewhat below the geometric horizon. This causes the apparent sunrise to be earlier than

(*Continues*)

[†] One should not confuse *insolation*, from the Latin *sol* = *sun*, with the word of essentially the same pronunciation, *insulation*, from the Latin *insula* = *island*.

(*Continued*)

the geometric one and the apparent sunset to be later. This effect can be partially corrected by using a solar zenith angle at sunrise and sunset of 90.833° instead of the geometric 90°.

However, it must be remembered that the refraction correction is latitude dependent and becomes much larger near the polar regions. Much more detailed information on the position of the sun can be found at www.srrb.noaa.gov/highlights/sunrise.

4. The insolation data assume perfectly transparent atmosphere. Meteorological conditions do, of course, alter the amount of useful sunlight in a major way.

In this part of this book, the position of the sun is characterized by the **zenith angle**, χ (the angle between the local vertical and the line from the observer to the sun), and by the **azimuth**, ξ, measured clockwise from the north. This is a **topocentric** system—the observer is at the origin of the coordinates. In Appendix B, we will use two different points of view: a **geocentric** system (origin at the center of Earth) and a **heliocentric** system (origin at the center of the sun). In our topocentric system, both χ and ξ are functions of

1. the local time of day, t,[†]
2. the day of the year, d, and
3. the latitude of the observer, λ.

Observe that the time, t, in our formulas is not the time shown in your watch. These times differ by the **time offset**, which has two components—one related to the difference in longitude of the place of interest from that of the center of the time zone and one owing to the **equation of time, EOT** (see Appendix B). To get a better feeling about these times, do Problem 14.24. In our formulas, t is exactly 12:00 when the *mean* sun crosses the local meridian—that is, when the solar zenith angle is at a minimum. The real sun is in general either ahead or behind the mean sun by an amount called the EOT.

The time of day is represented by the **hour angle**, α, a usage borrowed from astronomers, who in the past worked mostly at night and thus preferred to count a new day from noon rather than from midnight. They define the hour angle as

$$\alpha \equiv \frac{360}{24}(t-12) \qquad \text{degrees } (t \text{ in hours, 24-hr clock}) \qquad (12.4)$$

$$\alpha \equiv \frac{2\pi}{86400}(t-43200) \qquad \text{radians } (t \text{ in seconds, 24-hr clock}) \qquad (12.4a)$$

[†] For comments on time measurement, please refer to Appendix A to this chapter.

The day of the year or "season" is represented by the solar **declination**, δ, that is, by the latitude of the sun.

The solar declination can be found, for any day of a given year, in the **Nautical Almanac** or by consulting the NOAA or the Naval Observatory Web Site. It also can be estimated with sufficient precision for our purposes by the expression:

$$\delta = 23.44 \sin\left[360\left(\frac{d-80}{365.25}\right)\right] \quad \text{degrees,} \tag{12.5}$$

where d is the day number.

The solar zenith angle and the solar azimuth are given by

$$\cos\chi = \sin\delta\sin\lambda + \cos\delta\cos\lambda\cos\alpha \tag{12.6}$$

and

$$\tan\xi = \frac{\sin\alpha}{\sin\lambda\cos\alpha - \cos\lambda\tan\delta}, \tag{12.7}$$

where λ is the latitude of the observer.

To find the value of ξ, we have to take $\arctan(\tan\xi)$. Notice, however, that $\arctan(\tan\xi)$ is not necessarily equal to ξ. Consider, for instance, the angle 240° whose tangent is 1.732. A calculator or a computer will tell you that $\arctan 1.732 = 60^\circ$ because such devices yield the **principal value** of $\arctan\xi$, which, by definition, lies in the range from -90° to 90°. The following rule must be observed when obtaining ξ from Equation 12.7:

Sign(α)	Sign($\tan\xi$)	ξ
+	+	$180^\circ + \arctan(\tan\xi)$
+	−	$360^\circ + \arctan(\tan\xi)$
−	+	$\arctan(\tan\xi)$
−	−	$180^\circ + \arctan(\tan\xi)$

An alternative formula for determining the solar azimuth is

$$\cos(180^\circ - \xi) = -\frac{\sin\lambda\cos\chi - \sin\delta}{\cos\lambda\sin\chi}.$$

At both sunrise and sunset, $\chi = 90^\circ$[†]; thus, $\cos\chi = 0$. From Equation 12.6,

$$\cos\alpha_R = \cos\alpha_S = -\tan\delta\tan\lambda, \tag{12.8}$$

where $\alpha_{R,S}$ is the hour angle at either sunrise or sunset. The hour angle α_R, at sunrise, is negative, and α_S, at sunset, is positive.

$$\alpha_R = -\alpha_S. \tag{12.9}$$

[†] This will yield the geometric sunrise. As explained before, to correct for atmospheric refraction, χ at sunrise and sunset is taken as 90.833°.

12.2.2 Insolation on a Sun-Tracking Surface

If a flat surface continuously faces the sun, the daily average insolation is

$$<P> = \frac{1}{T}\int_{t_R}^{t_S} P_S dt \quad \text{W m}^{-2}, \tag{12.10}$$

where t_R and t_S are, respectively, the times of sunrise and sunset, T is the length of the day (24 hours), and P_S is the solar power density, which, of course, depends on the time of day and on meteorological conditions. Assuming (unrealistically) that P_S is a constant from sunrise to sunset, the average insolation, $<P>$, in terms of the hour angle, is

$$<P> = \frac{1}{\pi}\alpha_S P_S \quad \text{W m}^{-2}. \tag{12.11}$$

At the equinoxes, $\delta = 0$, and, consequently, $\alpha_S = \pi/2$, and

$$<P> = \frac{1}{2}P_S \approx 500\,\text{W}\,\text{m}^{-2} \approx 43.2 \quad \text{MJ}\,\text{m}^{-2}\text{day}^{-1}. \tag{12.12}$$

12.2.3 Insolation on a Stationary Surface

Instantaneous insolation on surfaces with elevation, ϵ, azimuth, ζ, is

$$P = P_S[\cos\epsilon\cos\chi + \sin\epsilon\sin\chi\cos(\xi - \zeta)]. \tag{12.13}$$

Care must be exercised in using Equation 12.13. The elevation angle, ϵ, is always taken as positive. See Figure 12.2. It is important to check if the sun is shining on the front or the back of the surface. The latter condition would result in a negative sign in the second term inside the brackets. Negative values are an indication that the surface is in its own shadow and that the insolation is zero.

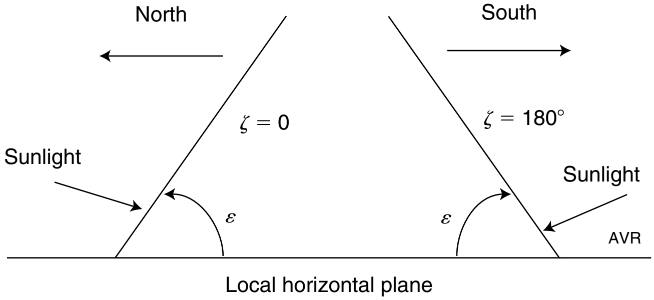


Figure 12.2 The two surfaces above have the same elevation but different azimuths.

The daily average insolation is

$$<P> = \frac{1}{T}\int_{t_R}^{t_S} P dt = \frac{1}{2\pi}\int_{\alpha_R}^{\alpha_S} P d\alpha. \tag{12.14}$$

For the general case, the preceding integral must be evaluated numerically. Figures 12.3 through 12.5 show some of the results. Figure 12.3 shows the insolation on south-facing surfaces located at a latitude of 40° north and with various elevation angles, all as a function of solar declination. A horizontal surface ($\epsilon = 0$) receives a lot of sunlight during the summer ($\delta = +23°$). At the height of summer, it receives more light than at the equator where the average (normalized) insolation is, by definition, 1, independent of the season. As the seasons move on, the insolation diminishes, and, in winter, it is less than 40% of that in summer.

A vertical surface facing the equator receives more insolation in winter than in summer. There is an optimum elevation that yields maximum year-round insolation and, incidentally, minimum seasonal fluctuation. For the latitude of Figure 12.3 (40°), the optimum elevation is 42°—that is, 2° more than the latitude.

The difference between the optimum elevation and the latitude is plotted, as a function of latitude, in Figure 12.4.

The latitude has small influence on the annual insolation for a surface at optimum elevation angle, as can be seen from Figure 12.5. At 67°, the

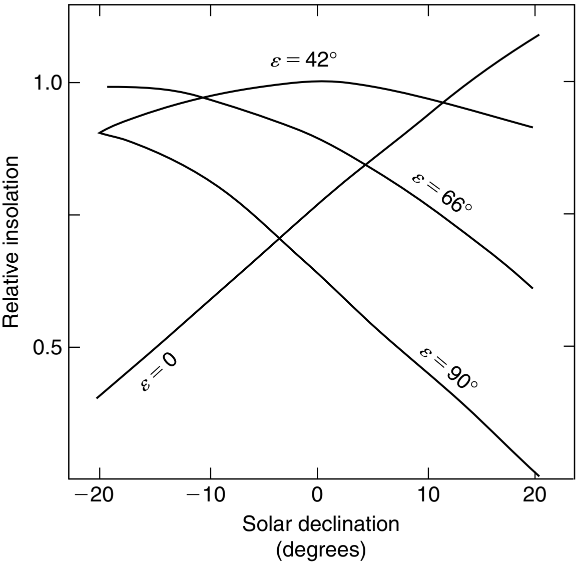


Figure 12.3 Relative insolation of surfaces with various elevations at a latitude of 40°. Each curve has been normalized by comparing the insolation with that on a horizontal surface at the equator.

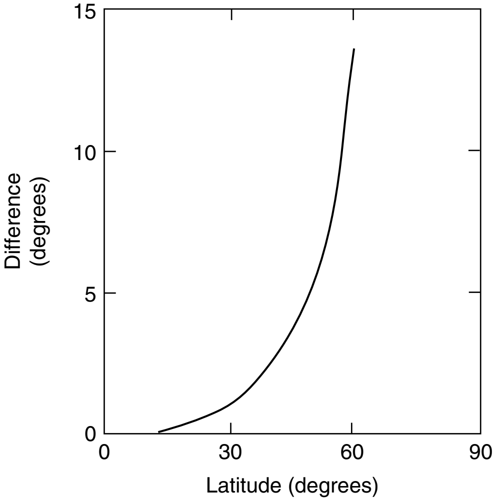


Figure 12.4 Difference between the optimum elevation angle of a solar collector and its latitude.

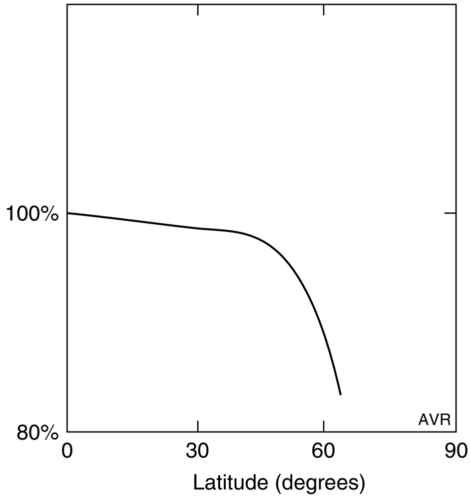


Figure 12.5 Effect of latitude on the annual insolation of a solar collector at optimum elevation.

highest latitude at which the sun comes up every day, the yearly insolation is still well above 80% of that at the equator. This assumes an atmosphere whose transparency does not depend on season and on elevation angles. Actually, the farther north, the larger the atmospheric absorption owing to the smaller average solar elevation angle.

Table 12.2 Estimated Insolation South-Facing Surface Elevation Equal to Latitude

City	**Average Insolation W/m^2**
Bangor, ME	172
Boston, MA	177
Buffalo, NY	161
Concord, NH	171
Hartford, CT	149
Honolulu, HI	230
Los Angeles, CA	248
Newark, NJ	186
New York, NY	172
Philadelphia, PA	185
Phoenix, AZ	285
San Francisco, CA	246
Tucson, AZ	286

On the other hand, in many equatorial regions, such as the Amazon valley, the cloud cover is so frequent that the average insolation is only some 60% of that under ideal meteorological conditions.

It can be seen that the yearly average insolation depends on the orientation of the surface, the latitude, and the prevailing meteorological conditions.

Table 12.2 shows the yearly average insolation in different cities of the United States for south-facing surfaces with an elevation equal to the local latitude. The table was adapted from one published in *IEEE Spectrum*, October 1996, page 53, whose source was "Optimal BIPV Applications," Kiss and Co., Architects, November 1995.

Clearly, these estimated insolation values vary from year to year owing to the variability of the cloud cover.

12.2.4 Horizontal Surfaces

For horizontal surfaces ($\epsilon = 0$), Equation 12.13 reduces to

$$P = P_S \cos \chi. \tag{12.15}$$

Consequently,

$$\begin{aligned} <P> &= \frac{1}{2\pi}\int_{\alpha_R}^{\alpha_S} P_S \cos \chi d\alpha \\ &= \frac{P_S}{2\pi}[\sin\delta \sin\lambda(\alpha_S - \alpha_R) + \cos\delta \cos\lambda(\sin\alpha_S - \sin\alpha_R)] \end{aligned}$$

$$= \frac{P_S}{2\pi} \cos\delta \cos\lambda (2 \sin\alpha_S + 2\alpha_S \tan\delta \tan\lambda)$$

$$= \frac{P_S}{\pi} \cos\delta \cos\lambda (\sin\alpha_S - \alpha_S \cos\alpha_S). \tag{12.16}$$

At the equinoxes, $\delta = 0$, $\alpha_S = \pi/2$; therefore

$$<P> = \frac{P_S}{\pi} \cos\lambda. \tag{12.17}$$

At the equator, regardless of δ, $\alpha_S = \pi/2$; therefore

$$<P> = \frac{P_S}{\pi} \cos\delta. \tag{12.18}$$

12.3 Solar Collectors

Methods for collecting solar energy for the production of either heat or electricity include

1. appropriate architecture,
2. flat collectors,
3. evacuated tubes,
4. concentrators, and
5. solar ponds.

12.3.1 Solar Architecture

Proper architecture is an important energy-saving factor. Among others, the following points must be observed:

12.3.1.1 Exposure Control

Building orientation must conform to local insolation conditions. To provide ambient heating, extensive use can be made of equatorward-facing windows protected by overhangs to shield the sun in the summer. Reduction or elimination of poleward-facing windows will diminish heat losses. Shrubs and trees can be useful. Deciduous trees can provide shade in the summer while allowing insolation in the winter.

12.3.1.2 Heat Storage

Structures exposed to the sun can store heat. This may be useful even in summer—the heat stored may be used to pump cool air by setting up convection currents.

Roof ponds can contribute to both heating and cooling. Any part of the building (walls, floor, roof, and ceiling) can be used for heat storage.

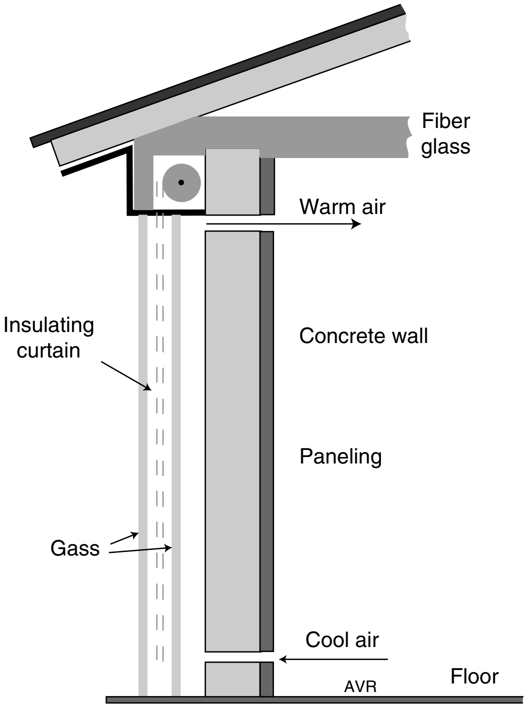


Figure 12.6 Heat storing wall. (Concept Construction, Ltd.)

Figure 12.6 shows an elaborate wall (proposed by Concept Construction, Ltd., Saskatoon, Saskatchewan) that acts as a heat collector. It consists of a 25-cm-thick concrete wall in front of which a glass pane has been installed leaving a 5-cm air space.

During the summer, the warmed air is vented outside, and the resulting circulation causes cooler air from the poleward face of the house to be taken in.

Instead of a concrete wall, the heat-storing structure can be a large stack of "soda" cans full of water (or, for that matter, full of soda). This takes advantage of the high heat capacity of water. The glazing is placed in contact with the cans so as to force the air to seep through the stack, effectively exchanging heat with it. (You may as well paint the cans black.)

12.3.1.3 Circulation

Heat transfer can be controlled by natural circulation set up in a building by convection currents adjusted by vents. Circulation is important from the health point of view. Attempts to save energy by sealing houses may cause an increase in the concentration of radon that emanates from the ground in some places but that is normally dissipated by leakages. Other undesirable gases accumulate, one being water vapor leading to excessive moisture. In addition, noxious chemicals must be vented. This is particularly true of formaldehyde used in varnishes and carpets.

To reduce heat losses associated with air renewal, air-to-air heat exchangers or **recuperators** are used. In this manner, the outflowing air pre-heats (or pre-cools) the incoming fresh air. About 70% of the heat can be recovered.

12.3.1.4 Insulation

The heat power, P, conducted by a given material is proportional to the heat conductivity, λ, the area, A, and the temperature gradient, dT/dx:

$$P = \lambda A dT/dx. \tag{12.19}$$

In the SI, the unit of heat conductivity is $\mathrm{WK^{-1}m^{-1}}$. Many different nonmetric units are used in the United States. Commonly, insulating materials have their conductivity expressed in BTU $\mathrm{hr^{-1}}$ $\mathrm{ft^{-2}}$ $(\mathrm{F/inch})^{-1}$, a unit that is a good example of how to complicate simple things. To convert from this unit to the SI, multiply it by 0.144.

Under a constant temperature gradient, $dT/dx = \Delta T/\Delta x$, where ΔT is the temperature difference across a material Δx units thick.

One has

$$P = \frac{\lambda A \Delta T}{\Delta x} = \frac{A\Delta T}{R}, \tag{12.20}$$

where

$$R \equiv \Delta x/\lambda. \tag{12.21}$$

R has units of $\mathrm{m^2K\ W^{-1}}$ or, in the United States, hr $\mathrm{ft^2F/BTU}$. Again, to convert from the American to the SI, multiply the American by 0.178.

Insulating materials are rated by their R-values. Fiberglass insulation, 8 cm thick, for instance, is rated R-11 (in the American system).

Consider a house with an inside temperature of 20 C and an attic temperature of 0 C. The ceiling has an area of $100\,\mathrm{m^2}$ and is insulated with R-11 material. How much heat is lost through the attic?

In the SI, R-11 corresponds to $11 \times 0.178 = 2\ \mathrm{m^2\ K\ W^{-1}}$. Thus, the heat loss under the assumed 20 K temperature difference is

$$P = \frac{100 \times 20}{2} = 1000 \quad W. \tag{12.22}$$

Thus, if the only heat losses were through the ceiling, it would take little energy to keep a house reasonably warm. There are, of course, large losses through walls and, especially, through windows. A fireplace is a particularly lossy device. If the chimney is left open, warm air from the house is rapidly syphoned out. If a metallic damper is used to stop the convection current, then substantial heat is conducted through it.

12.3.2 Flat Collectors

Flat collectors work with both direct and diffused light. They provide low-temperature heat (less than 70 C) useful for ambient heating, domestic hot water systems, and swimming pools. This type of collector is affected by weather, and its efficiency decreases if large temperature rises are demanded.

For swimming pools in the summer, when only a small temperature increase is needed, flat collectors can be over 90% efficient. It is necessary to operate them so that large volumes of water are only slightly heated rather than heating small amounts of water to a high temperature and then mixing them into the pool.

Simple collectors are black plastic hoses exposed to the sun. More elaborate collectors use both front and back insulation to reduce heat losses.

Collectors may heat water directly or may use an intermediate heat transfer fluid.

Figure 12.7 shows a cross section through a typical flat collector. Light and inexpensive aluminum is used extensively; however, it tends to be corroded by water. Copper is best suited for pipes. If an intermediate heat exchange fluid is employed, aluminum extrusions that include the channels for the liquid are preferred.

Some panels use a thin copper sleeve inserted into the aluminum tubing.

Panels can be black-anodized or painted. There is some question about the lifetime of paints exposed to solar ultraviolet radiation.

The front insulation can be glass or plastic. Glass insulation is fragile, but the plastic does not withstand ultraviolet well. To avoid heat losses through the back of the panel, insulation such as fiberglass mats or polyurethane foam is used. Polyurethane foam imparts good rigidity to the panel, allowing a reduction in the mass of the material employed.

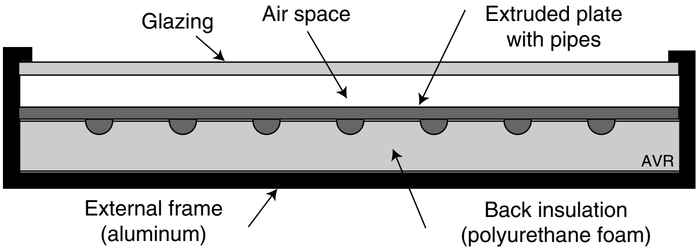


Figure 12.7 Cross section through a typical flat collector.

12.3.3 Evacuated Tubes

This type of collector consists of two concentric cylinders, the outer one of glass and the inner, a pipe through which the liquid flows. They bear a superficial resemblance to fluorescent lamps. A vacuum is established between the two cylinders, reducing the convection heat losses.

Evacuated tubes are nondirectional and can heat liquids to some 80 C. They are usually employed in arrays with spacing equal to the diameter of the outer tube. It is customary to place a reflecting surface behind the array.

12.3.4 Concentrators

Concentrators can be of the nonimaging or the focusing type. Either can be line concentrators (2-D) or point concentrators (3-D).

A solar collector consists of a concentrator and a receiver. The concentrator may be of the refracting (lens) or the reflecting (mirror) type. The receiver may be thermal or photovoltaic.

Two important parameters describe the collector performance:

1. the concentration, C, and
2. the acceptance angle, θ.

The concentration can be defined as either the ratio of the aperture area to the receiver area or as the ratio of the power density at the receiver to that at the aperture. These definitions are not equivalent; the latter concentration is preferable.

The acceptance angle is the angle through which the system can be misaimed without (greatly) affecting the power at the receiver (see Figure 12.10 later in this chapter).

There is a theoretical relationship between the concentration and the acceptance angle for the ideal case:

$$C_{ideal} = (\sin\theta)^{-1} \qquad \text{for a 2-D concentrator.} \tag{12.23}$$

$$C_{ideal} = (\sin\theta)^{-2} \qquad \text{for a 3-D concentrator.} \tag{12.24}$$

It is instructive to calculate the maximum temperature that a receiver can attain as a function of concentration. An ideal receiver will work in a vacuum (no convection losses) and will be perfectly insulated (no conduction losses). Nevertheless, radiation losses are unavoidable. They will amount to

$$P_r = \sigma\epsilon T^4 \qquad \mathrm{Wm}^{-2}, \tag{12.25}$$

where ϵ is the emissivity (taken as unity)[†] and σ is the Stefan–Boltzmann constant (5.67×10^{-8} W m^{-2}K^{-4}). See Chapter 6.

[†] It is, of course, desirable to have a low emissivity at the temperature at which the receiver operates and a high absorptivity at the region of the spectrum dominated by the incident sunlight.

The power density at the receiver (assuming no atmospheric losses) is

$$P_{in} = 1360\,C \qquad \text{Wm}^{-2}. \tag{12.26}$$

In equilibrium, $P_r = P_{in}$,

$$\sigma T^4 = 1360\,C \tag{12.27}$$

or

$$T = (2.4 \times 10^{10} C)^{1/4} = 394\ C^{1/4}. \tag{12.28}$$

With unity concentration (flat plate collector), $T = 394\,\text{K}$ or 120 C.

When the concentration is raised to 1000, the maximum temperature theoretically attainable is 2200 K. Were it possible to construct a collector with a concentration of 1 million, the formula would predict a receiver temperature of 12,400 K. This would violate the second law of thermodynamics because heat would be flowing unaided from the cooler sun (6000 K) to the hotter receiver. Clearly, the upper bound of the receiver temperature must be 6000 K.

One can arrive at the same conclusion by considering the expression for C_{ideal}. The solar angular radius, as seen from Earth, is 0.25°. Thus, the minimum acceptance angle that allows collection of light from the whole sun is 0.25°, and this leads to a $C_{max} = 52{,}000$ for a 3-D collector and 230 for a 2-D collector as calculated from Equations 12.23 and 12.24. A concentration of 52,000 corresponds to a T_{max} of 5900 K (Equation 12.28), just about right.

Numerous reasons cause the concentration (in terms of power densities) to be less than ideal:

1. Reflector shape and alinement errors
2. Less than perfect reflector surface reflectivity
3. Tracking errors
4. Atmospheric scatter
5. Atmospheric absorption

12.3.4.1 Holographic Plates

Just as a flat plastic sheet with appropriate grooves (a **Fresnel** lens) can concentrate light, a holographic plate can be fashioned to do the same. The advantage of the holographic approach is that the plate can simultaneously diffract and disperse, creating a rainbow of concentrated light with each region of the spectrum directed to a collector optimized for that particular color range. This is a significant advantage when using photovoltaic cells. This technology, which looks promising, is discussed in Chapter 14.

12.3.4.2 Nonimaging Concentrators

The simplest nonimaging concentrator is a cone. In a properly designed cone, all rays parallel to its axis will be reflected into the exit aperture. See ray A in Figure 12.8. The acceptance angle, however, is small. Ray B makes it through the exit, but ray C, though parallel to B, does not—it bounces around a number of times and finally returns to the entrance.

The performance of nonimaging concentrators is improved by the use of a **compound parabolic concentrator** (CPC). This device consists of parabolic surfaces as shown in Figure 12.9 (from Welford and Winston 1978). Notice that the section is not that of a truncated parabola but rather that of two independent parabolas, mirror images of one another. The CPCs in Figure 12.9 have identical exit apertures. The largest has the biggest collecting area and consequently the largest concentration; the acceptance angle, however, is small compared with that of the CPCs with smaller concentration.

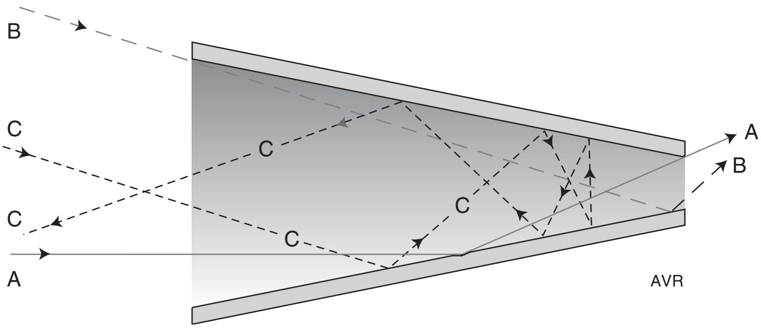


Figure 12.8 Ray paths in a conical concentrator.

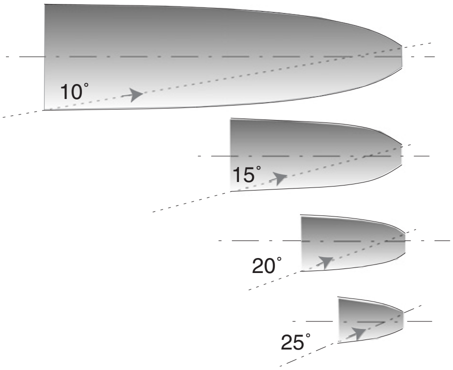


Figure 12.9 For the same exit aperture, the larger the entrance aperture, the smaller the acceptance angle.

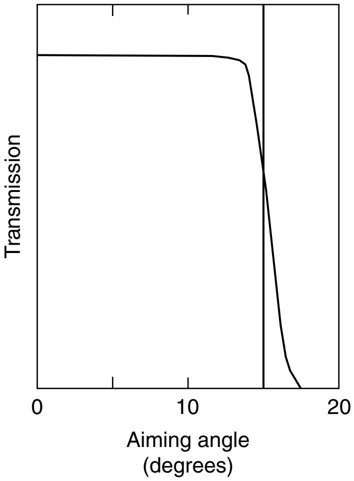


Figure 12.10 A CPC approaches the ideal transmission versus aiming angle characteristics.

The area of the reflecting surfaces of a CPC is much larger than that of a focusing paraboloid of the same concentration. This makes CPCs heavy, expensive, and difficult to mount.

Ideally, a plot of the normalized light power density that leaves the exit aperture as a function of the **aiming angle**, Θ, should be a rectangle: unity for $\Theta < \Theta_i$ and zero for $\Theta > \Theta_i$. Here, Θ_i is the acceptance angle. This ideal characteristic is poorly approached by a conical concentrator, but reasonably well achieved by a CPC. See Figure 12.10.

2-D nonimaging collectors with concentrations up to 2 or 3 need not track the sun. Devices with larger concentrations require a once a month re-aiming to compensate for variations in solar declination.

12.4 Some Solar Plant Configurations

12.4.1 High-Temperature Solar Heat Engine

A straightforward method of generating electricity from solar energy is to use concentrators to produce high temperatures that can drive either a Stirling or a Rankine (steam) engine.

Southern California Edison Company had a 10-MW facility in Barstow, in the Mojave Desert. It cost \$140 million—that is, \$14,000/kW. Since fossil-fueled plants may cost around \$1000/kW, it can be seen that this particular solar thermal installation can be justified only as a development tool. Since the “fuel” is free, it is important to determine the plant

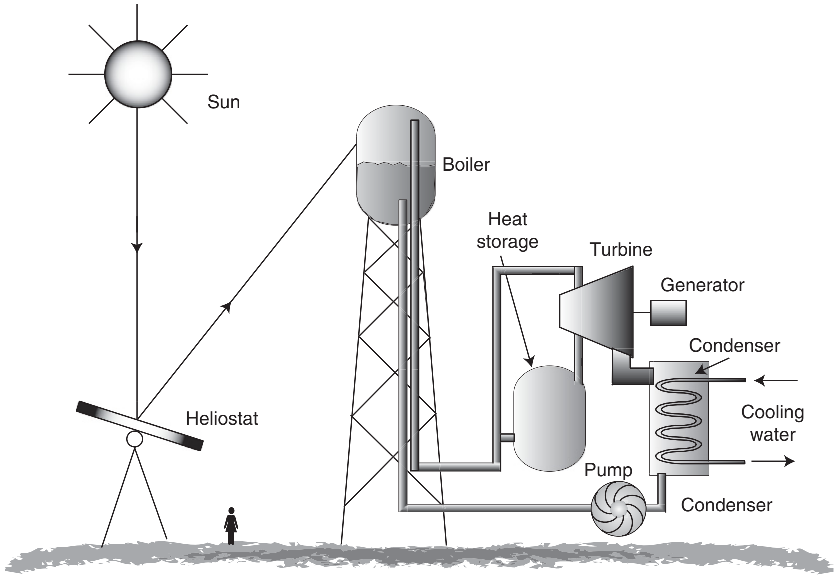


Figure 12.11 Solar One power plant of the Southern California Edison Co., in the Mojave Desert.

life and the operation and maintenance cost so as to be able to compare the cost of electricity over the long run with that from traditional sources.

The plant operated in the manner indicated schematically in Figure 12.11. The collecting area was roughly elliptical in shape and covered some 300,000 m^2 (30 hectares). Insolation, averaged over 24 hours, is probably around 400 W/m^2 for sun-tracking collectors, leading to an efficiency or the order of 8%.

The collector consisted of 1818 sun-tracking flat mirrors forming a gigantic heliostat capable of focusing the sun's energy on a boiler.

The boiler was a cylinder 7 m in diameter and 14 m in height. It was operated at 788 K (516 C) and 10.7 MPa (105 atmospheres).

A thermal storage system was incorporated having a 100 GJ (electric) capacity permitting the plant to deliver some 7 MW for 4 hours during nighttime.

Solar One was decommissioned, and an updated installation, Solar Two, took its place starting operation in July 1996. The new plant cost $48.5 million but inherited considerable assets from the previous effort (among others, the many heliostats), so that its true investment cost is hard to estimate.

Solar Two has the same 10-MW rating as Solar One. The main difference is the use of an intermediate working fluid—a $NaNO_3/KNO_3$ mix containing 40% of the potassium salt—that is heated in the solar tower and transfers its energy to the turbine operating steam by means of a heat exchanger. The salt leaves the tower at 565 C and, after delivering part of its heat energy to the steam, returns to the tower at 288 C.[†]

The salt mixture is somewhat corrosive, requiring low-grade stainless steel in pipes and containers. At the operating temperature, the mixture is quite stable and has a low vapor pressure.

The mirrors are made of a sandwich of two glass panes with a silver layer between them. This protects the reflecting layer from corrosion. When not in use, the mirrors are placed in a horizontal position to protect them from the destructive action of windstorms. This also reduces abrasion from wind-carried sand.

12.4.2 Solar Tower

A circular tent with 121-m radius made mostly of plastic material (and, partially, of glass) was built in Manzanares, Spain. The height of the tent is 2 m at the circumference and 8 m at the center where a 194-m tall chimney has been erected.

The air under the tent is heated by the greenhouse effect and rises through the 10-m diameter chimney driving a 50-kW turbine.

The installation, owned by the Bundesministerium für Forschung und Technologie, Bonn, Germany, was built by Schlaich Bergermann und Partner. It operated from 1989 through 1996 and served to demonstrate the principle.

In 2002, the Australian ministry for Industry, Tourism, and Resources gave its support to the firm EnviroMission to work on a much larger plant of the Manzanares type. The plant would generate at peak 200 MW of electricity and would cost \$800 million. This corresponds to \$4000/kW, a reasonable amount for a development project (typically, electric-generating plants cost about \$1000/kW). The tower is supposed to generate 650 GW-hours per year (2.3×10^{15} J/year). This represents an optimistic 36% utilization factor.

The proposed plant would have a collecting tent of 7-km diameter—that is, a collecting area of some 38×10^6 m^2. At a peak insolation of 1000 W/m^2, the efficiency would be about 0.5%.

Since the efficiency of the system depends on the height of the chimney, the design requires a 1000-m high structure. Such a tall structure will

[†] Both sodium nitrate and potassium nitrate go by the common name of **niter** or **saltpeter**. Sodium nitrate (a common fertilizer and oxidant) melts at 306.8 C, and potassium nitrate (used in the manufacture of black powder) melts at 334 C. Usually, alloys and mixtures melt at a lower temperature than their constituents. **Eutectic** mixture is the one with the lowest melting point. Solar Two uses less potassium than the eutectic because the potassium salt is more expensive than the sodium one. The mixture used in the plant melts at 288 C.

certainly be a challenge to civil engineers because (among other factors) it would be subject to enormous wind stresses. The tallest existing tower (excluding radio towers) is the 553-m Canada's National Tower in Toronto.

12.4.3 Solar Ponds

The OTEC principle discussed in Chapter 4 can be used to generate electricity from water heated by the sun. Surprisingly high temperatures can be obtained in some ponds.

In shallow ponds with a dark-colored bottom, the deep layers of water are warmed up and rise to the surface owing to their lower density. This causes mixing that tends to destroy any temperature gradient. In such ponds the temperature rise is modest because most heat is lost through the surface by the evaporation of water.

The solution is, of course, to cover the pond with an impermeable light-transmitting heat-insulating layer as is done in swimming pools with plastic covers.

It is interesting to observe that the insulating layer can be the water itself. If a vertical salinity gradient is created in the pond, so that the deeper layers contain more salt and become correspondingly denser, it is possible to impede convection and achieve bottom temperatures as high as 80 C.

Working against a 20-C cold sink, the Carnot efficiency of an OTEC would be 20%, and practical efficiencies of 10% do not seem impossible.

Difficulties involve

1. mixing owing to wind action and other factors,
2. development of turbidity owing to collected dirt and the growth of microorganisms.

Such difficulties may be overcome by using **gel ponds** in which the water is covered by a polymer gel sufficiently viscous to impede convection. Such gel must be

1. highly transparent to sunlight,
2. stable under ultraviolet radiation,
3. stable at the operating temperature,
4. insoluble in water,
5. nonbiodegradable,
6. nontoxic,
7. less dense than the saline solution, and
8. inexpensive.

E. S. Wilkins and his colleagues (1982) at the University of New Mexico (Albuquerque) claim to have developed such a gel.

To keep the surface clean, a thin layer of water runs on top of the gel, sweeping away dirt and debris.

Appendix A: The Measurement of Time

The Duration of an Hour[†]

How long is an hour?

In Roman times, the hour was defined as 1/12 of the time period between sunrise and sunset. Since this interval varies with seasons, the "hour" was longer in the summer than in the winter.

At the latitude of Rome, about 42° N, one hour would last anywhere between 44.7 modern minutes (in the end of December) and 75.3 minutes (in mid-June). This variability was a major problem for clockmakers who had to invent complicated mechanisms to gradually change the clock speed according to the season of year. See *On Architecture* by Vitruvius (Marcus Vitruvius Pollio, a book still being sold).

Much of the variability is eliminated by defining the hour as 1/24 of the interval between two consecutive noons, that is, two consecutive solar crossings of the local meridian. Unfortunately, this also leads to an hour whose length varies throughout the year, albeit much less than the Roman one. (See a detailed explanation in Appendix B.) The obvious solution is to define a **mean solar hour** as the average value of the **solar hour** taken over a one-year interval. But again, owing to the very slow changes in astronomical constants (eccentricity, semi-major axis, argument of perihelion, etc.), this definition of an hour will not be constant over long periods of time. The final solution is to define arbitrarily an **ephemeris hour** referred to an atomic clock. At present, the ephemeris hour is very close to the mean solar hour.

Astronomy and chronology, being ancient sciences, have inherited ancient notions and ancient terminology. The division of the hour into minutes and seconds is one example.

Using the Babylonian sexagesimal system, the hour was divided into "minute" parts (*pars minuta prima*, or first minute part, or simply "minutes") and then again into *pars minuta secunda*, or simply "seconds." The latter is the official unit of time for scientific purposes and has a value of 1/86,400 of a mean solar day.

The times obtained from the approximate formulas in this book are the mean solar times and may differ by as much as ±15 minutes from the true solar times.

Time Zones

The local mean solar time is not a convenient measure of time for everyday use because it depends on the longitude of the observer. It varies by 1 hour for every 15° of longitude. This means, of course, that the time

[†] For more information on the history of the hour, read Dohrn-van Rossum (1996).

in San Francisco is not the same as the time in Sacramento. The use of time zones, 1 hour or 15° wide, circumvents this difficulty. In each zone, the time is the same regardless of the position of the observer. At the zone boundaries, the time changes abruptly by 1 hour. The center meridian of any time zone is a multiple of 15°; the first zone is squarely centered on the zeroth meridian, that of Greenwich, and the time there is called **Greenwich mean time, GMT** (or, to astronomers, **universal time, UT**). The zone time is called **standard time** (such as, for instance, PST, for Pacific Standard Time, the −8 time zone centered at 120° W).

Time Offset

The true solar time, t_{true}, at any given longitude, L, can be found from

$$t_{true} = t_{local\ mean} + t_{offset}, \tag{12.29}$$

where t_{true} and t_{local} are expressed in hours and minutes, but t_{offset} is in minutes only,

$$t_{offset} = EOT - 4L + 60t_{zone} \quad \text{minutes}, \tag{12.30}$$

where EOT is the equation of time (in minutes), discussed in Appendix B, L is the longitude in degrees (east > 0, west < 0), and t_{zone} is the number of hours of the local time zone referred to the UT (east > 0, west < 0).

Here is a simple example:

What is the true solar time on February 20, at Palo Alto, California (125° W) when the local mean time is 12:00 (noon)?

The EOT for February 20 (scaled from Figure 12.19) is +14 minutes. The time zone of Palo Alto is Pacific Standard Time; that is, it is −8 h. We have

$$t_{offset} = 14 - 4 \times (-125) + 60(-8) = 34 \quad \text{minutes} \tag{12.31}$$

$$t_{true} = 12^{\mathrm{h}} : 00^{\mathrm{m}} + 34^{\mathrm{m}} = 12^{\mathrm{h}} : 34^{\mathrm{m}}. \tag{12.32}$$

The Calendar

A few recurring astronomical features serve quite obviously as measurements of time. There is the daily rise of the sun that leads to the definition of a **day** and its divisions (hours, minutes, and seconds). There are also the phases of the moon, which are repeated (approximately) every 28 days, leading to the notion of **month** and its subdivision, the **week**.

And then there is the time it takes the Earth to complete one orbit around the sun, which leads to the definition of **year** and to the recurrence of the seasons.

Unfortunately, the number of days in a year or in a month is not an exact integer, and this complicates the reckoning of the date. If one month were exactly 4 weeks (28 days), and if the year were exactly 12 months (336 days), there would be no difficulty. However, the year is closer to thirteen 28-day months. This leads to 364 days per year. The extra day could be declared a universal holiday. The trouble with this scheme is that it does not lend itself to an easy division in quarters. Thus, 12 months per year is the choice.

The first month of the Roman calendar used to be *Martius*, our present March, the fifth month of the year was *Quintilis*, the sixth, *Sextilis*, the seventh, *September*, and so on. From 153 B.C. on, *Ianuarius* was promoted to first place, leading to September being the ninth month, not the seventh as before.

The exact date of the equinox could be easily measured (as was probably done at Stonehenge and at other much older observatories), and thus, one could easily establish that the vernal equinox should always occur at the same date (March 21, for instance). This meant that occasionally one additional day would have to be added to the year. In this respect, the Romans were a bit sloppy and let the slippage accumulate until it became painfully obvious that the seasons were out of phase. If a given crop had to be planted at, say, the first day of spring, you could not assign a fixed date for this seeding day.

By the time the error became quite noticeable, the pontifex maximus[†] would declare a *mensis intercalaris*, an intercalary month named *Mercedonius*, and stick it somewhere toward the end of February, which was then at the end of the year. With time, the position of pontifex became a sinecure, and the calendar adjustment became quite erratic and subject to political corruption. To correct matters, in 46 B.C, Julius Caesar declared a year 445 days long (three intercalary months were added). He also decreed a new calendar, establishing that each year would be 365 days long, and, to account for the extra 1/4 day, every four years an additional day would be added. He commemorated this achievement by naming Quintilis after himself—thus introducing the name July. Not to be outdone, his nephew Octavian (Augustus) insisted on changing Sextilis to August. Of course, both July and August are 31-day months. What else?

The corrections worked for a while but not perfectly (because the year is not *exactly* $365\frac{1}{4}$ days). In March 1582, Gregory XIII introduced a fix—the Gregorian calendar used today. A number of European nations immediately adopted the new calendar; others resisted. England, always

[†] Chief bridge builder, possible in charge of bridge maintenance in old Rome.

bound by tradition, only adopted the Gregorian on September 2, 1752. The next day became September 14, 1752. So the answer to the trivia question "What important event occurred in the United States on September 10, 1752?" is "Nothing!" By the way, Russia only converted to the Gregorian calendar on February 1, 1918, and this is why the "October Revolution" actually occurred in November.

The Julian Day Number

In many astronomical calculations, it proves extremely convenient to have a calendar much simpler than any of those in common use. The simplest way to identify a given day is to use a continuous count, starting at some arbitrary origin in the past, totally ignoring the idea of year, month, and week. The astronomical Julian day number is such a system. We will define it as the day count starting with the number 2,400,000 corresponding to November 16, 1858. Thus, for instance, the next day, November 17, 1858, has a Julian day number of 2,400,001. The Julian day number starts at noon.

To determine the Julian day number corresponding to a given Gregorian date, it is sufficient to count the number of days after (or before) November 16, 1858 and add this to 2,400,000. This is easier said than done. It is a pain to count the number of days between two dates. We suggest the following algorithm taken from http://webexhibits.org/calendars/calendarchristian.html:

$$\begin{gathered} JD = d + \mathrm{INT}((153m + 2)/5) + 365y + \mathrm{INT}(y/4) \\ - \mathrm{INT}(y/100) + \mathrm{INT}(y/400) - 32045, \end{gathered}$$

where y stands for the year (expressed in four figures), m is the month, and d is the day.

Notice that there is no year zero in the Julian or Gregorian calendars. The day that precedes January 1, A.D. 1 is December 31, 1 B.C. If you use the formula above to determine a Julian day number of a B.C. date, you must convert to negative year numbers, as, for instance, 10 B.C., which must be entered as -9.

If you are dealing with ancient dates, you must realize that they are given, most frequently, in the Julian calendar, not the Gregorian, even if they are earlier than the year (45 B.C.) when the Julian calendar was established. This use of a calendar to express dates before it was established is an anticipation, or prolepsis, and is called a **proleptic** calendar. Do not confuse the Julian calendar with the Julian day number (different Juliuses, almost certainly).

For more information on Julian day numbers and on algorithms to convert Julian day numbers to Gregorian or to Julian dates, please refer to the URL given earlier.

Appendix B: Orbital Mechanics

Sidereal versus Solar

The most obvious measure of time is the interval between two consecutive **culminations** of the sun, called a **solar day**. The sun culminates (reaches the highest elevation during a day—the smallest zenith angle) when it crosses the local meridian and is, consequently, exactly true south of an observer in the northern hemisphere. As will be explained later, there is an easily determined moment in the year when equinoxes occur. The time interval between two consecutive vernal equinoxes—that is, the length of the **tropical year**—has been measured with—what else?—astronomical precision. It is found that during one tropical year, the sun culminates 365.24219878 times (call it 365.2422 times).† In other words, there are, in this one-year interval, 365.2422 solar days. We define the **mean solar hour** as 1/24 of a mean solar day (a day that lasts 1/365.2422 years). Unfortunately, the length of a solar day as measured by any reasonably accurate clock changes throughout the year. The change is not trivial—the actual solar day (the time between successive culminations) is 23 seconds shorter than the mean solar day on September 17, and it is 28 seconds longer on December 22. These differences accumulate—in mid-February the sun culminates some 14 minutes after the mean solar noon and in mid-November, some 15 minutes before mean solar noon. Later, we will explain what causes this variability, which renders the interval between consecutive culminations an imprecise standard for measuring time.

If we were to measure the time interval between successive culminations of a given star, we would find that this time interval is quite constant—it does not vary throughout the year. We would also find that a given star will culminate 366.2422 times in a tropical year. The number of "star," or **sidereal**, days in a year is precisely one more day than the number of solar days.

The discrepancy between the sidereal and the solar time is the result of the Earth orbiting *around* the sun. (Refer to Figures 12.12 and 12.13.)

In a planetary system in which a planet does not spin, its motion around the sun causes an observer to see the sun move *eastward* throughout the year, resulting in one apparent day per year. If, however, the planet spins (in the same direction as its orbital motion) at a rate of exactly 360 degrees per year, then the sun does not seem to move at all—the planetary spin exactly cancels the orbitally created day.

As explained, the completion of a full orbit around the sun will be perceived from the surface of a nonspinning planet as one complete day—the sun will be seen as moving in a complete circle around the planet.

†The length of a tropical year is decreasing very slightly, at a rate of $169 \times 10^{-10}\%$ per century. This means that there is a very small secular reduction in the orbital energy of Earth.

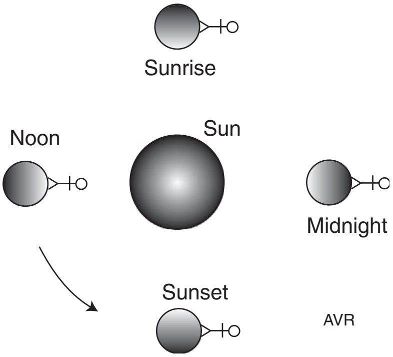


Figure 12.12 In a planetary system in which the planet does not spin, the orbital motion introduces one apparent day per year.

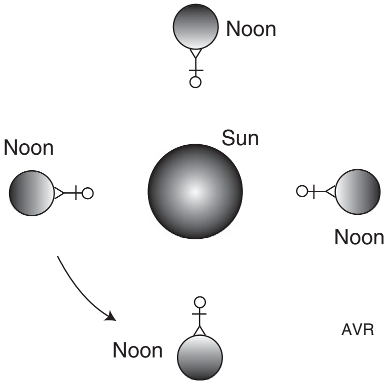


Figure 12.13 If the planet spins exactly once per year, the sun appears not to move.

If we count 24 hours/day, then each 15° of orbital motion causes an apparent hour to elapse, and 1° of orbital motion corresponds to 4 minutes of time.

Thus, in the case of the planet Earth, which spins 366.2422 times a year, the number of solar days is only 365.2422 per year because the orbital motion cancels one day per year.

We collected accurate definitions of time units in Table 12.3.

Orbital Equation

The angle *reference point–sun–planet* is called the **true anomaly** (see Figure 12.14). Notice the quaint medieval terminology frequently used in astronomy.

In a circular orbit, there is no obvious choice for a reference point. However, since most orbits are elliptical (or, maybe, hyperbolic), the periapsis

Table 12.3 Time Definitions

Year (tropical)	Interval between two consecutive vernal equinoxes
Mean solar day	1/365.24219878 years
Mean solar hour	1/24 mean solar days
Minute	1/60 mean solar hours
Second	1/60 minutes
Sidereal day	1/366.24219878 years
Mean solar hours/year	8,765.81
Minutes per year	525,948.8
Seconds per year	31.5569×10^6
Length of sidereal day	23:56:04.09

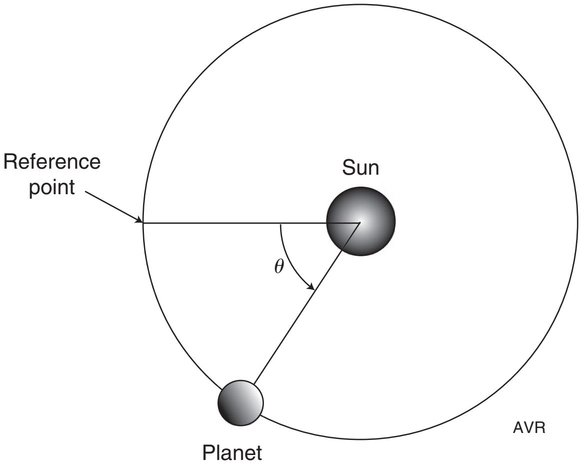


Figure 12.14 The angular position of a planet in its orbit in the ecliptic plane is called the true anomaly, θ.

(nearest point to the attracting body)[†] is a natural choice. The apoapsis is not nearly as convenient because, in the case of long period comets, it cannot be observed.

Thus, the origin for the measurement of the true anomaly is the periapsis. The anomaly increases in the direction of motion of the planet.

Consider a planetary system consisting of a sun of mass, M, around which orbits a planet of mass, m. Assume that the orbital velocity of the planet is precisely that which causes the centrifugal force, $mr\omega^2$ (where ω is

[†] Here we have a superabundance of terms with nearly the same meaning: **periapsis, perihelion, perigee, periastron, pericenter**, and **perifocus** and, **apoapsis, aphelion, apogee, apastron, apocenter**, and **apofocus.**

the angular velocity of the planet in radians/second) to equal the attracting force, GmM/r^2, where G is the gravitational constant (6.6729×10^{-11} $\mathrm{m^2 s^{-2} kg^{-1}}$). In this particular case, these two forces, being exactly of the same magnitude and acting in opposite directions, will perfectly cancel one another—the distance, r, from sun to planet will not change, and the orbit is a circle with the sun at the center:

$$mr\omega^2 = G\frac{Mm}{r^2}, \tag{12.33}$$

from which

$$r = \frac{GM}{\omega^2}^{1/3}. \tag{12.34}$$

For the case of the sun whose mass is $M = 1.991 \times 10^{30}$ kg and the Earth whose angular velocity is $2\pi/365.24$ radians per day or 199.1×10^{-9} radians per second, the value of r comes out at 149.6×10^9 m. This is almost precisely correct, even though we know that the orbit of Earth is not circular.

In fact, it would be improbable that a planet had exactly the correct angular velocity to be in a circular orbit. More likely, its velocity may be, say, somewhat too small, so that the planet would fall toward the sun, thereby accelerating and reaching a velocity too big for a circular orbit and, thus, would fall away from the sun decelerating, and so on—it would be in an elliptical orbit.

The orbital equation can be easily derived, albeit with a little more math than in the circular case.

We note that the equations of motion (in polar coordinates) are

$$m\left[-\frac{d^2r}{dt^2} + r\left(\frac{d\theta}{dt}\right)^2\right] = G\frac{mM}{r^2} \tag{12.35}$$

and

$$m\left(2\frac{dr}{dt}\frac{d\theta}{dt} + r\frac{d^2\theta}{dt^2}\right) = 0 \tag{12.36}$$

for, respectively, the radial and the tangential components of force.

Equation 12.36 can be multiplied by r/m yielding

$$2r\frac{dr}{dt}\frac{d\theta}{dt} + r^2\frac{d^2\theta}{dt^2} = 0. \tag{12.37}$$

Note that Equation 12.35 becomes Equation 12.33 for the circular case (in which d^2r/dt^2 is zero). In this case Equation 12.36 reduces to $d^2\theta/dt^2 = 0$ (because dr/dt is also zero) showing, as is obvious, that the angular velocity is constant.

Note also that

$$\frac{d}{dt}\left(r^2\frac{d\theta}{dt}\right) = 2r\frac{dr}{dt}\frac{d\theta}{dt} + r^2\frac{d^2\theta}{dt^2} = 0. \tag{12.38}$$

Integrating Equation 12.38,

$$r^2\frac{d\theta}{dt} = \text{constant}. \tag{12.39}$$

It is possible to show that Equations 12.35 and 12.36 represent, when t is eliminated between them, an equation of a conical section of eccentricity, ϵ (which depends on the total energy of the "planet"). When $\epsilon < 1$, the path or orbit is an ellipse described by

$$r = \frac{a(1-\epsilon^2)}{1+\epsilon\cos\theta}. \tag{12.40}$$

In the preceding, a is the semi-major axis and θ is the true anomaly. Given the position, θ, of the planet in its orbit, it is easy to calculate the radius vector, provided the major axis and the eccentricity are known.

Of greater interest is the determination of the position, θ, as a function of time. From Equations 12.39 and 12.40,

$$\frac{d\theta}{dt} \propto \frac{(1+\epsilon\cos\theta)}{a^2(1-\epsilon^2)^2} = A(1+\epsilon\cos\theta)^2. \tag{12.41}$$

If the orbit of Earth were circular, that is, if $\epsilon = 0$, then the angular velocity, $d\theta/dt$, would be constant;

$$\frac{d\theta}{dt} = \frac{360°}{365.2422\text{ days}} = 0.98564733 \quad \text{degrees/day}. \tag{12.42}$$

This is the mean angular velocity of Earth. Thus,

$$A = 0.98564733 \quad \text{degrees/day}. \tag{12.43}$$

and because the eccentricity of Earth is, at present, approximately 0.01670,

$$\frac{d\theta}{dt} = 0.98564733(1+0.0167\cos\theta)^2. \tag{12.44}$$

Consequently, when $\theta = 0$, that is, at the moment of perihelion passage, the angular velocity of Earth is 1.01884 degrees/day, while at aphelion the angular velocity is down to 0.95300 degrees/day.

A description of the motion of Earth—a table of θ as a function of time—can be obtained by evaluating

$$\theta_i = \theta_{i-1} + A(1+\epsilon\cos\theta_{i-1})^2\Delta t. \tag{12.45}$$

θ_0 is made equal to zero, corresponding to the perihelion.

Good accuracy is achieved even when the time increment, Δt, is as large as 1 day.

It turns out that the perihelion is not a very useful point of reference. It is more convenient to use a more obvious direction, the vernal equinox, as the initial point. For this, an **ecliptic longitude** is defined as the angle, measured along the ecliptic plane, *eastward* from the vernal equinox. The tabulation resulting from Equation 12.45 can be used, provided the **longitude of the perihelion** also called the **argument of perihelion**, is known. This quantity varies slowly but is, at present, close to −77°.

The heliocentric polar coordinate system used to derive the orbital motion of Earth is not very convenient for describing the position of a celestial body as seen from Earth. For this, the latitude/longitude system used in geography can easily be extended to astronomy. The various positions are described by a pair of angles: the **right ascension** equivalent to longitude and the **declination** equivalent to latitude. We are back to the old geocentric point of view, although we recognize, as Aristarchus of Samos did back around 200 B.C., that we are not the center of the universe. For such a system, the reference is the spin axis of the planet, a direction perpendicular to the **equatorial plane**, which passes through the center of Earth. Only by extreme coincidence would the equatorial plane of a planet coincide with the **ecliptic** (the plane that contains the orbit of the planet). Usually, these two planes form an angle called **obliquity** or **tilt angle**, τ. It is possible to define a **celestial equator** as a plane parallel to the terrestrial equator but containing the center of the sun, not that of the Earth.

Celestial equator and ecliptic intersect in a line called the **equinoctial line**. When Earth crosses this line coming from south to north (**the ascending node**), the **vernal equinox** occurs. At the descending node, when Earth crosses the line coming from the north, the **autumnal equinox** occurs.

The vernal equinox is used as a convenient origin for the measurement of both the ecliptic longitude and the right ascension. Remember that the ecliptic longitude is an angle lying in the ecliptic plane, while the right ascension lies in the equatorial plane.

The time interval between two consecutive vernal equinoxes is called the tropical year, referred to at the beginning of this appendix. In our derivation, we used the perihelion as the origin for measuring the true anomaly. The time interval between two consecutive perihelion passages is called the **anomalistic year**, which, surprisingly, is slightly longer than the tropical year. How can that be? The reason for this discrepancy is that the line of apsides slowly changes its orientation, completing 360° in (roughly) 21,000 years. The corresponding annual change in the longitude of the perihelion is $360/21{,}000 = 0.017$ degree per year.

Since the orbital angular velocity of Earth is roughly 1 degree per day, this means that the anomalistic year will be about 0.017 day (25 minutes) longer than the tropical one, which is 365.242 days long. Hence, the anomalistic year should be about 365.259 days long (more precisely, 365.25964134 days). The relation between the perihelion and the equinoctial time is illustrate in Figure 12.15.

Relationship between Ecliptic and Equatorial Coordinates

Celestial longitude is measured along the ecliptic, while right ascension—which is also a measure of longitude—is measured along the equatorial plane. Clearly, a simple relationship must exist between these coordinates.

Consider a right-handed orthogonal coordinate system with the center of Earth at the origin, in which the x-y plane coincides with the equatorial plane and the y-axis is aligned with the equinoctial line. The z-axis points north.

Let $\vec{s}$ be a unit vector, starting from the origin and pointing toward the sun,

$$\vec{s} = \vec{i}s_x + \vec{j}s_y + \vec{k}s_z. \tag{12.46}$$

This vector must lie in the ecliptic plane—that is, it must be perpendicular to the spin axis whose unit vector is

$$\vec{u} = -\vec{i}\sin\tau + \vec{k}\cos\tau. \tag{12.47}$$

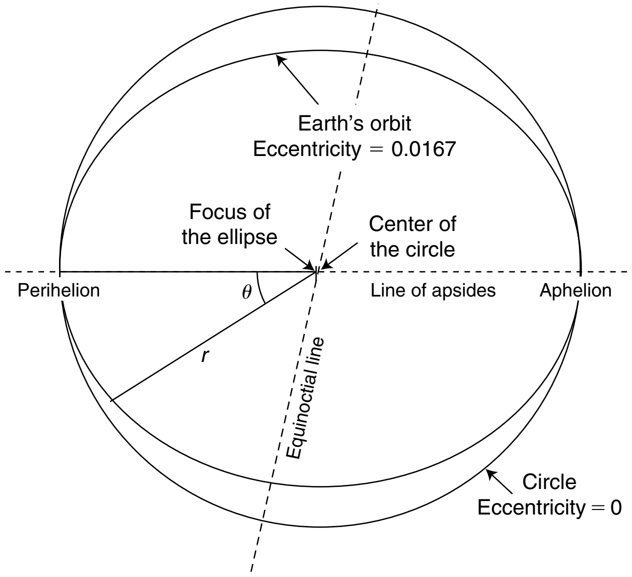


Figure 12.15 The orbit of Earth is slightly elliptical. The view is normal to the ecliptic.

If $\vec{s}$ is perpendicular to $\vec{u}$, then their dot product must be zero:

$$\vec{s}\cdot\vec{u} = 0 = -s_x \sin\tau + s_z \cos\tau, \tag{12.48}$$

$$s_z = s_x \tan\tau. \tag{12.49}$$

The reference for measuring longitude and right ascension is the direction of the vernal equinox, which, in our coordinate system has a unit vector $-\vec{j}$. Hence, the longitude, Λ, is given by

$$\cos\Lambda = -s_y. \tag{12.50}$$

Since $\vec{s}$ is a unit vector,

$$s_x^2 + s_y^2 + s_z^2 = s_x^2 + \cos^2\Lambda + s_x^2 \tan^2\tau = 1, \tag{12.51}$$

$$s_x^2(1 + \tan^2\tau) = 1 - \cos^2\Lambda, \tag{12.52}$$

$$s_x = \sin\Lambda\cos\tau. \tag{12.53}$$

But on the equatorial plane, the right ascension, $\Re$, is given by

$$\tan\Re = \frac{s_x}{-s_y} = \frac{\sin\Lambda\cos\tau}{\cos\Lambda} = \cos\tau\tan\Lambda. \tag{12.54}$$

$$\Re = \arctan(\cos\tau\tan\Lambda). \tag{12.55}$$

As usual, when one takes reverse trigonometric functions, the answer is ambiguous—the calculator or the computer gives only the principal value, and a decision must be made of what the actual value is. In the present case, the following Boolean statement must be used:

$$\begin{gathered}\text{If } \Lambda >= 90^\circ \text{ AND } \Lambda < 270^\circ \text{ THEN } \Re = \Re + 180^\circ \\ \text{ELSE IF } \Lambda >= 270^\circ \text{ THEN } \Re = \Re + 360^\circ.\end{gathered} \tag{12.56}$$

This relates the longitude of the sun to its right ascension.

The declination of the sun is

$$\sin\delta = s_z = s_x\tan\tau = \sin\Lambda\cos\tau\tan\tau = \sin\Lambda\sin\tau. \tag{12.57}$$

The Equation of Time

We now need to explain why the time between successive solar culminations varies throughout the year:

Assume you have on hand two instruments:

1. An accurate clock that measures the uniform flow of time calibrated in mean solar time. From one vernal equinox to the next, it registers the passage of $365.2422 \times 24 \times 60 \times 60 = 31.55692 \times 10^6$ seconds.

2. A sundial that can measure the solar time with a resolution of at least one minute.[†]

Set your clock to start at exactly noon (as seen in the sundial) on any arbitrary date. It will be noted that, although by the same date of the next year, the clock and the sundial are again synchronized, throughout the year, the sundial seems sometimes to be slow and, at other periods of the year, to be fast. The difference between the sundial time and the clock time—between the solar time and the mean solar time—may reach values of up to 15 minutes fast and 15 minutes slow. This difference is called the **Equation of Time (EOT)**.

Figure 12.19 shows how the EOT varies along the year. For planning solar collectors, it is sufficient to read the value of the EOT off the figure. For such an application, the empirical formula below (Equation 12.59) is an overkill.[††] It yields the EOT in minutes when the day of the year, d, is expressed as an angle, d_{deg}:

$$d_{deg} \equiv \frac{360}{365} d \quad \text{degrees.} \tag{12.58}$$

Greater precision may be useless because the EOT varies somewhat from year to year with a four-year period, owing to the leap years.

$$\begin{aligned} EOT = & -0.017188 - 0.42811 \cos(d_{deg}) + 7.35141 \sin(d_{deg}) \\ & + 3.34946 \cos(2d_{deg}) + 9.36177 \sin(2d_{deg}) \quad \text{minutes} \end{aligned} \tag{12.59}$$

We recall that there are two measures of solar longitude, both increasing eastward from the direction of the vernal equinox: one measure is along the ecliptic and is called the ecliptic longitude; Λ, the other, is measured along the equatorial plane and is called the right ascension, $\Re$. These two longitudes are exactly the same at the equinoxes and at the solstices, but are different anywhere in between. See Table 12.4.

The discrepancy between Λ and $\Re$ increases with the obliquity of the orbit. If the obliquity of Earth's orbit were zero, then $\Lambda = \Re$ under all circumstances.

Assume that you are on the surface of Earth at an arbitrary latitude but on a meridian that happens to be the one at which the sun culminates at the exact moment of the vernal equinox. At this moment $\Lambda = \Re = 0$.

Then, $23^{\text{h}} : 56^{\text{m}} : 04^{\text{s}}.09$ or 23.93447 h later, Earth has completed a full rotation and your meridian faces the same direction it did initially.

[†] Owing to the finite angular diameter of the sun (0.5°), it is difficult to read a sundial to greater precision than about 1 minute of time.

[††] http://www.srrb.noaa.gov/highlights/sunrise/program.txt. is the source of the NOAA formula mentioned.

Table 12.4 Difference between Longitude and Right Ascension

Season	$\mathbf{\Lambda - \mathfrak{R}}$
Spring	> 0
Summer	< 0
Autumn	> 0
Winter	< 0

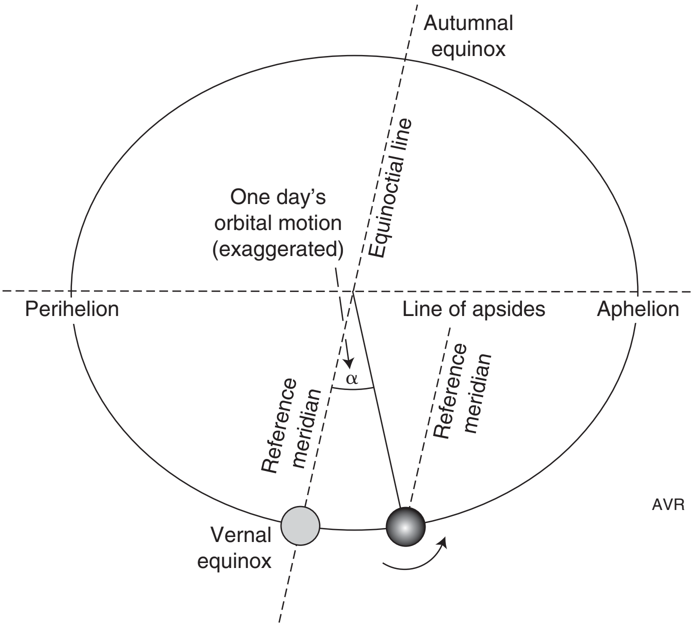


Figure 12.16 As the Earth moves in its orbit, the sun appears to move eastward.

However, the sun is not exactly at culmination. The reason is that the orbital motion of Earth has caused an apparent eastward motion of the sun. The Earth has to spin another α degrees for the reference meridian to be facing the sun again. See Figure 12.16.

If the orbit were circular and the obliquity zero, then the uniform eastward motion of the sun would be $360°/365.2422 = 0.985647$ degrees/mean solar day. In more technical terms, the rate of change of the anomaly would be constant and the **mean anomaly** would be

$$<\theta> = 0.985647t, \tag{12.60}$$

where t is the time (in mean solar days) since perihelion passage. The ecliptic longitude changes at the same rate as the anomaly. Since the spin rate of Earth is 360° in 23.93447 hours or 15.04107°/hour, or 0.2506845° per minute, it takes 0.985647/0.2506845 = 3.931823 minutes (0.06553 hours) to rotate the 0.985647° needed to bring the reference meridian once again under the sun—that is, to reach the next culmination or next noon. Not surprisingly, 23.93447 + 0.06553 = 24.00000 hours. This means that in this simple case, consecutive noons are evenly spaced throughout the year and occur exactly 24 hours apart, a result of the definition of the solar mean hour.

In reality, noons do not recur at uniform intervals—that is, the sun dial time does not track exactly the clock time. Twenty-four hours between noons is only the value averaged over one year. The difference between the clock time and the sun dial time, as stated before, is called the equation of time, EOT. Two factors cause this irregularity: the eccentricity of the orbit (leading to $EOT_{eccentricity}$) and the obliquity of the orbit (leading to $EOT_{obliquity}$).

Orbital Eccentricity

If the orbit is not circular, the rate of solar eastward drift (the rate of change of the anomaly) is not constant. It changes rapidly near perihelion and more slowly near aphelion. Hence, the true anomaly differs from the mean anomaly so that, after each 24 mean solar hours period, the Earth has to spin an additional $\theta - <\theta> \equiv C$ degrees. Here the difference, C, between the true anomaly and the mean anomaly is called the **equation of center**, another example of medieval terminology.

Since the spin rate of Earth is very nearly 1 degree in 4 minutes of time, the time offset between the true noon and the mean noon owing to the eccentricity of the orbit is

$$EOT_{eccentricity} = 4C \text{ minutes of time.} \tag{12.61}$$

In the preceding, the angle, C, is in degrees.

The eccentricity component of the equation of time varies throughout the year in a sinusoidal fashion with zeros at the perihelion and aphelion and with extrema of 8 minutes of time midway between these dates. Figure 12.17 shows these variations.

The value of C for any time of the year can be found by calculating θ using Equation 12.45 and subtracting the mean anomaly obtained from Equation 12.60. For many applications, it may prove more practical to calculate C using the following empirical equations:[†]

$$<\theta>= 357.52911 + 35999.05029T - 0.0001537T^2 \tag{12.62}$$

[†] The formulas are from a NOAA program: http://www.srrb.noaa.gov/highlights/sunrise/program.txt.

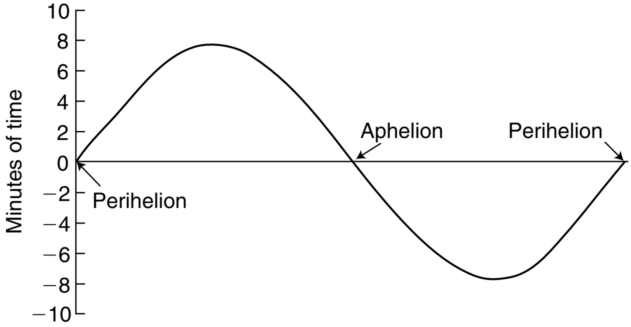


Figure 12.17 That part of the equation of time resulting from the ellipticity of Earth's orbit has a value (in minutes of time) equal to $4C$, where C is the equation of center—that is, the difference between the true anomaly and the mean anomaly. Observe that this function has a zero at both perihelion and aphelion.

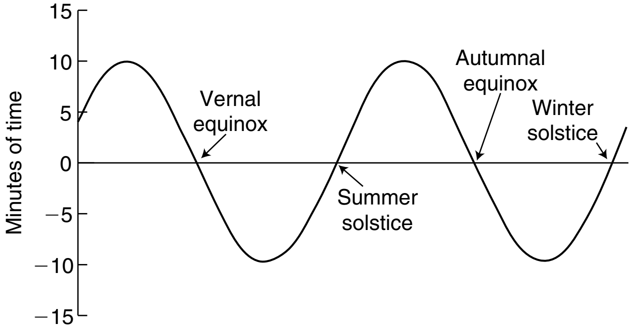


Figure 12.18 The part of the equation of time resulting from the obliquity of Earth's orbit has a value (in minutes of time) equal to $4(\Lambda - \Re)$—that is, four times the difference between the solar longitude and its right ascension. Zeros occur at the equinoxes and solstices.

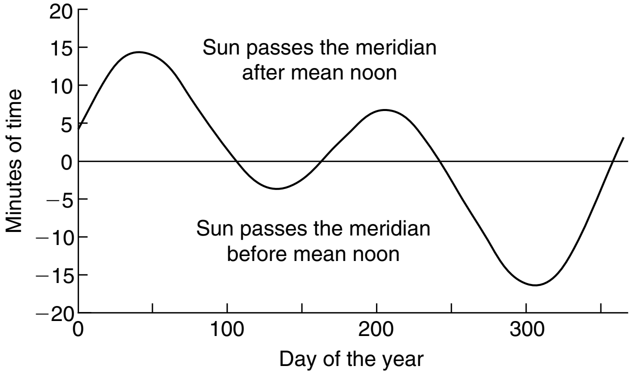


Figure 12.19 The observed equation of time is the combination of the effects owing to ellipticity and obliquity of Earth's orbit.

$$C = (1.914602 - 0.004817T - 0.000014T^2)\sin <\theta> \\ + (0.019993 - 0.000101T)\sin(2 <\theta>) \\ + 0.000289\sin(3 <\theta>) \text{ degrees.} \quad (12.63)$$

Here, T, is the number of Julian centuries since January 1, 2000. See farther the explanation of Julian dates a few pages back.

12.4.4 Orbital Obliquity

If the orbit is circular but the obliquity is not zero, then although the rate of solar ecliptic longitude increase (or the rate of the anomaly increase) is constant, the rate of change of the right ascension is not. The moment of solar culmination is related to the right ascension.

On the day after the vernal equinox, as seen by an observer on the reference meridian on Earth, the sun is at a right ascension of

$$\Re = \arctan(\cos\tau\tan\Lambda) = \arctan(\cos(23.44°)\tan(0.985647°)) \\ = \arctan(0.91747 \times 0.017204) = \arctan\big(0.015785\big), \quad (12.64)$$

$$\Re = 0.904322°. \quad (12.65)$$

For the sun to culminate, the Earth has to spin an additional 0.904322° rather than 0.985647° as in the case of zero obliquity. Thus, noon will occur somewhat earlier than in the zero obliquity situation. In fact, it will occur $4 \times (0.985647 - 0.904322) = 0.325$ minutes earlier.

Generalizing,

$$EOT_{obliquity} = 4(\Lambda - \Re). \quad (12.66)$$

The obliquity component varies, as does the eccentricity one, in a sinusoidal fashion but completes two cycles rather than one in the space of one year. The zeros occur at the equinoxes and solstices instead of at the perihelion and aphelion as they do in the eccentricity case. The amplitude of $EOT_{obliquity}$ is 10 minutes. This behavior is depicted in Figure 12.18, while the behavior of the full EOT (the sum of the eccentricity and the obliquity components) is depicted in Figure 12.19.

It is important not to confuse perihelion (when the Earth is closest to the sun) or aphelion (when it is farthest) with the solstices which occur when the solar declination is an extremum—that is, when $\delta \pm 23.44°$. As it happens, the dates of the solstices are near those of perihelion and aphelion, but this is a mere coincidence. There are roughly 12 days between the (summer) solstice and the aphelion and between the (winter) solstice and the perihelion. See Table 12.5.

Table 12.5 Dates of Different Sun Positions

	Approx. Date	**Approx Day No.**
Perihelion	January 2–4	2
Vernal equinox	March 20–21	80
Solstice (summer, n. hemisphere)	June 21	172
Aphelion	July 4–5	184
Autumnal equinox	September 22–23	262
Solstice (winter, n. hemisphere)	December 21–22	355

References

Johnson, Francis S., *Solar radiation, in Satellite Environment Handbook (second edition)*, editor Francis S. Johnson, Stanford University Press, **1965**.

Duffie, John A., and William A. Beckman, *Solar Energy Thermal Processes*, John Wiley, **1974**.

Vitruvius, Marcus Vitruvius Pollio, The *Ten Books on Architecture*, translated by Morris Hicky Morgan, *Dover*, New York, **1960**.

Welford, W. T., and R. Winston, *The Optics of Non-imaging Concentrators*, Academic Press, **1978**.

Wilkins, E. S., et al., Solar gel ponds, *Science 217*, p. 982, September 10, **1982**.

Further Reading

Dohrn-van Rossum, Gerhard, *History of the Hour: Clocks and Modern Temporal Orders*, University of Chicago Press, **1996**.

PROBLEMS

12.1 A time traveler finds himself in an unknown place on Earth at an unknown time of the year. Because at night the sky is always cloudy, he cannot see the stars, but he can accurately determine the sunrise time as well as the length of a shadow at noon. The sun rises at 0520 local time. At noon, a vertical mast casts a shadow 1.5 times longer than its height. What is the date, and what is the latitude of the place? Is this determination unambiguous?

12.2 An astronaut had to make an emergency de-orbit and landed on an island in the middle of the ocean. He is completely lost but has an accurate digital watch and a copy of *Fundamentals of Renewable Energy Processes*, which NASA always supplies.[†] He carefully times the length of the day and discovers that the time between sunrise and sunset is 10:49:12. He knows the date is January 1, 1997. He can now figure his latitude. Can you?

12.3 A building in Palo Alto, California (latitude 37.4 N) has windows facing SSE. During what period of the year does direct light from the sun enter the window at sunrise? Assume no obstructions, good weather, and a vanishing solar diameter.

What is the time of sunrise on the first day of the period? And on the last day? What is the insolation on the SSE-facing wall at noon at the equinoxes?

12.4 Consider an ideal focusing concentrator. Increasing the concentration causes the receiver temperature to increase—up to a point.

Beyond certain maximum concentration, the temperature remains constant. What is the maximum concentration that can be used on Mars for a 2-D and for a 3-D case? Some data:

Radius of the orbit of Mars is 1.6 AU.

1 AU is 150 million km.

The angular diameter of the moon, as seen from Earth, is 0.5 degrees.

12.5 Consider the arbitrary distribution function

$$\frac{dP}{df} = f - \frac{1}{2}f^2.$$

Determine for what value of f this function is a maximum.
Plot dP/df as a function of f for the interval in which $dP/df > 0$.
Now define a new variable, $\lambda \equiv c/f$, where c is any constant.
Determine for what value of f the distribution function $dP/d\lambda$ is a maximum.
Plot $|dP/d\lambda|$ as a function of f.

[†] Just kidding!

12.6 An expedition to Mars is being planned. Let us make a preliminary estimate of the energy requirements for the first days after the expedition lands.

Landing date is November 15, 2007, which is Mars day 118. At the landing site (17.00° N, 122° E) it will be just after local sunrise. The five-person landing team has all of the remaining daylight hours to set up the equipment to survive the cold night.

Prior to the manned landing, robots will have set up a water extraction plant that removes the liquid from hydrated rocks by exposing them to concentrated sunlight. Assume an adequate (but not generous) supply of water.

Power will be generated by photovoltaics and will be stored in the form of hydrogen and oxygen obtained from water electrolysis.

The photovoltaics are blankets of flexible material with 16.5% efficiency at 1 (Mars) sun. No concentrators will be used. These blankets will be laid horizontally on the Mars surface.

The electrolyzers operate at 90% efficiency.

MARS DATA (relative to Earth)

Mean radius of orbit	1.52
Gravity	0.38
Planetary radius	0.53
Length of day	1.029
Length of year	1.881
Density	0.719

The inclination of the plane of the Martian equator referred to the plane of its orbit is 25.20°.

The inclination of the plane of the Martian orbit referred to the ecliptic is 1.85°.

The average daytime Martian temperature is 300 K (just a little higher than the average daytime Earth temperature of 295 K). The Martian night, however, is cold! Average temperature is 170 K (versus 275 K, for Earth).

The vernal equinox occurs on Mars day 213.[†]

Define a Martian hour, h_m, as 1/24 of the annual average of the period between consecutive sunrises.

1. How long does the sunlit period last on the day of arrival?
2. Determine the available insolation on a horizontal surface (watts m^{-2}, averaged over a Martian day—that is, over a 24 h_m period).

[†] Not really. Although the other data are accurate, this date was pulled out of a hat.

3. Estimate the O_2 consumption of the five astronauts. Consider that they are on a strictly regulated diet of 2500 kilocalories per Martian day. Assume that all this is metabolized as glucose. Use 16 MJ per kg of glucose as combustion enthalpy.
4. How much energy will be required to produce the necessary amount of oxygen from water by electrolysis?
5. What area of solar cell blanket must be dedicated for the production of oxygen?
6. Assume that the mean temperatures of Mars are the actual temperatures at the astronaut's settlement. Assume further that the temperature of the Martian air falls instantaneously from its daytime 300 K to its nighttime 175 K and vice versa.

   The astronauts are housed in a hemispherical plastic bubble 10 m in diameter. The wall material is rated at R-12 (in the American system) as far as its thermal insulation is concerned. No heat is lost through the floor.

   The interior of the bubble is kept at a constant 300 K. During the night, stored hydrogen has to be burned to provide heat. The inside wall of the bubble is at 300 K, while the outside is at 175 K. Assume that the effective emissivity of the outside surface is 0.5.

How much hydrogen is needed per day? Express this in solar cell blanket area.

12.7 How long was the shadow of a 10-m tall tree in Palo Alto, California (37.4 N, 125 W) on March 20, 1991 at 0200 PM (PST)? Desired accuracy is ± 20 cm.

12.8 You are on a windswept plain on the planet Earth, but you know neither your position nor the date. There are no hills, and you can see the horizon clearly. This allows you to time the sunset accurately, but unfortunately your watch reads Greenwich time. Take a straight pole and plant it exactly vertically in the ground (use a plumb line). You have no meter- or yardstick, but you assign arbitrarily 1 unit to the length of the pole above ground. Observe the shadow throughout the day: at its shortest (at 08:57 on your watch) the shadow is 2.50 units long. Sun sets at 13:53. From these data alone, determine the date, latitude, and longitude unambiguously (i.e., decide if you are in the northern or the southern hemisphere).

12.9 Calculate the azimuth of a vertical surface that results in the maximum annual average relative insolation under the conditions below:

The surface is situated at latitude of 40° N in a region in which there is always a dense early morning fog (insolation = zero) up to 10:00 and then the rest of the day is perfectly clear.

Relative insolation is defined here as the ratio of the insolation on the surface in question to that on a *horizontal* surface at the equator on the same day of the year.

12.10 At what time of day is the sun due east in Palo Alto?

The latitude of Palo Alto is 37.4° north; the longitude is 122° west of Greenwich.

Do this for two days of the year: August 11 and November 15.

12.11 What is the insolation (W m^{-2}) on a surface facing true east and with 25° elevation erected at a point 45° N at 10:00 local time on April 1, 1990?

Assume that the insolation on a surface facing the sun is 1000 W m^{-2}.

12.12 The average food intake of an adult human is, say, 2000 kilocalories per day. Assume that all this energy is transformed into a constant uniform heat output.

Ten adults are confined to a room 5 × 5 × 2 meters. The room is windowless and totally insulated with R-11 fiberglass blankets (walls, door, ceiling, and floor). Outside the room the temperature is uniformly at 0 C. The air inside the room is not renewed. Assume that temperature steady state is achieved before the prisoners suffocate. What is the room temperature?

12.13 Here is a way in which a person lost in a desert island can determine her or his latitude with fair precision, even though the nights are always cloudy (Polaris cannot be seen).

Use a vertical stick and observe the shadow. It will be shortest at local noon. From day to day, the shortest noon shadow will change in length. It will be shortest on the day of the summer solstice: this will tell you the solar declination, δ.

On any day, comparing the length of the horizontal shadow with the length of the stick will let you estimate the solar zenithal angle, χ, quite accurately, even if no yard- (or meter-) stick is on hand.

If we know both δ and χ, it is possible to determine the latitude. Or is it?

Develop a simple expression that gives you the latitude as a function of the solar declination and the noon zenithal angle. No trigonometric functions, please! Is there any ambiguity?

12.14 As determined by a civilian-type Trimble GPS receiver, the latitude of my house in Palo Alto, California, is 34.44623° N.

At what time of the year are consecutive sunrises exactly 24 hours apart?

Call Δt the time difference between consecutive sunrises. At what time of the year is Δt a maximum?

What is the value of this maximum Δt? Express this in seconds.

12.15 An explorer in the Arctic needs to construct a cache in which to store materials that cannot stand temperatures lower than −10 C.

The surface air over the ice is known to stay at −50 C for long periods of time.

The cache will be built by digging a hole in the ice until the bottom is 0.5 meter from the water. The roof of the cache is a flat surface flush with the ice surface. It will be insulated with a double fiberglass blanket, each layer rated at R-16 (in the American system).

The heat conductivity of ice is 1.3 W $\mathrm{m^{-1}K^{-1}}$.

Assume that the area of the cache is so large that the heat exchange through the vertical walls can be neglected.

Estimate the temperature inside the cache when the outside temperature has been a steady −50 C for a time sufficiently long for the system to have reached steady state. Assume that the temperature inside the cache is uniform.

12.16 What is the solar azimuth at sunset on the day of the summer solstice at 58° north?

12.17 A battery of silicon photocells mounted on a plane surface operates with an efficiency of 16.7% under all conditions encountered in this problem. It is installed at a location 45° north. The time is 10:00 on April 1, 1995.

When the battery faces the sun directly, it delivers 870 W to a resistive load. How much power does it deliver if the surface is set at an elevation of 25° and faces true east?

12.18 Consider a mechanical heat pump whose coefficient of performance is exactly half of the ideal one. It uses an amount W of mechanical energy to pump Q_C units of energy from an environment at −10 C into a room at 25 C where an amount $Q_A = Q_C + W$ of energy is deposited. How many joules of heat are delivered to the room for each joule of mechanical energy used?

12.19 You have landed on an unknown planet, and, for some obscure reason, you must determine both your latitude and the angle, ι, of inclination of the planet's spin axis referred to the its orbital plane.

To accomplish such a determination, all you have is a ruler and plenty of time. Erect a vertical pole exactly 5 m tall and observe the length of the shadow at noon as it varies throughout the year.

The shortest length is 1.34 m, and the longest is 18.66 m.

12.20 The smallest zenithal angle of the sun was, on January 1, 2000, 32.3°. At that moment, the sun was to your south. What is your latitude?

12.21 You are 30° north. The day is September 15 of a nonleap year. It is 12:44 true solar time. A 10-m-long rod tilted due west is planted in the ground making a 30° angle with the vertical.

Calculate the position of the sun.

What is the length of the shadow of the rod?

12.22 On which day of the year does the sun appear vertically overhead at the three locations below. Determine the time (in standard time) in which this phenomenon occurs. Disregard the Equation of Time. Also assume that the time zones are those dictated by purely geographic considerations, not by political ones.
Locations:

Palo Alto, California (USA): 37° 29′ N, 122° 10′ W.

Macapá, Amapá (Brazil): 0.00, 51° 07′ W.

Brasília, Federal District (Brazil): 15° 53′ S, 47° 51′ W.

12.23 If you consult the URL http://mach.usno.navy.mil/ and follow the instructions, you will find that for San Francisco, California (W122.4, N37.8), on February 19, 2002, the sunrise occurred at 06:55 and the sun crossed the meridian at 12:24.

1. Using the information in the textbook, verify the meridian crossing time.
2. Still using the formulas in the textbook, calculate the sunrise time. If there is any discrepancy, indicate what causes it.

12.24 Aurinko (pronounced OW-rin-ko, where the stressed syllable, OW, has the same sound as in "how") is a (fictitious) unmanned airplane designed to serve as a repeater for radio signals replacing expensive satellites. It is equipped with 14 electric motors, each of which can deliver 1.5 kW. Its cruising speed is 40 km/hr (slightly faster than a man can run in a 100- or 200-m dash). It operates at 30 km altitude.

Wingspan: 75.3 m

62,120 solar cells

Maximum electric output of the cells: 32 kW when full sunlight is normal to the cells

1. When Aurinko is in its orbit (30 km above sea level), how far is its geometric horizon?[†] The geometric horizon differs from the radio horizon because the radio horizon is somewhat extended by atmospheric refraction.
2. What is the area on the ground that is reached by the direct rays from Aurinko. Disregard atmospheric refraction.
3. Assume that the airplane orbits a point 37.8° north. On the day in which the sunlight hours are the least, how long is the sunlit period as seen from the airplane at 30-km altitude? Disregard atmospheric refraction.

[†] This is not a space orbit around a planet. It is an orbit (as the term is used in aviation) around a point 30 km above sea level.

4. Assuming that the solar cells mounted on top of the wing of the airplane are always horizontal, what is the insolation averaged over the sunlit period on the day of the previous question? Since the airplane is above most of the atmosphere, the solar constant can be taken as 1200 W/m^2.

5. Assume that the total power consumption of the airplane while in orbit is 10 kW. (This includes propulsion, housekeeping, and communications.) The electric energy obtained from the photovoltaic array is in part directly used by the load, and the excess is stored to be used during the period when the array output is less than the load demand. The storage system has a turnaround efficiency of $\eta_{turnaround}$—that is, only a fraction, $\eta_{turnaround}$, of the energy fed into the system can be retrieved later.

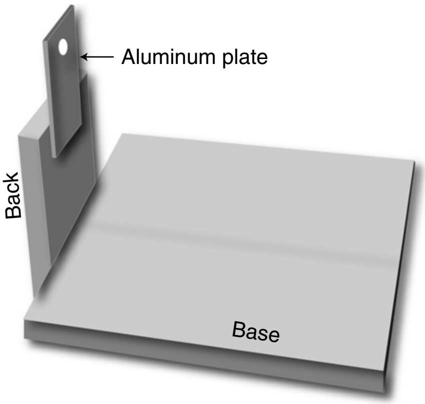
   Assume, for simplicity, that $\eta_{turnaround} = 1.0$. Assume also that the efficiency of the photovoltaic collectors is 20% independently of the sun power density.

   The solar array covers all the wing surface except for a rim of 20 cm. More clearly, there is a space of 20 cm between the leading edge and the array and the same space at the trailing edge and the wing tips. Consider rectangular (nontapered) wings.

6. What must the chord of the wing be (chord = distance between leading and trailing edges)?

12.25 In the solar spectrum, what fraction of the total power density is absorbed by silicon? *Hint:* Use Table 12.1.

12.26 This is an experiment that is technically easy to carry out. Unfortunately, it takes 365 days to complete. Hence, we are going to invert the procedure, and, from the theory developed in the textbook, we will calculate what the results of the experiment would be. If you have not been exposed to it, you may be surprised by what you get.



To set up the experiment, you would have to build the simple device illustrated in the figure. It consists of a wooden base large enough to support a standard piece of paper. A vertical piece of wood ("back") is attached to the base, and mounted on it is a thin aluminum rectangle with a small hole in it. The aluminum must be thin enough (say, 1 mm or less) so that the sun can shine through the hole even when it is far from perpendicular to the plate.

In the model we used, the hole was 129 mm above the base, but this is not a critical dimension. Orient the device so that the hole faces equatorward. The noon sun will cast a shadow, but shining through the hole will cause a little dot of light to appear on the base. What we want to do is to follow the path this dot of light will trace out. The exact time in which the observations are made is important. You have to start at the astronomical noon on one of the following days (in which the equation of time, EOT, is zero): April 15, June 14, September 2, or December 25.

To follow the dot of light, place a sheet of paper on the base. Immobilize it by using some adhesive tape. With a pen, mark the center of the dot of light at the moment the sun culminates. Since on the suggested starting dates the equation of time is zero, the sun will culminate (i.e., cross the local meridian) at 1200 Standard Time corrected by the longitude displacement. This amounts to 4 minutes of time for each degree of longitude away from the longitude of the meridian at the center of the time zone. The center of the time zone is (usually) at exact multiples of 15° of longitude. For instance, the Pacific Time Zone in North America is centered on 120° W. If you happen to be in Palo Alto (125° W, your time offset (on days when the EOT is zero) is $15° \times (120° - 125°) \times 4 = -20$ minutes. Thus, the sun will culminate at 1220 PST.

The next observation must be made 240 hours (or an exact multiple of 24 hours) late—that is, if you are in Palo Alto, it must be made exactly at 1220 PST. After one year, an interesting pattern will emerge.

The assignment in this problem is to calculate and plot the position of the light dot every 10 days over a complete year. Start on April 15 and do not forget the equation of time. Assume that you are in Palo Alto, California: 125° W 37.4° N.

12.27 The time interval between two consecutive full moons is called **a lunar month**, or a **lunation**, and lasts, on average, 29.53 days. Explain why this is not the **length of the lunar orbital period**—that is, the time it takes for the moon to complete one full orbit around the Earth.

Calculate the length of the lunar orbital period.

12.28 A supraluminar (faster than light) probe was sent to reconnoiter an Earth-like extrasolar planet whose characteristics are summarized in the accompanying table. All units used are terrestrial units, except as noted here. It was determined that the planet had a very tenuous atmosphere, totally transparent to the solar radiation. As the probe materialized in the neighborhood of the alien sun, it measured the light power density, which was 11 kW/m^2. The measurement was made when the probe was at exactly 50 million km from the sun. It was also determined that the planet was in an essentially circular orbit.

"Day" is defined as the time period between consecutive solar culminations (noons). It differs from the terrestrial day.

"Year" is the length of the orbital period—the time period between successive vernal equinoxes. It differs from the terrestrial year.

The "date" is designated by the day number counting from Day Number 1—the day of the vernal equinox.

Orbital radius	132×10^6	km
Orbital period	28.98×10^6	sec
Orbital inclination	26.7	degrees
Planetary radius	5800	km
Spin period	90,000	sec

The direction of the spin is the same as the direction of the orbital motion.

The probe will land at a point with a latitude of +45°.

1. At the probe landing site, what is the length, in seconds, of the longest and of the shortest daylight period of the year?
2. What are the dates of the solstices?
3. What is the daily average insolation on a horizontal surface on day 50?

Chapter 13
Biomass

13.1 Introduction

Wikipedia, the free online encyclopedia, defines **biomass** as “all nonfossil material of biological origin.” We will focus almost exclusively on vegetable matter that can lead to the production of useful energy, thus excluding most animal biomass.[†]

Any biomass-based energy process begins with the capture of sunlight and production of a chemical compound. This complicated step, called **photosynthesis**, leads basically to glucose. Subsequent biochemical transformations result in the creation of a very large number of compounds, some of very great commercial value.

At best, photosynthesis proceeds with efficiencies of less than 8%. By the time the final product is available for consumption, large chunks of energy have been spent in cultivation, fertilizing, harvesting, transporting the raw biomass, removing the excess water, and extracting the desired fuel. The overall efficiency is usually a fraction of 1%. This is a prime example of practical energy processes that, though extremely inefficient, are of commercial interests mainly because the economic and ecological aspects are favorable.

13.2 The Composition of Biomass

A plant can be considered as a structure that supports specialized organs. The structure consists of wood, and the specialized organs include the leaves that perform photosynthesis and the roots that collect water and nutrients. Materials are transported from one site to another by sap. Fruits perform the function of sexual reproduction. Energy in a plant is stored in roots and tubers, in the sap, and in the fruit. Structural parts of the plant also represent an accumulation of energy.

Plant tissue consists of 50% to 95% water.

Leaves contain, among other substances, proteins and much of the minerals taken up. Good plantation management frequently involves returning leaves to the ground as mulch and fertilizer. Leaves can be transformed into **biogas** (a mixture of methane and carbon dioxide) by anaerobic digestion,

[†] Some animal biomass is used in the production of biodiesel when the raw material is waste cooking oil.

leaving a residue of sludge, which is valuable as fertilizer because it retains most of the minerals used by the plant.

Wood, the structural part of plants, contains few minerals (typically, wood ashes represent less than 1% of the dry mass that was burned). Wood is made of cellulose, hemicellulose, and lignin in variable proportions (say, 50%, 30%, and 20%, respectively). Cellulose, a polyhexose, and hemicellulose, a polypentose, are carbohydrates (see Subsection 13.2.1). Lignin has a phenolic structure highly resistant to microorganisms.

The plant stores energy in the form of carbohydrates, hydrocarbons, and esters. Carbohydrates consist of sugars and their polymers such as starches, cellulose, and hemicellulose. Sugars can be stored as such in sap (sugarcane), tubers (sugar beets), and fruits. Starches are mostly found in fruit (grain), tubers (potatoes), and roots (manioc).

The hydrocarbons found in plants are generally polyisoprenes (terpenes)—that is, polymers of the alkyne hydrocarbon, isoprene, C_5H_8. They are found in some euphorbia such as the rubber tree.

Vegetable oils are esters, chemically quite different from mineral oils that are hydrocarbons. Usually, the oil from the pulp of the fruit is different from that obtained from the seed or kernel. The olive is one exception; its fruit and seed yield the same type of oil.

Whereas the technology of ethanol production is somewhat mature, the use of vegetable oil, though quite ancient, has not yet been efficiently extended to the fuel area; hence it still has great potential for improvement.

A large number of plants yield oil (116 species native to the Amazon region have been identified), but much of the oil currently being extracted is for food or for other nonfuel applications. Vegetable oils include babaçu, castor, jatropha, olive, palm, peanuts, rapeseed (canola), soy, and sunflower. Palm oil is by far the most productive—up to 6 tons of oil per hectare per year—while most of the other sources yield about one order of magnitude less. Owing to the dual use of vegetable oil as food and fuel, there is a problem in properly allocating the best usage of the material.

Vegetable oils can be used as fuels by themselves or can be transformed into **biodiesel**.

13.2.1 A Little Bit of Organic Chemistry

Hydrogen is the most abundant element in the universe, but carbon is the most versatile. By itself it forms graphite and diamond and those fantastic and useful molecules: graphene, nanotubes, and buckyballs. Carbon can form an enormous number of compounds whose study constitutes a separate branch of chemistry called **organic chemistry**.

13.2.1.1 Hydrocarbons

Hydrocarbons are compounds that contain only carbon and hydrogen atoms. They are either aliphatic or aromatic. Aromatic compounds are characterized by a very peculiar carbon–carbon bonding discussed in the

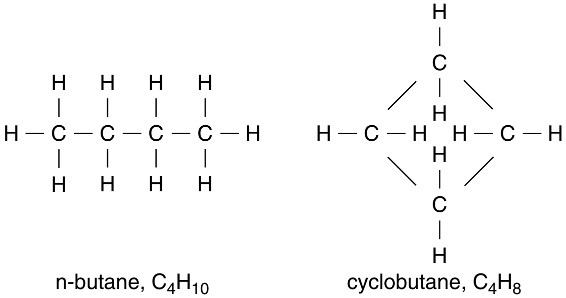


Figure 13.1 Both n-butane (an alkane) and cyclobutane (a cycloalkane) are saturated hydrocarbons.

Sub-subsection 13.2.1.7, "Heterocyles." By default, aliphatic hydrocarbons are hydrocarbons that are not aromatic.

Hydrocarbons are among the simplest carbon compounds, and, of these, the simplest is methane, CH_4. A whole slew of methane-like compounds exist having the general form C_nH_{2n+2}. They constitute the family of **alkanes** known also as **paraffin** hydrocarbons. Other series exist such C_nH_{2n} (**alkenes** or **olefins**), C_nH_{2n-2} (**alkynes** or **acetylenes**).

In alkanes, the bond between two carbons is a single bond, and alkane-derived compounds are said to be **saturated**. In alkenes there are double carbon–carborn bonds[†] and in alkynes, triple bonds, making these hydrocarbons **unsaturated**. Somewhat counterintuitively, multiple bonds are weaker than single ones, causing unsaturated compounds to be less stable than saturated ones. This is significant, as we shall see when we discuss **biodiesel** in Sub-subsection 13.2.1.3 on esters.

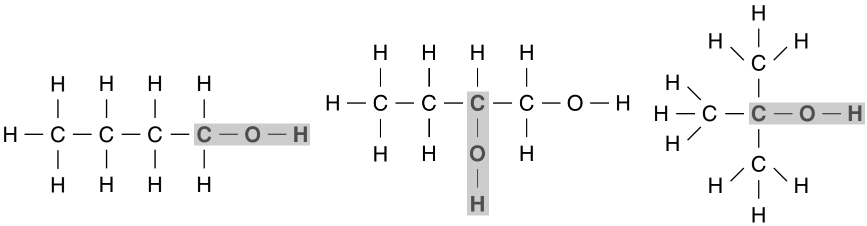
Not all C_nH_{2n} hydrocarbons are alkenes; it is possible to arrange the atoms of carbon in a circle, maintaining saturation even though two hydrogens are missing. Such a molecule is a **cycloalkane**, an example of which is cyclobutane, C_4H_8. Figure 13.1 shows the normal butane molecule, which accommodates 10 hydrogens, and cyclobutane, which is saturated with only 8 hydrogens.

Since organic chemistry may be said to start with hydrocarbons, it is instructive to investigate some of their functional derivatives.

13.2.1.2 Oxidation Stages of Hydrocarbons

Let us examine the successive stages of oxidation of a methane-series hydrocarbon. We chose propane because it is the simplest compound in which not all the carbons are at the end of the chain as in methane and ethane. The first oxidation step consists of replacing a hydrogen atom by an hydroxyl, OH. The result is an **alcohol**, and here the complications start: a surprising

[†] A hydrocarbon may have more than one double carbon–carbon bond. If two, it is called a **diene**, if three, a **triene**, and, in general a **polyene**.



Primary butanol

The carbon to which the functional group, OH, is attached is bound to a single other carbon.

Secondary butanol

The carbon to which the functional group, OH, is attached is bound to two other carbons.

Tertiary (tert) butanol

The carbon to which the functional group, OH, is attached is bound to three other carbons.

Figure 13.2 The three subsets of alcohol.

number of different alcohols can be formed. To begin with, there can be three subsets of alcohols, each with their peculiar chemistry. If the carbon to which the hydroxyl is attached is bound to a single other carbon, we have a **1-alcohol** or **primary** alcohol (1-propanol, in this case); if it is the carbon in the middle of the chain, that is, if the carbon to which the hydroxyl is attached is bound to two other carbons, we have a **2-alcohol** or a **secondary** alcohol (2-propanol, in this case). There is also a **3-alcohol** or **tertiary** alcohol when there are three other carbons bound directly to the hydroxyl-bearing carbon (as in the case of *tert*-butanol). See Figure 13.2.

In addition, more than one hydrogen can be replaced. If a single hydrogen is replaced by an hydroxyl, we have a **simple alcohol** no matter what subset it belongs to. If more than one hydrogen is replaced, we have a **polyol**—two hydroxyls are **diols**, three are **triols**, four are **tetrols**, and so on. Hence, when three hydrogens in propane are replaced by OHs, we have **propanetriol**, one of the most important alcohols in biochemistry, because all oils and fats are derived from this alcohol.

Figure 13.3 shows the structural formula of propane and its five alcohols. The center column shows the condensed structural formula, while the last column lists the technical name. As is common in organic chemistry, compounds tend to have a plurality of names. Propanetriol, for instance, is better known as **glycerol**, **glycerin**, and sometimes as **1,2,3-trihydroxypropane**, **glyceritol**, and **glycyl alcohol**.

The next oxidation step consists of the loss of two hydrogens: one in the OH and one previously attached to the carbon atom to which the OH was bonded. The remaining oxygen is now doubly bonded to the carbon. If the carbon was at the end of the chain, that is, if the alcohol was a primary alcohol, then a functional group O=C–H is formed, characteristic of **aldehydes**. If the carbon was in the middle of the chain (a secondary alcohol), then a **ketone** is formed.

The third step consists of substituting the H of the C=O–H group by an OH, creating a **carboxyl** group, O=C–O–H, characteristic of a

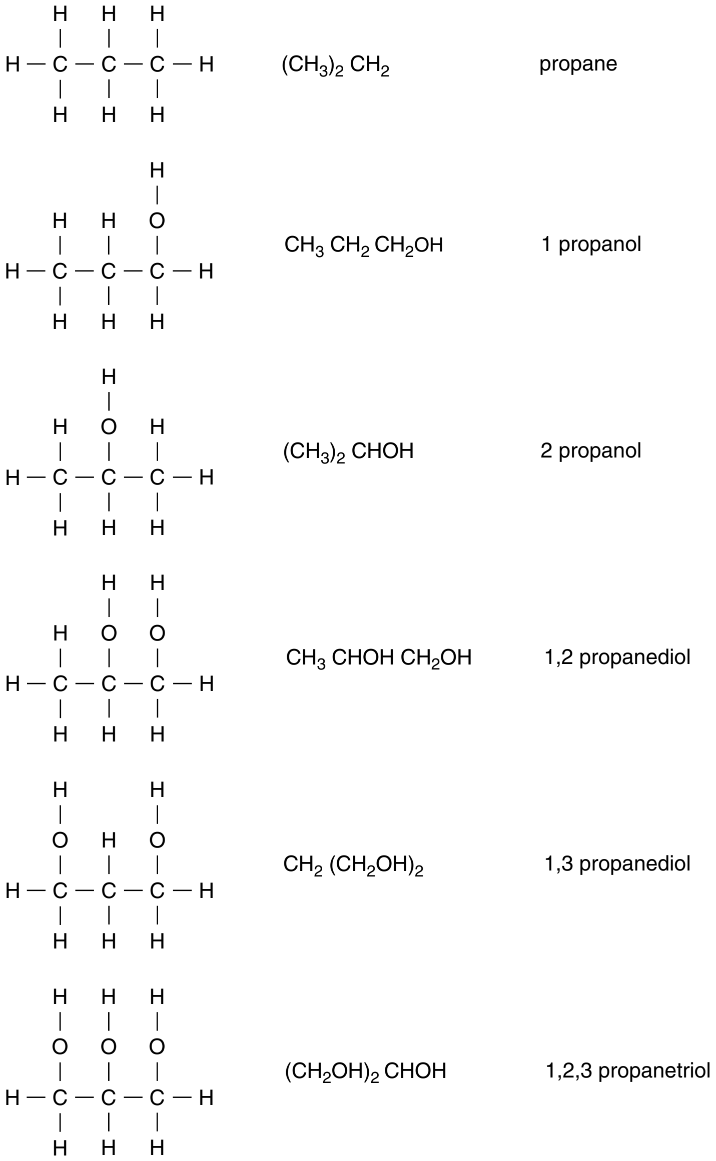


Figure 13.3 Propane and its alcohols.

carboxylic acid. Notice that acids are derived from aldehydes, not from ketones.

The final step in the oxidation of either an acid or a ketone is the formation of carbon dioxide and water.

In the sequence illustrated in Figure 13.4, the OH, OCH, and OCOH are called **functional groups**, of which there are many in organic

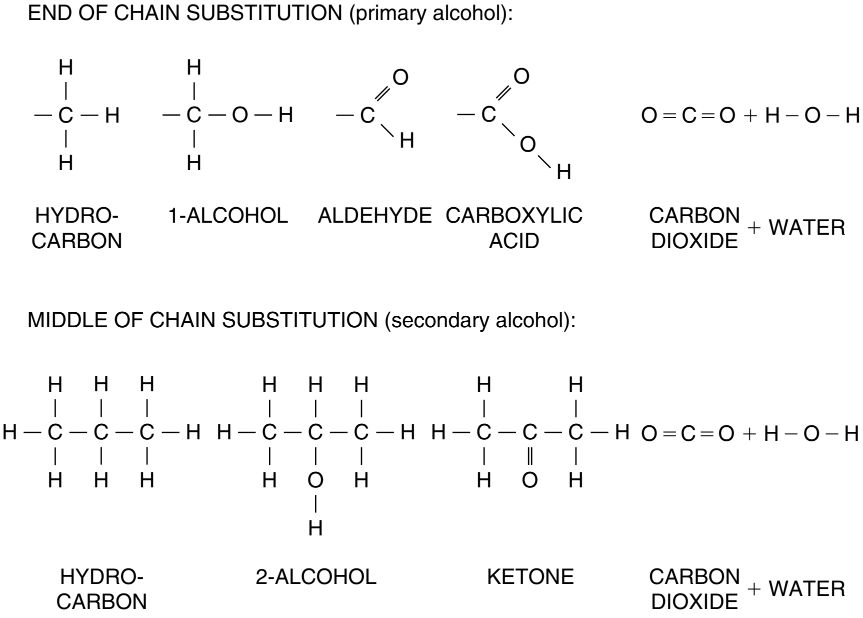


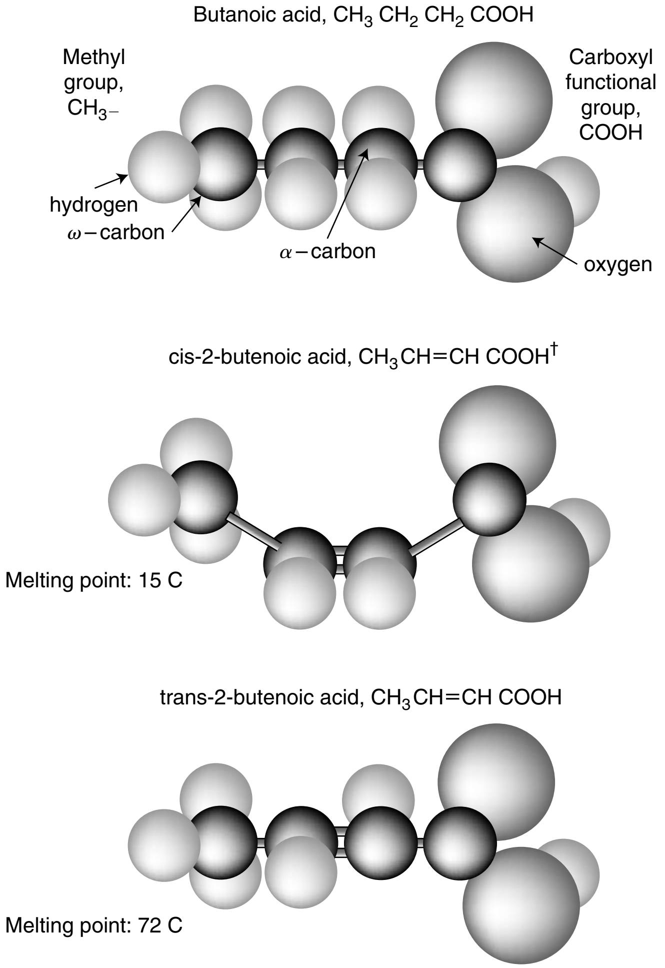
Figure 13.4 Oxidation sequence of a methane-series hydrocarbon.

chemistry. They determine the characteristic chemical reaction of the molecule. That part of the molecule that is not the functional group is frequently represented by the generalized symbol, R. Thus, for example, any alcohol may be written as ROH, and any carboxylic acid as RCOOH.

Some long-chain carboxylic acids are called **fatty acids**, a somewhat loose definition. Acetic acid, having two carbons, is not a fatty acid, but butanoic acid, with four carbons, is. As explained before, if all carbon–carbon bonds are single bonds, the acid is said to be **saturated**, if one carbon has a double bond to another carbon, it is **monounsaturated**, and if more than one double carbon–carbon bond exists, then it is **polyunsaturated**.

Unsaturated acids are more reactive (less stable) than saturated ones, and have a lower melting point (see Problem 13.9). That is why fish have only highly unsaturated fats in their body; otherwise they would become rigid when their fats solidified in icy waters. Since unsaturated fats are unstable, dead fish will quickly decompose and reek, not having the stability of mammal meat.

Fatty acids are designated by the letter "C" followed by the number of carbon atoms, then by a colon, and finally by the number of double bonds. Thus, for example, palmitic acid is C16:0 (saturated, with 16 carbons), stearic acid is C18:0 (saturated, with 18 carbons), oleic acid is C18:1 (monounsaturated, with 18 carbons), and linoleic acid is C18:2 (polyunsaturated, with 18 carbons). The carbon next to the COOH group is the α



Butanoic acid, $CH_3\,CH_2\,CH_2\,COOH$

cis-2-butenoic acid, $CH_3CH{=}CH\,COOH$[†]

trans-2-butenoic acid, $CH_3CH{=}CH\,COOH$

Figure 13.5 Butanoic acid is saturated, while butenoic acid is monounsaturated. In the *cis* conformation, the lopsidedness of the CHCH group in the middle of the molecule causes it to bend. Not so in the *trans* conformation. Notice that the melting point of the *cis* acid is lower than that of the *trans* because the tighter stacking of the molecules of the latter require higher temperature to disrupt.

carbon (the first letter of the Greek alphabet) while the carbon on the other end (forming a methyl group, CH_3) is the ω carbon (the last letter of the Greek alphabet).

The top row of Figure 13.5 shows butanoic acid (butyric acid). All carbon–carbon bonds are single; that is, the acid is saturated. In the middle row, we have a similar acid that has one double bond in the middle of

[†] Condensed formulas do not usually show bonds. Here, for emphasis, we have explicitly shown the double bond in the middle of the molecule.

the molecule. This requires shedding two hydrogens because a carbon can make only four bonds. The group, CHCH, in the middle of the molecule is lopsided, causing a kink on the molecule, as shown. This is the cis-2-butenoic acid. If the position of one of the hydrogens in this midchain group is altered, the molecule straightens out, forming the trans-2-butenoic acid.

13.2.1.3 Esters

Esters are the result of a reaction of an acid (frequently, a carboxylic acid) and an alcohol, with the liberation of a molecule of water. The oxygen in the water comes from the acid.

Figure 13.6 illustrates the **esterification** resulting from the reaction of the alcohol, ethanol (C_2H_5OH), with butanoic acid (C_3H_7COOH), yielding ethyl butyrate and water. As all esters, this one can be represented by the general formula, $R_{ac}COOR_{al}$, where R_{ac} is the acid group and R_{al} is the alcohol group. See "functional group" earlier in this sub-subsection.

Esters are responsible for many scents of fruits and flowers as well as for the aroma of wine. For example, pineapple, banana, and wintergreen smells are due to, respectively, ethyl butyrate, amyl[†] acetate, and methyl salicylate.

Of particular interest to the energy engineer are the esters of glycerin because they are oils or fats. Since glycerin is a triple alcohol (see 13.2.1.2), it can undergo multiple esterification with acids that can be identical or different. Thus, there are monoglycerides, diglycerides, and triglycerides. Most vegetable and animal oils and fats are a complex mixture of triglycerides. Some of them can be used directly as fuel in Diesel engines, but most usually require modifications of these engines to adapt them to the particular fuel.

Vegetable oils offer a wide palette of fatty acids, ranging roughly from C6 to C24. Some oils are a mixture of many different acids— sunflower

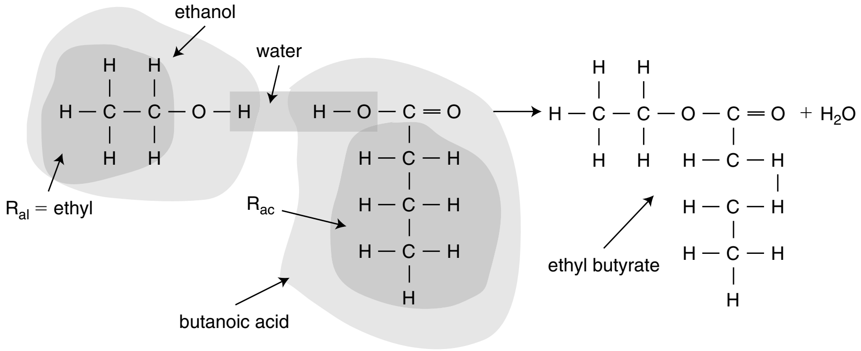


Figure 13.6 Ethanol, an alcohol, can combine with butanoic acid forming the ester, ethyl butyrate, and water.

[†] Amyl is the same as pentyl.

has at least 10 significant acids, dominated by linoleic; castor oil is 90% ricinoleic acid, which does not occur in most other oils. Olive oil is well known for its unsaturated acids (mainly oleic), while palm oil, widely used in the Third World for cooking, is notorious for its saturated palmitic acid composition. The degree of saturation is of great dietetic concern, as is the cis-trans isomerism. Figure 13.7 tabulates the fatty acid profiles of some selected vegetable oils.

Observe that fatty acids produced by plants and animals have, almost exclusively, an even number of carbons.

The use of straight vegetable oil (SVO) as Diesel fuel requires that the oil specifications approach that of petrodiesel. Important are

a. Cetane number.
b. Lubricity.[†]
c. Viscosity.
d. Cloud point. Mixed triglycerides do not have a clear freezing point—some components freeze before others, leaving a cloudy mass. It is best to talk about **cloud point** instead of **freezing point**.

Cetane number is a measure of how quickly a fuel will ignite in a compression ignition engine. Cetane ($C_{16}H_{34}$), also called hexadecane, is very quickly ignited, and its number is set to 100. However, the fuel does not need to contain cetane, to have a cetane number. The cetane number of the fuel should be above 41. Numbers larger than 50 do no harm, but they do no good either. Most vegetable oils have an adequate cetane number, but, if necessary, additives can be mixed with the oil. Small amounts of acetone have been used.

Diesel engines invariably employ injection pumps that need lubrication, which is usually provided by the fuel itself. Formerly, Diesel oil used to be adequate for the purpose, but modern fuel from which much of the sulfur has been removed may have insufficient lubricity and require a certain amount of added oil, frequently vegetable oil. SVOs by themselves have more than adequate lubricating properties, outperforming petrodiesel.

It is in the viscosity that the characteristics of vegetable oils depart most markedly from Diesel oil specifications. SVOs have viscosities too high for direct use in unmodified engines and can cause damage to the injection equipment. The viscosity of petrodiesel is 4 to 5 mm^2/s, while that of vegetable oils is 30 mm^2/s or more (canola oil has a viscosity of 37 mm^2/s), at 40 C. The viscosity here is the **kinematic viscosity**, η, which is related to the **dynamic viscosity**, μ, by the $\eta = \mu/\rho$ where ρ is the density of the oil.[††]

[†] Do not confuse lubricity (the capacity to protect rubbing surfaces) with viscosity (which has to do with the ease of flow).

[††] For a brief discussion on viscosity, see Section 15.8.

Figure 13.7

Fatty Acid Profiles of Selected Vegetable Oils[†].

	C8:0	C10:0	C12:0	C14:0	C16:0	C16:1	C18:0	C18:1	C18:1	C18:2	C18:3	DHSA	C20:0	C20:1	C20:2	C22:0	C22:1	C24:0
	caprylic	capric	lauric	myristic	palmitic	palmit-	stearic	oleic	ricin-	linoleic	lin-	dihydr-	arachidic	eicose-	eicosa-	behenic	erucic	lignoceric
						oleic			oleic		olenic	oxystearic		noic	dienoic			
Canola		1			3.8		1.9	62.4		20.1	8.4			1.5				
Castor					1.0		1.0	3.0	89.5	4.2	0.3	0.7		0.3				
Coco-nut	6.4	5.6	45.5	18.8	10.1		4.3	7.5		1.8								
Corn		4			10.3		2	25.5		59.3	1.1							
Cotton			1	0.8	21.9		2.3	16.6		56.4								
Flaxseed					4.8		3	21.4		15.2	54.2			0.4	0.4			
Hemp							2	13		70	20							
Jatropha					14–15	1.3–2.0	5.1–5.8	42–44		34–38	0.2–0.5							
Jojoba					0.5–3.0	0.3–05	0.1–0.2	5.0–12					0.1	66–74		0.2–1.0	9.0–19	0–0.5
Olive		7			8.7		3.5	76.3		8.6	0.8							
Palm				1.1	42.7		4.6	39.4		10.6			0.4			0.6		
Palm kernel	3.4	3.2	46.1	16.2	8.7		2.3	16.5		2.8								
Peanut			1		9.4		2.7	48.7		31.1		1	1.4	1.4		3.1		1.7
Safflower					7		2	13		78								
Sesame					12		4	25		52	8							
Soy					9.9		3.9	21.4		56.0	7.6							
Sunflower					5.7		4.8	15.3		71.2	0.5		0.3	0.2		1.2		0.3

[†]Oils are mixtures of triglyceridex. Jojoba “oil” is not a glyceride, it is a mixture of esters of higher (>3C) fatty alcohols—it is a true liquid wax.

To get around the problem of excessive viscosity one can

a. blend the vegetal oil with Diesel oil or kerosene.
b. preheat the oil to a temperature (65 C?) at which its viscosity has fallen to an acceptable value.
c. since SVOs are triglycerides (i.e., an ester of propanetriol), replace the propanetriol by either methanol or ethanol, an operation known as **transesterification**, which produces **biodiesel**.

The use of methanol in preparing biodiesel is currently popular, probably owing to the wide availability of this alcohol. It may turn out that ethanol is the better choice because it is produced from biomass and leads to a biodiesel with somewhat superior properties (it has a lower cloud point and slightly higher cetane number). Methanol is usually obtained from petroleum (see Chapter 10), although it can be produced from biomass. A final consideration in favor of ethanol is the toxicity of methanol.

One difficulty that straight vegetable oils share with biodiesel has to do with the degree of unsaturation. As food, saturated fats appear less desirable than unsaturated ones; in fuels the situation is reversed. Unsaturated oils are more unstable, tending to oxidize and to polymerize. That is why linseed oil, used as a vehicle for pigments in paints, dries. In engines, such drying leads to the formation of undesirable gums that clog filters and injectors. On the other hand, saturated acids make oils that have a high melting point and thus tend to solidify in the winter. In most of Brazil, this does not constitute a problem, but in northern United States and Canada, for instance, this freezing can seriously interfere with engine operation.

There is a simple test used to determine the degree of unsaturation. Unsaturated compounds have one or more double carbon–carbon bonds and consequently have less hydrogen than saturated compounds. Iodine will, like hydrogen, attach itself to the chain at a double bond site—two atoms per double bond. **Iodine value** is the number of grams of iodine taken up by 100 g of oil and is used as a measure of unsaturation. One technique for measuring the iodine value is the **Wijs method**[†]. Google it!

Transesterification starts with cleaving the ester bond, thus separating the fatty acid radicals from the glycerine. Most of the radicals will then combine preferentially with the methanol (or ethanol), not with the glycerine, forming the desired methyl or ethyl ester. The agent used to cleave the original bond is usually sodium or potassium hydroxide, caustic substances that tend to saponify free fatty acids (see the sub-subsection on 13.2.1.4, "Saponification"). This is, of course, undesirable and requires keeping the caustics to a minimum. Nevertheless, some free fatty acids will survive the transesterification and should be neutralized because they can corrode the engine. Note that free fatty acids (FFAs) may result from some acid radicals combining with water during the transesterification, or they may

[†] Pronounced approximately like "Weiss" in German or "vice" in English.

have been present all along in the oil used in the process. This is frequently the case when waste vegetable oils (WVOs), left over from cooking, are used because the cooking process generates such free acids (the longer and the hotter the cooking, the more acid is formed). To limit the amount of caustic added, it is necessary to perform an accurate titration of the oil used.

A solution of sodium or potassium hydroxide in methanol is prepared with the correct proportion of the hydroxide. The literature refers to such solution as methoxide, which, fortunately, it is not. The methoxide ion is a strong base resulting from the removal of a hydrogen atom from methanol. It is CH_3O^- and is a dangerous substance—poisonous and corrosive—and it is indeed produced when an alkali base reacts with methanol, *provided water is excluded*. In the presence of water, we have simply a solution of the hydroxide in the alcohol without production of significant amounts of sodium (or potassium) methoxide.

After the transesterification, excess alcohol, water, soaps, and catalyst must be removed from the final product. The by-product of the operation is glycerin (glycerol), a material that used to have industrial importance. However, the growing production of biodiesel has led to an excess production of glycerin, causing its market value to fall by some 90% between 2004 and 2007. It is becoming a waste product rather than a co-product; that is, the manufacturer has to pay for its disposal. Since for every 10 kg of biodiesel produced, roughly 1 kg of glycerin is obtained, it is attractive to ferment the glycerin into ethanol by using selected *Escherichia coli*, as suggested by Ramon Gonzales (Gonzales and Yazdani 2007). There is no lack of information on how to produce biodiesel: in 2006, over 20,800,000 entries were available from Google!

Raw material for biodiesel is either waste vegetable oil (WVO) or straight vegetable oil (SVO). WVO is attractive to small producers (such as individuals), but large-scale plants may have difficulties with feedstock supplies because soap manufacturers use a lot of waste oils.

A plethora of vegetable oils is available. The choice of those most suited for biodiesel production is a complicated subject much debated at present. Europeans are enthusiastic about rapeseed oil.[†] From purely the productivity point of view, nothing beats palm oil with its yield of up to 6000 kg of oil per hectare per year, compared with the best yield of rapeseed oil, which is 900 kg per hectare per year. Actually, there are two classes of palm oil, one extracted from the pulp (containing some 43% palmitic acid and 39% oleic) and one extracted from the kernel (46% lauric acid and 16% oleic). See Figure 13.7.

If the use of biomass as fuel aims at reducing our strategic dependence on foreign oil, then an important performance index is the ratio of energy

[†] Rapeseed oil, owing to its high erucic acid content, is unsuited for human consumption. Canola is a special rapeseed oil from plants selected to have less than 1% of the acid and is popular as a cooking oil.

obtained from the biofuel to the fossil fuel energy used in producing it. For palm oil, the ratio is 9.6:1; for Brazilian sugarcane, it is 8.3:1[††]; and for American corn ethanol, it is estimated at 1.4:1. This last figure is being disputed: a number of researchers believe it is actually less than 1:1, which would mean that more fossil energy is used than produced by the biomass.

Currently, there is growing enthusiasm for developing microalgae for oil production. However, these efforts are still in their initial phases. Algae that have a high oil yield tend to be less hardy than those with low yields. It appears that there will be an advantage in using closed systems in which the desired strains are better protected. Alga systems lend themselves to the sequestration of carbon dioxide produced by fossil fuel power plants.

13.2.1.4 Saponification

Esterification consists, as we saw, of the combination of an alcohol with an acid accompanied by the shedding of one molecule of water. The reverse is the separation of the alcohol from the acid with the addition of a water molecule: it is the **hydrolysis** of the ester, and it leads to the formation of **free fatty acids (FFAs)**. A strong base (an alkali metal hydroxide, for example) will cleave the ester bond, liberating the alcohol and producing an alkali metal salt of a carboxylic acid. It is a **saponification** reaction and has been in use for over two millennia. Usually, the ester is a triglyceride (an animal or a vegetable oil, more commonly, the latter) which is made to react with either sodium or potassium hydroxide. If, for instance, the oil used is that of olive, the result will be glycerol and, predominantly, sodium or potassium oleate which is a soap. Sodium soaps are "hard," while potassium soaps are "soft." The hard sodium soaps are insoluble in a strong sodium chloride solution, which is used to separate the soap from the glycerol.

The soap molecule has one end (the carboxylic end) that is hydrophillic, while the opposite end is lipophillic (or, at least, hydrophobic). This imparts to soap its surfactant and cleansing properties.

13.2.1.5 Waxes

It is convenient to make a somewhat arbitrary distinction between esters of the triple alcohol, glycerol, which are usually called **oils** or **fats**, and esters of other alcohols, called **waxes**. Indeed, while land animals use mostly fats in their metabolism, sea mammals produce abundant waxes such as **whale oil** and **spermaceti**. The spermaceti is mostly cetyl palmitate, an ester of palmitic acid and cetyl alcohol—a single alcohol derived from hexadecane, $C_{16}H_{32}$. Spermaceti, once used in candle making, was obtained from the head cavities of sperm whales. The ban on whaling has prompted the use of a botanical substitute: jojoba "oil"—a true liquid wax—that has great similarities to spermaceti.

[††] Some Brazilian authors quote values as high as 10:1.

Many waxes are esters of the double alcohol, ethanediol (ethylene glycol, or simply glycol). Much of what was said about triglycerides—a triple ester—applies to the glycol diesters.

13.2.1.6 Carbohydrates

One can think of a plant as a organism that converts carbon dioxide, water, and sunlight into glucose and, then, performs some clever chemistry on the molecule. Two of the fundamental requirements of a plant—storage of energy and creation of a structure—are satisfied in a straightforward fashion simply by polymerizing glucose two different ways, that is, using different types of linkage to stitch the molecules together.

Starch is mechanically weak—it can perform no structural function, and it is used only for energy storage. It is soluble in water and can assume both a crystalline and an amorphous configuration—it becomes a gel (amorphous) when in hot water. Starch is a polymer that can be taken apart with little difficulty and, thus, can be easily transformed into the glucose it came from and then fermented into ethanol. It is quite digestible. On the other hand, cellulose is crystalline and mechanically very strong; for this reason it is used to form vegetable cell walls giving rigidity to wood. It is insoluble and hard to decompose, hence the difficulty in transforming it into ethanol. Except for a group of termites, cellulose cannot be directly digested by animals. Other termites and all ruminants eat cellulosic matter but must use symbiotic microorganisms to provide the necessary enzymes. Although insoluble, cellulose is highly hydrophilic (cotton, which is 98% pure cellulose, easily soaks up water). Hence, cellulose is not adequate for making water-conducting ducts. Nature solves this problem by sheathing cellulose ducts with **lignin**, a phenolic substance (not a carbohydrate) that is very resistant to biodegradation.

Carbohydrates are sugars or condensation polymers of simple sugars. A condensation polymer is one in which, when the monomers combine, a small molecule, such as H_2O, is eliminated in a manner similar to that indicated in Figure 13.6: a hydroxyl from one monomer and a hydrogen from the next combine to form water. The reverse—depolymerization—reaction requires the addition of water and is called **hydrolysis**.

Simple sugars, known as **monosaccharides**, are compounds of carbon, hydrogen, and oxygen containing from 3 to 8 carbon atoms: 5-carbon sugars are called **pentoses**, 6-carbon sugars **hexoses**, and so on.

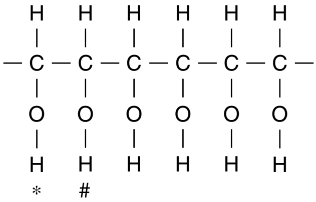
Some hexoses and dihexoses are fermentable: that is, they can biologically be transformed directly into ethanol. These include glucose, fructose, sucrose, maltose, and maltotriose. If the plant contains free sugar, its juices can be fermented directly (as is the case of sugarcane). More complex carbohydrates must first be hydrolyzed into simple sugars before fermentation.

Carbohydrate hydrolysis can be accomplished by weak acids or by enzymes. Yields obtained with inexpensive acids, such as sulfuric, are low (some 50% of the potential glucose is made available). Enzymes are expensive but lead to larger yields. Both processes are energy intensive.

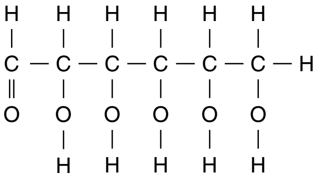
It may be possible to use microorganisms to perform both hydrolysis and fermentation in a single step.

Hexoses have the empirical formula $(CH_2O)_6$. A number of hexoses have this same formula; some are aldehydes (**aldohexoses**), such as glucose, galactose and mannose, and some are ketones (**ketohexoses**), such as fructose. To distinguish one from another, it is necessary to examine their structural arrangement. Hexoses can exist in either a linear chain or in a cyclic structure; the cyclic structure is more common and the only one found in aqueous solutions. To simplify the explanation that follows, we will use the linear formula.

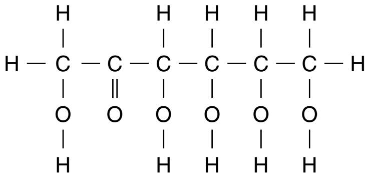
A simple way to arrange the atoms that constitute a hexose would be:



The preceding structure satisfies the empirical formula but leaves two dangling valences at the ends of the chain. To correct this, one can remove the H marked with an asterisk * and place it at the other end. The formula then represents correctly the (linear) structure of **glucose**:



Notice that glucose has an O=C–H group: it is an aldehyde, while another common hexose called **fructose** is a ketone because the H that was moved is the one marked with a # and is not at the end of the chain:

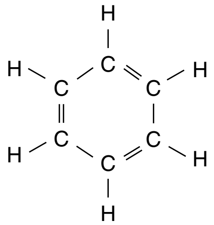


Sucrose is a dimer of the two preceding sugars. It consists of a glucose and a fructose molecule attached to one another with loss of a water molecule. Similarly, **maltose** is a glucose/glucose dimer, and **lactose** is a glucose/galactose dimer.

Another important polymer found in plants is hemicellulose whose main monomer is the pentose **xylose**. Hemicellulose contains, in addition to pentoses, some glucose units.

13.2.1.7 Heterocyles

One can classify organic compounds as **aliphatic, aromatic, or heterocyclic**. To understand this classification, it is best to start by examining the **benzene ring**. We observe that carbon, being tetravalent, can make, with itself, a single bond as in ethane, an **alkane**, a double bond as in ethene (ethylene), an **alkene**, or a triple bond as in ethyne (acetylene), an **alkyne**. The compound may be straight, branched, or cyclic. What is relevant here in our discussion is that a double bond (C=C, which measures 134 pm), is shorter than a single bond (C–C, which measures 153 pm). Now, take benzene, commonly represented by the structural formula,



This conventional representation would suggest that the ring consists of alternate double and single bonds. Yet, the carbons actually form a perfect hexagon; that is, all bonds are of equal length, a length of 140 pm, very nearly the average of double and single bonds. Thus, the bonds between the six carbon atoms of benzene are neither single nor double—they are a hybrid. We must learn a bit more about chemical bonds, and we will start with a very simple, purely geometrical description of two fundamental types of chemical bonds: the σ- and the π-bonds.

Sigma and Pi Bonds

Assume you want to interconnect two marbles so as to transform them into one single rigid body. One way to do this is to use a short rod and glue it to the two marbles as shown in the top sketch of Figure 13.8. You get a dumbbell-like object. The interconnecting rod *is exactly between the two marbles.* Let us call this kind of arrangement a ***σ*-bond**.

It is also possible to interconnect the two marbles by using two rods glued tangentially to the marbles, one above them and one below them (second sketch from the top, in the figure). Notice that *the rods are not directly between the marbles.* Notice also that two rods are used

(*Continues*)

(*Continued*)

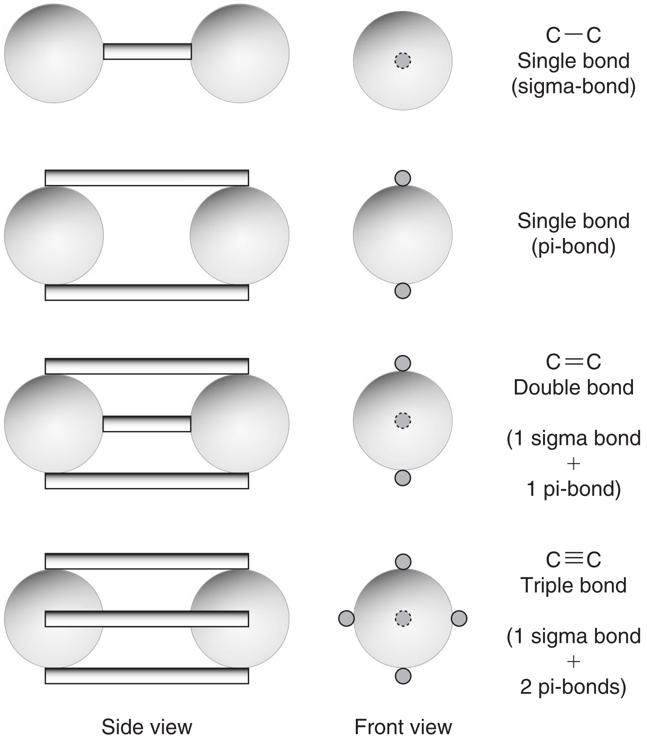


Figure 13.8 Sigma-bonds are located between two atoms while pi-bonds straddle the two atoms, half of the bond being above the pair and half below. Multiple bonds consist of one sigma-bond and one or two pi-bonds.

to make one single bond (otherwise, we would have an awkward mechanically lopsided arrangement). Call this a **π-bond**.

The position of the bond (rods) captures the essential difference between σ- and π-bonds.

If there is a double bond, then one of them is a σ-bond and the other is a π-bond. If there is a triple bond, then one is still a σ-bond and the other two are π-bonds. The two π-bonds do not interfere with one another because they are in mutually perpendicular planes. See the last sketch in the figure.

The analogy between the interconnection of marbles and the binding of atoms into molecules is, of course, extremely crude. We need a modest acquaintance with the concept of orbitals before we can become familiar with the idea of bonds.

Orbitals and ...

An object moving in the gravitational field of an attracting body describes a well-defined orbit, which, in most cases, is an ellipse. The orbit lies on a plane and is perfectly deterministic—; one can predict the position of the object at any given time with great precision.

In quantum mechanics, the situation is fundamentally different. There is no orbit in which the electron moves around a nucleus, and there is no way of predicting the exact position of the electron. A three-dimensional map of the probability of finding the electron in a given position relative to the nucleus is called the **orbital** of the electron.

Take a hydrogen atom. Its lone electron has a probability density (probability of finding the electron in an elementary volume of space) that depends only on the radial distance to the nucleus; it is independent of direction. The maximum probability occurs at a distance equal to that predicted by classical mechanics for the radius of a circular orbit, but there is a finite probability of finding the electrons at different distances from the nucleus. This kind of orbital is spherical (it is called an **s-orbital**[†]) and can be visualized as a spherical cloud. It is designated $1s^1$, where the first "1" refers to the first energy level or shell, and the exponent indicates that there is a single electron in this shell.

Helium has two electrons in the same energy shell as hydrogen (configuration: $1s^2$). Because two electrons is all that the 1s shell can hold, helium (as all full-shell elements) is chemically inert. Lithium has a full 1s shell and an extra electron in an s-orbital in the second energy shell (its electron configuration is $1s^2 2s^1$). In beryllium this second **s-subshell** is filled, but the second **shell** can accommodate a total of eight electrons. The last six electrons of this second shell are in p-orbitals that look quite different from the s-orbitals.

Each p-orbital resembles two elongated balloons facing one another with the nucleus between them. The probability of finding an electron on the plane perpendicular to the axis of the orbitals is zero. Observe that the axis of the p-orbitals has a direction (horizontal, in the illustration). There can be three identical, mutually perpendicular p-orbitals in each energy level above the second level, and, since each orbital can accommodate two electrons, a total of six p-electrons can exist per level.

[†] Although convenient to associate the letter "s" with the word "spherical," it actually stands for "sharp," an expression derived from spectroscopy. The letters "p," "d," and "f" used to designate higher orbitals have a similar origin ("principal," "diffuse," and "fine," respectively). However, the letter "g," which designates the highest orbitals, is simply the one that, follows "f" alphabetically.

(*Continues*)

(*Continued*)

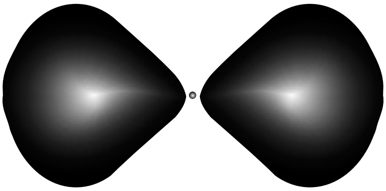


Figure 13.9 Each p-orbital resembles two elongated balloons. There are three mutually perpendicular sets.

Additional atomic orbitals (d, f, and g) differ in shape, but all we need here is to become familiar with s- and p-orbitals. The orbitals discussed above are **atomic** orbitals. When atoms congregate to form molecules, the various **molecular** orbitals that result are substantially different.

. . .bonds.

The hydrogen molecule consists of two mutually repelling protons. The two electrons constitute the "glue" that holds them together. The two spherical atomic orbitals fuse, forming a vaguely ellipsoidal molecular orbital in the space between the nuclei. The resulting negative charge attracts both nuclei trying to bring them together. At a given internuclear separation, the repulsion between the nuclei balances the attraction exerted by the negative charge of the orbital. A bond is formed tying the two nuclei to one another. This type of bond, characterized by having the bonding molecular orbital aligned between the atoms, is called a ***σ*-bond**. σ-bonds do not necessarily, have to be derived from s-orbitals; they can also be derived from **hybrid orbitals** as explained in what follows.

Consider the carbon atom. It has a total of six electrons. The first two in the 1s orbital are too tightly bound to the nucleus to take part in chemical reactions; the last four are chemically active (carbon is tetravalent). Among the last four electrons, two are in the 2s orbitals, while the other two are in one of the three available 2p orbitals; carbon's configuration is $1s^2 2s^2 2p^2$. Although there are only four valence electrons, there are a total of four orbitals available for occupancy (each holding two electrons). This all is true for isolated atoms. When carbon forms a compound, the molecular orbitals differ from the atomic ones and depend on the exact nature of the compound. We will examine several examples.

Let us start with the simple molecule, methane, CH_4.

(*Continues*)

(Continued)

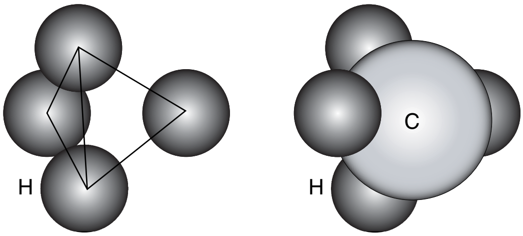
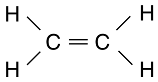


Figure 13.10 The figure on the left depicts four hydrogen atoms at the vertices of a regular tetrahedron. The methane molecule has such an arrangment with the addition of a carbon atom at the center of the tetrahedron as shown on the right. A σ-bond ties each hydrogen atom to the central carbon atom.

It can be shown that all four carbon–hydrogen bonds are identical. This requires that the hydrogens surround the carbon, forming a regular tetrahedron. The angle between adjacent bonds is 109.5°. The four molecular orbitals of carbon in methane are, of course, identical and, hence, have the same energy. They reach out in three dimensions along the edges of a tetrahedron and are called **sp^3 hybrids.**[†] Note that the bond between the carbon and each hydrogen is still a σ-bond formed by the fusion of the hydrogen s-orbital and one carbon sp^3 hybrid.

The situation is different in the case of ethene, C_2H_4, in which there is a double bond between the two carbons,



The four molecular orbitals consist of one standard 2p-orbital and three **sp^2 hybrids** that form a starlike flat figure (trigonal planar) with 120° between them. Two of the orbitals form individual σ-bonds with the s-electron of each hydrogen, while the third arm forms a σ-bond with one of the sp$^{\mathbf{2}}$ hybrids of the other carbon. This provides one of the carbon–carbon bonds. The other is a π-bond between the p-orbitals left over. To visualize this, remember that one of the p-orbital is perpendicular to the plane of the trigonal sp^2 hybrids. Looking along this plane, the p-orbital has one lobe above and one below the trigonal plane. The π-bond forms from the interaction between the lobes above

[†] It is best not to inquire where this quaint name comes from.

(Continues)

(Continued)

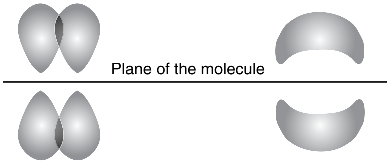


Figure 13.11 Two overlapping p-orbitals (left) will fuse, forming a pi-bond between the two carbon atoms in ethene (right). Observe that the bond consists of charges above and below the plane of the molecule, but noting is directly between them.

the plane with each other and the similar interaction between the lobes below the plane with each other. Notice that a π-bond results from molecular orbitals that do not lie directly on the line uniting the two atoms. The double bond consists of one σ-bond and one π-bond.

Look at ethyne, C_2H_2 with its triple carbon–carbon bond.

$$H-C\equiv C-H$$

Here the four molecular orbitals of carbon are two **sp hybrids** and two 2p orbitals. One of the sp hybrids together with the s-orbital of the hydrogen forms the σ-bond of the H–C connection.

The other sp hybrid forms one of the three carbon/carbon bonds. The other two bonds are the π-bonds formed between the remaining 2p orbitals. Remember that the two 2p orbitals are perpendicular to one another and also perpendicular to the axis of the ethyne molecule. The triple bond between the carbons consists of a σ-bond and two π-bonds.

Both the σ and the π bonds result from electron sharing between two atoms and, to this extent, the electrons are delocalized, that is, not bound to a single atom. These bonds, which are very strong, are called **covalent**. Other types of bonds exist. The most obvious one is the **ionic** bond in which a metal completely transfers one of its electrons to an electronegative atom, so that two ions are formed, a metallic positive ion and a negative one, which strongly attract one another. A good example is sodium chloride.

Another bonding mechanism is the **metallic bond**. In a metallic crystal the lattice is formed by ions, and some electrons are completely delocalized, free to move to any point of the crystal and thus able to conduct electricity.

(Continues)

(*Continued*)

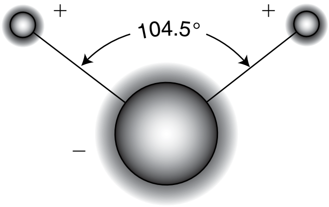


Figure 13.12 In the water molecule, the electrons in the two hydrogen atoms move very close to the oxygen atom. As a consequence, the oxygen becomes negatively charged and the hydrogens, positively charged. The molecule becomes polarized—it acts as dipole.

A much weaker but very important bond is the **hydrogen bond**. To explain how it works, we will take a look at water, which has remarkable properties: it has the highest surface tension and dielectric constant of all liquids, the highest heat of vaporization of all substances, and the second highest heat of fusion and heat capacity of all substances, just behind ammonia.

The water molecule is held together by a couple of covalent bonds between the oxygen atom and the two hydrogen atoms. The exact nature of the bond is not entirely clear. Remember that the electron configuration of oxygen is $1s^2 2s^2 2p^4$—it has four electrons in p-orbitals and can form 4 sp^3 molecular orbitals, which, as we have seen, usually have an angle of 109.5° between them. Two of these orbitals are part of the two covalent bonds with hydrogen, and from this one would expect that the angle, H–O–H, should be 109.5°. However, X-ray spectroscopy measures an angle of only 104.5°.

Here is a somewhat hand-waving explanation: Of the six p-electrons in oxygen, only two are engaged in binding to hydrogen, the other two pairs are in the remaining two sp^3 orbitals by themselves. They form no bonds, but they distort the position of the two bonding orbitals squeezing the hydrogens a bit together so that the molecule is bent.

Although the water molecule as a whole is uncharged, it does have localized charges (which cancel one another): a negatively charged oxygen and two positively charged hydrogens. This is a polarized molecule, and it is able to attract (bond with) other polarized molecules because a positive hydrogen in one molecule attracts a negative oxygen in another. This bonding imparts to water its peculiar properties (among others, its ability to remain in the liquid phase at temperatures in

(*Continues*)

(*Continued*)

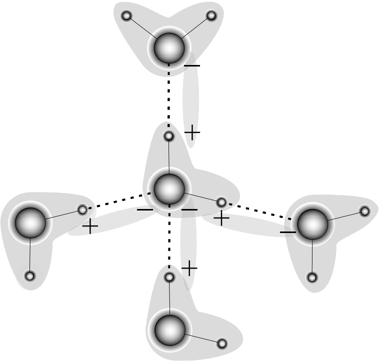


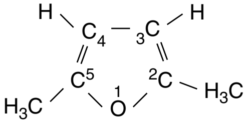
Figure 13.13 The water molecule in the center of the illustration can make four hydrogen bonds (dotted lines) with neighboring molecules.

which such small molecules are usually in the gaseous state). Hydrogen bonds are weak compared with covalent ones (about 1/20th as strong), yet they are of great importance in biochemistry.

Now we can go back to the benzene molecule. The six carbons form a plane partially held together by single σ bonds. In addition, each carbon has a p-orbital above and below the plane. The p-orbitals of adjacent carbons overlap, forming π-bonds. p-electrons are not tied to individual atoms; they are **delocalized** being "smeared" over the whole molecule. Delocalized electrons in π-bonds, as described, are the characteristics of aromaticity. This arrangement does not occur in aliphatic compounds.

Heterocyclic compounds have a carbon ring structure (aromatic or not) containing, in the ring itself, other atoms such as nitrogen, oxygen, or sulfur.

Furan is an aromatic heterocyclic 4-carbon compound with oxygen as the "foreign" element. Its empirical formula is C_4H_4O. Of interest is 2,5 dimethyl furan (DMF), which holds some promise as a biomass-derived fuel.



The advantages of DMF as a fuel, compared with ethanol, include a much higher volumetric energy density (about the same as gasoline), lower volatility (boiling point 93 C, versus 78 C for ethanol), and insolubility in water, which keeps it from absorbing moisture from the atmosphere.

However, there is little experience with DMF as a fuel, and one must be prepared for unpleasant surprises such as unexpected toxicity. For a discussion of the route for production of DMF from biomass, see Schmidt and Dauenhauer (2007).

Another substance being considered as biofuel in lieu of ethanol is butanol, an alcohol of high volumetric energy density and of low solubility in water. The problem here is the difficulty in economically producing butanol from biomass by fermentation.

13.3 Biomass as Fuel

There are numerous ways to use biomass as fuel:

1. Biomass can be burned as harvested. Wood, sawdust, corncobs, rice husks, and other agricultural residue can be burned directly in appropriate furnaces. The heat of combustion of most dry (carbohydrate) biomass lies in the 15- to 19-MJ/kg range.
2. The utility of wood as a fuel has traditionally been improved by transforming it into charcoal.
3. Wood can be gasified to drive vehicles and to fuel industries.
4. Methanol can be made from wood.
5. Vegetable oils can be used in Diesel engines.
6. The sap of some plants is so rich in hydrocarbons that it also can be used directly as Diesel fuel.
7. Sugars and starches can be fermented into ethanol.
8. Biomass residues can be digested into a methane-rich biogas.

No matter which technology is used, relative to fossil fuels, biomass has the important advantage of not contributing to the increase of CO_2 in the atmosphere. Some plausible scenarios foresee, for the not too far future, a CO_2-driven change of the planetary mean temperature leading to the melting of the polar ices and flooding of coastal areas, and, possibly, to modifications in ocean currents that could bring extreme cold to Europe.

Regardless of whether or not a near-future climate catastrophe will occur, it is undeniable that the CO_2 concentration in the atmosphere is growing alarmingly thanks to the emissions from our burning of fossil fuels.

The American daily consumption of gasoline alone is now up to 7.3 million barrels per day, spewing into the atmosphere 220 million tons of carbon every year. This is such a large number that it is difficult to visualize: our carbon consumption from the use of gasoline is equivalent to burning 35 million giant trees every year. If all the other oil used up is considered, the above numbers become three to four times larger. There is also a vast amount of natural gas and coal being burned (56% of all the electricity in the United States of America is derived from coal). All these fuels generate unbelievable amounts of carbon dioxide.

When Earth was young and devoid of life, its atmosphere consisted mostly of nitrogen and carbon dioxide, with no free oxygen. Early life was completely anaerobic (more accurately, anoxygenic). But microorganisms began to learn how to use sunlight to extract free hydrogen from the water in which they lived. The available hydrogen then became the energy source to drive further chemistry to build carbohydrates and proteins and, of course, DNA (or perhaps RNA). However, when water is split, in addition to hydrogen, one also gets an extremely reactive and poisonous gas—oxygen, a gas in which the bacteria and their descendants, the plants, had little interest. Oxygen was released into the oceans and then, into the atmosphere as a pollutant, while carbon dioxide was removed and, in part, retained and fossilized, creating the fossil fuels we now use. With the passing of time, most of the CO_2 was scrubbed off the air and stored underground. Life, being adaptable, learned not only to tolerate oxygen but actually to benefit from its availability. Thus, animals were created. At present we are busily burning fossil fuels in a serious attempt to restore the atmosphere to its pristine, anoxygenic days.

When fuel produced from biomass is burned, all the carbon released was removed from the air the previous growing season, not carbon that has been stored underground for eons. Thus, the net amount of atmospheric carbon dioxide does not change. If the technology for sequestration of furnace-emitted carbon dioxide becomes economical, biomass-burning will actually reduce the total amount of carbon dioxide in the atmosphere.

An additional advantage of biomass-derived fuel is that it is low in sulfur, one of the causes of acid rain.

Using biomass as fuel is not the only way to reduce our dependence on fossil fuels. Biomass can also supply many feedstock for organic chemicals. This is nothing new: in the past many such chemicals were indeed produced from biomass, a usage that became less popular as the price of sugar climbed much faster than that of petroleum. The situation is now reversed and biomass may again, become the raw material of choice.

There are two major paths for transformation of biomass into various chemicals. One is the Fisher–Tropsch process discussed in Chapter 10; it yields syngas (a mixture of hydrogen and carbon monoxide, which is the starting point for the synthesis of numerous organic molecules. The other is fermentation. Usually, the word describes the **anaerobic** transformation of carbohydrates (sometimes proteins) into lactic acids (and other acids) and into ethanol (and other alcohols). The process requires special enzymes, normally supplied by microorganisms—yeasts or bacteria.

13.3.1 Wood Gasifiers

Fuel shortages during World War II motivated the development of gasifiers (using charcoal or wood) for fueling automobiles and tractors. Gasifiers operate by destructive distillation of a solid fuel using part of this very fuel to generate the necessary heat.

After the war, advanced gasifiers made their appearance. Their output is a mixture of hydrogen, carbon monoxide, carbon dioxide, and nitrogen (from the air used in the partial combustion process). When a gasifier is fed oxygen instead of air, its product gases can generate high flame temperatures, making these devices useful in the steel industry.

The syngas produced by the gasifiers can be fed directly to existing oil furnaces, provided they are equipped with special burners. Conversion from oil burning to wood burning does not require a completely new furnace.

A good wood gasifier will deliver to the furnace 85% of the combustion energy of the wood. Part of this energy is the chemical energy of the generated fuels, and part is the heat content of the hot gases. If the gases are allowed to cool prior to burning, only 65% of the wood energy is recovered.

Wood contains little sulfur and, from this standpoint, is a clean fuel. Its burning does, however, generate tars that tend to pollute the exhaust gases and to clog passages in the equipment. To avoid this problem, most modern gasifiers are of the downdraft type. This causes the recirculation of the generated gases through the hottest region of the flame, cracking the tars and yielding a surprisingly clean burning fuel. Usually, the gases are circulated through a cyclone to remove the particulates.

Eucalyptus trees that grow quickly even in poor, dry land are one of the often used sources of wood. In the United States, a variety of hybrid willow bush that grows over 3 m per year is being investigated as a fuel source for big power plants. Willow shrubs (that do not look like the usual willow trees) will, like eucalyptus, grow in marginal land. Edwin H. White is the dean of research of the College of Environmental Science and Forestry of the State University of New York where the experimental work is being done. The original idea of using willow shrubs is from Sweden.

13.3.2 Ethanol

13.3.2.1 Ethanol Production

Ethanol can be mixed with gasoline in any desired proportion. When used by itself, owing to its high antiknock properties, it is more efficient than gasoline as an automotive fuel because higher compression ratios can be tolerated (11:1 versus 9:1 for gasoline). This partially compensates the smaller volumetric energy density of ethanol. Based only on the relative volumetric energy content, alcohol should be competitive with gasoline at the price of 71% of that of gasoline. In the end of 2005, the price of ethanol at a Brazilian gas station was only 51% that of gasoline—this, without any government subsidy. Thanks to progressive advances in technology and in management, the price of 1 m^3 of Brazilian ethanol fell from US$400 in 1985 to US$100 in 2006 (constant 2002 dollars). Since few cars in the United States are capable of using pure ethanol, this fuel is being used only as an additive to gasoline, improving the octane rating of the latter. In Brazil, half of the vehicles sold in mid-2005 had **flex-fuel** engines capable of automatically adjusting themselves to use any fuel, from pure gasoline to pure ethanol. This solution, popular in Brazil, is not used in the United States because of the lack of wide distribution of pure ethanol in gas stations.

The easiest way to make ethanol from biomass is to use sugar-producing plants such as sugarcane. Such sugars are directly fermentable. The next easiest way is to use plants that produce starches that can economically be transformed enzymatically into sugars and then fermented (manioc roots). Finally, the most difficult way is to use cellulosic matter that is widely abundant but much more difficult to hydrolyze than starches. Economic production of ethanol from cellulose would make a vast supply of biomass available for transformation into liquid fuel. This explains the great effort that is being made to reduce the cost of enzymes capable of hydrolyzing cellulose. In the last few years, this effort has, resulted in a more than order of magnitude reduction in the price of these enzymes.

It is not sufficient, however, to have access to appropriate raw material; a substantial amount of energy is needed for the ethanol production, especially for distillation. If this additional energy comes from fossil fuels, then there may be scant advantage in the process. To insure greater independence from fossil fuels, it is necessary that the extra energy itself come from biomass. The advantage of sugarcane is that it produces both the sugar in its sap and the fuel for its processing from its bagasse (the squeezed-out stalks) and from the straw.

At present, 96 tons of raw[†] sugarcane are produced annually per hectare in Brazil, yielding 6400 liters of ethanol per hectare per year.[††] This is energetically equivalent to 25 barrels of oil per hectare per year. The country currently has a sugarcane-planted area of 5 million hectares and could produce the equivalent of some 125 million barrels of oil per year. However, 57% of the cane production is earmarked for sugar production.

Harvesting used to be done manually after the cane was burned in the field and then cropped leaving the roots for regrowth. This was labor intensive (generating employment), facilitated transportation of the cane stalks to the mill, destroyed pests, and returned ashes to the plantation, acting as fertilizers. Mechanical harvest allows the use of the abundant straw as an additional source of biomass and avoids the atmospheric pollution resulting from field burning. The mechanized harvest has now been adopted by the majority of planters, at least in the south of the country.

[†] Includes 15 tons of straw.

[††] Some mills report 7700 liters ha^{-1} $year^{-1}$.

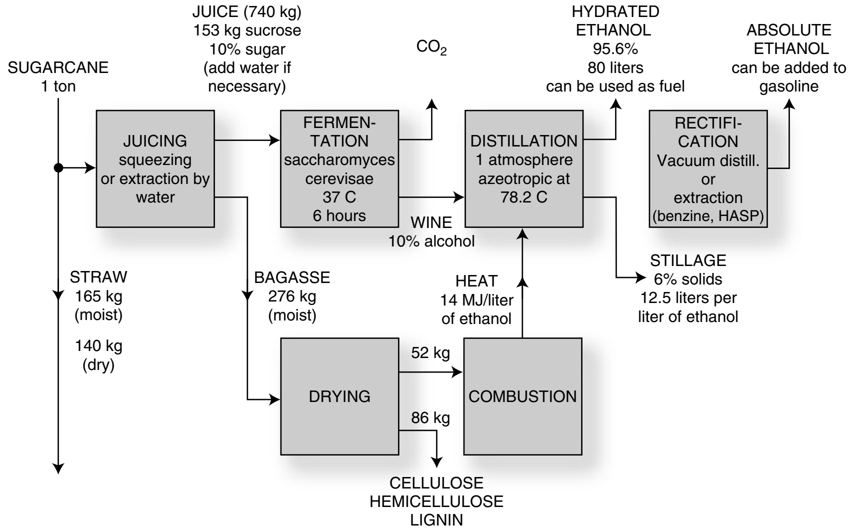


Figure 13.14 Flowchart of a typical Brazilian ethanol plant.

Each ton of cane yields 153 kg of sugars, mostly sucrose (2.54 GJ at 16.6 MJ/kg), 276 kg of moist bagasse, or 138 kg of dry bagasse (2.5 GJ at 18.1 MJ/hg), and 165 kg of moist straw or 140 kg of dry straw (2.14 GJ at 15.3 MJ/kg). It can be seen that the sugars represent only roughly one-third of the available energy and that another third used to be lost by burning the straw in the field, a procedure that is being phased out. The 80-liter ethanol output (1.87 GJ) represents a 74% efficiency in the fermentation/distillation step. However, if the straw and bagasse not used for generating heat and electric power for the mill is submitted to the Dedini Fast Hydrolysis Process (Dedini Hidrólise Rápida, DHR), then an additional 70.6 liters of alcohol is produced, bringing the total to 151 liters per ton of raw sugarcane. In addition, the distillation leaves behind about 13 liters of vinasse[†] per liter of alcohol produced. The vinasse contains much of the plant nutrients absorbed by the cane and a certain amount of organic matter that can be transformed into methane at the rate of 0.1 m^3 of the gas per liter of alcohol produced. This adds 0.5 GJ to the energy output of the mill, which will then total about 4 GJ per ton of cane.

Hydrolyzing straw and bagasse is, by no means, the only way to increase the energy output of a sugarcane mill. Since only a fraction of the bagasse is needed to power the stills, the rest can be burned, together with the straw, generating steam to drive turbogenerators. The amount of electricity thus made available far exceeds the internal demand of the mill, and can be sold to external consumers. In 2006, the installed capacity of Brazilian sugarcane mills was about 3.3 GW, the equivalent of three

[†] The after-distillation residue. Portuguese: *vinhaça*.

nuclear power plants. Updating the boiler/turbine to operate at 80 atmospheres will add another 5 GW in the next five years.

If neither cellulosic alcohol nor cogeneration is used, then the excess cellulose can find many applications as, for example, a raw material for production of acoustic paneling. From the hemicellulose, furfural can be extracted. Furfural is a solvent used to refine lubricating oils. It is also used in the preparation of furfuric resins used in sand casting. Lignin can be the raw material for production of some phenolic compounds. It also can be transformed into metallurgical coke.

Industrial ethanol production starts with the separation of juice from bagasse. This is done by squeezing or by extraction with water. The juice, suitably diluted to permit fermentation, is placed into vats to which a yeast culture (*Saccharomyces cerevisiae*) has been added.[†] Within a short time, carbon dioxide begins to evolve. The process lasts 4 to 6 hours,[††] after which the wine containing about 10% alcohol is transferred to the distillery. The carbon dioxide can be collected and compressed for sale. It is much better for use in fizzling drinks than CO_2 produced from fossils because it does not have bad-tasting impurities.

Ethanol is miscible with water; that is, the two substances can be dissolved in one another in any proportion without settling out. They form an azeotrope[†††] boiling point of 78.2 C at normal pressure (pure ethanol boils at 78.5 C). Water and alcohol co-distill with a fixed composition of 95.6% alcohol. Atmospheric pressure distillation yields **hydrated** alcohol that contains too much water to be added to gasoline to make gasohol. Nevertheless, it is perfectly adequate as a fuel when used all by itself.

To obtain **anhydrous** alcohol, the hydrated form must be **rectified**. This can be done by "vacuum" distillation (at 95 mm Hg, the azeotrope consists of 99.5% ethanol), or as is currently done, by using substances that either absorb the alcohol or the water. Benzene, for instance, forms a ternary azeotrope, benzene-water-alcohol that has a lower boiling point than anhydrous alcohol. Alternatively, water can be extracted by using appropriate molecular sieves.

13.3.2.2 Fermentation

The production of ethanol as we have discussed is based on the fermentation of glucose. Glucose can be degraded by burning—an aerobic process—releasing a substantial amount of energy. The end products are the carbon dioxide and the water the plant used to synthesize this sugar. It is also possible to degrade glucose *anaerobically* with only a minor release of energy so that the final product, which can be either ethanol or lactic acid, still

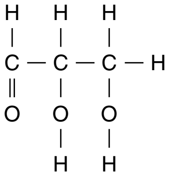
[†] Another alcohol producing microorganism is the bacterium *Zymomona mobilis*, which can ferment more concentrated sugar solutions than *Saccharomyces*.

[††] Twenty-five years ago, fermentation lasted 24 hours.

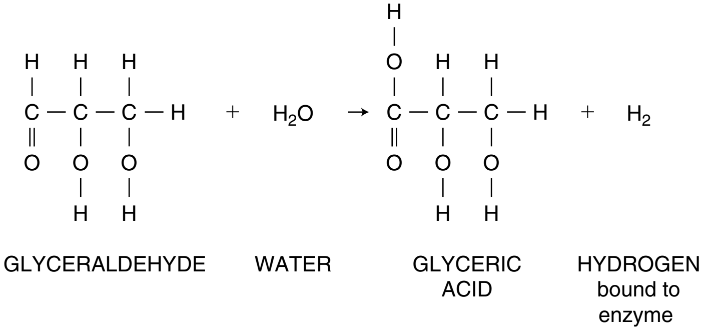
[†††] A mixture whose composition cannot be changed by distillation.

has significant energy left. This process is called **fermentation**, and is mediated by microorganisms.

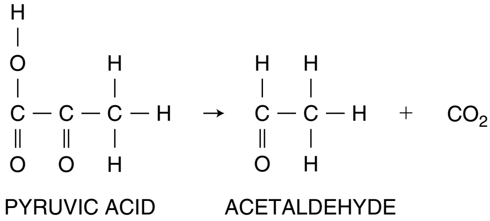
The first step in the fermentation of glucose is its enzymatic splitting into two identical, 3-carbon, glyceraldehyde molecules:



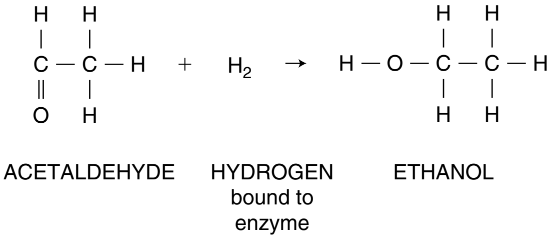
The glyceraldehyde reacts with water and is oxidized into glyceric acid by an enzyme that removes and binds two H atoms:



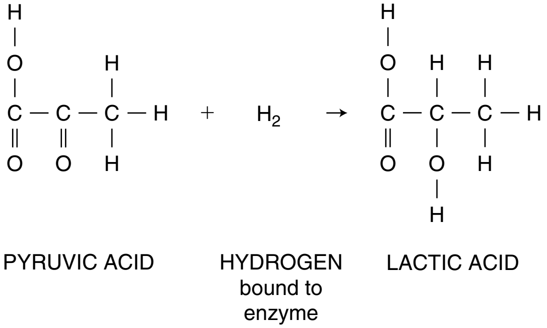
Glyceric acid loses one water molecule and is transformed into pyruvic acid that decomposes into carbon dioxide and acetaldehyde. Thus, the 6-carbon glucose is transformed into the 2-carbon acetaldehyde, which is the number of carbons in ethanol:



Finally, the acetaldehyde interacts with the hydrogen-carrying enzyme and is reduced to ethanol, regenerating the enzyme:



This sequence is one of several ways in which organisms use glucose. It is the metabolic path used by *Saccharomyces cerevisiae* (*sakcharon* = sugar + *mykes* = fungus + *cerevisiae* = of beer), the yeast commonly used in the fermentation of sugars into alcohol. The utilization of glucose by muscles follows exactly the same path up to the production of pyruvic acid. In muscles, this acid can be directly reduced by the hydrogen-carrying enzyme (skipping the acetaldehyde step) and by producing lactic acid:



The overall reaction for the fermentation of glucose is

$$(CH_2O)_6 \rightarrow 2CO_2 + 2C_2H_5OH. \tag{13.1}$$

The heat of combustion of glucose is 15.6 MJ/kg and, since its molecular mass is 180 daltons, its heat of combustion per kilomole is 2.81 GJ. Ethanol has a heat of combustion of 29.7 MJ/kg and a molecular mass of 46 daltons. This amounts to 1.37 GJ/kilomole. When burned, the 2 kilomoles of ethanol in Equation 13.1 produce 2.73 GJ of heat. Consequently, the transformation of glucose into ethanol proceeds with the surprisingly high ideal efficiency of 2.73/2.81 = 0.975.

13.3.2.3 Drawback of Ethanol

Notwithstanding the current popularity of ethanol, the wide adoption of this fuel presents some disadvantages. The most obvious, of course, is the low volumetric energy density, discussed previously. But there are others.

When using ethanol as a gasoline additive, there is the serious possibility of phase separation if the unavoidable water content of the mixture exceeds a certain percentage that depends on temperature. Even when the original ethanol used in the blend is perfectly anhydrous (and, hence, more expensive), water will be collected by condensation from moist air, especially during a cold night. When water is separated from gasoline, it will sink to the bottom and can cause rusting of pipelines and storage containers. The transportation of ethanol-blended gasoline by pipelines is generally avoided. However pure ethanol can probably be economically transported in segregated pipelines.

Bioethanol is produced by fermentation that typically leads to a 10% ethanol-in-water solution. The subsequent separation of the alcohol from the water is done by distillation, an energy-intensive process. On the other hand, the fermentation of higher alcohols, which have a limited solubility in water, could lead to an automatic phase separation from which the desired product can be skimmed at a low cost in energy. Butanol, for example, will settle out from solution when its concentration exceeds some 8%. Unfortunately, butanol and other higher alcohols become quite toxic to the fermenting microorganism at concentrations way below these settling-out levels. The trick is to find microorganisms that are resistant to such toxicity and are sufficiently specific in the production of the desired fuel, and that do not produce a whole spectrum of undesired chemicals. An interesting effort in this direction is described by Atsumi, Hanai, and Liao (2008), who have genetically altered *E. coli* to produce efficiently several C4 and C5 alcohols. They point out that similar results might be achievable with *S. cerevisiae.*

In addition to these technical disadvantages, there are serious political considerations militating against the promotion of ethanol as fuel. The main justification for the use of bioethanol as fuel is the reduction of petroleum consumption. This can effectively be accomplished with ethanol from sugarcane, but is questionable when ethanol is derived from corn. Although politically attractive, corn ethanol incentives may be counterproductive as far as replacing fosil fuels is concerned. In additionaly, diverting corn for fuel production has a demonstrated adverse effect on food prices.

13.3.3 Dissociated Alcohols

The efficiency of alcohols, both methanol and ethanol, as automotive fuel can be boosted by dissociating them prior to burning. The endothermic dissociation can be driven by the waste exhaust heat.

Three different factors contribute to the efficiency increase:

1. The dissociated products have larger heats of combustion than the original alcohol.
2. The dissociated products have large resistance to "knocking" and tolerate higher compression, leading to better thermodynamic efficiency.
3. Leaner mixtures can be used because hydrogen, the major component of the dissociation, has a wide flammability range.

To avoid transporting gases, the dissociation of the alcohol has to be accomplished in the vehicle. Alcohols can be dissociated from hydrogen and carbon monoxide or, if water is introduced for the shift reaction, from hydrogen and carbon dioxide. The direct dissociation of methanol occurs at 250 to 350 C, while steam reforming can be performed at some

200 C. Equilibrium tends toward almost complete dissociation. Undesirable side-reactions constitute a problem requiring careful choice of catalysts.

The system involves a catalytic converter, a vaporizer, a superheater, a reactor, and a gas cooler. Methanol is evaporated by the heat from the engine coolant and leaves the vaporizer at 80 C. It enters the superheater where engine exhaust heat raises its temperature to 250 C. The hot alcohol vapor is then fed to the reactor where additional exhaust heat drives the dissociation in the presence of a catalyst. To improve volumetric efficiency, the mixture of hydrogen and carbon dioxide (accompanied by small amounts of impurities) is cooled to 100 C before returning to the engine.

Test results with a 65 kW Chevrolet Citation[†] (1980) engine showed 48% more efficiency than with gasoline. The compression ratio of the engine was changed from its original 8.3:1 to 14:1. See Finegold and McKinnon (1982).

13.3.4 Anaerobic Digestion

Fermentation transforms a limited number of substances (sugars) into an ethanol/water mixture from which the alcohol has to be recovered by means of an energy-intensive distillation. In contrast, anaerobic digestion, transforms many different vegetable and animal substances into methane, a gas that, being of rather low solubility in water, evolves naturally and can be collected with a minimal expenditure of energy. Fermentation is well understood, whereas digestion has been less studied because, until recently, important industrial applications were few, mostly for sewage treatment.

Most of the energy in the raw material appears in the methane produced; only little is used in creating microbial cells. Digestion is, therefore, an efficient way to refine biomass.

Consider the digestion of glucose:

$$(CH_2O)_6 = 3CH_4(g) + 3CO_2(g). \tag{13.2}$$

One kilomole of glucose has a higher heat of combustion of 2.81 GJ. The 3 kilomoles of methane have a heating value of $3 \times 0.89 = 2.67$ GJ. Although the reaction is exothermic, little energy is lost: 95% of it appears in the methane. The enthalpy change is -0.14 GJ per kmole of glucose. On the other hand, the free energy change is much larger: -0.418 GJ per kmole of glucose. Since $\Delta S = (\Delta H - \Delta G)/T = 927$ kJ/K per kmole of glucose, it can be seen that the reaction leads to a substantial increase in entropy—the thermodynamic driving force is, consequently, large.

[†] General Motors exhibits sheer genius in naming some of its cars. In addition to a car named Citation, and another, Impact (the early name of the EV-1), it also had the Nova, which in Spanish means "It won't go."

Anaerobic digestion proceeds in four distinct steps:

1. *Hydrolysis* A group of different reactions mediated by several types of fermentative bacteria degrade various substances into fragments of lower molecular mass (polysaccharides into sugars, proteins into peptides and amino acids, fats into glycerin and fatty acids, nucleic acids into nitrogen heterocycles, ribose, and inorganic phosphates). This step solubilizes some normally insoluble substances such as cellose.
2. *Acidogenesis* Acid-forming bacteria produce carbon dioxide and a series of short-chain organic acids and alcohols.
3. *Acetogenesis* Further degradation is promoted by acetogenic bacteria that convert the alcohols and higher acids into acetic acid, hydrogen, and carbon dioxide.
4. *Methanogenesis* The acetic acid, hydrogen, and carbon dioxide produced in Steps 1, 2, and 3 are used by methanogenic archaea[†] to produce methane and carbon dioxide from the acid and methane and water from the hydrogen and carbon dioxide.

Table 13.1 shows some of the reactions that take place in each of these four steps.

Reaction 1 in the table is the hydrolysis of cellulose, always a difficult step. Nature has chosen cellulose as a structural material because of its stability. The fact that lignin tends to protect cellulose makes it that much more difficult to attack. Removing lignin would accelerate the process but may prove noneconomical.

Reactions 2, 3, and 4 transform glucose (from Reaction 1) directly into (ionized) acids, respectively, acetic, propanoic, and butanoic with liberation of carbon dioxide and hydrogen (in the case of the last two reactions).

Reaction 5 hydrolyzes glucose into carbon dioxide and hydrogen.

Reactions 6 though 10 produce acetic acid from some of the acids resulting from Reactions 2, 3, and 4.

Finally, Reactions 11, 12, and 13 lead to the final product: methane and carbon dioxide (biogas).

The result of the digestion process is a methane/carbon dioxide mixture called **biogas**, typically containing 65% of methane, by volume.

Steps 1, 2, and 3 are mediated both by **obligate** anaerobes and by **facultative** anaerobes that can operate in the absence of oxygen but

[†] Modern taxonomy classifies life forms into three domains: Eubacteria (true bacteria whose genetic material is not collected into a nucleus), archaea, and eucarya (eucarya have a nucleus). Some archaea exhibit unusual behavior, thriving at surprisingly high temperatures (above 100 C); they are **hyperthermophiles**. Their enzymes, stable at high temperatures, are used industrially when operation at elevated temperatures is necessary. Notice the omission of viruses, which, according to this view, are not considered alive.

Table 13.1 Estimated Free Energy Changes of Some Reactions in Aerobic Digestion (after D. L. Klass 1984)

	Reaction	Conditions: 25 C, pH = 7.	Free energy (MJ/Kmole)
	Fermentative bacteria		
1.	$C_6H_{10}O_5 + H_2O \rightarrow (CH_2O)_6$		−17.7
2.	$(CH_2O)_6 \rightarrow 3CH_3CO_2^- + 3H^+$		−311
3.	$(CH_2O)_6 + 2H_2O \rightarrow CH_3CH_2CO_2^- + H^+ + 3CO_2 + 5H_2$		−192
4.	$(CH_2O)_6 \rightarrow CH_3CH_2CH_2CO_2^- + H^+ + 2CO_2 + 2H_2$		−264
5.	$(CH_2O)_6 + 6H_2O \rightarrow 6CO_2 + 12H_2$		−25.9
	Acetogenic bacteria		
6.	$(CH_2O)_6 + 2H_2O \rightarrow 2CH_3CO_2^- + 2H^+ + 2CO_2 + 4H_2$		−216
7.	$CH_3CH_2CO_2^- + H^+ + 2H_2O \rightarrow CH_3CO_2^- + H^+ + CO_2 + 3H_2$		+71.7
8.	$CH_3CH_2CH_2CO_2^- + H^+ + 2H_2O \rightarrow 2CH_3CO_2^- + 2H^+ + 2H_2$		+48.3
9.	$CH_3CH_2OH + H_2O \rightarrow CH_3CO_2^- + H^+ + 2H_2$		+9.7
10.	$2CO_2 + 4H_2 \rightarrow CH_3CO_2^- + H^+ + 2H_2O$		−94.9
	Methanogenic archaea		
11.	$CH_3CO_2^- + H^+ \rightarrow CH_4 + CO_2$		−35.8
12.	$CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$		−131
13.	$HCO_3^- + H^+ + 4H_2 \rightarrow CH_4 + 3H_2O$		−136

are not poisoned by this gas. Step 4 is mediated by strictly anaerobic archaea, which do not survive the presence of oxygen. Digestion is the result of the cooperation of many microorganisms—a complete ecological system.

In a batch mode, the rate-limiting steps are Steps 1 and 4 (the one that involving methanogens)—Steps 2 and 3 proceed rapidly, while Step 4 proceeds at a much slower rate.

Digestion can be easily demonstrated in the laboratory by mixing organic matter (manure, for instance) with water and placing it in a sealed jar as suggested in Figure 13.15. The center opening is for loading the mixture, the one on the left is for removing of samples of the liquid (mostly for pH determination), and the one on the right is the biogas outlet. The gas produced by the digester can be washed by bubbling it through a water-filled beaker after which it is collected for chromatographic analysis.

To expedite the process, the mixture should be inoculated with sludge from another digester (the local sewage plant is a convenient source). The jar should be kept at about 37 C. After a few days, the pH of the mixture drops markedly, indicating that the acid-forming step is in progress.

Initially, little gas evolves, but sampling indicates that oxygen is being consumed. Soon, gas starts bubbling out of the liquid, and the chromatograph will show it to be mostly carbon dioxide. In some 10 days, methane

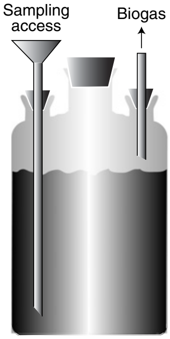


Figure 13.15 A laboratory digester.

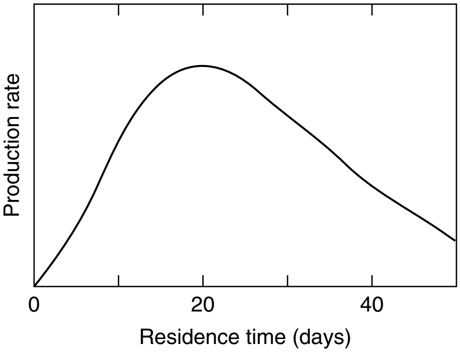


Figure 13.16 Typical biogas production rate in batch digesters.

starts appearing in the output and its relative concentration builds up rapidly to the final level.

The rate of gas production—that is, the digestion rate, varies as shown in Figure 13.16. The integrated production is shown in Figure 13.17. Practical digesters operate either in the batch or in the continuous mode. Batch digesters are loaded with an input slurry and then sealed. The different microorganisms must then grow, and the four digestion phases will occur more or less in sequence. Steady state is never achieved, and long residence times (30 to 60 days) are required.

Continuous or semicontinuous digesters receive slurry at one end of the tank; spent sludge is withdrawn from the other. The material slowly moves through regions where the different reactions occur in nearly steady-state conditions. A shorter residence time is achieved (10 to 20 days). Long residence times translate into small **loading rates**—that is, into the need for large tanks for a given gas production rate, increasing investment cost.

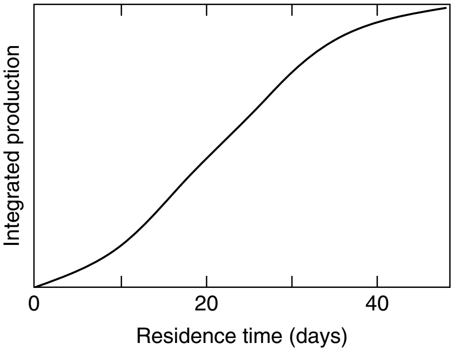


Figure 13.17 Integrated biogas production in batch digesters.

The influent to the digester consists of a watery slurry containing chopped pieces of organic matter. Part of the solids, whether in suspension or in solution, can (ideally) be digested and part cannot. Those that can be digested are called **volatile solids** (VS) and those that cannot are called **fixed solids** (FS).

Methane yield, Y, is the amount of methane produced from 1 kg of VS. Depending on the influent, the ideal yield, Y_0, corresponding to a total digestion of the VS, ranges between 21 and 29 MJ of methane per kg of VS.

Digesters are able to use only a fraction, R, of the VS; their yield is $Y = RY_0$. R depends on the digester design, on the nature of the influent, and on the residence time. It may range from as low as 25% for grass clippings to as high as 50% for activated sewage sludge.

Actual gas production depends on the **allowable loading rate**, L, measured in kg (VS) per cubic meter of the digester per day.

The power density of the digester is

$$P = LRY_0 \quad \text{joules/m}^3 \text{ per day.} \tag{13.3}$$

For representative values, $Y_0 = 25$ MJ(CH_4)/kg(VS) and $R = 0.4$,

$$P = 10L \quad \text{MJ(CH}_4\text{)/m}^3 \text{ per day.} \tag{13.4}$$

or

$$P \approx 100L \quad \text{W(CH}_4\text{)/m}^3. \tag{13.5}$$

Batch digesters can handle loading rates up to 1.5 kg (VS) $\text{m}^{-3}\text{day}^{-1}$ and can produce methane at a rate of 150 W/m^3 of digester. Continuous digesters, owing to their loading rates of up to 6 kg (VS) $\text{m}^{-3}\text{day}^{-1}$, produce some 600 W/$\text{m}^3$. Larger power densities lead to more economical methane production. One way to increase power density is to decrease the residence time. This can be accomplished in several ways, including:

1. Operation at higher temperature,
2. Agitation,
3. Microorganism immobilization,
4. Strain selection,
5. Material addition,
6. Pretreatment of the influent.

Three temperature ranges favor digestion:

1. **Psychrophilic** range centered around 5 C
2. **Mesophilic** range centered around 37 C
3. **Thermophilic** range centered around 55 C.

The fastest digestion is thermophilic. However, only large installations can be economically operated at a high temperature owing to excessive heat losses in small digesters (because of their large surface-to-volume ratio).

Methane and carbon dioxide are only slightly soluble in water; microbubbles of these gases tend to surround the methanogenic organisms that produce them, reducing their contact with the liquid and retarding the reaction. Agitation will dislodge these bubbles, as will lowering the pressure and increasing the temperature. Agitation can sometimes be economically achieved by rebubbling the produced gas through the slurry.

The greatest gain in digestion speed is achieved by microorganism immobilization. The ecosystem that promotes digestion is slow in forming, and every time the residue is removed, so are the microorganisms, and new cultures have to be established. This is true for both batch and continuous digesters. One way to ensure the retention of the microflora is to use a **biological filter** in which the digester is filled with inert chips on which a bacterial slime is formed. The residue is removed while the chips remain in the equipment. For this arrangement to operate properly, it is necessary that the residue be sufficiently fluid, a condition difficult to achieve with many of the raw materials used. In fact, a problem in the operation of a digester is the clogging caused by the thick sludge formed. Thus, adequate pretreatment of the raw material may be essential for efficient operation.

To improve digestion rates, one can add to the influent certain materials such as enzymes, growth factors, and fertilizers. Since proper carbon-to-nitrogen (C/N) and carbon-to-phosphorus (C/P) ratios are necessary for efficient digestion, it may be important to add chemicals for the correction of these ratios when certain raw materials are to be processed. Good C/N ratios hover around 20:1 and C/P ratios, around 80:1.

If the biogas is to be shipped over long distances, it may be important to upgrade it to pipeline quality (by removing much of the CO_2), using any of the methods discussed in Chapter 10.

Biogas is produced spontaneously by garbage dumps and by wastewater treatment plants. The gas from garbage dumps is most often simply allowed to escape into the atmosphere. The methane in it is a potent greenhouse gas, and it may be accompanied by objectionable impurities such as hydrogen chloride. In some instances, wells are driven into garbage dumps, the biogas is collected and purified, and then injected in normal natural gas pipelines. Wastewater treatment plants frequently collect the gas and flare it, making no use of the energy it contains.

It is estimated that methane, corresponding energetically to some 5000 barrels of oil, is released daily by wastewater treatment plants. It is a modest amount of energy but one that should not be wasted. Some municipalities feed this gas to fuel cells to produce electricity and usable heat. This eliminates the need for separate emergency generators.

Currently, the main application of anaerobic digestion is in sewage treatment. Sewage and other liquid organic waste, from which, in a **preliminary treatment**, large solid objects have been removed, are piped to settling tanks (**primary treatment**) where a great deal of the solids precipitate out forming a **sludge** that is sent to an anaerobic digester. The output of the digester is methane gas and a type of sludge that can be sold to farmers as fertilizer (provided heavy metals and other toxic substances are not present). The liquid left over from this phase is sometimes sent to a **secondary treatment** tank where *aerobic* bacteria remove much of the pathogens still left in the system. The action in this tank can be expedited by supplying the oxygen the bacteria consume. This can be done either by aeration (at the cost of some energy) or by growing microalgae that photosynthetically produce the gas. The output of this stage is safe to release into the environment (again, provided it is not contaminated by toxic substances). Unfortunately, this effluent contains nitrates and phosphates that may promote algal bloom (**eutrophication**). Further improvement of the effluent can be accomplished in a **tertiary treatment**, which may yield pure drinking water. Many technologies can be used at this last stage, including reverse osmosis. Alternatively, the effluent can be sent to artificially created wetlands, where nature will take care of the purifying.

Sewage treatment does not necessarily include all of the preceding stages. It can be truncated, frequently right after the primary treatment.

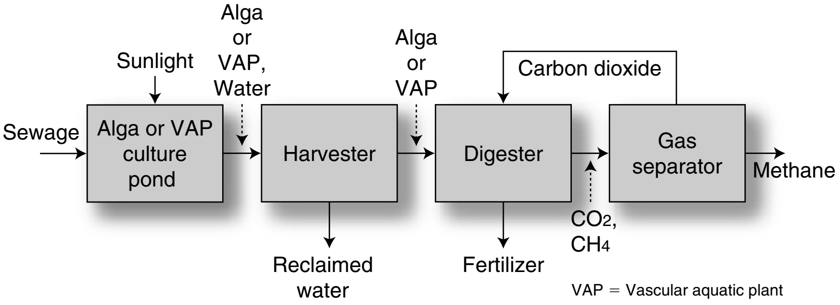


Figure 13.18 Combination sewage treatment–fertilizer–methane plant.

Tertiary treatment is not too commonly used. Sometimes, extreme truncation is the practice: the sewage is simply not treated at all.

The use of algae for aeration suggests that these could be harvested for further anaerobic digestion, thus producing additional methane. For this to be possible, selected algae must be used to make the harvesting practical. They must be relatively large so that they can be filtered out. The difficulty is in ensuring that the correct species remain in the system, otherwise they may be naturally replaced by species unsuitable for harvesting.

One possibility for producing additional methane and more fertilizer and for removing toxic substances is to allow certain vascular aquatic plants to grow in the sewage. Among the most attractive of these plants is the water hyacinth (*Eichhornia crassipes*). This is a floating plant of extraordinary productivity. Under favorable circumstances it will produce 600 kg of *dry* biomass per hectare per day (Yount 1964). This enormous vegetative capacity has transformed *E. crassipes* into a vegetable pest. It has invaded lakes and rivers in Africa and Asia and even the United States, interfering with the movement of boats, with fishing, and with hydroelectric plants. In addition, *E. crassipes* can harbor in its leaves the snail carrying the fluke that causes the serious disease, schistosomiasis. The problem was big enough for a specialized journal, the *Hyacinth Control Journal*, to be created. Only in its country of origin is the water hyacinth in (partial) equilibrium with its local predator—the Brazilian manatee, *Trichechus manatus manatus*, known locally as “peixe boi” (cow fish). *E. crassipes* is very useful in removing heavy metals and phenols from polluted waters.

Wolverton (1975) has demonstrated that the plant can remove, daily, some 0.3 kg of heavy metals and over 50 kg of phenol per hectare. To maintain this removal capacity, periodic harvest (every five weeks?) must be carried out. The collected material can be used for CH_4 production, but the sludge should not be used as fertilizer owing to the accumulation of toxins.

13.4 Photosynthesis

Photosynthesis is an extremely complicated process best left to plant physiologists and biochemists. Here, with an apology to these experts, we will make a somewhat unscientific attempt to look at a plant leaf as a system to be examined from the point of view of the engineer. This approach, though far from accurate, will, it is hoped, lead to some insight into the behavior of plants and of some of the mechanisms involved.

Formally, photosynthesis is a process through which carbon dioxide and water are transformed into carbohydrates and oxygen according to

$$n\mathrm{CO_2} \ + \ n\mathrm{H_2O} \rightarrow (\mathrm{CH_2O})_n + n\mathrm{O_2}. \tag{13.6}$$

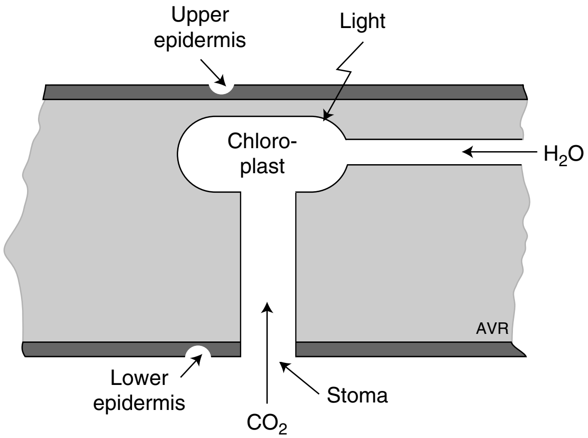


Figure 13.19 Schematic representation of the photosynthesizing unit in a plant leaf.

It is well known that this is a complex process. It is not a simple dissociation of the carbon dioxide with subsequent attachment of the freed carbon to a water molecule, as it might appear from an inspection of the chemical equation (13.6). By means of radioactive tracers, it can be shown, that the oxygen liberated comes from the water, not from the carbon dioxide. Figure 13.19 represents schematically the photosynthesizing unit in a plant leaf. Photochemical reactions occur in the compartment labeled **chloroplast**. This compartment must have a transparent wall (labeled **upper epidermis** in the figure) to allow penetration of light. Water and carbon dioxide must be conveyed to the chloroplast.

In this schematic, water is brought in through a duct, while carbon dioxide enters from the atmosphere via a channel whose opening is on the **lower epidermis**.

This opening is called **stoma** (plural: either stomas or stomata; from the Greek, "mouth"). A waxy coating renders epidermises impermeable to water. Thus, losses of water occur mainly through the stomata.

Carbon dioxide from the atmosphere moves toward the chloroplast by diffusion. Its flux, ϕ, is therefore proportional to the difference between the concentration, $[CO_2]_a$, of the carbon dioxide in the atmosphere and the concentration, $[CO_2]_c$, of this gas at the chloroplast.

The flux is also proportional to the **stomatal conductance**, V, a quantity that has the dimensions of velocity:

$$\phi = V\Big([CO_2]_a - [CO_2]_c\Big). \tag{13.7}$$

The carbon dioxide sink is, of course, the photosynthetic reaction in the chloroplast. Under steady-state conditions, the rate of carbon dioxide uptake per unit leaf area is equal to the incoming flux. This rate is proportional to the light power density, P (W m^{-2}), the dioxide concentration, $[\mathrm{CO_2}]_c$ (kmoles m^{-3}), and to the reaction rate, r (m^{3}J^{-1}). Thus,

$$\phi = r[\mathrm{CO_2}]_c P. \tag{13.8}$$

Eliminating $[\mathrm{CO_2}]_c$ from Equations 13.7 and 13.8, one obtains

$$\phi = \frac{r[\mathrm{CO_2}]_a P}{1 + (r/V)P}. \tag{13.9}$$

ϕ is measured in kmoles of $\mathrm{CO_2}$ uptake per second per square meter of leaf area.

Under conditions of low light power densities—that is, when $P << V/r$—the carbon dioxide uptake rate is proportional to the concentration of this gas in the atmosphere and to the light power density:

$$\phi = r[\mathrm{CO_2}]_a P. \tag{13.10}$$

When the power density is high—that is, when $P >> V/r$—the uptake rate becomes

$$\phi = V[\mathrm{CO_2}]_a. \tag{13.11}$$

At high light power densities, the carbon dioxide uptake rate is independent of P but is still proportional to $[CO_2]_a$.

The reaction rate, r, is a basic characteristic of photosynthesis, and one would expect it to be the same for all advanced plants. Measurements made by plant physiologists confirm this expectation for cases in which there is a normal concentration of oxygen in the air (Ehleringer and Björkman 1977).

r can be determined by measuring the $\mathrm{CO_2}$ uptake rate as a function of the light power density in the region in which Equation 10 is valid.

We have, then,

$$r = \frac{1}{[\mathrm{CO_2}]_a}\frac{\phi}{P}. \tag{13.12}$$

From the Ehleringer and Björkman data, it can be seen that the slope of the ϕ versus P plot in the linear region (see Figure 13.22 later in the chapter) is 178 nmoles/J $= 178 \times 10^{-12}$ kilomoles/J. The measurements were made at RTP with a carbon dioxide concentration of 330 ppm—that is, with a $\mathrm{CO_2}$ pressure of 330 μatm. At RTP the volume, $\forall$, occupied by 1 kilomole of any perfect gas is

$$\forall = \frac{RT}{p} = \frac{8314 \times 298}{1.0133 \times 10^5} = 24.45 \ \ \mathrm{m^3/kmole}. \tag{13.13}$$

Hence, 330 ppm correspond to

$$[\mathrm{CO_2}] = \frac{330 \times 10^{-6}}{24.45} = 13.5 \times 10^{-6} \ \ \mathrm{kmoles/m^3}. \tag{13.14}$$

Thus,

$$r = \frac{178 \times 10^{-12} \ \ \mathrm{kmoles/J}}{13.5 \times 10^{-6} \ \ \mathrm{kmoles/m^3}} = 13.2 \times 10^{-6} \ \ \mathrm{m^3/J}, \tag{13.15}$$

and

$$\phi = \frac{13.2 \times 10^{-6}[\mathrm{CO_2}]_a P}{1 + 13.2 \times 10^{-6} P/V} = \frac{178 \times 10^{-12} P}{1 + 13.2 \times 10^{-6} P/V}. \tag{13.16}$$

The second part of the equation is for a $\mathrm{CO_2}$ concentration of 330 ppm. This equation should (roughly) predict the carbon dioxide uptake rate as a function of the sunlight power density for any plant provided the stomatal velocity, V, is known. Arbitrarily assigning a value of 0.004 m/s to V, independent of the light power density, we would get the plot of Figure 13.20.

At this point, we can predict how the photosynthetic efficiency depends on light power density. The useful output of photosynthesis is the energy content of the created biomass, which is about 440 MJ per kilomole of carbon. This is more than the enthalpy of formation of carbon dioxide (393.5 MJ/kmole, in absolute value). The additional energy comes from the hydrogen that is always present in plant tissue.

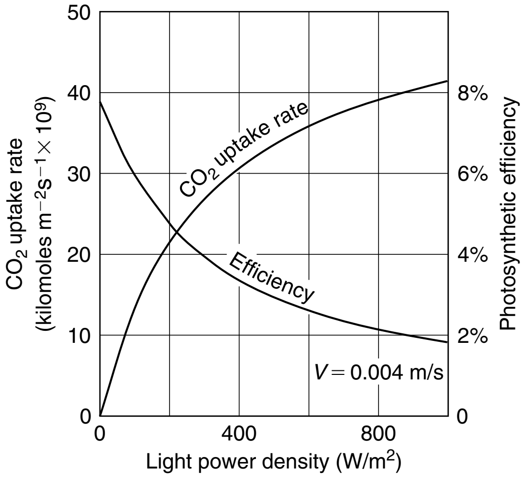


Figure 13.20 A very simple model of photosynthesis predicts a reduction of efficiency with rising light power density.

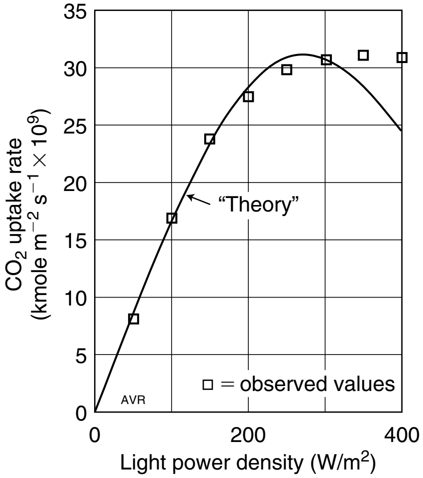


Figure 13.21 Using a light power density-dependent V, a better match between theory and observation is obtained.

Thus, the uptake of 1 kmole of CO$_2$ corresponds to the **fixation** of 440 MJ. The rate of fixation is therefore 440 ϕ MWm^{-2}, and the photosynthetic efficiency is

$$\eta_\nu \simeq 4.4 \times 10^8 \phi / P. \tag{13.17}$$

If this is true, then the largest possible photosynthetic efficiency is

$$4.4 \times 10^8 \times 178 \times 10^{-12} = 0.078. \tag{13.18}$$

In order to test the simple model that has been developed, we need measured data on carbon dioxide uptake rate as a function of light power density.

Consider the data of Björkman and Berry (1973) for *Atriplex rosea* (Common name: Redscale). We find that, although the general shape of the ϕ vs. P curve of predicted and measured values is roughly the same, the numerical values are quite different. It would be astounding if it were otherwise because we took the value, $V = 0.04$ m/s, out of a hat. In addition, we assumed, without justification, that V is independent of P. By experimentation, we find that if we make $V = 0.0395 \exp(-0.00664P)$,[†] we do get a better match between our "theory" and observation, at least in the lower light power density region. See Figure 13.21.

There is indeed a mechanism to regulate the stomatal conductance. The lips of the stomata consist of special cells that, when full of water,

[†] With a medieval twist of mind, we chose this exponential dependence because it is well known that nature, which abhors a vacuum, actually loves exponentials.

become turgid (swell up). When dry, these cells flatten out and close the opening. It is not impossible that the size of the stomatal opening is controlled, at least indirectly, by the light intensity on the leaf because the more light, the greater the desiccation. It is not unreasonable that V decreases as P increases. *Increasing the illumination may cause a reduction in the stomatal velocity accentuating the predicted decrease in efficiency.*

Björkman and Berry data are not only for *A. rosea* but also for another plant of the same genus but of a different species: *A. triangularis* (common name: Spearscale). Figure 13.22 shows that *A. rosea* is substantially more efficient than *A. triangularis* under the conditions of the experiment ($T = 25\,\mathrm{C}$ and $[\mathrm{CO_2}] = 300\,\mathrm{ppm}$). The CO_2 uptake rate for either plant peaks well below full sunlight ($1000\,\mathrm{W m^{-2}}$). Plants are optimized for the average light they receive (one leaf shading another), not for full sun.

To understand the reason for the difference in photosynthetic performance of the two species of *Atriplex*, one must become more familiar with the photosynthesis mechanism.

In most plants, photosynthesis is carried out by two pigments—**chlorophyll A** and **chlorophyll B**. These, when purified and in solution, are green in color—that is, they reflect green light that, therefore, cannot be utilized by the plant and, consequently would be wasted. In living plants, the situation may be less extreme, but leaves are still green and thus fail to use a band of the solar spectrum in which the power density is substantial. Figure 13.23 shows the absorption spectra for the two chlorophylls.

In all chlorophyll plants, transformation of CO_2 into carbohydrates is performed by a chain of reactions called the **Calvin–Benson cycle**. The

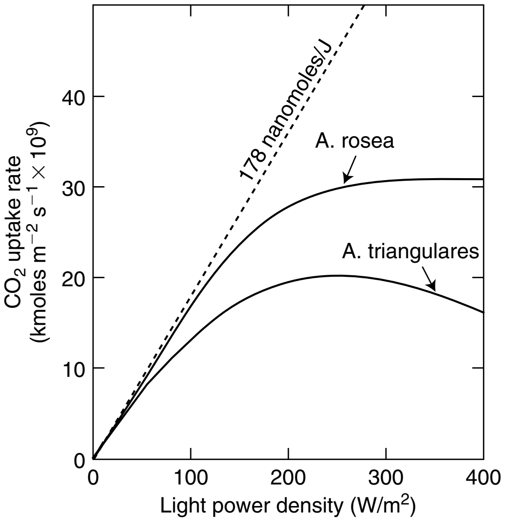


Figure 13.22 CO_2 uptake rate by two different varieties of *Atriplex*.

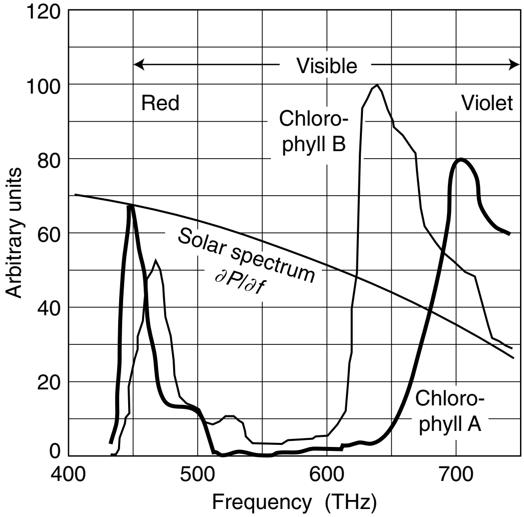


Figure 13.23 Absorption spectra of chlorophylls A and B.

mechanism involves the combination of CO_2 with a 5-carbon sugar called RuDP (ribulose-1,5-diphosphate) and the formation of two molecules of a 3-carbon substance—PGA (phosphoglyceric acid), which is then converted into carbohydrates and, part of it, into more RuDP to perpetuate the cycle.

The difference between *A. rosea* and *A. triangularis* resides not in the basic cycle but rather in the manner in which CO_2 is delivered to the cycle. In *A. triangularis*, atmospheric CO_2 is taken directly to the RuDP, while in *A. rosea*, there is an intervening ancillary "pumping" mechanism that delivers the CO_2 to the RuDP in a manner somewhat parallel to the delivery of oxygen to animal cells by the flowing blood. See Figure 13.24.

The "hemoglobin" of plants is PEP (phospho-enol-pyruvate) that has great affinity for CO_2. The end products of the PEP-CO_2 reaction are malic and aspartic acids, each containing four carbons per molecule. In the presence of appropriate enzymes, the two acids release their CO_2 to the RuDP and are eventually transformed back to PEP, closing the pumping cycle.

Plants that use PEP and consequently create a 4-carbon acid as the first product in the photosynthesis chain are called **C4** plants, whereas those whose first photosynthesis product is a 3-carbon acid, are called **C3** plants. As a general rough rule, almost all trees and most shrubs are C3 plants, while many tropical savanna grasses and sedges are C4.

Carbon dioxide is much more reactive with PEP than with RuDP. In addition, when the CO_2/O_2 ratio becomes small, the presence of oxygen interferes with the functioning of the RuDP and photosynthesis is halted. Thus, under low CO_2 concentrations, C3 plants interrupt their photosynthesis, while C4 plants continue to operate.

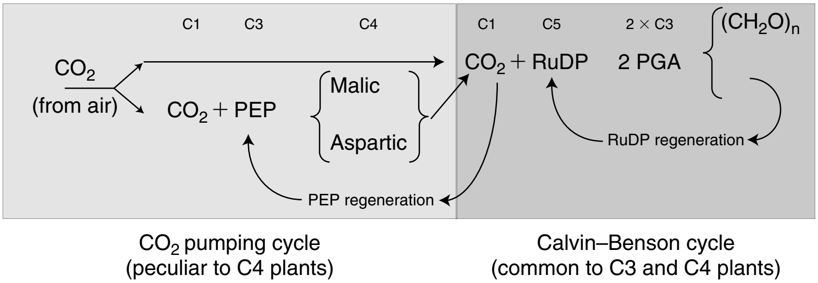


Figure 13.24 The difference between a C3 and a C4 plant is that the latter has an additional mechanism to extract CO_2 from the air and deliver it to the Calvin–Benson cycle. The symbols C3, C4, and so on indicate the number of carbon atoms in the molecule.

Low CO_2 concentrations at the chloroplast occur when the stomatal conductance is insufficient owing to the desiccation of the lipcells in the stoma. This saves water (which evaporates through the stomata) but also hinders the flow of CO_2. Nevertheless, the plant metabolism continues even though photosynthesis may be halted. The plant continues its respiration, burning carbohydrates and evolving CO_2, an activity exactly opposite to photosynthesis. This phenomenon can be detected in any plant in the dark. Because they interrupt their photosynthesis under high light power densities, C3 plants, are also net CO_2 *producers* under such conditions. This is called **photorespiration** and cannot be detected in C4 plants.

An interesting experiment demonstrates the difference between C3 and C4 plants. A C3 and a C4 plant are placed under a bell jar and are properly illuminated. Both may prosper as long as there is an adequate CO_2 supply. Soon, however, this gas is used up, and the C3 plant will go into photorespiration consuming its own tissues, shriveling up and producing the CO_2 that the C4 plant scavenges to grow.

When the CO_2 supply is adequate, C4 plants lose their advantage because of the no longer necessary "pumping" mechanism that uses up energy. Figure 13.25 shows that, at high CO_2 concentrations, *A. triangularis* is more efficient than *A. rosea* (data from Björkman and Berry 1973).

The advantages of the C4 plants over the C3 are maximum under conditions of high light power density, high temperatures, and limited water supply. These are the conditions found in semiarid tropical regions. Sugarcane, a tropical plant, for instance, is a C4 plant. C4 plants also outperform C3 plants when CO_2 concentrations are low. During the last glacial maximum (some 20,000 years ago), the CO_2 pressure in air fell to 190–200 μatm. This caused grasses to replace forests in many regions. See Street-Perrott et al. (1997).

Water plants, of course, do not have to save water. One would expect that their stomata do not have regulating valves (a water-saving adaptation). Indeed, leaves of water plants wilt quite rapidly when plucked. In

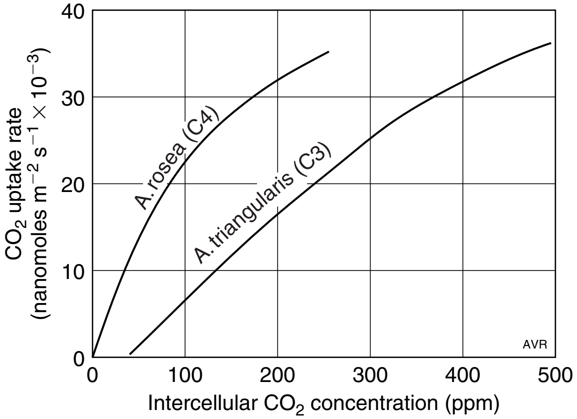


Figure 13.25 At high CO_2 concentrations, C3 plants are more efficient than C4.

such plants, the stomatal conductance will be independent of light level, and their CO_2 uptake rate will not fall off with increasing light power density.

Thus, vascular aquatic plants are potentially a more efficient biomass producer than land plants. This is confirmed by the high productivity of such water plants as the water hyacinth (*Eichhornia crassipes*).

References

Atsumi, Shota, T. Hanai, and J. C. Liao, Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451**, pp. 86–90, **2008**.

Björkman, O., and J. Berry, High-efficiency photosynthesis, *Scientific American* **229**, p. 8093, **1973**.

Ehleringer, James R., and O. Björkman, Quantum yields for CO_2 uptake in C_3 and C_4 plants, *Plant Physiology* **59**, pp. 86–90, **1977**.

Finegold, J., and T. McKinnon, Dissociated methanol test results, *Alcohol Fuels Program Technical Review*, U.S. Department of Energy (SERI), Summer **1982**.

Gonzalez, Ramon, and S. S. Yazdani, Anaerobic fermentation of glycerol: A path to economic viability for the biofuels industry. *Current Opinion in Biotechnology*, 17532205, May 24, **2007**.

Klass, Donald L., Methane from anaerobic fermentation, *Science,* **223**, p. 1021, March 9, **1984**.

Schmidt, Lanny, D., and Paul L. Dauenhauer, Hybid routes to biofuels, *Nature* **447**, pp. 914–915, **2007**.

Street-Perrott, F. Alayne, Y. Huang, R. A. Perrott, G. Eglinton, P. Baker, L. Ben Khelifa, D. D. Harkness, and D. O. Olago, Impact of lower atmospheric carbon dioxide on tropical mountain ecosystems, *Science*, **278**, 1422, November 21, **1997**.

Wolverton, B. C. (with coauthors), A Series of NASA Technical Memoranda, TM-X-727720 through 27, TM-X-727729 through 31, NASA National Space Technology Laboratories, Bay St. Louis, MS 39520, **1975**.

Yount, J. L., Report of the 35th annual meeting, Florida Anti-mosquito Association, p. 83, **1964**.

PROBLEMS

13.1 A plant leaf takes up 0.05 μmole of CO_2 per minute per cm^2 of leaf area when exposed to sunlight with a power density of 50 W/m^2 in an atmosphere containing 330 ppm of CO_2. When the light power density is raised to 600 W/m^2, the uptake is 0.36 μmole $min^{-1}cm^{-2}$. Assume that in the above range, the stomata do not change their openings.

What is the expected uptake at 100 W/m^2 in the same atmosphere and the same temperature as above? And at 1000 W/m^2?

Measurements reveal that the uptake at 1000 W/m^2 is only 0.40 μmole $min^{-1}cm^{-2}$. This reduction must be the result of a partial closing of the stomata. What is the ratio of the stomatal area at 1000 W/m^2 to that at 100 W/m^2?

What is the expected uptake at 100 W/m^2 if the CO_2 concentration is increased to 400 ppm?

13.2 An automobile can be fueled by dissociated alcohol. The energy necessary for such dissociation can come from waste exhaust heat. In the presence of catalysts, the process proceeds rapidly at temperatures around 350 C.

Consider liquid methanol that is catalytically converted to hydrogen and carbon monoxide. Compare the lower heats of combustion of methanol with those of the products. Do you gain anything from the dissociation?

Compare the entropy of gaseous methanol with that of the products. Does this favor the reaction?

Assume a gasoline engine with a 9:1 compression ratio fueled by

1. gasoline,
2. methanol, and
3. dissociated methanol.

Assuming that the three fuels lead to identical engine behavior, compare the energy per liter of the fuel compared with that of gasoline.

The compression ratio is now changed to the maximum compatible with each fuel:

gasoline:	9:1
methanol:	12:1
dissociated methanol:	16:1

Remembering that the efficiency of a spark-ignition is

$$\eta = 1 - r^{1-\gamma},$$

where r is the compression ratio and γ is the ratio of the specific heats (use 1.4), compare the new energy per liter ratios.

The gasoline and methanol molecules are complex, and this leads to a low value of γ. With hydrogen and carbon monoxide, γ is much higher. Change the above calculations using 1.2 for gasoline and methanol, and 1.7 for the dissociated methanol.

13.3 Under proper conditions, water hyacinths (*Eichhornia crassipes*), a floating plant, will grow at such a rate that their dry biomass increases 5% per day. The total water content of these plants is high (94%). Nevertheless, 400 kg of dry matter can be harvested daily from one hectare of plantation.

Consider a plantation with a 1-hectare area consisting of a long canal (folded upon itself). At the starting point (seeding end), the canal is 0.5 m wide. It expands gradually enough to just accommodate the growing plants that are slowly swept along by the current. Assume a current of constant speed such that the plants take 60 days to float from the seeding end to the harvesting end.

1. How wide must the canal be at the harvesting end?
2. How long must the canal be?
3. How much energy is harvested per day (express this in GJ and in barrels of oil)?
4. How many kilograms of wet plants must be used daily as seed?
5. Assume that 50% of the biomass energy is converted into methane through a digestion process. Estimate the methane yield in cubic meters per day.
6. If the methane is burned in a gas turbine or in a diesel engine with 20% efficiency, what is the average electric power that can be generated?
7. Assuming that the average depth of the canal is 1 m, what is the amount of water that has to flow in daily provided there is no loss by evaporation and infiltration?
8. Do you have any good ideas of how to ensure that the water velocity is kept reasonably constant notwithstanding the expanding canal?

13.4 Sugarcane is submitted to an illumination of 500 W/m^2. Assuming a stomatal velocity of 6 mm/s, what is the photosynthetic efficiency (defined as the ratio of the heat of combustion of the dry biomass generated to the incident solar energy)?

13.5 Here is a typical task that an energy consultant might tackle:

The operator of a large alcohol distillery wants to know if it makes economic sense to use the leftover bagasse as a further source of ethanol. In the traditional process, the amount of bagasse obtained from 1 ton of burned and cropped sugarcane is larger than the amount that has to be burned to drive the distillation process. The excess is either sold or used to generate electricity for the plant. The question is how much additional alcohol can be obtained by hydrolyzing all the polysaccharides (cellulose and hemicellulose) in the leftover bagasse. We will make the following simplifying assumptions:

a. The hydrolysis will yield 600 grams of sugars (glucose and pentoses) per kilogram of polysaccharides.
b. The hydrolysis requires no energy (not true!).
c. The glucose-to-ethanol and the pentose-to-ethanol yields are the same as the sucrose-to-ethanol yields of the traditional process.
d. The data for this problem are those discussed in Section 13.3.2 of the textbook.

Calculate the additional amount of alcohol that can be obtained from 1 ton of burned and cropped sugarcane. Comment.

13.6 A digester consists of a cylindrical stainless steel tank with a diameter, d, and a height, $2d$. The metal is 3 mm thick.

An R-5 (American system) fiberglass blanket completely covers the tank.

The contents of the digester are agitated so that they are, essentially, at a uniform temperature of 37 C. To simplify this problem, assume that the influent (the material fed in) is preheated to 37 C.

Stainless steel has a thermal conductivity of $\lambda = 60$ W m^{-1} K^{-1}.

We desire a net production rate of 1 kW of methane. The digester must produce, in addition, enough methane to fire a heater that keeps the material in it at a constant temperature (the digestion process itself generates negligible heat). The efficiency of the heater is 70%.

Loading rate is $L = 4$ kg of volatile solids per cubic meter of digester per day. Assume that 1 kg of volatile solids produces 25 MJ of methane and that 40% of all the volatile solids in the influent are digested.

The outside temperature is such that the external walls of the digester are at a uniform 20 C.

Estimate the diameter of the digester.

13.7 A hypothetical plant has perfectly horizontal leaves. The carbon dioxide uptake rate, ϕ, depends linearly (in the usual manner) on the solar light power density, P, provided $P \leq 150$ W/m^2. Above this value, ϕ is constant, independently of P.

Assume the insolation at normal incidence is 1000 W/m^2 during all daylight hours. The latitude is 45°N.

1. What is the amount of carbon fixed by each square meter of leaf area during the winter solstice day?
2. Estimate the number of kilograms of dry biomass the plant produces on the winter solstice day per hectare of leaves.

13.8 A hemispherical, perfectly transparent container has a 5-m radius. The bottom part is covered with moist soil on which a bush with horizontal leaves is planted. The leaves are arranged in such a way that they do not shade one another. The volume occupied by the plant by the plant is 0.75 m^3, and its leaf area is 4 m^2.

The "air" inside the container has the following composition:

O_2	52 kg
N_2	208 kg
A	2.6 kg
H_2O	26 kg
CO_2	0.78 kg

The temperature inside the container is a uniform 298 K.

The soil is moist enough to supply all the water needed by the plant. However, no water is ever exchanged directly between soil and air.

The plant has the usual value of r (13.2×10^{-6} m^3/J) and has a stomatal conductance of 10 mm/s.

Argon has a molecular mass of 40 daltons.

The plant has a carbon dioxide uptake rate proportional to the illumination power density, P, for values of $P < 200$ W/m^2 and independent of P for greater values. The light has the same spectral distribution as the sun.

1. What is the air pressure inside the container?
2. What is the *total* carbon dioxide uptake rate (kmoles/s and mg/s) under the above circumstances, when the illumination power density is 150 W/m^2?
3. Make a rough estimate of how long it will take to reduce the CO_2 concentration to the level of normal (outside) air.
4. Assume that the leaf area of the plant does not change. Calculate more accurately the time required to reduce the carbon dioxide concentration to the normal value.
5. What is the composition of the air when the carbon dioxide concentration reaches its normal level?

13.9 Consult Google for properties of vegetable oils. Tabulate melting point (or cloud points) vs. iodine value. Plot melting points versus iodine value, and do a linear regression. What is the degree of correlation between these variables. Does the melting point rise or drop with increasing iodine value? Give a reasonable mechanism for such a behavior.

Do this for at least 12 different oils so that you can draw acceptable statistical inferences.

13.10 Draw the structural formula for 2,3-pentanediol and for 2,2,4-trimethylpentane.

Also draw the corresponding condensed structural formulas.

13.11 What is the main reason to prefer biodiesel to straight vegetable oils (SVO) as fuel for diesel engines?

What should you do to make SVOs more acceptable as diesel fuel (other than making biodiesel out of them)?

13.12 During anaerobic digestion, glucose is transformed into methane through a series of steps. The overall reaction is

$$(CH_2O)_6 \rightarrow 3CH_4 + 3CO_2.$$

Calculate the percentage of methane (by both volume and mass) produced.

13.13 Biodiesel is produced by transesterification of vegetable or animal oils. Usually, the glycerine in the oil is replaced by methanol. Since this amounts to using a fuel (methanol) to produce a fuel (biodiesel), one is entitled to question what energy gain is achieved.

Invariably, vegetable oils are a mixture of many different triglycerides, but since one of the most common fatty acids found in vegetable oils is palmitic acid, $C_{16}H_{32}O_2$, we will assume that our raw material is a pure triglyceride consisting of three palmitic acid groups. Thus, the final product—the biodiesel—will be methyl palmitate.

a. Stoichiometrically, how many kg of methanol are required for each kg of vegetable oil?
b. If the methanol, were used directly as fuel, how much energy would be released?
c. If 1 liter of biodiesel is used as fuel, how much energy is released?
d. What is the methanol-to-biodiesel energy ratio?

Fuel	**Density (kg/m^2)**	**Higher heat of comb. (MJ/liter)**
Diesel	850	40.9
Biodiesel	885	36.6
Methanol		22.7

13.14 You have 1000 kg of a vegetable oil that happens to consist of a single triglyceride (not a mixture of triglycerides). The three acids in each molecule are all palmitic acid (C16:0).

1. What is the molecular mass of the triglyceride?
2. What is the proper chemical name of this triglyceride?
3. Estimate the number of kilograms of glycerine produced when the above vegetable oil is transesterified with ethanol.

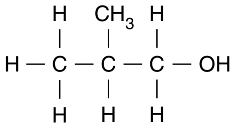
13.15 A carbon–carbon bond can be single, double, or triple. Which one is shorter: the single or the double?

13.16 What is the basic difference between a σ-bond and a π-bond?

13.17 Is amyl oleate a saturated or an unsaturated ester?

Amyl is the same as pentyl, the radical of the 5-carbon alkane, pentane.

13.18 The molecule below is being proposed as a biofuel (at least, some of its isomers are). Give two names for the molecule.



13.19 What is the iodine value of glyceryl trilinoleate? The molecular mass of the oil is 878 daltons, and the atomic mass of iodine is 127 daltons. Linoleic acid is C18:2.

13.20 What is the empirical formula of oleic acid? Don't just look it up in Google. Derive the formula starting from the original hydrocarbon and considering the progressive states of oxidation.

Chapter 14
Photovoltaic Converters

> Two plates of red copper, each of them superficially covered with a thin layer of cuprous oxide, are placed in a saturated solution of sodium chloride. If now one of these plates is exposed to light radiation and the other is kept in the dark, an electric current passes through the wire that connects the two electrodes. This current continues as long as the exposure lasts. The whole system returns to its original state in the dark. If the other electrode is illuminated, the electric current produced in the circuit flows in the opposite direction.
>
> *Science*, April 26, 1929

14.1 Introduction

Economics are crucial to the success of any energy utilization system, and photovoltaic converters are no exception. At present, development of such converters is at the stage in which prices are still coming down rapidly. Nevertheless, it is difficult to compare photovoltaic with other systems solely on the basis of investment cost. Although photovoltaics have low operating cost, since they consume no fuel, their peak power can be only realized on a clear day, with the converter facing the sun. The average power will be less than half the peak power for sun tracking systems (owing to nighttime) and less than one-fourth of peak power for nontracking systems.[†]

Because of the intermittence of the sunlight, storage systems or standby power generators are frequently required, substantially increasing investment costs. Exceptions include photovoltaics used to supply peak loads that coincide with periods of maximum insolation. Also, photovoltaics used directly to power irrigation systems need no storage provisions.

An important consideration in any type of power plant is the ratio of the *average* power delivered to the peak installed power. This is related to the plant utilization factor. In Chapter 1, we saw that, in the United States., nuclear power plants operate with an utilization factor of over 90%—that is, over 0.9 watts are (on average) actually delivered for each watt of installed plant capacity. Maintenance shutdowns are the main limit to full utilization of the installed power capacity. In the case of wind and solar power, the main limitation comes from the unreliability and intermittence of the wind

[†] Clearly, these limitations do not apply to systems that collect sunlight out in space.

or the sun, respectively. In Chapter 12, we listed representative average insolations for a number of U.S. localities. Remember that peak insolation is 1000 W/m^2. That list shows that the average is between one-fourth and one-sixth of the peak. One would think that for photovoltaic plants, this would be the expected ratio of average to peak power. However, practical experience is always somewhat disappointing when compared to expectations. Two German utility-scale photovoltaic installations for which data are available report their average-to-peak ratio as only 0.11. The Cal State Campus at Hayward, in sunny California, being blessed by a better climate, estimates an average-to-peak ratio of 0.16. Hence, one can roughly estimate that a photovoltaic system must have six times bigger peak capacity than a nuclear plant that delivers the same average power.

The efficiency of photovoltaic systems is low compared with that of traditional thermal or hydroelectric plants. It may be over 20% for sophisticated crystalline silicon systems and some 5% for some inexpensive thin-film ones. However, efficiency is not of primary interest in many photovoltaic installations. The cost per peak watt may be the important characteristic. At $0.20 per peak watt (if this figure is ever attained)[†], even low-efficiency photovoltaics would be attractive. Conventional hydroelectric and fossil-fueled plants cost around $1.00/W, while nuclear ones may cost over $5.00/W. But, again, these energy sources, unlike photovoltaics, can operate continuously and thus, on average, produce much more energy.

At the low price mentioned, it is possible that many buildings will have their external surfaces covered with photovoltaics. In some cases, the average energy generated may well exceed the needs of the building. However, it will be generated only on sunny days, not on rainy ones, nor will there be any generation at night. Consequently, adequate storage facilities must be available, especially in case of residences, where demand during the day may be small, while at night, requirements are higher.

If the building is **off-grid**, as some rural properties are—that is, if it has to be entirely self-sufficient, expensive batteries or some other storage scheme are needed. On the other hand, if the building is connected to the power grid, storage can be provided by the local utility company in what is called a **utility-tie** system. The excess energy generated by the customer can be sold to the utility for a price below that charged by the utility to the customer.[††] The price differential would pay for storage and distribution. A dual metering system can be used: one meter measures the outgoing power from the customer to the utility, and the other the power from the utility to the customer. A **utility-intertied inverter** must be used synchronize the customer-generated electricity with the grid.

[†] In mid-2008, photovoltaics cost about $2.00 per peak watt.

[††] The Public Utilities Regulatory Policy Act of 1978 fixes the so-called **avoided cost** that is the minimum amount that a utility has to pay an independent power producer. Frequently, this is only the cost of the fuel saved.

In areas of unreliable energy supply, subjected to frequent blackouts, a hybrid of the off-grid and the utility-tie system may be useful.

Such **building-integrated photovoltaic** (BIPV) systems will become progressively more popular as the price of solar collectors decreases. The land area of the building, the structure on which to mount the solar collectors (roof and external walls), the very roof, and the connection to the grid are all investments made even if no BIPV is used and thus should not be charged to the BIPV cost even though they must be included in the cost of centrally generated PV systems.

BIPV seems more appropriate for individual residences, which have a lot of roof area per inhabitant, and for any one- or two-floor structures such as those used by some factory or office complexes. Apartment buildings, with their much larger population density, would be at a disadvantage. Of course, shade trees would be contraindicated. Roofs would have to resist hailstorms and, in California, be immune to the pelting by avocados falling from overhanging trees.

In evaluating the performance of any particular grid-intertied building integrated photovoltaic installation, the first thing to do is to select the proper degree of optimism. At the current state of the art (2008), uncertainties are large and improvements are still coming fast: it is possible, by using reasonable assumptions, to prove either the effectiveness of the system or its complete undesirability.

Let us consider a typical one-family residence in California. It will perhaps have a 200-m^2 roof area. If electricity is used for cooking but not for heating, it will use an average of 50 kWh per day at a peak of less than 10 kW. The expected average yearly insolation is 250 W/m^2. If relatively inexpensive thin-film photovoltaic units are used, one can count on 5% efficiency, which means, under optimal conditions, each square meter of roof will generate 12.5 W on average, or 2500 W for the whole roof. Almost certainly both the tilt and the orientation of the roof will not be optimum. Let us say that one can count on 10 W/m^2 or 2000 W total (on a day-and-night) average. The peak production will be four times larger or 8 kW. This does not quite satisfy the 10 kW peak requirement, but the excess can come from the grid.

2 kW average correspond to 48 kWh per day, very close to the desired value of 50 kWh. Thus, the energy balance is acceptable.

At peak insolation, the photovoltaics will generate 40 W/m^2. If the cost is \$1.00/W (peak), the collectors will cost \$40/m^2, or \$8000 for the roof. We must now make a number of additional assumptions:

1. *Labor cost: \$3000*
2. *Ancillary equipment (controls, inverter, etc.): \$2000*

(*Continues*)

(*Continued*)

3. Cost of capital: $15%/year on a 10-year loan.
4. Longevity of equipment: > 10 *years.*

Under these assumptions, the initial investment is of $13,000, and the yearly cost is $1950. During one year, the system will generate $48 \times 365 = 17{,}500$ *$/kWh.*

Hence, the cost of the generated electricity is $1950/17{,}500 = 0.111$ *$/kWh, which, by coincidence, is almost precisely the price of electricity in northern California. Thus, at first glance, during the first 10 years, it makes no economic difference if the electricity comes from the utility or from the BIPV system. If there are no maintenance or repair costs, then, after 10 years, the electricity is free for as long as the equipment holds out.*

Of course, there are additional costs. For example, although the average power generated is equal to the average power used, the customer will sell his surplus at a price much lower than what he will pay for the energy received from the utility.

This example illustrates the difficulty of making a reasonable assessment of the economic possibilities of a system still in development. For instance, if the efficiency of the photovoltaic blankets is much higher than the value of 5% assumed and/or the cost per peak watt is much lower than the $1.00/W we used, then the system will suddenly be extremely attractive. One cannot dismiss the possibility that in the near future, efficiencies of 10% and costs of $0.50/W may become reality.

BIPV systems have not yet made a significant contribution to power generation, but **utility-scale** photovoltaic plants are increasing in number and so is, dramatically, the total installed power in the world, as can be seen from Figure 14.1. The vigorous growth depicted in the figure is greatly the result of the efforts of Germany[†] that, notwithstanding not being known as a sunny land, has 47% of the worlds (photovoltaic) sun collecting installations. Spain follows with a respectable 28%, while the United States, usually in the lead of innovative efforts, has only 15%. Surprisingly, Japan, with its large number of small solar plants, lags far

[†] Germany continues its photovoltaic expansion. In 2008, the Göttelborn plant was upgraded from 4.4 MW (2007) to 16.1 MW, and the Solarpark Waldpolenz is raising its output from 10 MW (2007) to a whopping 40 MW. This, however, pales in comparison with the plans for a very large solar park in the neighborhood of San Luis Obispo in Southern California, where OptiSolar is proposing a 550-MW plant—the Topaz Solar Farm—which will cover 24 km^2 and is estimated to cost about 1 billion. This amounts to a somewhat optimistic cost of 1800 $/kW.

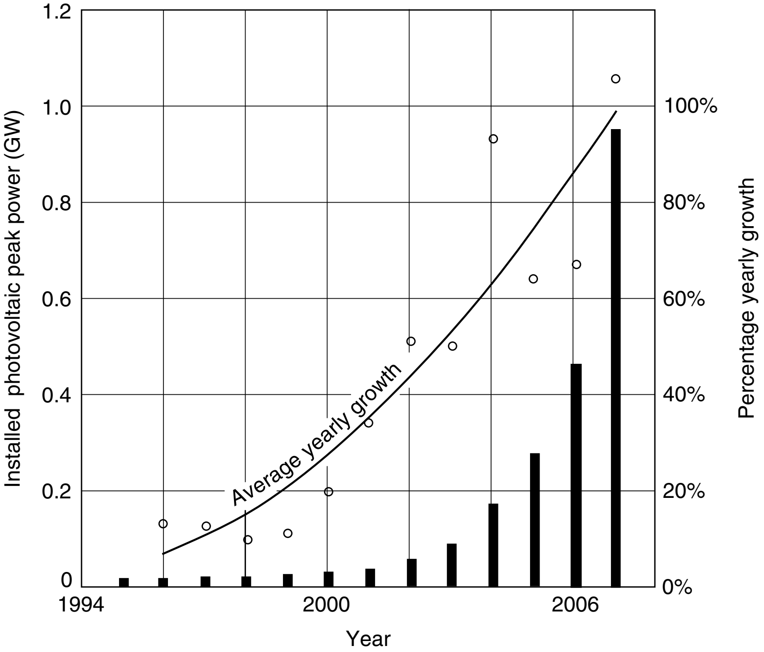


Figure 14.1 The growth of worldwide utility-scale photovoltaic installations was discouragingly slow up to about 2002 and then took off at an amazing pace (growth from 2006 to 2007 exceeded 100%!). Data from http://www.pvresources.

Table 14.1 Utility-Scale Photovoltaic Plants

Country	MW(peak)	Percentage (of worldwide capacity)
Germany	451.4	47%
Spain	266.7	28%
USA	147.2	15%
Italy	17.8	2%
Japan	13.3	2%
Portugal	11.8	1.2%

behind. Although Table 14.1 lists almost all of the worlds important PV installations, Great Britain, France, and Russia do not appear in it.

Even at the rarely achieved maximum of 1000 W/m^2, sunlight is a diluted form of energy. Combined with the modest efficiency of photovoltaic converters, this leads to a requirement for larger areas if one wants to generate significant amounts of electric energy. When rooftops are available, they constitute an obvious location for photovoltaic collectors. But there is not enough rooftop area available. In 2008, only 29% of the utility-scale collectors were on rooftops, 79% were directly on the ground, and 1% were in assorted other places such as walls.

A really excellent compilation of photovoltaic installations data can be found at *Large-Scale Photovoltaic Power Plants*, Cumulative and Annual Installed Power Output Capacity, <pvresources.com>.

Seeing the intense rhythm of photovoltaic development, one cannot suppress one's enthusiasm for these efforts. They are important and they deserve support, but they have to be put in perspective.

In 2008, the total installed capacity of large-scale photovoltaics in the whole world was about 1 GW, which equals the nominal power of one typical nuclear plant. But one cannot equate 1 GW of photovoltaics to 1 GW of nuclear. About 6 GW of photovoltaic are energetically equivalent to 1 GW of nuclear. And the world has over 400 commercial nuclear plants with a potential to produce 370 GW. Using our ratio, to replace these nuclear plants, one would need to increase the current photovoltaic capacity by a factor of nearly 2200.

There are application niches in which photovoltaics are well entrenched such as power sources for many remotely located devices and, of course, for a number of consumer electronics. One common use is in powering satellites and other spacecraft that do not stray too far away from the sun. Craft that venture into deeper space—Jupiter and beyond—must be powered by radioisotope thermal generators (RPGs) or, in the near future, by radioisotope Stirling generators (RSGs). See the chapters on thermoelectrics and on heat engines. It is not impossible that RSG will turn out to have advantages over photovoltaics in these space applications.

Decades of research have focused on achieving higher efficiencies in photovoltaics. Success in this endeavor has mostly been associated in greater production and operating costs—a hoped for dramatic reduction in the watts/dollar ratio has not materialized. More recently, some of the effort has been toward reducing cost even if that means accepting much lower efficiencies, provided there is a substantial gain in the ratio mentioned earlier.

Broadly, three distinct techniques are used in building solar cells:

1. Crystalline material, most frequently silicon. These are expensive but yield good efficiencies (>20%) if they are made of single crystals. Somewhat cheaper, but less efficient, are polycrystalline units.
2. Amorphous thin films (Si, GaAs, $CuInSe_2$, TiO_2, etc.). These have efficiencies of some 7% but are much less expensive. They can be flexible sheets.
3. Organic polymers, still in early development, which could easily become the best overall solution. They promise low cost, light weight, and the flexible. Photovoltaic blankets cheap enough to serve as roofing materials could replace shingles, tarpaper, or tiles. Their manufacturing methods will probably be less toxic than those of inorganic materials.

14.2 Theoretical Efficiency

In this section, we derive the theoretical efficiency of photocells without direct reference to the exact mechanism of their implementation except that we assume that all cells have to perform the functions of carrier generation and carrier separation. These functions can be carried out either in a same region of the cell or in separate ones.

In the general discussion of photocell efficiency, in this section, we assume that the carrier separation function is carried out without any losses and that one electron-hole pair is created for every incident photon that has an energy, $hf \geq W_g$. We will call W_g the **band-gap energy**, *even though in some cells the required energy is not associated with raising an electron from the valence to the conduction band.*

We also assume that the material is transparent to photons of energy less than W_g. These photons do not interact with the photosensitive material and thus have no photoelectric effect. Finally, we assume that all photons with energy above the band-gap contribute to the load an amount of electric energy exactly equal to W_g. The excess energy, $hf - W_g$, is simply transformed into heat and constitutes a loss.

An appropriate material—in general a semiconductor—will be transparent or not to a photon, depending on the frequency of the photon. The exact boundary between transparency and opacity depends on the type of material considered. Table 14.2 displays the data for some semiconductors. Diamonds, a form of carbon that crystallizes in the same manner as silicon and germanium, being highly resistant to heat and radiation, are a promising material for transistors that have to operate in hostile environments.

The few readers who need to refresh their basic knowledge of semiconductors may take advantage of Appendix B in this chapter, which offers a very elementary description of the band structure and of the formation of p-n junctions.

A structure that, exposed to light, generates electric energy constitutes a **photovoltaic** cell, or simply, a **photocell**. Photocells made of bulk semoconductors are referred to as **photodiodes**.

Table 14.2 Light Absorption Limits for Some Semiconductors

Material	**ν_0 (THz)**	**λ (nm)**	**W_g (eV)**	**Region in which transition from transparent to opaque occurs**
α-Sn	19.3	15,500	0.08	Far infrared
Ge	162	1850	0.67	Infrared
Si	265	1130	1.10	Infrared
GaAs	326	920	1.35	Near infrared
GaP	540	555	2.24	Visible
C	1300	230	5.40	Ultraviolet

Photovoltaic (PV) cells exposed to monochromatic light can theoretically achieve 100% efficiency converting radiation to electric energy. In the majority of cases, photocells are exposed to broad-band radiation—that is, to a stream of photons of different energies. Under such circumstances, the efficiency is limited by the two mechanisms discussed earlier:

1. Weaker photons (those with less than a given frequency) fail to interact with the material.
2. More energetic photons will deliver to the load only a part of the energy, the rest being thermalized.

In all cases, whether we are considering ideal or practical devices, their efficiency is defined as the ratio of the power, P_L, delivered to the load to the power, P_{in}, of the incident radiation,

$$\eta \equiv \frac{P_L}{P_{in}}. \tag{14.1}$$

The characteristics of broad-band radiation can be described by specifying the power density, ΔP, of the radiation in a given frequency interval, Δf, as was done for solar radiation in Table 12.1 (Chapter 12). Alternatively, taking the $\Delta P/\Delta f$ ratio to the limit, one writes an equation expressing the dependence of $\partial P/\partial f$ on f. The total incident power density is, then,

$$P_{in} = \int_0^\infty \frac{\partial P}{\partial f} df. \tag{14.2}$$

In the case of the black body, $\partial P/\partial f$ is given by **Planck's equation**,

$$\frac{\partial P}{\partial f} = A\frac{f^3}{e^{\frac{hf}{kT}} - 1}, \tag{14.3}$$

where A is a constant having the units of W m^{-2}Hz^{-4}. Hence,

$$P_{in} = A\int_0^\infty \frac{f^3}{e^{\frac{hf}{kT}} - 1} df. \tag{14.4}$$

Let $x \equiv \frac{hf}{kT}$, then

$$df = \frac{kT}{h}dx \quad \text{and} \quad f^3 = \left(\frac{kT}{h}\right)^3 x^3. \tag{14.5}$$

$$P_{in} = A\left(\frac{kT}{h}\right)^4 \int_0^\infty \frac{x^3}{e^x - 1}dx. \tag{14.6}$$

The definite integral, $\int_0^\infty \frac{x^3}{e^x-1}dx$ has the value $\pi^4/15$; therefore

$$P_{in} = A\left(\frac{kT}{h}\right)^4 \frac{\pi^4}{15} = aT^4, \tag{14.7}$$

where a (W m^{-2} K^{-4}) is also a constant.

When the temperature of a black body radiator increases, not only does the total power, P, increase (Equation 14.7), but, in addition, the peak radiation is shifted to higher frequencies as can be seen from Figure 14.2. There is a simple relationship between the frequency, f_{peak}, and the temperature, T.

The proportionality between the light power density and the fourth power of the temperature is related to the **Stefan–Boltzmann** law.

From Equation 14.3, we see that the shape of the distribution curve is determined by the factor, $\frac{f^3}{e^{\frac{hf}{kT}}-1}$. The peak occurs when

$$\frac{d}{df}\left(\frac{f^3}{\mathrm{e}^{\frac{hf}{kT}}-1}\right) = 0. \tag{14.8}$$

Making the $x \equiv \frac{hf}{kT}$ substitution and taking the derivative, we obtain

$$(3-x)\exp x - 3 = 0, \tag{14.9}$$

whose numerical solution is $x = 2.821$. From the definition of x,

$$f_{peak} = \frac{k}{h}xT = 59.06 \times 10^9 T. \tag{14.10}$$

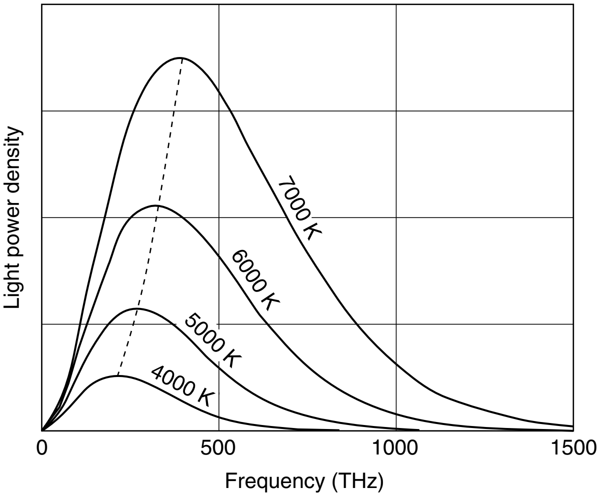


Figure 14.2 The peak of the p vs. f curve of a black body moves toward higher frequencies as the temperature increases.

For $T = 6000\,\text{K}$, $f_{peak} = 354\,\text{Thz}$.

The relation between f_{peak} and T is the **Wien's displacement law**.

It is useful to relate the total flux, ϕ, of photons that, given a specified spectral distribution, corresponds to a power density, P_{in}. Consider a small frequency interval, Δf, centered on the frequency f. Since each photon has energy hf, the power density of radiation in this interval is

$$\Delta P = \Delta\phi\; hf\ \text{W/m}^2, \tag{14.11}$$

where $\Delta\phi$ is the photon flux (photons $\text{m}^{-2}\text{s}^{-1}$) in the interval under consideration. In the limit, when $\Delta f \rightarrow 0$ (and dividing both sides by df),

$$\frac{d\phi}{df} = \frac{1}{hf}\frac{\partial P}{\partial f}, \tag{14.12}$$

and

$$\phi = \frac{1}{h}\int_0^\infty \frac{1}{f}\frac{\partial P}{\partial f}df. \tag{14.13}$$

Particularizing for the black body case and, once more, letting $x \equiv hf/kT$,

$$\phi = \frac{A}{h}\int_0^\infty \frac{1}{f}\frac{f^3}{e^{\frac{hf}{kT}}-1}df = \frac{A}{h}\int_0^\infty \frac{f^2}{e^{\frac{hf}{kT}}-1}df, \tag{14.14}$$

$$\phi = \frac{A}{h}\left(\frac{kT}{h}\right)^3\int_0^\infty \frac{x^2}{e^x-1}dx = 2.404\frac{A}{h}\left(\frac{kT}{h}\right)^3. \tag{14.15}$$

because the definite integral, in this case, has the value 2.404.

Still for black body radiation, we can find the ratio of the light power density to the corresponding photon flux. From Equations 14.7 and 14.15,

$$\frac{P}{\phi} = \frac{A\left(\frac{kT}{h}\right)^4\frac{\pi^4}{15}}{2.404\frac{A}{h}\left(\frac{kT}{h}\right)^3} = 37.28\times10^{-24}\,T. \tag{14.16}$$

It should be noted that formula 14.16, is valid only if the full spectrum is considered. For a truncated spectrum, for instance, one that has some regions removed by a filter, it is necessary to calculate separately the total power density, P, and the total flux of photons, ϕ, and form the ratio.

Not surprisingly, the ratio of total power to total photon flux increases proportionally to the temperature because, as we saw when we derived Wien's displacement law, the higher the temperature, the more energy the average photon has.

Example 14.1

What is the photon flux when light radiated from a 6000 K black body has a power density of 1000 W/m^2?

From Equation 14.16,

$$\phi = \frac{P}{37.28 \times 10^{-24}\,T} = \frac{1000}{37.28 \times 10^{-24} \times 6000}$$

$$= 4.47 \times 10^{21} \quad \text{photons m}^{-2}\text{s}^{-1}. \tag{14.17}$$

For the ideal case, the efficiency of the device is, of course,

$$\eta_{ideal} = \frac{P_{L_{ideal}}}{P_{in}}. \tag{14.18}$$

We now need to know $P_{L_{ideal}}$.

If broad-band radiation falls on a semiconductor with a band-gap energy, $W_g = hf_g$, the photons with frequency $f < f_g$ will not create carriers. A fraction

$$G_L = \frac{1}{P}\int_0^{f_g} \frac{\partial P}{\partial f} df, \tag{14.19}$$

of the total radiation power density, P_{in}, will be lost.

Let ϕ_g be the total flux of photons with $f > f_g$. Each photon creates a single electron-hole pair with energy hf. However, as stated, the energy in excess of W_g will be randomized and will appear as heat and each photon contributes only W_g joules to the electric output. The useful electric energy (the energy, P_L, delivered to a load) will be

$$P_L = \phi_g W_g \quad \text{W/m}^2. \tag{14.20}$$

The flux of photons with energy larger than hf_g (adapting Equation 14.13) is

$$\phi_g = \frac{1}{h}\int_{f_g}^{\infty} \frac{1}{f}\frac{\partial P}{\partial f} df. \tag{14.21}$$

The useful power is

$$P_L = hf_g\phi_g = f_g\int_{f_g}^{\infty} \frac{1}{f}\frac{\partial P}{\partial f} df, \tag{14.22}$$

and the efficiency is

$$\eta_{ideal} = \frac{P_L}{P_{in}} = f_g \frac{\int_{f_g}^{\infty} \frac{1}{f}\frac{\partial P}{\partial f} df}{\int_0^{\infty} \frac{\partial P}{\partial f} df}. \tag{14.23}$$

Observe that η_{ideal} depends only on the spectral distribution and on the W_g of the semiconductor. It completely ignores the manner in which the device operates. Unlike the efficiency of real photocells, η_{ideal} does not depend on the level of illumination. Again, for a black body,

$$\phi_g = \frac{A}{h}\int_{f_g}^{\infty} \frac{f^2}{\mathrm{e}^{\frac{hf}{kT}} - 1} df = \frac{A}{h}\left(\frac{kT}{h}\right)^3 \int_X^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx, \tag{14.24}$$

where $X = hf_g/kT = qV_g/kT$.

It should be obvious that the ratio $\sigma \equiv \phi_g/\phi$ depends only on the nature of the radiation considered, not on its intensity. The ratio is

$$\sigma \equiv \frac{\phi_g}{\phi} = \frac{\int_X^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx}{\int_0^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx} = \frac{\int_X^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx}{2.404} = 0.416 \int_X^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx. \tag{14.25}$$

For 6000-K black body radiation, the ratio is a fixed 0.558 if $W_g = 1.1\,\mathrm{eV}$, the band-gap energy of silicon. The ideal efficiency of a photodiode is then

$$\eta_{ideal} = \frac{15}{\pi^4}\left(\frac{h}{k}\right)^4 \frac{f_g}{T^4} \int_{f_g}^{\infty} \frac{f^2}{\mathrm{e}^{\frac{hf}{kT}} - 1} df. \tag{14.26}$$

It is more convenient to work with the band-gap voltage, V_g, instead of the corresponding frequency, $f_g = \frac{q}{h} V_g$,

$$\eta_{ideal} = \frac{15}{\pi^4}\left(\frac{h}{k}\right)^4 \frac{q}{h}\frac{V_g}{T^4} \int_{\frac{qV_g}{h}}^{\infty} \frac{f^2}{\mathrm{e}^{\frac{hf}{kT}} - 1} df. \tag{14.27}$$

Letting $x \equiv \frac{hf}{kT}$ as before,

$$\begin{aligned}\eta_{ideal} &= \frac{15}{\pi^4}\left(\frac{h}{k}\right)^4 \frac{q}{h}\left(\frac{kT}{h}\right)^3 \frac{V_g}{T^4} \int_{\frac{qV_g}{kT}}^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx = \frac{15}{\pi^4}\frac{q}{k}\frac{V_g}{T} \int_{\frac{qV_g}{kT}}^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx \\ &= 1780 \frac{V_g}{T} \int_{\frac{qV_g}{kT}}^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx. \end{aligned} \tag{14.28}$$

The lower limit of the integral is that value of x corresponding to f_g.

There is no analytical solution to the preceding integral, but it can either be solved numerically or the table in Appendix A to this chapter can be used to determine the value of the definite integral (which is, of course, a simple number function of the lower limit of the integral).

Examaple 14.2

What is the flux of photons that have more energy than that of the silicon band gap (1.1 eV, i.e., $V_g = 1.1\,\mathrm{V}$) when light radiated from a 6000-K black body has a power density of $1000\,\mathrm{W/m^2}$?

Equation 14.25 gives us the ratio, σ, of ϕ_g to ϕ.

For the particular combination of this example ($V_g = 1.1\,\mathrm{V}$ and $T = 6000\,\mathrm{K}$), the ratio is 0.558, and from Example 14.1, $\phi = 4.47 \times 10^{21}$ photons $\mathrm{m^{-2}\ s^{-1}}$. Consequently,

$$\phi_g = \sigma\phi = 0.558 \times 4.47 \times 10^{21} = 2.49 \times 10^{21} \text{ photons } \mathrm{m^{-2}s^{-1}}. \tag{14.29}$$

Examaple 14.3

What is the ideal efficiency of the photocell under the circumstance of the previous example?

Using Equation 14.28,

$$\eta_{ideal} = 1780 \frac{1.1}{6000} \int_{2.125}^{\infty} \frac{x^2}{\mathrm{e}^x - 1} dx. \tag{14.30}$$

The lower limit of the integral is $X = hf_g/kT = qV_g/kT = 2.125$.

The value of the definite integral is 1.341 (by interpolation in the table in Appendix A to this chapter); hence,

$$\eta_{ideal} = 1780 \frac{1.1}{6000} 1.341 = 0.438. \tag{14.31}$$

Figure 14.3 shows how the ideal efficiency of a photocell depends on the band-gap energy when exposed to a black body at 6000 K (about the temperature of the sun). Our efficiency calculations, based on Equation 14.28, use a very simple model, that totally ignores the photocell itself, which is assumed to be 100% efficient. Its results are identical to the **ultimate efficiency** of Shockley and Queisser (SQ) (1961).

Perhaps one of the earliest calculations of theoretical efficiency as a function of a band gap is the work by Prince (1955). His model considers the best possible silicon cell made under the limitations of the then

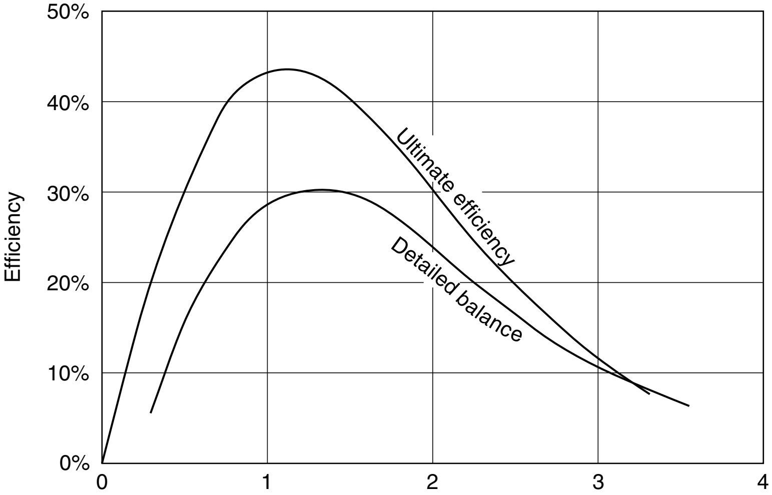


Figure 14.3 Dependence of the efficiency of a photodiode on its band-gap energy. Black body at 5800 K.

primitive technology. Specifically, it assumes values of lifetimes of minority carriers that have been vastly improved upon. Although the general shape of the efficiency versus band gap curve is roughly the same as that from Equation 14.28, the absolute values of estimated efficiencies are much lower. He sets the maximum theoretical efficiency at 21.7 and proceeds to explain why this value is unattainable.

Until 1961, there was no clear agreement as to what band-gap would yield (theoretically) the highest efficiency when exposed to sunlight. See Loferski (1956). In 1961, Shockley and Queisser published a much cited paper deriving the theoretical limits of solar cell efficiencies operating under certain assumptions, some of which we used in our derivation. One assumption we did not make was that the photocell involved a *p*-*n* junction, which implies irreducible radiative recombination of electron-hole pairs. For this reason, the SQ **detailed balance** model predicts somewhat lower efficiencies than the **ultimate efficiency** model in Figure 14.3.

Since the solar spectrum is not exactly that of a black body, the dependence is somewhat different from that shown in the figure. Also, the exact spectral distribution of sunlight in space differs from that on the ground owing to atmospheric absorption.

Notwithstanding all these limitations, efficiencies greater than these **black body spectrum efficiencies** can be achieved. This is done by creating situations in which one or both of the efficiency-limiting mechanisms discussed at the begining of this section are circumvented. Three techniques are discussed in the next three sections.

14.3 Carrier Multiplication

In the derivation of the ideal efficiency of photocells in the preceding section, we made the assumption that photons with energy greater than that of the band gap will yield a single **exciton** (electron-hole pair). The excess energy of the photon will appear as excess kinetic energy of the electron and of the hole and will be quickly be dissipated as heat: it is **thermalized**.[†] This assumption is almost precisely true for photocells made of bulk material.

Why does a photon with energy $> 2W_g$ not yield additional excitons? One reason is that a photon is indivisible; it cannot split into two parts and act on two different covalent bonds. One single exciton is generated from the interaction. However, the energetic carriers created could cause ionization of an atom by impact and thus create another exciton. This **impact ionization** occurs with low efficiency ($< 1\%$, in bulk Si) and makes a minute contribution to the number of excitons (and thus to the number of carriers) created.

Schaller and Klimov (2004) have demonstrated that carrier multiplication by impact ionization can proceed with nearly 100% efficiency if PbSe is in the form of nanocrystals (of about 5-nm size).

The excess photon energy, $W_{ph} - W_g$, is partitioned between the electron and hole according to their respective masses. Since in PbSe these particles have nearly the same mass, each particle will receive approximately the same excess energy: $(W_{ph} - W_g)/2$, which, for impact ionization, must exceed W_g. So, it is not surprising that additional carriers are created in lead selenide when $W_{ph} > 3W_g$. This establishes, for this particular case, the **threshold energy** necessary for impact ionization.

The process efficiency increases (initially) as the excess energy of the photon increases beyond the threshold value. See Figure 14.4, adapted from the paper by Schaller and Klimov. We fitted a straight line through the data to show the trend. Disregard the difference between circular and square data points—they represent two different measurement techniques.

Calculations show that, for lead selenide cells, the ideal efficiency of photodiodes can exceed 60%. To be sure, no such cells were tested in the Schaller and Klimov study, but Greenham, Peng, and Alivisatos (1996) demonstrated solar cells based on nanocrystals back in 1996.

PbSe is a material of great interest for solar cells because it can have its W_g adjusted from about 0.3 to 1.3 eV. It absorbs radiation strongly from near infrared all the way into the ultraviolet. Band-gap adjustment is discussed further in Subsection 14.4.1.

To better understand carrier multiplication, do Problem 14.26.

[†] The loss mechanism is through **electron–phonon interaction.**

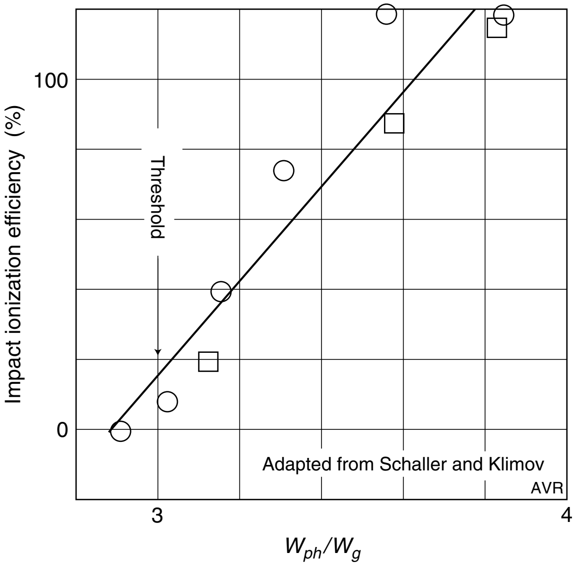


Figure 14.4 Impact ionization efficiency in PbSe as a function of the excess photon energy.

14.4 Spectrally Selective Beam Splitting

The broad-band solar spectrum can be split into slices of radiation whose frequencies match the properties of the semiconductors used in the photodiodes. Thus, for silicon, a narrow band beginning at 265 THz (in the infrared) would best match its 1.1 eV band-gap energy. Other spectral slices would be directed either to other photodiodes or to an absorber to generate heat.

14.4.1 Cascaded Cells

Conceptually, the simplest beam-splitting system is achieved when one superposes two photodiodes with different band gaps. Assume that Diode #1, with the larger band gap, W_{g1}, is on top (i.e., it is the one that receives the direct unfiltered light). It will absorb photons with energy above W_{g1}, but it is transparent to those with energies below this threshold. If the diode underneath (Diode #2) has a lower band gap, W_{g2}, it will absorb part of the photons that passed through Diode #1. Such a **cascade** arrangement can, for obvious reasons, yield efficiencies larger than those of single devices. See Problem 14.1.

For instance, if Diode #1 has a band gap of 1.8 eV and Diode #2 one of 1.1 eV, the ideal efficiency of the combination is 0.56 when exposed to sunlight in space. Compare this with the best possible single-diode efficiency of 0.43. One could have cascade cells using more than two semiconductors.

An infinite number of semiconductors would, of course, lead to a *theoretical* efficiency of 100%.

Numerous practical problems limit the actual efficiency of cascade cells, and progress is almost invariably accompanied by increase in cost. One problem that arises when one attempts to cascade diodes is that of interconnection. Normally, the nonilluminated face of the device is metallized to act as a current collector. This makes it opaque and, consequently, useless in a cascade configuration. The solution is to build the two (or more) diodes from a single semiconductor crystal with the individual junctions in series. To do this, the materials of the two diodes must have (nearly) the same lattice constant, but, of course, different band gaps. By varying the stoichiometry of quaternary compounds such as AlInGaAs, it is possible to tailor the band gap to the desired value while keeping the lattice constant about the same. One practical device uses Ga$_{0.5}$In$_{0}$.5P/GaAs junction, a combination that achieves a lattice match but does not optimize the band-gap ratio. This particular combination provides band-gaps of 1.85 and 1.43 eV.[†]

Wladek Walukiewicz (Lawrence Berkeley National Laboratory), among others, describes (2002) a solution to the cascaded cells problem. For years, light-emitting diodes (LEDs) emitted red light (GaAsP) or, at most, green (GaP). The quest for blue LEDs lasted a number of years until, in the 1980s, Isamu Akasa and Shuji Nakamura (see, for instance, H. Amano et al and S. Nakamura) developed the technique for growing thin films of gallium nitride whose large band-gap energy (3.4 eV) caused the generation of near ultraviolet light. Doping with indium allowed the changing of W_g so as to obtain blue light. Although the films had an enormous amount of defects, the material displayed great tolerance to imperfection. GaAsP and GaP LEDs with the same density of defects would emit no light. The Lawrence Berkeley team grew indium gallium nitride crystals of high quality and was able to show that by varying the relative amount of indium, the band gap could be tuned from 3.4 eV (pure GaN) all the way down to 0.7 eV (pure InN). This 5:1 range covers most of the solar spectrum. Cascade cells can thus be built choosing the best W_g combination such as 1.7/1.1 eV.

The difference in the lattice constant of the different layers of cascade cells—that is, the mismatch of the geometry of the crystal—introduces strains that, at best, create imperfections detrimental to the performance of the device and can even cause cracks in the material. Here it is where the great tolerance to defects favors the various In$_x$Ga$_{1-x}$N layers, which, presumably, can work well even under substantial strain.

[†] Dopants have to be found to obtain both p- and n-varieties of the above semiconductors.

The $In_xGa_{1-x}N$ can be doped with silicon to produce n-type material, but, at present, there are difficulties in finding an adequate dopant for the needed p-type material. It should be pointed out here that the $In_xAl_{1-x}N$ system has an even wider (0.7 to 6.2 eV) range than the $In_xGa_{1-x}N$ system.

14.4.1.1 Multiband Semiconductors

The cascade cells make better use of the solar spectrum by using more than one p-n junction. The production of such cells is complicated and expensive. A different solution is to use a single p-n junction in a material that has more than one band-gap. Instead of simply having one valence band and one conduction band, such materials have one valence and *two* separate conduction bands, as indicated in Figure 14.5.

This configuration was investigated before it was discovered how to fabricate multiband materials. Luque and Martí (1997), using the procedure described in the Shockley and Queisser article, calculated the theoretical efficiency of multiband cells as a function of the lower of the two band-gaps (the V → I transition). The results are displayed in Figure 14.6.

Not obvious a priori is the higher efficiency of the multiband cell compared with cascade cells. Unfortunately, when Luque and Martí wrote the article, there were no useful multiband materials available. Since then, efforts have been made to create them. One approach is to create appropriate superlattices, a technique that leads to very expensive cells.

Yu et al. (2003) describes a new class of semiconductors called **highly mismatched alloys** whose properties can be dramatically altered by substitution of a small number of atoms by very different elements. They can be made into multiband semiconductors. Examples are III-V (such as GaAs) and II-VI (such as ZnTe) alloys in which some of the V anions are replaced by nitrogen or some of the group VI anions are replaced by oxygen. In particular, Yu et al. described manganese doped zinc telluride in which

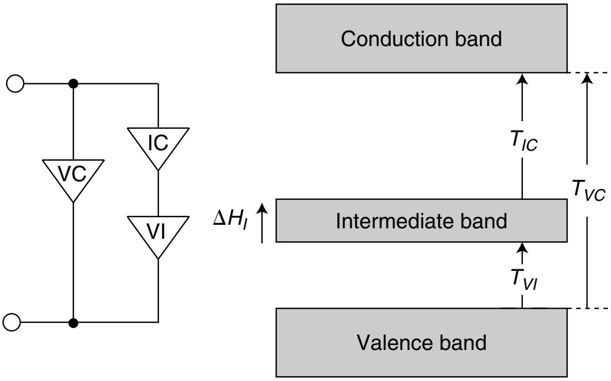


Figure 14.5 Single-junction, multiband photocell.

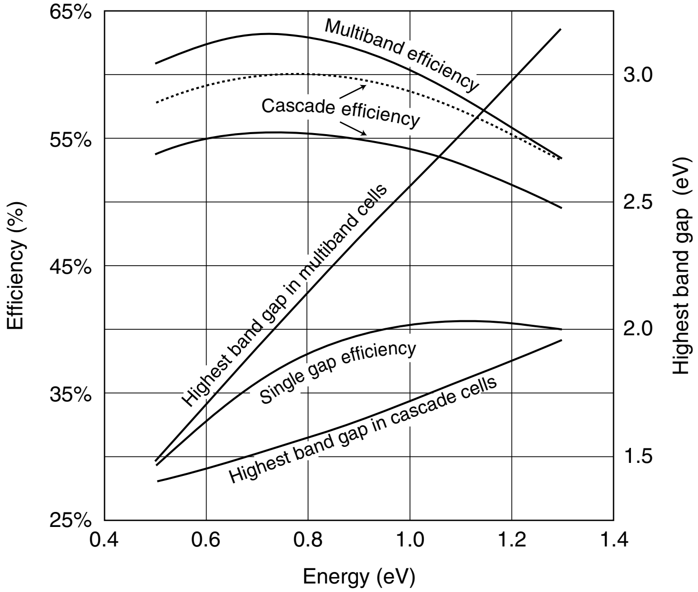


Figure 14.6 Comparison of the performance of single-gap, cascade and multiband solar cells. Solid lines were computed using the Shockley and Queisser "detailed balance" procedure. The dotted line was calculated using the "ultimate efficiency" procedure. For each value of the lower band gap, the optimum upper band-gap was calculate and its value is plotted as a (nearly) straight upward slanting solid line. All data, other than the dotted line, are from the article by Luque and Martí (1997).

some tellurium atoms were replaced by oxygen.[†]: $Zn_{1-y}Mn_yTe_{1-x}O_x$. This material has an intermediate narrow band in the band-gap. The catch is that the amount of oxygen required for the creation of this intermediate band by far exceeds the limit of solubility of oxygen in the alloy. However, by pulsed laser melting the alloy into which oxygen was implanted followed by fast annealing allows the freezing of the oxygen in the required concentration.

The original alloy had a band-gap of 2.32 eV; that is, the conduction band started 2.32 eV above the top of the valence band. After "oxygenation," the new conduction band rose to 2.56 eV, and a narrow intermediate band formed at 1.83 eV, exactly what is needed for a multiband photocell.

In 2008, Rose Street Labs set up a facility in Phoenix, Arizona, for the development of the multiband photocell technology the company had acquired from the Lawrence Berkeley Lab.

[†] In all cases, the replacing impurity has the same valence as the replaced atom—it is an **isovalent substitution**.

Table 14.3 Transmittance of a Cobalt Sulfate Filter (1 g/liter)

f (THz)	T
200	0.020
273	0.29
333	0.15
375	0.88
429	0.93
500	0.90
600	0.30
750	0.92

14.4.2 Filtered Cells

Concentrated sunlight can be filtered through cells containing cobalt sulfate solution that will absorb part of the energy while transmitting the rest to underlying silicon diodes.

Hamdy and Osborn (1990) measured the transmittance of 5-cm-thick cobalt sulfate filters with different concentrations. For a concentration of 1 gram per liter, the transmittances at different light frequencies were as shown in Table 14.3.[†] The absorbed energy is used to generate useful heat. This kind of filter has a window in the 350- to 550-THz range and thus passes radiation that matches silicon characteristics reasonably well. The band-gap energy of silicon corresponds to 265 THz.

14.4.3 Holographic Concentrators

Perhaps the most attractive beam-splitting technique is the holographic concentrator. A holographic plate is prepared in such a way that it acts as an extremely dispersive cylindrical lens. Sunlight is concentrated by factors of 50 to 100 and dispersed into a rainbow as indicated in Figure 14.7 Two (or more) photodiodes are mounted perpendicular to the holographic plate and positioned in such a way as to be exposed to that part of the spectrum at which they are most efficient.

The far infrared, which is of little interest for photovoltaic conversion, is directed to a region where the corresponding heat can be dissipated without affecting the photodiode, which, as discussed later, must operate at low temperature to avoid loss in efficiency.

One should remember here the definition of efficiency of a hologram: it is the ratio of the light power diffracted to the incident light power. Most

[†] The transmittance value for 750 THz is not a typo. The cobalt sulfate filter is quite transparent to near UV.

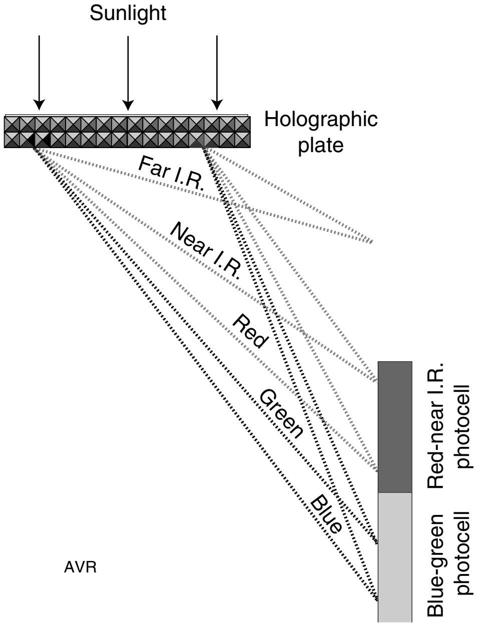


Figure 14.7 A holographic plate serves as both a concentrator and a spectrum splitter.

of the undiffracted light just passes through the hologram. This efficiency is frequency dependent and, thus, a hologram has a given bandwidth.

For photovoltaic applications, a holographic plate must have high efficiency (some 90%) and a large bandwidth, because the solar spectrum has significant energy over a larger than 2:1 frequency range.

Northeast Photosciences has demonstrated holographic photovoltaic systems with over 30% efficiency.

14.5 Thermophotovoltaic Cells

Another scheme to increase the efficiency of solar converters recirculates the photons that, owing to the transparency of the photodiodes, failed to be absorbed. The body that emits the radiation must be in the immediate vicinity of the diodes. Consider Figure 14.8 (left) in which a radiator (heated by the sun through concentrators, by flames from a combustion unit, by a nuclear reactor, or by radioactive decay) illuminates a photodiode with a band gap, W_g. Photons with energy above W_g interact with the semiconductor (solid lines) and generate an electric current. Those with less than W_g pass through the semiconductor and, on being reflected by the mirror, are returned to the radiator (dashed lines).

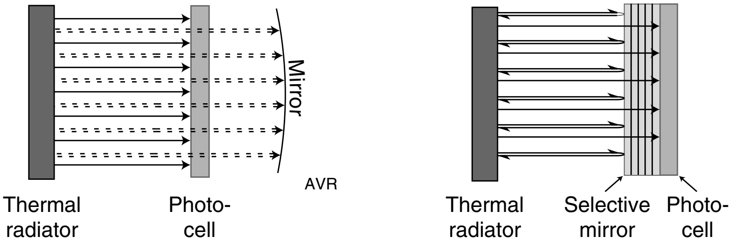


Figure 14.8 Two configurations of thermophotovoltaic converters.

The mirror can be the back electrode of the diode. In devices developed by Richard Swanson at Stanford, the mirror was a layer of silver separated from the diode by a small thickness of oxide. The electrical connection between the mirror and the diode was assured by a polka dot pattern of contacts that obscure only 1% of the mirror. Silicon was used and, owing to its relatively large band-gap, the radiator had to operate at the high temperature of 2200 to 2300 K. The choice of radiator material is difficult: not only does it have to be refractory and chemically stable, but it must also have a very low vapor pressure at the operating temperature to avoid being distilled onto the surface of the cool photocell. To reduce convection losses, it should operate in a vacuum, but that worsens the distillation problem; hence a small pressure of inert gas (argon) can be used as a compromise.

The solution is to operate at lower temperatures using semiconductors with a smaller band-gap. GaAs (0.7 eV) is the current semiconductor of choice. With radiator temperatures between 1200 and 1500 K, the whole operation shifts into the infrared region. Considerable spectrum manipulation is used to bring the efficiency of the device up to an acceptable level. This is accomplished by using an optical filter in front of the photocell instead of a mirror behind it. The filter (see the box, "Photonic Devices") is built right on top of the cells—providing a certain protection—and passes radiation above the band-gap of the semiconductor, while reflecting the longer wavelengths back to the emitter. In addition, the radiator has its surface modified to enhance emissions at the frequencies absorbed by the photocell.

Photonic Devices

Oil slicks and rainbows unfold sunshine into a spectrum of colors, which, despite their common source, look quite different.

A thin oil layer on top of a water surface gives rise to two different interfaces: one from air to oil and one from oil to water. See

(*Continues*)

(*Continued*)

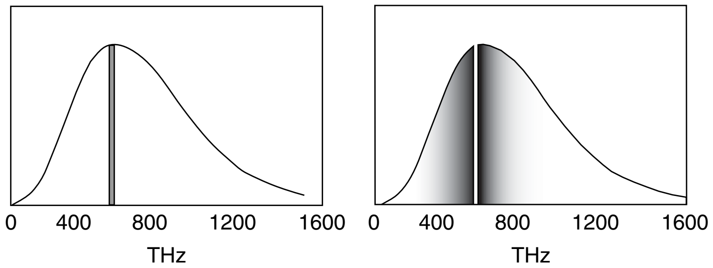


Figure 14.9 A rainbow (or a prism) will separate the white light into its individual colors. Each narrow slice is essentially monochromatic (left). An oil slick will show all colors simultaneously except one narrow slice that is canceled out by interference (right).

Figure 14.9. Owing to the difference between the indices of refraction, a small part of the incident light is reflected by each interface and, for a given light frequency, the two reflections cancel one another by destructive interference, removing a slice of the spectrum from the otherwise white reflected light. The exact frequency of cancellation depends of the thickness of the slick and the angle of view, leading to the well-known iridescence of oil slicks. This is the mechanism that generates the color of a blue jay's feather, the wings of some butterflies, and the shimmering of an opal. It is also the glare reducing mechanisms of blued optical surfaces.

If one builds up many layers of transparent media of alternating refractive indices, one can create an interference filter with prescribed band-pass and band-stop regions. Such filters can be made with rather sharp skirts, that is, with fairly abrupt changes of transmittance and reflection as a function of frequency. The device so constructed constitutes an **1D photonic crystal**. See Figure 14.10 (left).

More complex devices can be built, as, for example the 2-D and 3-D filters sketched in the figure.

Readers adept at physics can discover the analogy between electrons in a crystal and photons in a region of regularly spaced fluctuations in refractive indices. Electrons, influenced by periodical potential changes, can encounter forbidden energy levels such as the band-gap in semiconductors. Photons, influenced by periodic refractive index fluctuations, can find forbidden frequency bands in which they cannot propagate, thus acting as filters.

To learn more about photonics, read Joannopoulos et al. (2008).

(*Continues*)

(*Continued*)

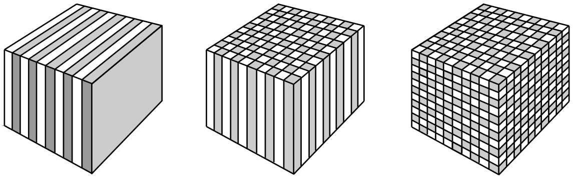


Figure 14.10 A sandwich of many layers of alternating dieletric constant materials constitutes a 1-D photonic crystal (left). 2-D (center), and 3-D (right) crystals can also be constructed.

MIT did considerable work in perfecting thermophotovoltaic devices. A group led by John Kassakian did the work we (very superficially) describe here. The system used a tungsten emitter operating at temperatures between 1200 and 1500 K, a band-pass filter, and a gallium arsenide photocell. Figure 14.11 shows the characteristics of the 1-D photonic filter, which had a pass band between 0.7 to 1.6 eV (170 to 390 THz). The filter was made of five 170-nm-thick layers of silicon alternating with 390-nm-thick layers of silicon dioxide deposited directly on the GaAs. In the figure, we sketched in a shaded area representing an idealized (eyeballed) band pass.

The filter passed a slice of the *black body* spectrum that contains 52% of the total radiated energy. Owing to the 82% transmissivity, only 43% actually reaches the photocell. The ideal conversion to electricity would be 21%. This number results from adapting Equation 14.28 to the truncated spectrum and multiplying by the average transmittance (0.82):

$$\eta_{ideal} = 0.82 \times 1780 \frac{V_g}{T} \left(\int_{\frac{qV_L}{kT}}^{\infty} \frac{x^2}{e^x - 1} dx - \int_{\frac{qV_U}{kT}}^{\infty} \frac{x^2}{e^x - 1} dx \right), \qquad (14.32)$$

where V_L and V_U are, respectively, the lower and upper cutoffs of the filter expressed in electron volts. For a black body at 6000 K, a semiconductor band-gap of 0.7 eV, and the filter being discussed, the ideal efficiency of the photocell is 17%. This is not too good, but one must remember that a great deal (but not all) of the energy that fails to be transmitted to the photocell is reflected back to the infrared emitter and is reused.

Additional improvements can be introduced. Our rough estimate was based on black body radiation. However, the emission from a hot object does not, in general, have a black body spectral distribution. It depends on how its emissivity varies with frequencies. Camping lights use a flame that would radiate lots of infrared and little light, a fact corrected by using a

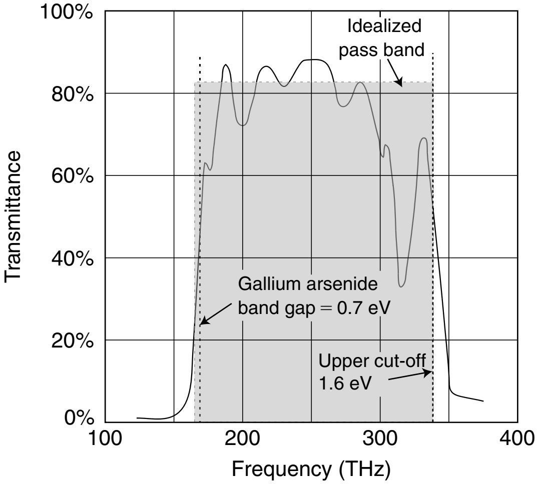


Figure 14.11 The above 1-D photonic filter made at MIT was designed to pass a large fraction of the radiation captured by GaAs photocells.

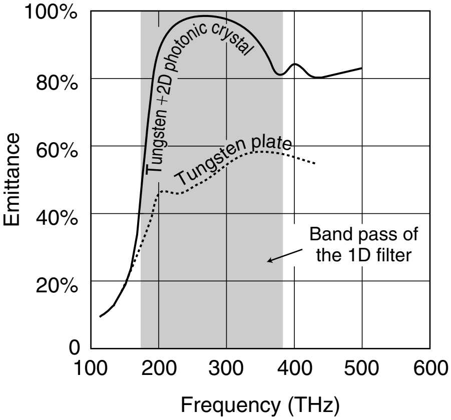


Figure 14.12 Spectral distribution of the radiation of a heated body modified (using a 2-D photonic filter) to emphasize regions of interest.

metal oxide[†] mantle having much higher emissivity in the visible than in the infrared region. It shines with an intense white light.

In the MIT experiment, the tungsten radiator had its emitted spectrum altered so as to enhance the frequencies absorbed by the photocell. A 2-D photonic filter was used to this end. Figure 14.12 sketches the spectral distribution of the treated radiator and compares it with the performance of

[†] Usually, thorium oxide with 1% cerium oxide, or, for those who fear the slight radioactivity of thorium, yttrium oxide with 3% cerium oxide.

a bare tungsten radiator. It can be seen that, in the frequencies of interest, there is a doubling of the power density emitted.

Achieved efficiencies of thermophotovoltaic devices are still modest. Orion Morrison of the Vehicle Research Institute (Western Washington University) developed a system that had an efficiency of 7%. It was used in an experimental hybrid vehicle, the Vicking 29, built at that university.

14.6 The Ideal and the Practical

> *"Silicon solar energy converters should be able to be made with efficiencies as high as 10 percent. At present, the best units are 6 percent efficient."*
>
> *M. B. Prince, J. Appl. Phys., May 1955.*

We have discussed the ideal photocell without specifying any particular implementation. An ideal cell is transparent to part of the spectrum and opaque to the rest. If there is no carrier multiplication, each absorbed photon gives rise to a single electron and a single hole, which, when created in the same region of the device, must be separated. In a *p-n* photodiode, this is accomplished by the junction potential. The resulting electron delivers to the external load exactly W_g units of energy regardless of the energy of the photon that created it. The excess energy is thermalized.

The efficiency of ideal cells exposed to monochromatic photons with energy just above W_g is 100%. When exposed to broadband radiation, the efficiency depends only on the nature of the radiation and the value of the band gap. This is the **simple ideal spectrum efficiency** of the cell.

It is possible to exceed the simple ideal efficiency by:

1. making use of photons that, having insufficient energy, failed to be absorbed by the cell. This is, for instance, the case of the beam-splitting system or of thermophotovoltaic arrangements.
2. making use of the excess energy imparted to the carriers during their generation. This is the case of the carrier multiplication systems.

Practical photocells fail to reach the ideal efficiency owing to a number of loss mechanisms:

1. Some of the incident photons are reflected by the cell instead of being absorbed or are absorbed by obstructions such as current collectors.
2. If the thickness of the photoactive material is insufficient, not all photons with energy above W_g are absorbed—the material may not be opaque enough.

3. Not all electron-hole pairs created live long enough to drift to the *p*-*n* junction. If their lifetime is small or if they are created too far from the junction, these pairs will recombine and their energy is lost. Electron-hole pairs may drift toward the surface of the device where the recombination rate is high and thus will not make it to the junction.
4. Carriers separated by the *p*-*n* junction will lose some of their energy while on the way to the output electrodes owing to the resistance of the connections. This constitutes the **internal resistance** of the cell.
5. Mismatch between photocell and load may hinder the full utilization of the generated power.

14.7 Solid-State Junction Photodiode

In this section we will consider *p*-*n* junctions and their use in converting light into electric energy.

For a more detailed discussion of *p*-*n* junctions, see da Rosa (1989). Those who require a more basic introduction to semiconductors can read Appendix B to this chapter.

Consider a single semiconductor crystal consisting of two regions in close juxtaposition—an ***n*-region** into which a minute amount of certain foreign atoms capable of easily releasing an electron has been introduced, and a ***p*-region** containing atoms capable of binding an electron. These foreign atoms or **dopants** are, respectively, called **donors** and **acceptors**. For semiconductors such as germanium and silicon, donors are Column V elements such as phosphorus, arsenic, and antimony, and acceptors are Column III elements such as boron, aluminum, gallium, and indium.

A donor that has released an electron becomes positively charged, but it remains firmly attached to the crystal and cannot move under the influence of an electric field—it constitutes an **immobile charge**. On the other hand, the released electron is free and being able to **drift** in an electric field becomes a negative **carrier** (of electricity). Correspondingly, acceptors, having acquired an extra electron, become negatively charged but are also immobile, whereas the hole—left over by the removal of the electron—can drift and become a positive carrier. A hole, although it is simply an absence of an electron where normally an electron should be (see Appendix B), can be treated as a particle with positive mass and a positive charge equal, in absolute value, to that of the electron.

The *n*-side of the junction contains positive donors and an abundance of free electrons, while the *p*-side contains negative acceptors and positive holes. Free electrons, more abundant in the *n*-side, diffuse toward the *p*-side, whereas holes from the *p*-side migrate to the *n*-side. If these particles were uncharged, the diffusion process would only stop when their concentration became uniform across the crystal. Being charged, their separation causes an electric field to be established, and a compensating drift current makes carriers flow back in a direction opposite to that of the diffusion current.

The drift current is driven by a **contact potential** created as follows: The migrating electrons not only transport negative charges to the p-side but also leave uncompensated positively charged donors in the n-side. The holes coming in from the p-side contribute to the accumulation of positive charges in the n-side and leave uncompensated negatively charged acceptors in the p-side. Thus the n-side becomes positive and the p-side negative.[†]

In an unbiased p-n junction (one to which no external voltage is applied), the overall current is zero, not because diffusion and drift currents are themselves zero but because these currents exactly balance one another. In other words, the equilibrium in a junction is **dynamic**, not static.

The equilibrium consists of four currents:

$j_{n_D} = -qD_n dn/dx$	(diffusion current of electrons from the n-side to the p-side).
$j_{n_E} = q\mu_n nE$	(drift current of electrons from the p-side to the n-side).
$j_{p_D} = -qD_p dp/dx$	(diffusion current of holes from the p-side to the n-side).
$j_{p_E} = q\mu_p pE$	(drift current of holes from the n-side to the p-side).

The symbols used in this section include the following:

n, concentration of electrons—number of free electrons per unit volume.
p, concentration of holes.
N_a, concentration of acceptors.
N_d, concentration of donors.
n_i, intrinsic carrier concentration.
D_n, diffusion constant of electrons.
D_p, diffusion constant of holes.
μ_n, mobility of electrons.
μ_n, mobility of holes.

The **intrinsic carrier concentration** is a temperature-dependent quantity that characterizes a given semiconductor. It is

$$n_i = BT^{3/2} \exp\left(-\frac{W_g}{2kT}\right), \tag{14.33}$$

[†] Contact potential is not an exclusive property of semiconductors. Two different metals, when joined, will also develop a contact potential between them.

where B can be calculated from knowledge of the **effective masses** of electrons and holes in the semiconductor in question. Usually, one does not find a tabulation of different values of B. What one finds are tabulations of n_i for a given semiconductor at a specified temperature. From that, B is calculated, and values on n_i at different temperatures can be found. The usual values of n_i, at 300 K, are $1.5 \times 10^{10}\,\mathrm{cm}^{-3}$ for silicon and $2.5 \times 10^{13}\,\mathrm{cm}^{-3}$ for germanium.

The total current, j, is the sum of the electron and the hole currents and must be equal to zero because there is no external connection to the device:

$$j = j_n + j_p = 0. \tag{14.34}$$

This equation can be satisfied by having either $j_n = -\,j_p \neq 0$ or $j_n = j_p = 0$. If either current were different from zero, holes or electrons would accumulate in one side of the junction—a situation that could not be sustained for any appreciable length of time. Hence, each of the currents must individually be zero. In other words, the drift of holes must (under equilibrium conditions) be exactly equal to the diffusion of holes. The same must be true for electrons.

To find the magnitude of the contact potential, one solves one of the above equations. Take, for instance, the hole current:

$$j_p = q\left(\mu_p p E - D_p \frac{dp}{dx}\right) = 0 \tag{14.35}$$

and, since $E = -dV/dx$,

$$\mu_p p \frac{dV}{dX} + D_p \frac{dp}{dx} = 0, \tag{14.36}$$

$$\frac{dp}{p} = -\frac{\mu_p}{D_p} dV = -\frac{q}{kT} dV. \tag{14.37}$$

Integrating this equation from deep into the p-side to deep into the n-side, one obtains

$$\ln \frac{p_p}{p_n} = [V(p_n) - V(p_p)] \frac{q}{kT}. \tag{14.38}$$

In the absence of an external current, there is no voltage drop across the undisturbed crystal (that part of the crystal far away from the junction). Thus, $V(p_n) - V(p_p)$ is the contact potential, V_C:

$$V_C = \frac{kT}{q} \ln \frac{p_p}{p_n} \approx \frac{kT}{q} \ln \frac{N_a N_d}{n_i^2}. \tag{14.39}$$

The approximation in Equation 14.39 is valid only when $p_n = n_i^2/N_d$.

Example 14.4

The doping concentrations of a particular silicon junction are $N_a = N_d = 10^{16}\,\text{cm}^{-3}$. The operating temperature is 300 K. What is the contact potential?

The values above are representative doping levels for many silicon devices. Since in each cubic centimeter of silicon one finds a total of 5×10^{22} silicon atoms, we can see that the impurity levels are very small (in relative terms): 1 impurity atom for each 5 million silicon atoms. From the chemical point of view, we are dealing here with super pure silicon! Since the diode is at 300 K, the intrinsic carrier concentration is $n_i = 10^{10}\,\text{cm}^{-3}$.

Using Equation 14.39,

$$V_C = \frac{kT}{q} \ln \frac{N_a N_d}{n_i^2}$$

$$= \left(\frac{1.38 \times 10^{-23} \times 300}{1.6 \times 10^{-19}}\right) \ln \left(\frac{10^{16} \times 10^{16}}{10^{20}}\right) = 0.72\,\text{V}.$$

Photocells and tunnel diodes, among others, use heavily doped semiconductors; consequently, their contact potential is somewhat higher.

With doping concentrations of $N_a = N_d = 10^{19}\,\text{cm}^{-3}$, the contact potential is $V_C = 1.08\,\text{V}$.

The contact potential cannot be measured by applying a voltmeter to a *p*-*n* junction because the potentials around a close circuit will exactly cancel out (provided the circuit is at uniform temperature).[†] If the potentials did not cancel out, then a *p*-*n* junction would drive a current through some external load delivering energy. This would constitute a heat engine delivering a useful output under a zero temperature differential—a thermodynamic impossibility.

Another way to see that the potentials in a close circuit must cancel out is to consider that any connection between the free ends of a *p*-*n* junction must involve at least one additional external junction whose voltage opposes that of the original one, as explained next. Figure 14.13 shows a *p*-*n* junction shorted out by a copper wire. In addition to the device junction, there are two external junctions, one between the *p*-material and the copper wire and one between the copper wire and the *n*-material. A conceptually simpler way to short out the device is shown in Figure 14.14 where the external

[†] If there is a temperature difference along the circuit, then thermoelectric effects will appear, as discussed in Chapter 5.

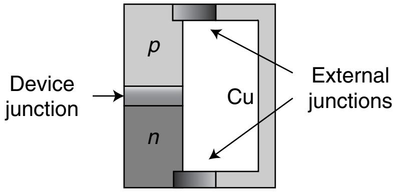


Figure 14.13 Two external junctions are formed when a photodiode is shorted out by a metallic wire.

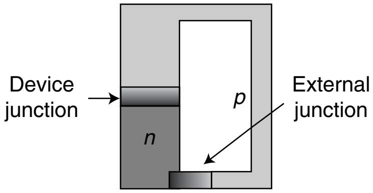


Figure 14.14 A single external junction is formed when a diode is shorted out by its own material.

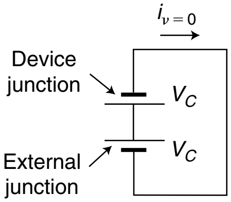


Figure 14.15 No current circulates in a shorted junction at even temperature and no illumination.

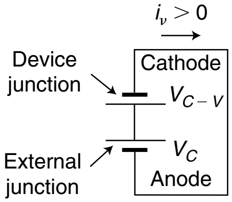


Figure 14.16 When light reduces the potential across the device junction, an external current circulates.

wire is made of the same material as the p-side of the diode. Now there is a single external junction between the p-wire and the n-material. Clearly, the two p-n junctions (the device junction and the external junction) have the same contact potential and oppose one another. Electrically, the situation is as depicted in Figure 14.15. Obviously, the external current is zero. When illuminated, the situation becomes that shown in Figure 14.16.

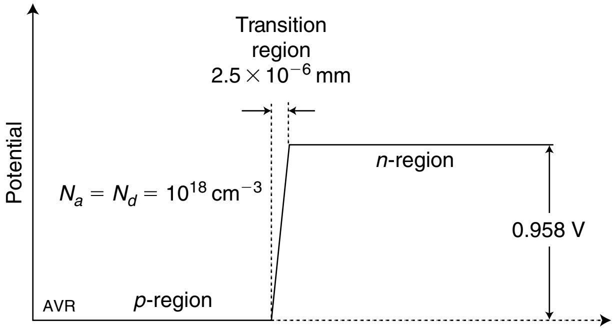


Figure 14.17 The electric field of a *p*-*n* junction is confined to a very narrow transition region.

To understand what happens when the device is exposed to light, examine Figure 14.17, which depicts the potential across the photodiode. The values shown correspond to a Si diode with equal doping in the two sides (10^{18} dopant atoms per cm^3). The contact potential at ambient temperature is 0.958 V, the n-region being positive with respect to the p-region. The potential in the device is uniform except in a narrow transition region only 2.5×10^{-6} cm wide where there is a strong electric field of 380,000 V/cm.

Light creates exitons that will recombine after a short **lifetime**, τ, unless an electric field separates the electrons from the corresponding holes. Owing to the narrowness of the transition region, not too many electron-hole pairs are created in the region itself. Most are created in the neighborhood of the region, and, because they move at random, they have a chance of stumbling into the high electric field where the separation occurs: the electrons created in the negatively charged p-region will drift to the n-side, causing it to become less positive. By the same token, holes created in the positively charged n-region will drift to the p-region causing the latter to become less negative. Thus, light has the effect of *reducing* the contact potential, which becomes, let's say, $V_c - v$, as depicted in Figure 14.16. Since, as we are going to explain, the external junction is not light sensitive and, hence, still has a contact potential, V_c, there is no longer an exact cancellation of the two contact potentials, and a net external voltage, v, will appear constituting the output of the photodiode. A current, I_ν, will flow through the short in the direction indicated.

For the electron-hole pairs created in the field-free region to reach the transition region, they must survive long enough—that is, their lifetime, τ, must be sufficiently large. If the lifetime is small, these pairs will recombine, contributing nothing to the output of the photodiode. The larger τ, the larger the efficiency of the device.

Thus, it is of paramount importance that the material from which the diode is made be as free from defects as possible because defects reduce τ. On the other hand, the external junction is, intentionally, made to have extremely low τ and thus is quite insensitive to light.[†]

The conventional direction of the external current is from the p-side of the device (the cathode) to the n-side (the anode). Its magnitude is

$$I_\nu = q\phi_g A \mathrm{A} \tag{14.40}$$

or

$$J_\nu = q\phi_g \, \mathrm{A/m^2}, \tag{14.40a}$$

where q is the charge of the electron, ϕ_g is the flux of photons with energy larger than the band-gap energy, W_g, and A is the active area of the junction. This assumes 100% quantum efficiency—each photon creates one electron-hole pair capable of reaching the potential barrier of the junction.

The short-circuit current is proportional to the photon flux and can be used for its measurement. There is a minute temperature dependence of I_ν, but only if the incident light is not monochromatic. This results from the small change in the band-gap energy with temperature. The higher the temperature, the smaller this energy and, consequently, the larger the fraction of the spectrum that has sufficient energy to disturb covalent bonds and, thus, to generate current.

A short-circuited photodiode can be represented by a shorted current source as suggested by Figure 14.18. If however, the short is replaced by a load, R_L, then a voltage, $V_L = I_L R_L$, will appear across it. This voltage will drive a current, I_D, through the diode. See Figure 14.19.

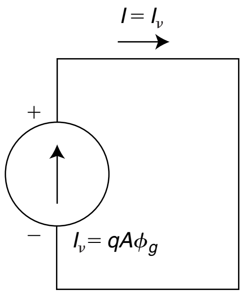


Figure 14.18 A short-circuited diode can be represented as a current source.

[†] The very act of soldering a wire to the diode will create enough imperfections as to reduce the τ at this junction to negligible values.

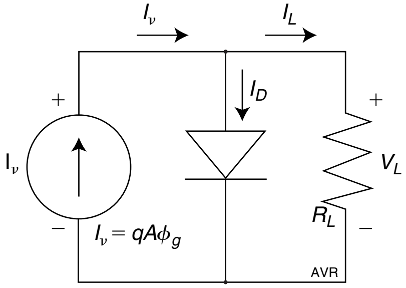


Figure 14.19 Model of an ideal photodiode with a resistive load.

Clearly,

$$I_L = I_\nu - I_D, \tag{14.41}$$

where I_D is a function of the voltage, V_L, across the diode. We need to know the functional relationship between I_D and V_L.

We saw that in an unbiased diode, there is a balance between the diffusion and the drift currents: they exactly cancel one another. When an external voltage (bias) is applied, the potential barrier at the junction is altered. If the bias reinforces the built-in barrier, only a minute current (the **reverse saturation current**), I_0, flows. When the bias is forward—that is, the potential barrier is lowered—a substantial current flows. This forward current occurs when a positive voltage is applied to the *p*-side of the device—that is, the positive (conventional) direction of the forward current is into the *p*-side. The diode behaves in a markedly asymmetric fashion with respect to the applied voltage.

As shown in any elementary text on electronics, the current through a *p*-*n* diode is

$$I_D = I_0\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{14.42}$$

Figure 14.20b illustrates the V-I characteristics of a *p*-*n* diode for small applied biases.

If $|V|$ is much larger than kT/q, and $V < 0$, the exponential term becomes negligible compared with 1 and I_D **saturates** at a value $I = -I_0$. That is, the current becomes independent of the applied voltage. Since at room temperature, $kT/q = 0.026\,\mathrm{V}$, the saturation occurs for voltages as small as $-0.1\,\mathrm{V}$.

If the applied voltage is positive and much larger than kT/q, then the exponential term dominates and the current increases exponentially with voltage. A positive bias of as little as 0.5 V causes the current through the diode to grow to $2 \times 10^8 I_0$. It can be seen that the reverse saturation

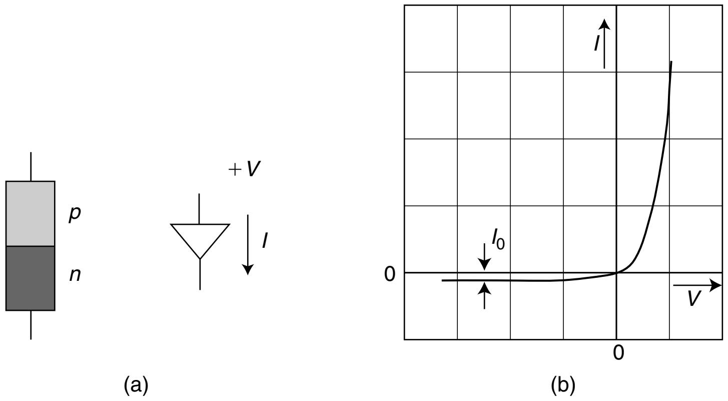


Figure 14.20 The V-I characteristics of a p-n diode.

current, I_0 is a quantity extremely small compared with the diode currents under modest forward biases.

Applying Equation 14.42 to Equation 14.41,

$$I_L = I_\nu - I_0\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{14.43}$$

The current, I_D, that flows through the diode diverts some current from the load. The smaller, I_D, the more current flows through the load, and, presumably, the more efficient the diode. Notice that I_D is proportional to I_0 (Equation 14.42). This explains the great effort made to produce photodiodes with the smallest possible reverse saturation current, I_0.

Inverting Equation 14.43,

$$V = \frac{kT}{q}\ln\left[\frac{I_\nu - I_L}{I_0} + 1\right]. \tag{14.44}$$

A real diode has an internal resistance, R_s, that causes the voltage across the load to drop to

$$V_L = \frac{kT}{q}\ln\left[\frac{I_\nu - I_L}{I_0} + 1\right] - I_L\,R_s. \tag{14.45}$$

To a first approximation, a photodiode acts as a current generator. It is a low-voltage, high-current device; several must be connected in series to generate practical voltage levels. There are difficulties in such connections.

The current through the series is the current generated by the weakest diode. A partial shadowing of a solar panel consisting of many diodes in series can drastically reduce the overall output.

One solution consists of building a larger number of minute diodes in a series string, which then are paralleled to produce useful current. Partial shadowing will then disable a limited number of diodes.

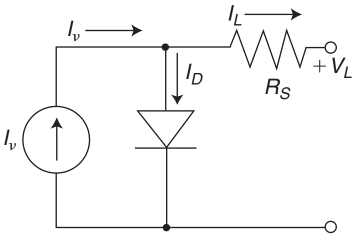


Figure 14.21 Circuit model of a real photodiode.

The open-circuit voltage of the photodiode is found by setting $I_L = 0$ in Equation 14.45:

$$V_{oc} = \frac{kT}{q}\ln\left(\frac{I_\nu}{I_o}+1\right) \approx \frac{kT}{q}\ln\left(\frac{I_\nu}{I_0}\right). \tag{14.46}$$

The voltage depends on the I_ν/I_0 ratio. Both currents are proportional to the cross section of the junction; their ratio, therefore, does not depend on the cross section. It depends on the photon flux (the light power density) and on the characteristics of the diode. Typically, this ratio is of the order of 10^7 at one sun ($1000\,\mathrm{W/m^2}$) in silicon. At 300 K, this corresponds to an open-circuit voltage of 0.42 V.

Owing to its logarithmic dependence, the open-circuit voltage varies slowly with changes in light power density and is not a convenient photometric measure.

The V-I characteristic of a photodiode can be obtained by plotting Equation 14.45 as shown in Figure 14.22. The currents were normalized by dividing by I_0. Three different illumination levels are displayed, corresponding to I_ν/I_0 of 10^6, 5×10^6, and 10^7.

For the highest level of illumination in the figure, in addition to the characteristics for diodes with no series resistance, we have also plotted the curve for a diode with normalized resistance of $R_s = 7.5\times10^{-9}/I_0$ ohms.

The power a photodiode delivers to a load depends on, among other things, the value of the load resistance. The optimum operating point occurs at a specific voltage, V_m, and a corresponding current, I_m (see derivation further on). These points are indicated by the small black circles in Figure 14.22.

The area under the curves has the dimension of power. The large darker gray rectangle represents the $V_m \times I_m$ product (the maximum power the diode can deliver when an ideally matched load is used), whereas the lighter gray rectangle represents the $V_{oc} \times I_\nu$ product. The ratio between these two areas (or the two powers) is called the **fill factor**, F_f. Both gray areas are for the case of $I_\nu/I_0 = 10^7$, which is the highest level of

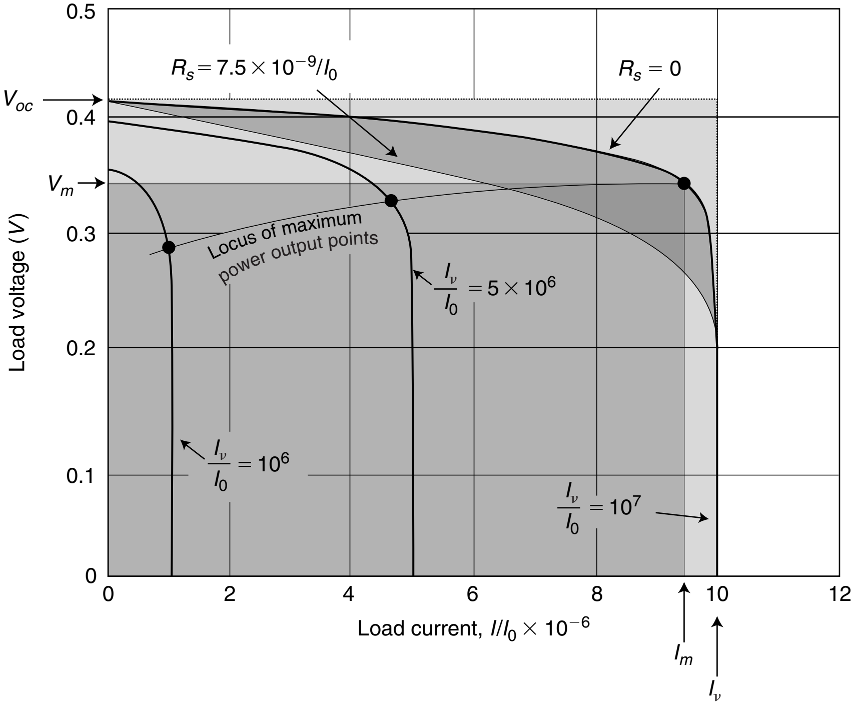


Figure 14.22 V-I characteristic of a typical photodiode.

illumination appearing in the figure. V_m and I_m are for a diode with no series resistance.

$$F_f = \frac{V_m \times I_m}{V_{oc} \times I_\nu}. \tag{14.47}$$

The gray region with curved boundaries corresponds to the difference in area between the characteristics of a diode with no series resistance and one with a given series resistance. This area represents the losses associated with this internal resistance. Clearly, the smaller the series resistance (i.e., the smaller the losses), the larger the fill factor. This factor indicates how closely a given diode approaches the power $V_{oc} \times I_\nu$.

It is important to inquire how the maximum deliverable power from a photodiode depends on a number of different parameters such as level of illumination and temperature. This can, of course, be done by numerical experimentation using the expressions for V_m and I_m. However, since the calculation of V_m involves an equation (Equation 14.51) that cannot be solved analytically, this procedure does not yield a simple formula from which the effects of the different parameters can be assessed.

On the other hand, the $V_{oc} \times I_\nu$ product can be calculated from a simple equation (Equation 14.49). For this reason, in the following

Table 14.4 Fill Factor as Function of the Level of Illumination

I_ν (A)	V_{oc} (V)	V_m (V)	I_m (A)	Fill factor $\frac{V_m \times I_m}{V_{oc} \times I_\nu}$
0.1	0.358	0.293	0.092	0.752
1	0.417	0.348	0.931	0.777
10	0.477	0.404	9.40	0.796
100	0.537	0.461	94.7	0.813

considerations, we inferred the behavior of $P_{L_{max}}$ from that of the $V_{oc} \times I_\nu$ product. This would be perfectly correct if these two quantities were strictly proportional to one another. However, this is not quite the case.

Table 14.4 displays values of the open-circuit voltage, V_{oc}, of the maximum-power voltage, V_m, of the maximum-power current, I_m, and of the fill factor F_f for a hypothetical resistanceless photodiode having a reverse saturation current of 10^{-7} A/m^2. It can be observed that, although the ratio is not constant, it varies relatively little when the illumination level changes by three orders of magnitude. Consequently, studying the behavior of the $V_{oc} \times I_\nu$ product leads to a simplified insight on the behavior of $P_{L_{\max}}$.

There are four important observations regarding the efficient operation of photovoltaic devices:

1. *All else being the same, the efficiency of a photodiode increases with an increase in the light power density.*

To gain a qualitative insight on the effect of the input light power density, P_{in}, on the efficiency, consider the following facts:

The *open-circuit* voltage, V_{oc}, is, as seen before,

$$V_{oc} = \frac{kT}{q} \ln\left(\frac{I_\nu}{I_0}\right). \tag{14.48}$$

The $I_\nu V_{oc}$ product is

$$I_\nu V_{oc} = I_\nu \frac{kT}{q} \ln\left(\frac{I_\nu}{I_o}\right). \tag{14.49}$$

Since I_ν is proportional to the input light power density, P_{in}, and V_{oc} grows logarithmically with I_ν, and, hence, with P_{in}, it follows that the $I_\nu V_{oc}$ product grows faster than linearly with P_{in}.

Accepting that the power output of the photodiode follows the behavior of the $I_\nu V_{oc}$ product, then, plausibly, an increase in P_{in} causes a more than proportional increase in P_{out}. That is, the efficiency increases with P_{in}.

One can reach the same conclusion more rigorously as shown below.

The power delivered by the diode is

$$P = VI = VI_\nu - VI_0\,[\exp(qV/kT) - 1]. \tag{14.50}$$

Maximum power flows when $V = V_m$ such that $dP/dV = 0$, which leads to a transcendental equation that must be solved numerically:

$$\left(1 + \frac{qV_m}{kT}\right) \exp\!\left(\frac{qV_m}{kT}\right) = J_\nu / J_0 + 1. \tag{14.51}$$

Here, we have switched from currents to current densities.

An empirical formula that closely approximates the solution of the above equation is

$$V_m = V_A \ln\!\left(\frac{P_{in}}{V_B J_0}\right), \tag{14.52}$$

where

$$V_A = 2.2885 \times 10^{-2} - 139.9 \times 10^{-6} \ln J_0 - 2.5734 \times 10^{-6} (\ln J_0)^2, \tag{14.53}$$

$$V_B = 4.7253 - 0.8939 \ln J_0. \tag{14.54}$$

An inspection of Figure 14.22 shows that the optimum operating point must lie near the knee of the V-I characteristics, as indicated by the solid circles in the figure.

From Equation 14.43,

$$J_m = J_\nu - J_0 \left[\exp\!\left(\frac{qV_m}{kT}\right) - 1\right], \tag{14.55}$$

where J_m is the current density when $V = V_m$.

The maximum power (per unit active area) the diode delivers to a load is

$$P_m = V_m J_m, \tag{14.56}$$

and the maximum efficiency (when the load is properly matched to the photodiode) is

$$\eta_m = \frac{V_m J_m}{P_{in}}. \tag{14.57}$$

In Section 14.2, we define a device-independent **ideal efficiency**, η_{ideal}, a function of only the spectral distribution and the band gap of the semiconductor, W_g. The **ideal-diode efficiency**, η_m, defined earlier, considers the case of an ideal photodiode that has a nonzero reverse saturation current, I_0. Designers strive to make I_0 as small as possible, but it cannot be made vanishingly small (see Subsection 14.7.1). If I_0 were taken as zero, then the diode would be modeled as a simple current generator (with no shunting diode), and no open-circuit voltage could be defined because current generators cannot deliver a current to an infinite resistance.

On the other hand, from the manner in which a p-n diode works, it can be seen that the output voltage is limited to the contact potential, V_C. One could then argue that as $I_0 \to 0$, $J_m \to J_\nu$, and $V_m \to V_C$ and the power delivered to the load (per unit active area) would be $P_L = I_\nu V_C$. The value of V_C can be obtained from Equation 14.39. Thus,

$$P_{L_{\max}} = J_\nu \frac{kT}{q} \ln \frac{N_a N_d}{n_i^2}. \tag{14.58}$$

But $J_\nu \propto P_{in}$. This relationship is demonstrated later in Example 14.5. Let γ be the constant of proportionality between J_ν and P_{in}, then

$$P_{L_{\max}} = \gamma P_{in} \frac{kT}{q} \ln \frac{N_a N_d}{n_i^2}, \tag{14.59}$$

and the efficiency would be

$$\eta_m = \gamma \frac{kT}{q} \ln \frac{N_a N_d}{n_i^2}. \tag{14.60}$$

The largest value that either N_a or N_d can have, in silicon, is around 10^{19} atoms/cm^3. Remember that silicon has an atomic concentration of about 5×10^{22} atoms per cm^3. This is roughly the limit of solubility of most dopants in silicon. A larger concentration of such dopants will cause some of them to settle out. Using $\gamma = 0.399$ (again, please refer to Example 14.5), for the case of a 6000-K black body radiation we get,

$$\eta_m = 0.399 \left(\frac{1.38 \times 10^{-23} \times 300}{1.6 \times 10^{-19}} \right) \ln \left(\frac{10^{19} \times 10^{19}}{10^{20}} \right) = 0.43. \tag{14.61}$$

Although this result is almost precisely the ideal efficiency of a silicon photocell exposed to 6000-K black body radiation, its precision is spurious. One reason is that zero reverse saturation currents cannot be realized even theoretically. In addition, there is nothing magical about the maximum level of doping, which is limited by the solubility of dopants in silicon and can, by no stretch of the imagination, be directly related to the ideal efficiency of a photocell.

2. *The efficiency of a photodiode decreases with an increase in the reverse saturation current.*

We have shown that the behavior of the $I_\nu V_{oc}$ product is a good indicator of the behavior of the $I_m V_m$ product and, consequently, of the efficiency of the photodiode. Referring back to Equation 14.49, it can be seen that an increase in I_0 results in a decrease in $I_\nu V_{oc}$ and, thus, plausibly, in the efficiency. Also, the numerical experimentation in Example 14.5 confirms this conclusion.

Good photodiodes are designed to minimize I_0. This, as explained in Section 14.8 can be accomplished by

1. using highly doped semiconductors and
2. striving for the largest possible minority carrier lifetimes.

Example 14.5

How does the efficiency of an ideal silicon photodiode vary with light power density and with the reverse saturation current density, J_0, when exposed to the radiation from a 6000-K black body?

To answer this question, one must solve a number of equations numerically. A spreadsheet is an excellent platform do to so.

Start by picking a value for P_{in} and one for J_0. Calculate the corresponding efficiency using

$$\eta_m = \frac{V_m J_m}{P_{in}}. \qquad (14.62)$$

Increment J_0 and recalculate η_m. After all desired values of J_0 have been used, select a new P_{in} and repeat the whole procedure.

To calculate η_m, we need to have the values of V_m and I_m as a function of the selected parameters P_{in} and J_0.

Calculate V_m

Given these parameters, calculate V_m from

$$V_m = V_A \ln\left(\frac{P_{in}}{V_B J_0}\right), \qquad [14.52]$$

where

$$V_A = 2.2885 \times 10^{-2} - 139.9 \times 10^{-6} \ln J_0 - 2.5734 \times 10^{-6} (\ln J_0)^2, \qquad [14.53]$$

$$V_B = 4.7253 - 0.8939 \ln J_0. \qquad [14.54]$$

Calculate I_m

Use

$$J_m = J_\nu - J_0 \left[\exp\left(\frac{qV_m}{kT}\right) - 1\right].$$

To do this we need J_ν as a function of P_{in}.

From Equation 14.25,

$$\sigma = 0.416 \int_{\frac{qV_g}{kT}=2.1256}^{\infty} \frac{x^2}{e^x - 1} dx = 0.416 \times 1.3405 = 0.5577. \qquad (14.63)$$

(*Continues*)

(*Continued*)

The value of the definite integral was from Appendix A. Still from Equation 14.25,

$$\phi_g = \sigma\phi = 0,5577\phi. \tag{14.64}$$

From Equation 14.16

$$\phi = \frac{P_{in}}{223.7 \times 10^{-21}}, \tag{14.65}$$

$$\phi_g = 0.5577\frac{P_{in}}{223.7 \times 10^{-21}}, \tag{14.66}$$

$$J_\nu = q\phi_g = 1.6 \times 10^{-19} \times 0.5577\frac{P_{in}}{223.7 \times 10^{-21}} = 0.399P_{in}, \tag{14.67}$$

$$J_m = 0.399P_{in} - J_0\left[\exp\left(\frac{qV_m}{kT}\right) - 1\right]. \tag{14.68}$$

Using these formulas, we calculated the ideal efficiency of silicon photodiodes exposed to the radiation from a 6000-K black body (roughly, the radiation from the sun).

This was done for various light power densities and different values of the reverse saturation current density. The results are tabulated in Table 14.5 and plotted in Figure 14.23.

Table 14.5

J_0(nA/m^2) →	10	20	50	100	200	500	1000
P_{in}(W/m^2) ↓	η_m	η_m	η_m	η_m	η_m	η_m	η_m
10	0.1637	0.1570	0.1482	0.1415	0.1349	0.1261	0.1195
20	0.1704	0.1637	0.1548	0.1482	0.1415	0.1327	0.1261
50	0.1793	0.1726	0.1637	0.1570	0.1503	0.1415	0.1349
100	0.1860	0.1793	0.1704	0.1637	0.1570	0.1482	0.1415
200	0.1928	0.1860	0.1771	0.1704	0.1637	0.1548	0.1482
500	0.2017	0.1950	0.1860	0.1793	0.1726	0.1637	0.1570
1000	0.2085	0.2017	0.1928	0.1860	0.1793	0.1704	0.1637
2000	0.2153	0.2085	0.1996	0.1928	0.1860	0.1771	0.1704
5000	0.2243	0.2175	0.2085	0.2017	0.1950	0.1860	0.1793
10000	0.2311	0.2243	0.2153	0.2085	0.2017	0.1928	0.1860
20000	0.2379	0.2311	0.2221	0.2153	0.2085	0.1996	0.1928
50000	0.2469	0.2401	0.2311	0.2243	0.2175	0.2085	0.2017
100000	0.2538	0.2469	0.2379	0.2311	0.2243	0.2153	0.2085

(*Continues*)

(Continued)

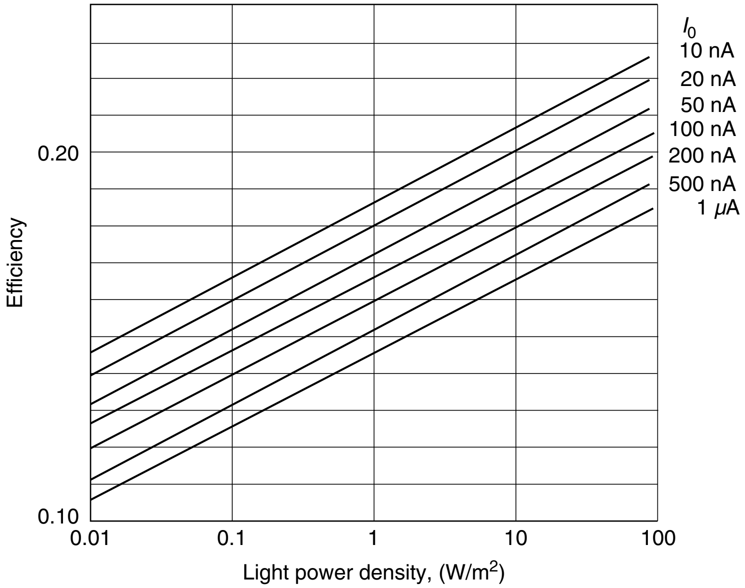


Figure 14.23 The efficiency of lossless silicon photodiodes exposed to 6000-K black body radiation increases logarithmically with light power density and decreases with increasing reverse saturation current.

It can be seen from Figure 14.23 that there is a logarithmic dependence of the efficiency on the level of illumination and that the efficiency decreases when the reverse saturation current increases.

3. *The efficiency of a photodiode decreases with an increase in the operating temperature.*

Equation 14.62 shows that the efficiency is proportional to V_m, and, since V_m increases when V_{oc} increases, it is sufficient to investigate the temperature dependence of the latter.

Referring to Equation 14.48, it is not immediately clear what happens to V_{oc} when the temperature changes. The kT/q factor causes the voltage to rise proportionally to T, but the $\ln(I_\nu/I_0)$ factor causes the temperature to influence the voltage in the opposite direction through its effect on I_0. A more careful analysis has to be carried out.

Using information from the next section, we can write

$$V_{oc} = \frac{kT}{q}\ln\left(\frac{J_\nu}{J_0}\right) = \frac{kT}{q}\ln\left(\frac{I_\nu}{qAn_i^2\left[\sqrt{\frac{D_p}{\tau_p}}\frac{1}{N_d} + \sqrt{\frac{D_n}{\tau_n}}\frac{1}{N_a}\right]}\right). \qquad (14.69)$$

Here, we inserted the expression for J_0,

$$J_0 = qn_i^2 \left[\sqrt{\frac{D_p}{\tau_p}} \frac{1}{N_d} + \sqrt{\frac{D_n}{\tau_n}} \frac{1}{N_a} \right], \tag{14.70}$$

where n_i is the **intrinsic concentration** of holes and electrons and N_d and N_a are, respectively, the donor and acceptor concentrations, D and τ are, respectively, the **diffusion constant** and the **lifetime** of the minority carriers.

The lifetime of minority carriers is a function of the quality of the material and of the heat treatment it has undergone. It increases with temperature (see Ross and Madigan 1957). The diffusion constant behavior depends on the doping level. In purer materials it decreases with temperature, and in heavily doped ones (as is the case of photodiodes) it may even increase with temperature. At any rate, the combined effect of temperature on D and τ is relatively small and, for this general analysis, will be disregarded. This allows us to write

$$J_0 = \Lambda n_i^2, \tag{14.71}$$

where Λ is

$$\Lambda \equiv q \left[\sqrt{\frac{D_p}{\tau_p}} \frac{1}{N_d} + \sqrt{\frac{D_n}{\tau_n}} \frac{1}{N_a} \right]. \tag{14.72}$$

The intrinsic carrier concentration—the number of either electrons or of holes per unit volume when the semiconductor is perfectly (and impossibly) pure—is given by

$$n_i = BT^{3/2} \exp\left(-\frac{qV_g}{2kT} \right), \tag{14.73}$$

where V_g is the band-gap voltage and B is a constant that varies from material to material.

The expression for the open-circuit voltage becomes

$$V_{oc} = \frac{kT}{q} \ln\left(\frac{J_\nu \exp \frac{qV_g}{kT}}{\Lambda B^2 T^3} \right), \tag{14.74}$$

in which all quantities, other than T are (completely or nearly) temperature independent. Remember that J_ν has only a minuscule dependence on T.

From expression 14.74,

$$\frac{dV_{oc}}{dT} = \frac{k}{q}\left\{\ln\left[\frac{J_\nu \exp\left(\frac{qV_g}{kT}\right)}{\Lambda B^2 T^3}\right] - 3 - \frac{qV_g}{kT}\right\}$$

$$= \frac{k}{q}\left\{\ln\left[\frac{J_\nu}{J_0(T)}\right] - 3 - \frac{qV_g}{kT}\right\}$$

$$= \frac{k}{q}\left\{\ln\left[\frac{300^3 J_\nu}{T^3 J_0}\right] - 3 - \frac{qV_g}{kT}\right\}. \qquad (14.75)$$

Here, $J_0(T)$ is the temperature-dependent reverse saturation current density, and it is equal to $(T/300)^3 J_0$ where J_0 is the reverse saturation current density at 300 K.

It can be seen that dV_{oc}/dT is negative (V_{oc} goes down with increasing temperature) as long as

$$\ln\left[\frac{J_\nu}{J_0(T)}\right] < 3 + \frac{qV_g}{kT}, \qquad (14.76)$$

or

$$\frac{J_\nu}{J_0(T)} < \exp\left(3 + \frac{qV_g}{kT}\right) \sim 200 \times 10^{12}. \qquad (14.77)$$

We used $qV_g/kT \sim 30$ as a representative minimum value.

We also recognize that for an excellent photodiode exposed to 1 sum (1000 W/m^2), the J_ν/J_0 ratio at 300 K is $\sim 10^8$. For the inequality above to become nonvalid, J_ν would have to reach the value of 2 million A/m^2, a patent impossibility. Thus, under all plausible circumstances, V_{oc} decreases with an increase in temperature.

The theoretical dependence of dV_{oc}/dT on the illumination level is plotted for three different semiconductors in Figure 14.24. The plot is based on Equation 14.75.

It is apparent that

1. all temperature coefficients of the open-circuit voltage are negative; that is, in all cases, an increase in T results in a decrease in V_{oc} and, consequently, in the power output and efficiency;
2. the larger the band gap, the larger (in absolute value) the temperature coefficient of the open-circuit voltage; and
3. the larger the illumination, the smaller the temperature coefficient.

Actual measured open-circuit voltages of a silicon photodiode are displayed in Figure 14.25 (see Rappaport 1959). The observed temperature coefficients are about 2 mV/K, roughly what is predicted by Equation 14.75.

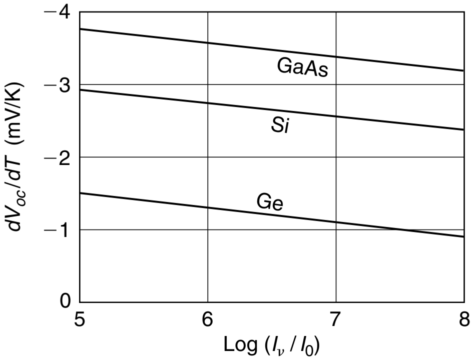


Figure 14.24 Temperature coefficient of the open-circuit voltage.

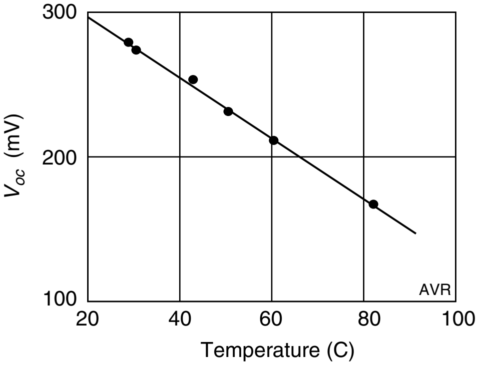


Figure 14.25 Observed temperature variation of the open-circuit voltage.

Unfortunately, there is no information about the temperature and the illumination level used in Rappaport's measurements, so a more precise comparison between prediction and observation cannot be made.

Large efficiencies are obtained with large concentration ratios (i.e., large light power densities). However, increasing the concentration tends to raise the operating temperature of the diode, degrading the performance. As a result, in order to profit from high concentration, it is necessary to provide adequate cooling. Here silicon with its larger thermal conductivity has an advantage over gallium arsenide.

Sometimes it is possible to operate diodes under a thin layer of water that removes the heat.

4. For maximum efficiency, a load must "match" the photodiode.

If good efficiency is desired, photodiodes must be operated near I_m. This means that for each level of illumination the optimum load must have a different resistance.

Consider a photodiode operating at the largest expected light power density, $P_{in_{max}}$. A properly matched load is one that causes the diode to operate at the maximum output power, that is, with a load voltage, V_m, defined by Equation 14.51. The corresponding load current, I_m, can be obtained from Equation 14.55. It should be noticed that, owing to the almost vertical characteristic of the diode at high currents, $I_m \approx I_\nu$.

The load must be adjusted to

$$R_L = \frac{V_{m_{max}}}{I_{m_{max}}}, \tag{14.78}$$

and the power delivered is

$$P_{L_{max}} = I^2_{m_{max}} R_L \approx I^2_{\nu_{max}} R_L. \tag{14.79}$$

Expose, now, the photodiode to a light power density $P_{in} = \lambda P_{in_{max}}$ where λ is a quantity smaller than 1, but maintain the same load resistance, R_L. The short-circuit current, I_ν, is strictly proportional to P_{in} and is now

$$I_\nu = \lambda I_{\nu max}. \tag{14.80}$$

The power delivered to the load is

$$P_L = I_L^2 R_L \approx \lambda^2 I_{\nu_{max}} R_L \tag{14.81}$$

because $I_L \approx I_\nu$.

In other words, with a constant load resistance the electric power delivered falls with the square of the ratio of the sun power densities. Thus, if the maximum expected illumination is, say, 1000 W/m^2 and the load is adjusted to match these conditions, then if the illumination falls to 100 W/m^2, the dc output falls all the way to 1% of the former value. In contrast, if the load resistance is changed to cause operation at the "matched" condition, the dc power will only fall to 10%.

Since non–sun-tracking solar systems are subject to widely fluctuating illumination throughout a day, it is important to use devices (**load followers**) that automatically adjust the load characteristics to match the requirements of the photodiodes.

14.7.1 The Reverse Saturation Current

The reverse saturation current in a diode is the sum of the saturation current owing to holes, I_{p0}, and that owing to electrons, I_{n0}. Simple *p*-*n* diode theory leads to the expression

$$I_0 = I_{p0} + I_{n0} = qA\left(\frac{D_p}{L_p} p_n + \frac{D_n}{L_n} n_p\right). \tag{14.82}$$

Notice that the ratio of the diffusion constant, D, to the diffusion length, L, has the dimension of velocity so that, when multiplied by the carrier concentration (p_n or n_p), it yields a diffusion flux. Multiplying all by qA (where A is the area of the junction) converts the flux into current.

L and D are related to one another by the lifetime, τ, of the minority carriers:

$$L = \sqrt{D\tau}. \tag{14.83}$$

Consequently,

$$I_0 = qA\left[\sqrt{\frac{D_d}{\tau_p}}p_n + \sqrt{\frac{D_n}{\tau_n}}n_p\right] \tag{14.84}$$

and since

$$p_n = \frac{n_i^2}{N_d}, \tag{14.85}$$

and

$$n_p = \frac{n_i^2}{N_a}, \tag{14.86}$$

it follows that

$$I_0 = qAn_i^2\left[\sqrt{\frac{D_p}{\tau_p}}\frac{1}{N_d} + \sqrt{\frac{D_n}{\tau_n}}\frac{1}{N_a}\right], \tag{14.87}$$

where n_i is the intrinsic concentration of holes and electrons and N_d and N_a are the donor and acceptor concentrations, respectively.

In the derivation of I_0, several simplifying assumptions were made.†

† Among the most important simplifications are the following:

1. All the recombination of minority carriers was assumed to occur in the bulk of the material, and none at the surface. The surface recombination velocity, s, was taken as zero or, at least, much smaller than D/L.
2. The width, w, of the diode, normal to the junction, was taken to be much larger than the diffusion length ($w >> L$).
3. It was assumed that no generation and no recombination of electron-hole pairs occurred in the transition region of the junction.

If assumption 1 is removed and assumption 2 is kept ($s >> D/L$)—that is, if recombination is dominated by surface effects, then I_0 is given by

$$I_0 = qAn_i^2\left[\sqrt{\frac{D_p}{\tau_p}}\frac{1}{N_d}\coth\frac{w_n}{L_p} + \sqrt{\frac{D_n}{\tau_n}}\frac{1}{N_a}\coth\frac{w_p}{L_n}\right].$$

If, on the other hand, 2 is removed and 1 is kept then I_0 is given by

$$I_0 = qAn_i^2\left[\sqrt{\frac{D_p}{\tau_p}}\frac{1}{N_d}\tanh\frac{w_n}{L_p} + \sqrt{\frac{D_n}{\tau_n}}\frac{1}{N_a}\tanh\frac{w_p}{L_n}\right].$$

If both assumptions a and b are removed, then the expression for I_0 becomes rather complicated (cf. Rittner 1977).

Table 14.6 Diffusion Constants in Silicon

	$N = 10^{16}\,\text{cm}^{-3}$	$N = 10^{18}\,\text{cm}^{-3}$	
D_n	27	8	cm^2/s
D_p	12	5	cm^2/s

An examination of Equation 14.87 reveals that the larger the doping (N_d and N_a) the smaller the saturation current and, thus, the larger the efficiency. However, the doping level also influences the values of the diffusion constant and of the lifetime of the minority carriers. The diffusion constant tends to decrease with increasing doping, thus helping in the reduction of I_0. Typically, for silicon at 300 K, we have the values shown in Table 14.6.

In Table 14.6, N represents either N_a or N_d.

The lifetime of the minority carriers goes down with increasing doping and thus partially counteracts the improvement in saturation current resulting from the other effects. There is no fundamental physical relationship between lifetime and doping level.

As the technology progresses, better lifetimes are achieved with a given impurity concentration. Empirically, it has been found that

$$\tau = \frac{\tau_0}{1 + N/N_0}, \tag{14.88}$$

where N is the doping level and τ_0 and N_0 are parameters that depend on the type of carrier. For holes, for example, τ_0 is given as 400 μs and N_0 as $7 \times 10^{15}\,\text{cm}^{-3}$. These data are somewhat old and may be outdated.

If N exceeds some $10^{17}\,\text{cm}^{-3}$,

$$\tau_p N_d \approx \tau_n N_a \approx 2.8 \times 10^{12}\text{sec cm}^{-3}. \tag{14.89}$$

Introducing this empirical relationship into our simplified formula for the reverse saturation current, we have, for silicon at 300 K,

$$\frac{I_0}{A} = 10^{-5} \times \left[\sqrt{\frac{D_p}{N_d}} + \sqrt{\frac{D_n}{N_a}}\right] \text{A cm}^{-2}. \tag{14.90}$$

Using this information, we can calculate I_0/A for the case of a lightly doped diode ($N_a = N_d = 10^{16}\,\text{cm}^{-3}$) and for a heavily doped one ($N_a = N_d = 10^{18}\,\text{cm}^{-3}$). We get, respectively, $0.9\,\text{pA cm}^{-2}$ and $0.05\,\text{pA cm}^{-2}$. The larger current is just a rough estimate because our approximation is not valid for such low doping. These extremely low saturation currents lead, of course, to high open-circuit voltages: 0.637 V and 0.712 V, respectively, for silicon at 1 sun.

Experimentally observed voltages at 1 sun have reached 0.68 V and, at 500 suns, 0.800 V . This corresponds to 28% efficiency.

14.7.2 Practical Efficiency

As is invariably the case, the practical efficiency of photodiodes departs substantially from the ideal one. The reasons for this include the following.

1. *Surface reflection.* Surface reflection is caused by the abrupt change in refractive index when the rays pass from air to the diode material. A more gradual (even though stepwise) change can be achieved by coating the diodes with a transparent medium having an index of refraction somewhere between that of air and that of silicon. This is identical to the "bluing" technique used in photographic lenses.
2. *Poor utilization of the available surface.* Some photodiodes are produced from circular wafers of the type employed in the semiconductor industry. When arrayed side by side, considerable empty space is left.
3. *Thinness of the cells.* Excessively thin diodes fail to interact with all the available light. This is more a problem with silicon than with gallium arsenide owing to gallium arsenide's greater opacity. Silicon cells must be over 100 μm thick versus only a few μm for GaAs. However, gallium is quite expensive (roughly one-fourth of the price of gold), and, in addition, it is difficult to create single crystals with more than a few square centimeters area. This limits crystalline GaAs to space use (where its high resistance to radiation damage becomes important) or as photocells in concentrator systems where the cost is transferred from the cell itself to that of the concentrator. In this application, GaAs's greater insensitivity to heat is an asset.
4. *Series resistance of the cells.* The material from which the diodes are made is not a good conductor of electricity. The path between the region where the current is generated and where it is collected must be minimized. On the other hand, these collectors, as explained before, should not obscure the source of light. To achieve a compromise between these two requirements, a grid of thin silver strips is built on the diode as illustrated in Figure 14.26. These strips, as produced by current technology, come out too flat and too thin—120 micrometers wide and only 10 micrometers thick. They not only obscure too much of the collecting area but have lots of voids that increase their resistance to the flow of the collected current.

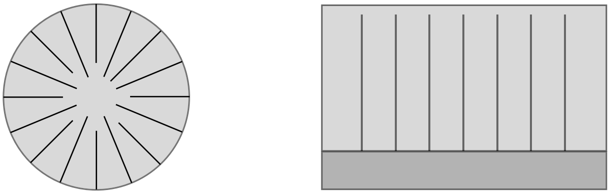


Figure 14.26 Typical current collector strips on photodiodes.

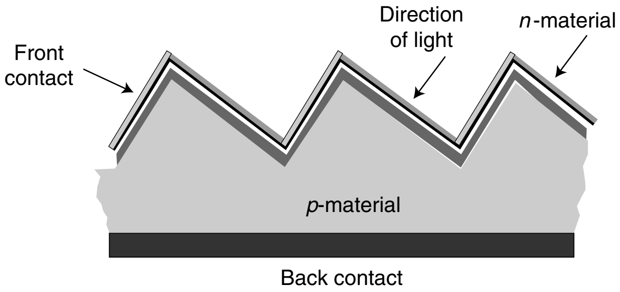


Figure 14.27 Configuration that reduces series resistance of photocells.

A company called 1366[†] Technologies has a proprietary process for depositing strips 20 by 20 micrometers in cross section. They block less light and have better conductance, in addition to saving silver.

Another ingenious solution for the problem, developed by Peter Borden, is shown in Figure 14.27. Grooves about 50 micrometers apart are created on the surface of the semiconductor. One face of these grooves is metallized and constitutes the current collector; the other face receives the illumination. Light is caused to shine slantwise on the cell in a direction perpendicular to the active face, fully illuminating it while leaving the contact face in shadow. Thus, the collectors, though large, do not interfere with the incoming light.

By choosing an appropriate crystallographic orientation, it is possible to create grooves whose cross section is not that of an isosceles triangle—that is, grooves that have the active face larger than the current collecting one. This reduces the required amount of tilt.

5. *Lifetime of the minority carriers.* Some electron-hole pairs are created too far from the potential barrier to survive the length of time it would take to reach it by diffusion. The longer the minority carrier lifetime, the larger the number of electrons and holes that can be separated by the barrier and, consequently, the larger the diode efficiency. This subject was discussed in more detail toward the end of the preceding section.

14.8 Dye-Sensitized Solar Cells (DSSCs)

In the solid-state junction photocells we have been discussing so far, the semiconductor absorbs the light, creates an electron-hole pair, and separates them from one another, making the electron available for circulation in the external circuit. Nature, through photosynthesis, performs each of

[†] Solar constant: about 1366 W/m^2.

these functions in separate regions of the leaf. This permits optimization of each individual process. Sometimes it pays to emulate Mother Nature.[†] That is what Professor Michael Grätzel of the Ecole Polytechnique de Lausanne (a Ph.D from the Technical University of Berlin) did in 1991 when he created the **Grätzel cell**.

Though still in their infancy, Grätzel cells hold great promise as possibly a very inexpensive way of harnessing sunlight. Currently, in 2008, their efficiency is modest (11%), and their life is not yet long enough. Correcting this limitation to, say, a life expectancy of some 20 years would by itself seem sufficient to ensure its widespread application. But, of course, efficiency will also be improved.

Grätzel cells do not *require* p-n junctions. Light interacts directly with a **sensitizing dye**, D, causing it to switch to an excited state:

$$D + photon \to D^*. \tag{14.91}$$

D^* is metastable and would soon return to its ground state, D, but before this happens, the excited dye injects an electron into the conduction band of a suitable semiconductor, most often, TiO_2 in its anatase form[††], on which the dye is adsorbed. A carrier is created in the semiconductor and diffuses to the top transparent electrode, becoming available to the external circuit.

By injecting an electron into a neighboring region, the dye becomes oxidized (a hole is created) and must be regenerated. This is accomplished by immersing the dye-stained TiO_2 into a redox electrolyte, almost invariably containing the iodide/triiodide couple, I^-/I_3^-, which reduces the dye and acquires a hole. The hole diffuses to the catalytic counter electrode where it combines with the electron coming back to the cell via the external circuit. A simple dye-sensitized solar cell is shown in Figure 14.28.

In its anatase form, titanium dioxide has a band-gap of 3.2 eV, being transparent to all visible radiation and to ultraviolet below 772 THz; it will absorb only 12% of the 6000 K black body spectrum.

When a thick layer of dye is applied, good light absorption is achieved, but essentially only the layer of dye molecules directly in contact with the titanium dioxide is capable of injecting an electron into the oxide. A monomolecular layer of dye intercepts little light (1%) because the cross section for light absorption of the dye is much smaller than the area of one dye molecule. A solution to this quandary is the use of extremely rough titanium dioxide surfaces. Such "fractal" surfaces have a huge actual surface as much as 1000 times bigger than the projected surface. Dye monolayers

[†] Not always. Airplanes do very well without flapping their wings like birds.

[††] In addition to TiO_2, other binary oxides have been investigated as elements of Graëtzel cells, such as In_2O_3, Nb_2O_5, SnO_2, and ZnO. Ternary oxides such as $BaSnO_3$, $SrTiO_3$, and Zn_2SnO_4 may provide the advantage of easy tuneability of their different characteristics.

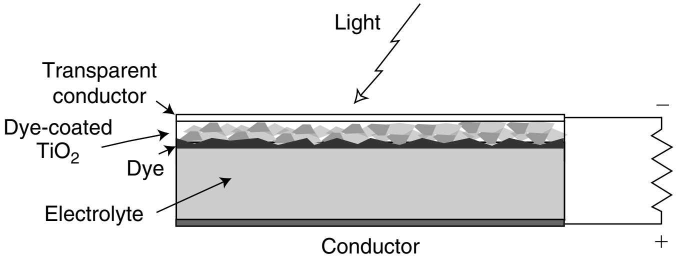


Figure 14.28 A simple dye-sensitized solar cell.

are formed, offering a multiple molecular path to the photons moving perpendicular to the projected surface.

The highly porous **mesoscopic**[†] titania layer is made of small (≈ 10 to 20 nm) grains lightly sintered together to ensure ohmic contact. To reduce resistance to diffusing electrons, the layer is made very thin (some 10 μm).

The **quantum efficiency**, η_I[††], that is, the ratio of electrons available to the external circuit to the number of photons falling on the cell, is strongly influenced by the **light harvesting efficiency**, η_{dye}, of the dye, that is, by the fraction of the photons absorbed by the dye. Although η_I and η_{dye} can be defined for a broad spectrum, both are actually a function of photon frequency and can be defined as monochromatic quantities.

Clearly, the choice of dye plays a major role in cell performance. Many Grätzel cells use ruthenium-based, dyes which, though expensive, yield excellent results as witnessed by the quantum efficiency (depicted in Figure 14.29) of cells using the ruthenium dye, N719. Data are from Grätzel.

Many dyes are being investigated, specially inexpensive porphyrins, molecules common in biology; they are part of both chlorophyll and hemoglobin. Chlorophyll and xanthophyll (a carotene, i.e., a terpene, not a porphyrin) have been tried as dyes in Grätzel cells. They potentiate each other as they do in a green leaf; they work best when intermixed than individually. Quantum dots are also being investigated as sensitizers.

A major effect that reduces the quantum efficiency is the light reflection by the glass front electrode. But there are other mechanisms. In an ordinary *p*-*n* diode, electrons and holes have to navigate through the bulk of the material before being separated by the potential cliff at the junction, and many pairs recombine before this separation occurs. To reduce this problem, crystals have to be grown as defect-free as possible, a requirement that results in expensive processes.

[†] Mesoscopic designates a size range in which neither quantum mechanics nor classical physics applies accurately. It could as well be called nanoscopic.

[††] Variously called **external quantum efficiency (EQE)** or **incident photon conversion efficiency (IPCE)**.

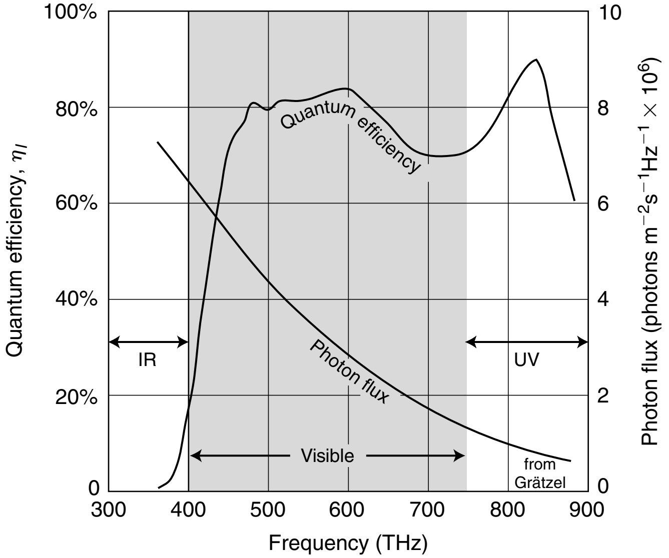


Figure 14.29 Quantum efficiency of a cell using titanium dioxide sensitized by the ruthenium dye, N719. Data from Grätzel. Photon flux based on a 6000 K black body with a power density of 1000 W/m^2.

In dye-sensitized solar cells (DSSC), holes and injected electrons are never in the same region of the material and cannot recombine directly. Nevertheless, some recombinations do occur because, when open-circuited, no carriers circulate externally, in spite of still being created by light.

Several mechanisms lead to the wasteful recombination of photoelectrons in DSSC. In the following description, we will adopt the notation of attaching a subscript to the letter "e" to indicate the region of the device in which the electron resides.

1. The recently injected electron in the titanium dioxide, e_{TiO_2}, may cross the interface back to the oxidized dye before it can be regenerated:

$$e_{TiO_2} + D^* \rightarrow D. \qquad (14.92)$$

   Fortunately, under short-circuit conditions, this reaction proceeds much more slowly than the dye regeneration reaction. As the cell voltage grows, the reaction rate rises. Thus, even though this mechanism barely affects the short-circuit current, it becomes important at near open-circuit conditions.
2. The electron in the titanium dioxide may recombine with the hole in the dye (before it is transferred to the electrolyte):

$$e_{TiO_2} + D^+ \rightarrow D. \qquad (14.93)$$

3. The electron in the titanium dioxide may react with the oxidized electrolyte, R^+,

$$e_{TiO_2} + R^+ \rightarrow R. \tag{14.94}$$

4. If the transparent front electrode contains tin dioxide, the electron in its transit through this electrode may recombine with the oxidized electrolyte that seeps through the titania layer and bathes the electrode:

$$e_{SnO_2} + R^+ \rightarrow R. \tag{14.95}$$

For a more complete discussion of these loss mechanisms (and other related topics), read Brian A Gregg, 2003.

The voltage delivered by a DSSC is the difference between the quasi-Fermi level of the illuminated titanium dioxide and the contact potential between the counterelectrode and the electrolyte. Titanium dioxide, being a n-semiconductor, has a Fermi level near the conduction band. The dye must have a Fermi level such that its excited states are at higher potential than the bottom of the dioxide conduction band so that electrons can be injected in the dioxe. Typically, the voltage of the cell is about 0.7 V, see Figure 14.30.

One can make a rough calculation of the cell's output by determining the quantum efficiency (from which the load current can be estimated) and multiplying by the assumed voltage of 0.7 V. For example, the cell of

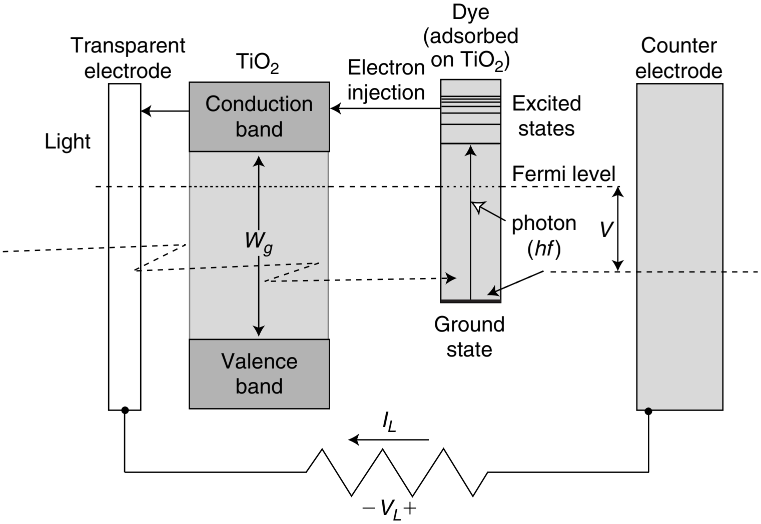


Figure 14.30 Energy levels in Grätzel cells. The external voltage is the difference between the quasi-Fermi level of the TiO_2 and the Nernst potential of the electrolyte.

Figure 14.29, when exposed to 1000 W/m^2 of 6000-K black body radiation, will receive a total photon flux (in the visible region of the spectrum) of 1.4×10^{21} photons s^{-1} m^{-2}, which, at the observed 80% quantum efficiency will result in a load current of 180 A m^{-2}, yielding an electric output of 126 Wm^{-2}. Since we assumed a light input of 1000 W m^{-2}, the estimated efficiency would be 12.6%. This is suspicioulyⁱ close to the actual measured efficiency of 11% reported by Grätzel.

At the moment, DSSCs look very promising, with a few weak spots. One of the less serious weaknesses is the use of glass instead of plastics, and the other, of greater concern is the use of liquid electrolytes that leads to proneness to leakage, to freezing in cold weather, and to causing increased internal pressures in the summer.

Conducting Electrodes

Transparent conductors find use in low emissivity, automatically dimming, or-self defrosting windows, in display panels, and as front electrodes of some solar cells. For this last-named application, a number of oxides and nitrides have been proposed, but the most popular is indium tin oxide, $\text{In}_{2-x}\text{Sn}_x\text{O}_3$, or ITO.

Because indium is more costly than silver, alternatives are being investigated. Fluorine doped zinc oxide, ZnO:F, and cadmium stannated, Cd_2SnO_4, are promising alternatives.

A transparent coat of conducting material must have a high electrical conductivity, σ, and a small light absorbtionn coefficient, α; hence, the ratio between these two quantities is a representative figure of merit of the materials performance:

$$S = \frac{\sigma}{\alpha}. \tag{14.96}$$

The manufacturer of the transparent electrode will experimentally select the thickness, t, that provides an optimal performance. This thickness depends, of course, on the material used. Given this thickness, the measurement of interest to compare performances is the sheet resistance, $\Re$:

$$R = \rho\frac{L}{A} = \rho\frac{L}{Wt} \equiv \Re\frac{L}{W}, \tag{14.97}$$

which defines $\Re \equiv \rho/t$. A is the cross-sectional area, and t is the thickness of the coating. Notice that the dimensions of $\Re$ are the same as of

(*Continues*)

[†] Suspiciously because our estimate is extremely rough!

(*Continued*)

Table 14.7 Properties of Some Transparent Conductors

Material	Sheet resistance $\Re$ (Ω)	Absorption coefficient α	Figure of merit S (Ω^{-1})
ZnO:F	5	0.03	7
Cd_2SnO_4	7.2	0.02	7
In_2O_3:Sn	6	0.04	4

R; that is, $\Re$ is measured in ohms. Table 14.7, taken from the article by Gordon (2000), lists some pertinent characteristics of a few coatings.

A good overall discussion of transparent conductors can be found in the Gordon (2000).

14.9 Organic Photovoltaic Cells (OPC)

In 2008, the jury was still out regarding the future of organic photodiodes. The promise of extremely inexpensive collectors, albeit of modest efficiency, was pitted against the disadvantages of a short lifetime owing to UV damage.

When one thinks of organic polymers or plastics, one thinks of nonconductors. This is correct in the vast majority of cases and only in 1963 was the existence of highly conducting plastics demonstrated by a number of Australian scientists (see articles by R.McNeill et al, B. A. Bolto and D. E. Weiss, and B. A. Bolto et al.)

In 2000, the Nobel prize in Chemistry was awarded to Alan J.Heeger, Alan G. MacDiarmid, and Hideki Shirakawa, “for the discovery and development of conducting polymers”, even though they were not the authors of the Australian papers mentioned above.

Conducting polymers opened up a large field, which only now is beginning to be industrialized. The production of inexpensive photovoltaic sheets is one possibility. Most conducting polymers act as semiconductors, having an occupied and an unoccupied allowed band separated by a forbidden band. This corresponds to the valence, conducting, and forbidden bands of inorganic semiconductors. Polymer band-gaps range from 0.5 eV to as high as 2 eV, covering the near infrared to near ultraviolet parts of the spectrum.

What is the mechanism of electrical conduction in solids, and what gives rise to the band structure mentioned above?

A gas consisting of nonionized atoms does not conduct electricity because the electrons are confined to each atom and little interaction with neighboring electrons takes places. Conduction requires that some electrons be **delocalized**[†], that is, be able to move relatively freely through the bulk of the material. A solid with no delocalized carriers cannot conduct electricity.

Consider the molecular bonds that tie carbon atoms to other atoms in a molecule. There are two major types of *molecular* orbitals (Chapter 13) :

1. One, called an **sp^3 orbital** occurs in methane, CH$_4$. Four **σ-bonds**[††] connect the carbon to each of the hydrogens. All four bonds are identical, meaning that the bonding orbitals stick out in three dimensions, forming a 109.5° angle between themselves in a tetrahedral arrangement

   An isolated carbon atom has six electrons in the electronic configuration

   $$\underset{1s}{\underline{\uparrow\downarrow}}\ \underset{2s}{\underline{\uparrow\downarrow}}\ \underset{2p}{\underline{\ \uparrow\ }}\ \underset{2p}{\underline{\ \uparrow\ }}\ \underset{2p}{\underline{\quad}}$$

   in which the vertical arrows represent up or down spins of the electron.

   If the four valence electrons (two in the 2s subshell and one each in two of the 2p subshells) are to make *four identical molecular orbitals*, the electrons (in the carbon *molecule*) must rearrange themselves as

   $$\underset{1s}{\underline{\uparrow\downarrow}}\ \underset{sp^3}{\underline{\ \uparrow\ }}\ \underset{sp^3}{\underline{\ \uparrow\ }}\ \underset{sp^3}{\underline{\ \uparrow\ }}\ \underset{sp^3}{\underline{\ \uparrow\ }}.$$

   This creates four sp^3 orbitals. In the case of methane, all carbon valence electrons are engaged in identical σ-bonds and, being directly between atoms, are well protected from external fields and, thus, firmly localized.

2. In double-bonded molecules (ethene, C_2H_4, for example), the C=C bond consists of one σ- and one π-bond. Each carbon has two hydrogens attached to it via a σ-bond. Thus, each carbon forms three σ-bonds and one π-bond. The 3 σ-bonds, which lye on a flat plane, are spaced 120° apart and are formed by the fusion of the s-orbital of the two hydrogens and of the remaining carbon with the **sp^2**

[†] The expression, "delocalized" is being used somewhat loosely: electrons can be delocalized in a covalent bond, for instance, without becoming carriers.

[††] In a σ-bond, the bonding orbital resides directly between the bound atoms.

orbitals in an arrangement such as

$$\underset{1s}{\underline{\uparrow\downarrow}}\;\underset{sp^2}{\underline{\;\uparrow\;}}\;\underset{sp^2}{\underline{\;\uparrow\;}}\;\underset{sp^2}{\underline{\;\uparrow\;}}\;\underset{p}{\underline{\uparrow}}.$$

In this case, in addition to the σ-bonds, we also have a π-bond with electrons, relatively unprotected, above and below the plane of the molecule. If a number of such molecules are attached to one another, the electrons in the π-bonds may fuse together and become delocalized. This may permit the assembly to conduct electricity. Diamond in which bonds are formed from sp^3 orbitals is an insulator, while graphite, in which bonds are made of sp^2 orbitals, is a conductor. In a similar manner organic polymers, such as polyethylene, in which the atoms are held together by sp^3-derived bonds, is an insulator, while **conjugated polymers**[†] such as polyacetylene, which have sp^2-derived bonds, may be conductors. Indeed, iodine doped polyacetylene is a better conductor than mercury.

If we are interested in semiconducting polymers, we must inquire into what mechanism causes the appearance of forbidden energy bands.

14.9.0.1 Band Structure in Inorganic Semiconductors

One of the earliest successes of quantum mechanics was the explanation of the discrete frequencies radiated by an excited atom. Spectroscopists had observed that, unlike solid bodies, isolated atoms, when heated to a high temperature, do not radiate a continuum of frequencies. Rather, only a series of well-defined frequencies or spectral lines are generated. Take hydrogen which, when heated, or better, when electrically excited, emits in the ultraviolet at 2457.8, 2926.3, 3086.4, 3292.1,... THz (Lyman series), in the visible at 457.1, 617.2, 691.1, 822.8,... THz (Balmer series), and at infrared 160.0, 234.1, 274.3, 365.8,.... THz (Paschen series).

Classical mechanics gets into trouble when it tries to model the hydrogen atom as a miniature planetary system: an electron orbiting a proton. The reason is that the orbit of an electron would be a circle or an ellipse; it would not be a straight line. To keep it curving, the electron must be accelerated inward (as the Earth is accelerated inward toward the sun). But the electron has a net charge, and when a charge is accelerated, it will radiate energy; that is, it will lose energy. Hence, classical mechanics predicts a spiraling electron—a rapidly decaying orbit. Here is where an ad hoc assumption was made by Niels Bohr. He assumed, arbitrarily, that the angular momentum of the electron can only have certain discrete values, integer multiples of $h/2\pi$,

$$mr^2\omega = n\frac{h}{2\pi}, \tag{14.98}$$

where m is the mass of the electron and n is any positive integer.

[†] Conjugated polymers have alternating single and double bonds as in the benzene ring.

To be in a circular orbit, the centrifugal force must exactly balance the electrostatic attraction,

$$mr\omega^2 = \frac{Zq^2}{4\pi\epsilon_0 r^2}, \tag{14.99}$$

where Z is the charge (the number of protons) of the nucleus, and ϵ_0 is the permittivity of free space.

Let us solve both equations above for ω^2. From Equation 14.98,

$$\omega^2 = \frac{n^2h^2}{4\pi^2r^4m^2}, \tag{14.100}$$

and from Equation 14.99,

$$\omega^2 = \frac{Zq^2}{4\pi\epsilon_0 r^3 m}. \tag{14.101}$$

From Equations 14.100 and Equation 14.101, solving for r,

$$r = \frac{\epsilon_0 n^2 h^2}{\pi Z q^2 m} = 5.3 \times 10^{-11}\frac{n^2}{Z} \quad \text{meters}. \tag{14.102}$$

The total energy of the electron in its orbit is the sum of its kinetic and potential energies,

$$W = \frac{mv^2}{2} + \int_{\infty}^{r} \frac{Zq^2}{4\pi\epsilon_0 r^2} dr. \tag{14.103}$$

Energy has no absolute value; we can only calculate the change of energy between two states. Assume the energy of the electron infinitely far from the attracting center is zero (it is actually a maximum; thus, at smaller r, the total energy of the electron is smaller than zero, that is, a negative number,

$$W = \frac{m\omega^2 r^2}{2} - \frac{Zq^2}{4\pi\epsilon_0 r} = -\frac{q^4 m}{8\epsilon_0^2 h^2}\frac{Z^2}{n^2} \tag{14.104}$$

$$W = -2.19 \times 10^{-18}\frac{Z^2}{n^2} \tag{14.105}$$

and, for hydrogen ($Z = 1$),

$$W = -2.19 \times 10^{-18}\frac{1}{n^2}. \tag{14.106}$$

When an excited atom, say, $n = 2$, decays to the ground state ($n = 1$), the atom loses an amount of energy,

$$\Delta W = W_2 - W_1 = -2.19 \times 10^{-18}\left(\frac{1}{2^2} - \frac{1}{1^2}\right) = 1.64 \times 10^{-18} \quad \text{joules.} \tag{14.107}$$

This lost energy appears as a photon of frequency $f = 1.64 \times 10^{-18}/h = 2480$ Thz. This is the first line of the Lyman series (Lyman-α).

Using Equation 14.107, we can calculate the frequencies of all spectral lines. The Lyman series results from an excited atom returning to its ground state ($n = 1$); the Balmer series results from its return to $n = 2$; and the Paschen series results from a return to $n = 3$.

The above derivation works well with hydrogen but is a very primitive approach to the determination of the discrete energy levels of the electron in an atom. It fails with more complicated atoms, which require quantum mechanics to solve for the probability of finding an electron at a given position relative to the nucleus. The map of such a probability is called the **orbital**, while the classic approach we used describes the **orbit** of the electron from which we calculate the energy levels shown in Figure 14.31.

Thus, in isolated atoms (or in a gas), the electrons can only be in discrete orbitals (see Chapter 13, subsection on heterocycles); that is, electrons can only have certain discrete energy levels. The question to answer is what happens to these levels when the gas condenses into a solid crystal in which atoms do not act in isolation—they interact with one another determining the crystallization form. It turns out that in a crystals, the

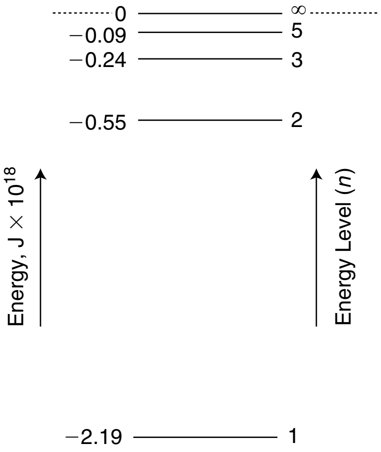


Figure 14.31 Some of the allowed energy levels of the electron in a hydrogen atom.

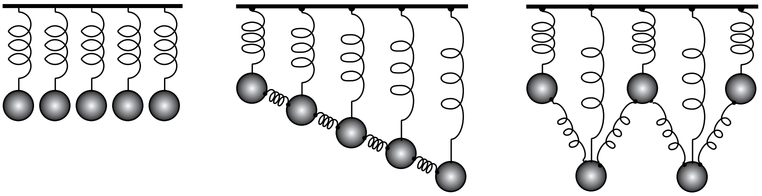


Figure 14.32 Coupled mechanical oscillators have a frequency that depends on the mode of excitation.

electrons are confined to *bands* of energy in which they can have a number of continuously varying energy levels. To understand the reason for the appearance of such energy bands in solids, consider a number of identical mechanical oscillators consisting of a mass or a bob attached to a spring and constrained to only vertical motion, as depicted in Figure 14.32. Assume that the oscillators are lossless and, since they are identical, they will all oscillate with the same frequency. If the system is excited by lifting all bobs simultaneously and then releasing them, they will all move up and down in phase.

Assume now that a weak coupling spring interconnects the bobs. As long as the system is excited in phase, the coupling spring does not change its length and, therefore, plays no role. The frequency is still exactly that of an isolated oscillator. If they are excited in any other manner, the coupling spring will change its length during oscillation introducing additional restoring forces and, thus, altering the frequency. The degree of frequency alteration depends on the manner of excitation. The middle illustration of Figure 14.32 causes only a minor frequency increase, while that on the right-hand side of the figure, in which the bobs are in phase opposition, causes maximum frequency change. It is easy to understand that the system has as many different frequencies as there are oscillators. The number of atoms in a cubic centimeter of silicon is more that 10^{23}. No wonder the bands can be treated as a continuum of frequencies (or of energy). Let us return to the energy levels of the electrons in a given atom. If we have a number of isolated atoms, each one is, of course, allowed the same energy levels as any other. As the material condenses, the atoms approach one another and interact: they begin to be coupled to each other. The more densely packed, the higher the degree of coupling and the larger the energy spread, that is, the wider the allowed energy band. This phenomenon is schematically represented in Figure 14.33, where two energy levels are plotted as a function of the lattice constant (notice that the lattice constant—a measure of the interatomic spacing—increases toward left). As this constant becomes smaller, the interaction becomes more marked, resulting in wider energy bands. It also turns out that the mean band energy becomes progressively smaller up to a point and then increases. The crystal is formed with a lattice constant that corresponds to minimum energy.

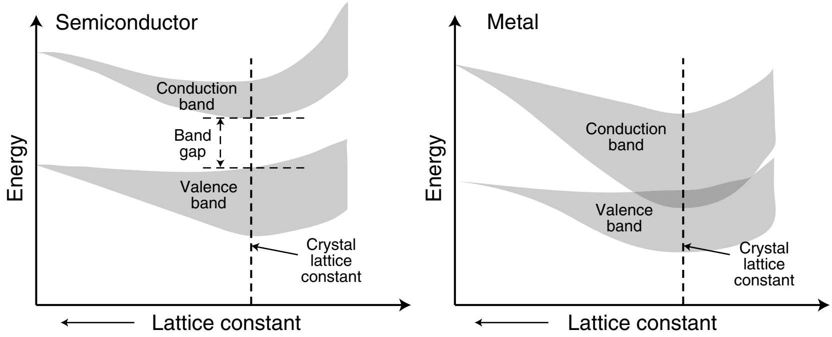


Figure 14.33 As the lattice constant of a crystal is reduced, single-energy levels spread out into bands. In semiconductors there is a gap between the bands; in metals there is an overlap.

Consider two distinct energy levels in a free atom: one is the level (actually, a collection of closely spaced levels) occupied by the valence electrons, in the unexcited atom, and the other is the next higher allowable level, normally unoccupied. When the material coalesces into a solid, each level is transformed into an energy band. Two cases can be distinguished: in one, the higher band (conduction band) overlaps the lower (valence band) (right-band side of Figure 14.33) and we have a metal. In the other, no such overlap occurs: there is a gap between the top of the valence band and the bottom of the conduction band (left-hand side of the figure), and we have a semiconductor, or, if the band-gap is very large, an insulator.

There is an equivalence between the **highest occupied molecular orbital, (HOMO)** of organic chemistry, and the valence band in inorganic materials, and the **lowest unoccupied molecular orbital, (LUMO)** and the conduction band. Polymer cells, discussed in the next subsection, usually utilize heterojunctions. When a *p*-*n* junction is created, say, in silicon, both sides are made of the same material, albeit with different doping—we have a **homojunction**. When the material of the two sides of a junction is different, we have a **heterojunction**. Wikipedia has a good article on heterojunctions.

14.9.0.2 Polymer Solar Cells (PSCs)†

There are crucial differences in the mechanisms of inorganic and organic solar cells. See Figure 14.34. Both absorb photons creating an **exciton**—a bound electron-hole pair. Think of an exciton as an electron orbiting its companion hole, somewhat like an electron in a hydrogen atom. The ionization potential of hydrogen is rather high (13.6 eV), whereas the energy necessary to dissociate an exciton is much smaller and depends on the

†Extensive and very clear discussions of polymer solar cells can be found in the dissertations by Klaus Petritsch (2000) and Lambert Koster (2007).

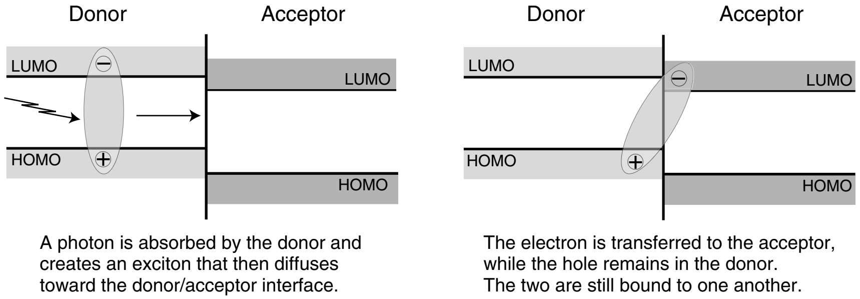

Donor
Acceptor
LUMO
LUMO
HOMO
HOMO
A photon is absorbed by the donor and creates an exciton that then diffuses toward the donor/acceptor interface.
Donor
Acceptor
LUMO
LUMO
HOMO
HOMO
The electron is transferred to the acceptor, while the hole remains in the donor. The two are still bound to one another.

Figure 14.34 An early polymer cell use a heterojunction (a p- and a n- conducting plastic, juxtaposed).

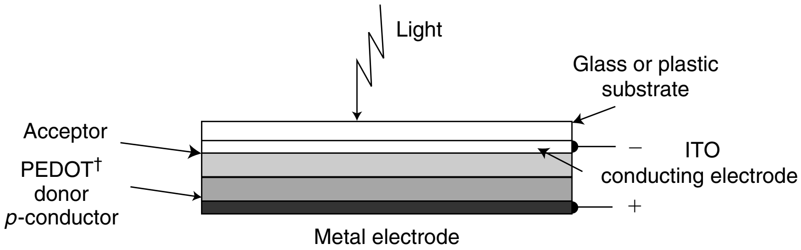


Figure 14.35 The basic mechanism of polymer solar cells. After the creation of the bound electron-hole pair, the two carriers separate and each migrates to the proper electrode. After Koster (2007).

screening effect of the dielectric constant of the medium. The high dielectric constant of inorganic semiconductors (silicon has dielectric constant of about 12) causes the binding energy of an exciton to be about 0.1 eV, while the low constant of plastics (say, 2) results in binding energies of the order of 1 eV making dissociation difficult.

One can get an idea of how a polymer solar cell *should* work by examining Figure 14.35. The cell consists of a transparent substrate (glass or plastic) upon which a thin transparent conducting layer (usualy ITO) is deposited. Next comes a layer of the p-type semiconducting plastic known by its acronym, PEDOT,† which has a band-gap in the useful range and is supposed to absorb photons and create bound electron-hole pairs called **excitons**. The exitons migrate, transporting energy but not

†PEDOT is currently the most widely employed conducting polymer. It is a p-type semiconductor, used, for instance, in plastic LEDs and as a transparent anti-static coating. PEDOT stands for Poly(3,4-ethylenedioxythiophene). Unfortunately, owing to its insolubility, it is difficult to apply in thin layers. This problem is circumvented by copolymerizing it with poly(4-styrenesulphonate), PSS. The combination has good film-forming properties and lends itself to spin coating, screen printing, ink jetting, and so on. PEDOT:PSS is chemically stable, quite transparent to visible radiation (it has a low band-gap of about 1.8 eV), and is a good electrical conductor.

charge (because they have no net charge) to the heterojunction (between the PEDOT and an n-type semiconductor) where the built-in electric field will dissociate them, the electron being injected into the n-material and eventually making its way to the reflective[†] metal electrode, while the hole remains in the PEDOT and moves to the ITO electrode.

When such a device is tested, it generates electricity at a very disappointing level of efficiency for the following reasons:

1. The diffusion length of the excitons is only some 10 nm—excitons created much farther than this from the donor/acceptor interface will recombine producing heat instead of delivering charges to the output. Consequently, to prevent useless recombinations, the active thickness of the donor layer must be very thin, comparable with the exciton diffusion length.
2. Because the photoactive layer must be thin, it may be too transparent, intercepting only a fraction of the incoming photon flux, in spite of the much higher absorption coefficient of plastics compared to silicon.
3. The exciton must be dissociated so that the individual charges are separated and head to their proper electrodes. Such separation occurs under a number of circumstances, one of which is at the interface between donors and acceptors. For efficient charge separation, the materials must have sufficiently different electron affinities and ionization potentials. See the accompanying box, "Cations and Anions." Some interfaces allow the exciton to jump into the material with the lower band-gap without being taken apart and thus making no contribution to the output current.
4. Once separated, the charges must be transported to their electrodes. Since they have finite mobilities, this transport dissipates energy and appears as a series resistance of the cell. Also, charges may recombine while being transported and will thus be lost to the output current.
5. Upon arrival at their terminals, charges must overcome barriers before they can be transferred to the electrodes.
6. Part of the light shining on the cell is reflected back, failing to reach the photoactive layer. As soon as other problems with polymer cells become less severe, attention must be paid to reducing surface reflections in these cells.
7. Current cells do not use the optimum band-gap for sunlight. This is another area in which eventual improvements will take place.

[†] Many animals, mainly predators, improve their night vision by having a reflective layer, **tapetum lucidum** (frequently) behind the retina. The tapetum reflects back to the retina light that failed to be absorbed in the first pass. The reflective metal collector in certain solar cells has a similar function.

Cations and Anions

There are two ways of transforming a neutral atom into an ion:

1. An electron (or more than one) can be *removed* from the atom, leaving it with a positive charge—**a cation**. The energy to remove the electron is called the **ionization potential**. Sodium atoms require little energy to become ionized (5.139 eV), although francium holds the record for least ionization energy (3.83 eV), while the record for highest ionization potential goes to helium with 24.587 eV.
2. An ion can be created by *attaching* an electron to a neutral atom, resulting in a negative ion or **anion**. The energy absorbed in this process is the **electron affinity**. Chlorine absorbs −3.62 eV when it attaches an electron; that is, this process releases 3.62 eV—the resulting chlorine ion is more stable than the atom itself. Some elements such as alkali metals and noble gases do not form stable negative ions.

A major improvement in efficiency can be obtained by abandoning the planar interface between the donor and the acceptor and resorting to **bulk heterojunctions** in which donor and acceptors are blended together forming a layer with intermixed regions or each material. Each region is small enough (a few nm) so that excitons created in the donor are never too far from the donor/acceptor interface. This provides a very large interface area but presents technical difficulties in obtaining arrangements with enough contact between regions of a like composition so that the carriers can percolate effortlessly to the correct electrode.

Currently, the best polymer solar cells have demonstrated nearly 6% efficiency with a promise of reasonable improvement in the future. The other major problem with these cells is that of achieving a sufficient life span.

14.10 Solar-Power Satellite

Science, science fiction, and magic are three areas of human endeavor separated by fuzzy and changing boundaries. Magic today may be science tomorrow. Arthur C. Clarke expressed this idea in his third law: “Any sufficiently advanced technology is indistinguishable from magic.”

(*Continues*)

(*Continued*)

Serious research, advanced enough to resemble science fiction, if not magic, may occasionally lead to practical results: landing on the moon. Often, it leads to a dead end, sometimes because of impossible technological obstacles but, more frequently because of impossible economics. Nevertheless, valuable knowledge can be generated.

NASA's Solar Power Satellite (SPS) project was a study that ran into insurmountable financial barriers, even though it could have led to the generation of low-cost electricity. Its dependence on the economy of scales was such that to make it economical, an investment of trillions of dollars would be required.

Critics pointed out valid difficulties in the project. It was shown that the energy needed to fabricate the photocells exceeded the total energy the system would provide over its lifetime. This was true enough if the photocells were produced by the Czochralski method, then universally used by the semiconductor industry. Yet, modern polymers may one day yield photovoltaic systems that are much lighter (and easier to deploy) than the single-crystal system envisioned and at attractive energy payback ratios.

It also looked impossible to launch into orbit the huge masses required. This was again true using current technology. But another science fiction idea could come to the rescue: the space elevator—the dangling of a rope from a geostationary satellite, its other end anchored to the surface of Earth.

"Climbers" would crawl up and down this rope delivering their cargo to outer space at a cost much lower than by using rockets. What made this proposal a piece of science fiction is that it is easy to show that even the strongest steel cannot sustain its own weight from geostationary orbit. But then, at the time the SPS was being investigated, no one had ever hear of carbon nanotubes (discovered in 1991 by Sumio Iijima). It now seems that ribbons made of a composite of polymers and nanotubes can do much more than hold their own weight.

All told, the SPS study did not lead to any major practical application, but one can learn from it. We will attempt to do so in this section.

A drawback of photovoltaic power generation is the intermittent and unreliable nature of insolation. To compensate, one needs either large storage systems or a source of backup power. Both solutions substantially increase the required capital investment. A further disadvantage is that load centers tend to be concentrated in regions where good insolation is unavailable. This translates into the need of long transmission lines.

Transmission lines can be made quite short if the power is beamed by means of microwaves directly to the neighborhood of consuming centers. Solar collectors in Arizona could feed New York or Chicago through microwaves generated on the ground and reflected by satellites. Power management would be simplified by simply switching the beam from one user to another.

To get around the unreliability of insolation, the photodiodes must be placed in space, in a geostationary orbit. It is then possible to achieve almost constant exposure to the sun.[†]

Originally proposed by Peter Glazer of the Arthur D. Little Corporation, the SPS concept was investigated by NASA but met considerable opposition.

The proposed configuration would use satellites, each one capable of delivering 5 GW to the power grid on Earth.

The development of the SPS involved four major elements:

1. Energy conversion in space
2. Energy transmission to Earth
3. Space transportation
4. Space construction

14.10.1 Beam from Space

Power generated in geostationary orbit must somehow be sent to the consumer on Earth. One of the surprising results brought out by the SPS study is that a microwave beam can transmit energy more efficiently than a physical transmission line of comparable length. In fact, the calculated efficiency from the input of the transmitting antenna on the satellite to the receiving antenna output on ground is 74%, more than double the efficiency realizable with a metallic transmission line. In addition, the power a beam can carry is much larger than the carrying capacity of the largest existing transmission line (3.25 GW transmission lines between Itaipu and São Paulo, Brazil).

Numerous constraints exist in designing such a microwave beam. The frequency must be such as to minimize ionospheric and atmospheric absorption, and it must be within one of the bands allocated to industrial heating. This led to the choice of the "microwave oven" frequency of 2.45 GHz.

There is also a constraint regarding the maximum power density of the beam as it transverses the ionosphere. A maximum level of 230 W/m^2 was deemed acceptable. Ionospheric nonlinearities would presumably cause power densities much above this limit to generate harmonics of the 2.45 GHz signal capable of interfering with other radio services.

[†] A short eclipse would still occur near local midnight during the equinoxes. The duty cycle of a solar-power satellite (in geostationary orbit) is about 99%.

Concerns with interference also prompted the requirement that the beam power level be quite low outside the designated receiving area. Because it is physically impossible to abruptly truncate an electromagnetic beam, the one coming down from the satellite was to have a gaussian radial distribution.

Another major concern had to do with the proper aiming of the beam. An ingenious solution was found for this problem. See the subsection on the radiation system.

14.10.2 Solar Energy to DC Conversion

A thermomechanical-electric conversion system using Rankine engines was considered but was abandoned in favor of photovoltaics. Two different photovoltaic materials were studied—silicon and gallium aluminum arsenide—leading to two different structures, as shown in Figure 14.36.

The total surface area of both configurations was nearly the same (5.2 km by 10.4 km for the silicon version and 5.25 km by 10.5 km for the gallium arsenide one), a total of some 55 km^2. Of this, 96% is a sun-collecting area so that, at 1360 W/m^2, a total of 72 GW would be collected under the most favorable conditions.

The silicon cells were to be exposed to unconcentrated sunlight, while the gallium arsenide cells would work with a concentration ratio of 2, achieved by means of a simple through-type reflector. This is possible because the gallium arsenide cells can be operated efficiently at a higher temperature (125 C) than the silicon ones (36 C). At these temperatures cell efficiency would be 18.2% for the GaAs and 16.5% for the Si. Owing to a 8.5% reflection loss in the concentrator, both cases would deliver about the same 10.5 GW of dc power.

A secular degradation of the SPS was expected owing, in part, to the progressively decreasing efficiency of the cells in space. The decay can be

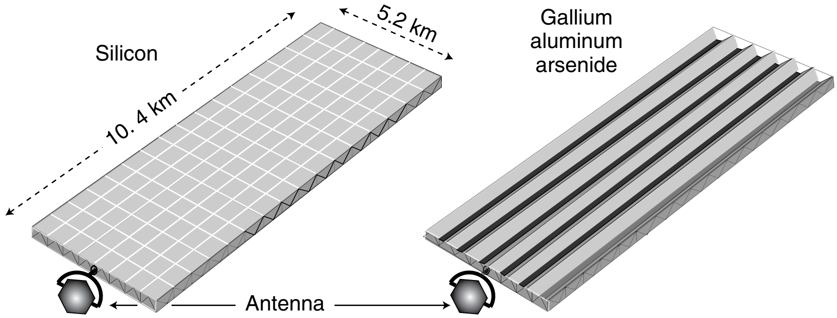


Figure 14.36 Two proposed configurations for the 5-GW power satellite. The one on the left uses silicon cells, and the one on the right, gallium aluminum arsenide.

reduced by *in situ* annealing of the silicon cells carried out periodically by means of lasers mounted on a roving structure. GaAs cells, operating at a higher temperature, would be self-annealing. Decay caused by micrometeorite damage was expected to amount to only 1% in 30 years.

Of the 10.5 GW, only 8.2 GW would reach the microwave generators, the rest being lost in the feeding conductors. These losses may seem excessive. Consider, however, that there is a limit to the dc voltage transmitted. The microwave generators can use, at most, 40,000 volts. Also, a much higher voltage would cause a breakdown in the (very) tenuous atmosphere at geostationary heights; 10.5 GW at 40 kV correspond to 262,000 amperes! It does not take a large resistance to cause huge I^2R loses.

14.10.3 Microwave Generation

Several microwave generators were considered, including

1. magnetrons,
2. transistors,
3. amplitrons, and
4. klystrons.

Magnetrons would have difficulty in delivering the required spectral purity necessary to avoid interference with other services. Remember that, even if spurious radiation is kept 60 dB below the carriers, this amounts to nearly 6 kW of undesired signal.

For use in the SPS, transistors needed considerable more development. Given enough time, it might be possible to create appropriate transistors.

Amplitrons would have the following advantages:

1. Low mass: 0.4 kg/kW compared with 0.7 kg/kW for klystrons.
2. High efficiency: 88% (possibly, 90%) compared with 85% for klystrons.
3. Almost infinite cathode life: amplitrons operate by secondary emission and could use platinum cathodes.

The advantages of klystrons are as follows.

1. Per tube, klystrons have an order of magnitude more output than a single amplitron.
2. Klystrons use high anode voltage (40 kV versus 20 kV for amplitrons). This reduces the currents and consequently the mass of the dc power distribution system, which represents a substantial portion of the mass of the satellite;
3. Klystrons have much higher gains than amplitrons. The resulting lower rf excitation power facilitates phase control essential for forming and aiming the beam.

NASA was inclined to use high-efficiency klystrons (85%) delivering 7 GW to the transmitting antenna. Since the power output of each klystron is 50 kW, a total of 140,000 tubes would be needed. Assuming a mean lifetime between failure of 5000 hours, one would expect a failure every 2 minutes! This points out the necessity of using tubes with extremely long life and of having automated procedures for diagnosing and repairing defects.

14.10.4 Radiation System

The microwave power is transferred to the antenna (the small hexagonal structure in Figure 14.36) through a rotary joint. The rotary joint is necessary to keep the antenna aimed at Earth, while the collector points toward the sun. Thermal considerations limit the power density of the transmitting array to 22 kW/m^2. Based on this figure, an area of 300,000 m^2 of transmitting antenna would be required—a circle with 600-m diameter. But the power density across the antenna aperture cannot be constant because this would generate a beam whose shape does not satisfy the ground safety requirements. Some regions of the array must have larger power densities than others (still observing the maximum of 22 kW/m^2). This results in a transmitting antenna 1 km in diameter.

The antenna was a planar-phased array consisting of a large number of radiators, grouped in subarrays, each of which must be exactly at the same distance from the ground target. Since the transmitted wavelength is 12 cm (2.45 GHz), the mechanical alignment of the individual radiators would have to be correct within less than 1 centimeter over the whole 80 hectares of antenna in order to form a beam properly directed to the ground station. This is impossible to achieve. Of course, it is not the geometric distance that counts, it is the electric distance. Hence, the unavoidable mechanical misalignment can be compensated by changing the phase of individual subarrays. To this end, a transmitter on the ground sends up a "pilot" radio signal at a frequency slightly different from the one beamed down. If all radiators, now acting as receivers, were in their correct position, the phase of the received pilot signal would be the same from subarray to subarray. Any mispositioning will appear as a phase difference (relative to a reference phase received at the center of the antenna). The phase of the transmitted signal is altered by exactly the negative of the received phase error. This ensures that the phase of all radiators in the plane of the antenna now has the correct value to focus on the spot on the ground from where the pilot signal came. An accidental loss of phase control defocuses the beam, spreading it out harmlessly over a wide area.

As mentioned before, the surprisingly good efficiency of 74% can be achieved in the transmission link. Thus, 5.15 GW would be available at the output of the ground antenna, and, of these, 5 GW would be delivered to the grid.

The beam width of the satellite antenna was such that, if the illuminated area on the ground were at the equator, it would have the shape of a circle 10 km in diameter.† This is equivalent to an average power density of a little over 80 W/m^2. However, the beam is not uniform: it has a gaussian shape with a peak power density of 230 W/m^2. This shape specification is important to keep the radiation level outside the collecting area on the ground low enough to avoid health hazards and interference.

14.10.5 Receiving Array

The ground receiving system consisted of a number of rectennas, that is, antennas equipped with their individual rectifiers. This scheme avoids the necessity of adjusting the phase of the antenna current so that their output can be added up.

The proposed antennas were half-wave dipoles. The effective area of such dipoles is

$$A = 1.64\frac{\lambda^2}{4\pi,} \qquad (14.108)$$

where the factor, 1.64, is the gain of a dipole relative to an isotropic radiator. Since the wavelength is 12.2 cm, each antenna sees an area of 20 cm^2. The total number of antennas in the 10-km diameter circle is 50 billion! To build this many antennas in one year, 1500 antennas have to be built per second, a clearly challenging task.

The antenna problem is complicated by the fact that the rectifier attached to it is necessarily a nonlinear device. Strong harmonic generation will take place, and this must be kept from being reradiated. Each antenna must be equipped with an appropriate filter. The antenna–filter rectifier combination must cost a small fraction of a dollar; otherwise the cost of the ground collecting system would be prohibitive.

The dc output from all antennas was to be added up and fed to inverters for conversion to 60 Hz ac and distribution to customers.

14.10.6 Attitude and Orbital Control

The many factors that perturb both the orbit and the attitude of the SPS include the following.

1. Solar and lunar gravitational pull.
2. Lack of spherical symmetry in the geogravitational field.

†At 40° latitude, the footprint is a 10- by 14-km ellipse.

3. Solar radiation pressure.
4. Microwave recoil.
5. Rotary junction friction.
6. Aerodynamic drag.
7. Interaction with the geomagnetic field.
8. Gravity gradient torques.

Some of the orbital and attitudinal corrections were to be made by argon ion rockets. As much as 50,000 kg of argon would be needed annually for this purpose.

14.10.7 Space Transportation and Space Construction

Both size and mass of the solar power satellites present a major challenge to the astronautical engineer. Clearly, the satellites would have to be assembled in space. It is estimated that each would require 850 man-years of space labor, even using automatic assembly machines. Large crews living long months in orbit would be needed.

The transport vehicle for the materials would have to be one order of magnitude larger than the space shuttle. For each satellite, some 400 launches would be needed. If the construction rate were one satellite per year, this would translate into more than one launch per day. To be economical, some 50 satellites must be placed in orbit; the one per year rate would require a sustained effort for half a century.

14.10.8 Future of Space Solar Power Projects

Early enthusiasm for space-based solar power projects was dampened when more realistic cost estimates were made. The DOE estimated the cost of research and construction of one demonstration satellite with its ground-based rectenna at over \$100 billion. Each additional unit would cost 11.5 billion. In 1981, the National Research Council pushed the cost estimate to some \$3 trillion and the time of completion to 50 years. Many groups opposed the whole idea. The ground solar people did not want such a grand project sucking in most of the funds for solar energy research. Fusion proponents also hated the idea of a major competitor. To counter this, SPS advocates maintain that fusion power is the power of the future and always will be.

The cost of launching the components into geosynchronous orbit and of assembling a solar power satellite in space are staggering. This has led some to propose the construction of the satellite on the moon using lunar-mined materials. A launch from the moon requires substantially less energy than from Earth. Alternatively, the solar collectors and microwave generators could be permanently based on the moon if an economical way of beaming the energy to Earth could be found.

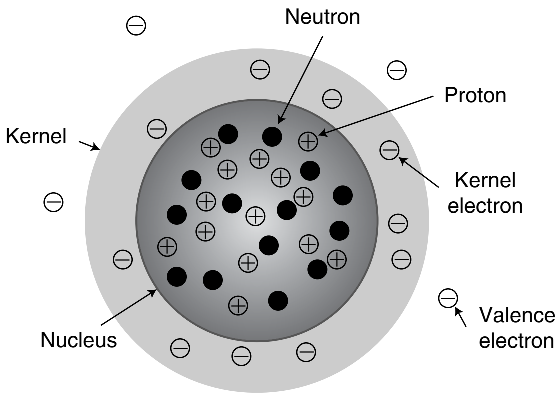
Appendix A: Values of Two Definite Integrals Used in the Calculation of Photodiode Performance.

X	$\int_X^\infty \frac{x^2}{e^x-1}dx$	$\int_X^\infty \frac{x^3}{e^x-1}dx$	X	$\int_X^\infty \frac{x^2}{e^x-1}dx$	$\int_X^\infty \frac{x^3}{e^x-1}dx$
0.0	2.4041	6.4935			
0.1	2.3993	6.4932	2.6	1.0656	4.5094
0.2	2.3855	6.4911	2.7	1.0122	4.3679
0.3	2.3636	6.4855	2.8	0.9605	4.2259
0.4	2.3344	6.4753	2.9	0.9106	4.0838
0.5	2.2988	6.4593	3.0	0.8626	3.9420
0.6	2.2576	6.4366	3.1	0.8163	3.8010
0.7	2.2115	6.4066	3.2	0.7719	3.6611
0.8	2.1612	6.3689	3.3	0.7293	3.5226
0.9	2.1073	6.3230	3.4	0.6884	3.3859
1.0	2.0504	6.2690	3.5	0.6494	3.2513
1.1	1.9911	6.2067	3.6	0.6121	3.1189
1.2	1.9299	6.1363	3.7	0.5766	2.9892
1.3	1.8672	6.0579	3.8	0.5427	2.8622
1.4	1.8034	5.9719	3.9	0.5105	2.7381
1.5	1.7390	5.8785	4.0	0.4798	2.6171
1.6	1.6743	5.7782	4.1	0.4507	2.4993
1.7	1.6096	5.6715	4.2	0.4231	2.3848
1.8	1.5452	5.5588	4.3	0.3970	2.2737
1.9	1.4813	5.4408	4.4	0.3722	2.1660
2.0	1.4182	5.3178	4.5	0.3488	2.0619
2.1	1.3561	5.1906	4.6	0.3267	1.9613
2.2	1.2952	5.0596	4.7	0.3058	1.8642
2.3	1.2356	4.9254	4.8	0.2861	1.7706
2.4	1.1773	4.7886	4.9	0.2675	1.6806
2.5	1.1206	4.6498	5.0	0.2501	1.5941

A reasonable approximation of the integral $\int_X^\infty \frac{x^2}{e^x-1}dx$ is given by

$$\begin{aligned}\int_X^\infty \frac{x^2}{e^x-1}dx &\approx 2.4164 - 0.086332X - 0.37357X^2\\ &\quad + 0.099828X^3 - 0.0078158X^4. \end{aligned} \tag{14.109}$$

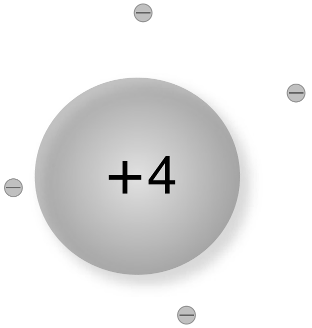
Appendix B: A Semiconductor Primer

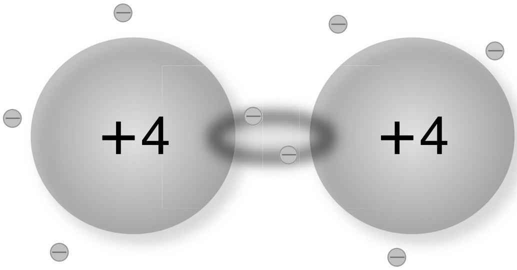


A silicon atom consists of a **nucleus** (dark sphere), containing 14 **protons** (heavily rimmed circles with a "+" sign) and 14 **neutrons** (solid black), surrounded by 14 **electrons** (circles with a "−" sign). The nucleus carries a charge of +14, but the atom (unless it is ionized) has no net charge owing to the 14 electrons swarming around it.

Observe that 10 of the 14 electrons are tightly bound to the nucleus and are difficult to remove. However, four electrons (called **valence electrons**) can be easily removed (ionizing the atom) and are, therefore, able to take part in chemical reactions. Consequently, silicon is **tetravalent**.

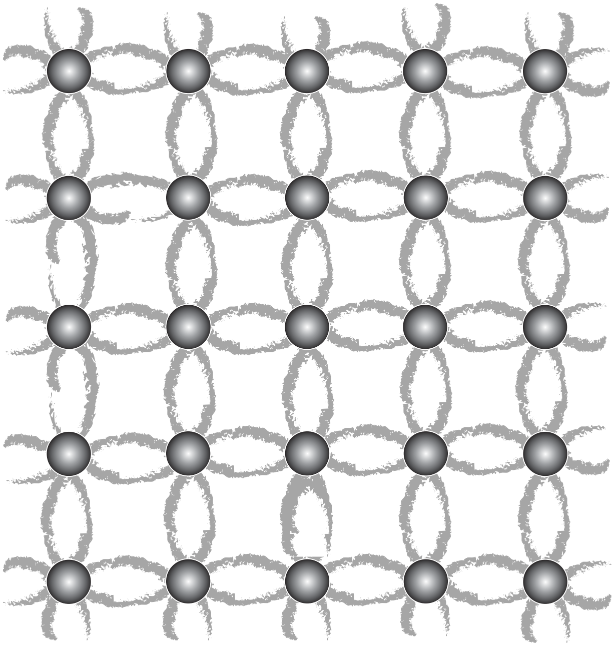
It proves convenient to represent a silicon atom as consisting of a **kernel** surrounded by four valence electrons. Because the kernel has 14 positive charges from the nucleus and 10 tightly bound electrons, it has a net charge of +4. The atom, as a whole, is of course neutral because the charge of the four valence electrons cancels that of the kernel.





Two silicon atoms can be bound one to the other by exchanging valence electrons. Such a bond is called a **covalent bond**.

Because it has four valence electrons, each silicon atom can make *four* covalent bonds, attaching itself to four neighboring atoms. This might lead to a lattice structure as depicted in the accompanying graphic. Indeed, it does so in the case of carbon, which crystallizes in a two-dimensional fashion when it forms graphite. Silicon (and carbon in the diamond form) has a three-dimensional crystal that is difficult to depict in a flat drawing. For simplicity, we will continue to use the flat picture.

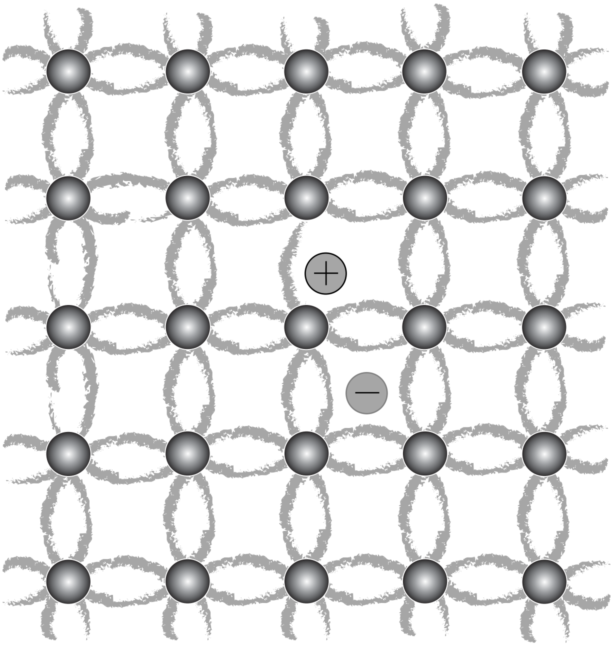


At 0 K, all valence electrons are engaged in covalent bonds and are, therefore, unavailable as **carriers**—that is, as transporters of electric charge. No current can flow through the crystal; it is an **insulator**.

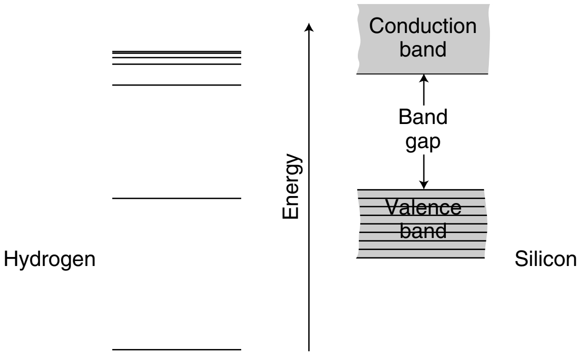
However, if a bond is disrupted (by thermal agitation of the lattice or through the impact of a photon or a high-speed free electron), then one of the valence electrons is ejected from the bond and becomes free to carry electricity, leaving behind an incomplete bond, one in which a **hole** exists into which an electron from a neighboring bond can fall. This causes the hole to move to a new place. Thus, the disruption of a bond creates a *pair* of carriers—an electron and a hole imparting some degree of conductivity to the material.

It is clear from our picture, that, in this particular case, the number of free (**conduction**) electrons is exactly equal to the number of holes. Such materials are called **intrinsic**.

One can now understand why semiconductors usually have an electric conductivity that increases when the temperature increases: the warmer the material, the greater the number of carriers.



Another way of looking at this situation is to realize that, just as in isolated atoms, electrons in a crystal can only occupy certain discrete energy levels. In the case of solids, electrons are confined to certain **bands** of energy levels. In semiconductors, such as silicon, one band, called the **valence band**, is completely occupied and electrons cannot move. The next band (**conduction band**) is completely empty. The energy difference between the top of the valence band and the bottom of the conduction band is called the **band gap**. In silicon, the band-gap is 1.1 eV.



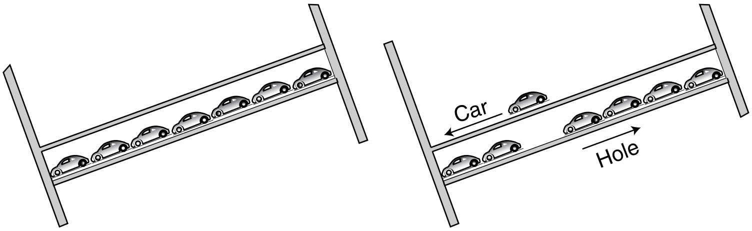
Whereas the free electron is a negative carrier of charge, the hole can be treated as a positive carrier. In the presence of an electric field, their motion is in opposite directions; yet the resulting current is in the same direction because of the opposite charges they carry.

A simple analogy to the motion of electrons and holes is that of the motion of VW beetles and the corresponding "holes."

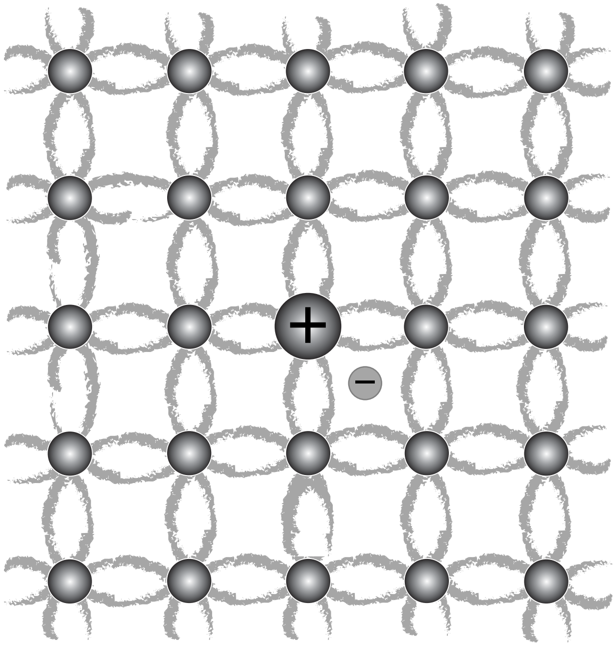
In a garage in San Francisco with a floor fully occupied by Volkswagen beetles (in neutral and with no brakes on, so they can roll freely), the cars cannot move even if an earthquake tilts the garage floor.

If a car is hoisted to the next floor, then, with a tilted garage, the car will roll to the left. Simultaneously, some cars in the lower floor will also move, in effect causing the "hole" in the row of cars to move *up*.

The energy necessary to lift a car from one floor to the next is equivalent to the band-gap energy in semiconductors.

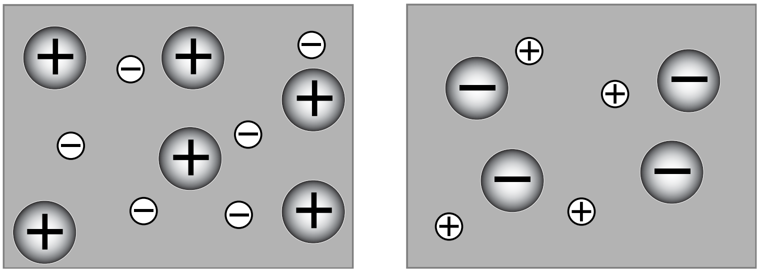


So far, we have dealt with absolutely pure silicon. In fact, all silicon contains some impurity that may dramatically alter the properties of the material. In the picture on the next page, a phosphorous atom has replaced one silicon atom. Since phosphorus has a valence of 5, it has enough electrons to complete the four covalent bonds that anchor the atom in the lattice. In addition, there is one electron left over, which then acts as a carrier. Electron-hole pairs (not shown in the picture) are still being created by the thermal agitation of the lattice, but the number of



electrons now exceeds that of holes—the material is ***n*-silicon**, one in which the dominant carrier is negative. The phosphorus kernel has a +5 charge, and the crystal site has only four covalent bond electrons. Thus, the site has a +1 positive charge, which being immobile (because it is tied to the lattice) does *not* constitute a carrier. It is called a **donor**.

If instead of a pentavalent atom such as phosphorus, the impurity is a trivalent one such as aluminum, then there is an insufficiency of electrons and only three of the covalent bonds are satisfied, leaving one incomplete—that is, leaving a free hole—the material is ***p*-silicon**, one in which the dominant carrier is a positive hole. Similarly to the case of phosphorus, the aluminum atom represents a −1 immobile charge—it is an **acceptor**.



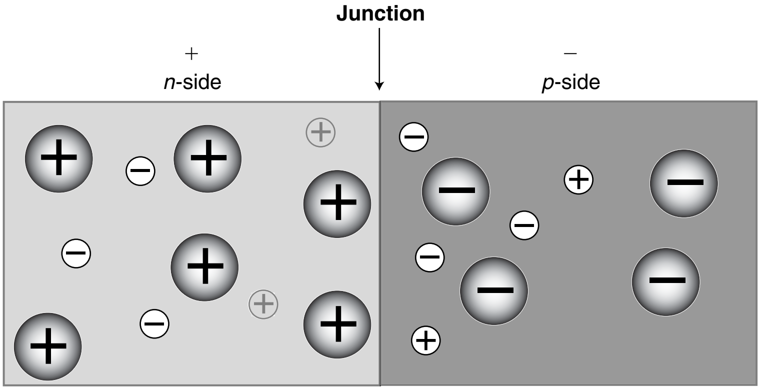
From now on, for clarity, we will omit most of the background of silicon atoms and covalent bonds from our drawings and show only the carriers and their corresponding donors or acceptors as in the figure in the preceding page. It must be emphasized that the donors and acceptors are not carriers because they cannot move.

The left-hand drawing represents an n-material, and the right-hand one, a p-material. Both are electrically neutral, having an equal number of positive and negative charges.

The introduction of certain impurities into the mass of silicon is called **doping**. Typically, the amount of doping is small, ranging from 1 dopant atom for every 10,000 silicon atoms (extremely heavy doping) to 1 dopant atom for every 100 million silicon atoms (very light doping).

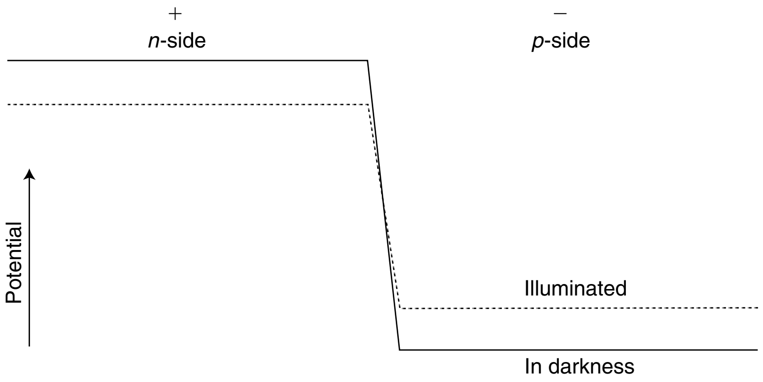
Things start becoming really interesting when a crystal has a p-region juxtaposed to an n-region as shown below.

Electrons, more abundant in the n-side, tend to diffuse to the p-side, while holes tend to diffuse into the n-side. The donors and acceptors, of course, cannot move. The net effect of these diffusions is that the n-side becomes positive (in the illustration, there are eight positive and only three negative charges). By the same token, the p-side becomes negative (seven negative and two positive charges). Thus, a **contact potential** is created, which in silicon at room temperature can be around 1 V, depending on the doping.



The potential across most of the n-side and p-side of the device is constant (no electric field). All the field is concentrated across a narrow transition region. Owing to the narrowness of this region (a few tens of a nanometer), the electric fields developed can be enormous—tens of millions of volts/meter.

When light shines near a p-n junction, electron-hole pairs may be created on either side. If they are very far from the transition region, they will recombine after a few microseconds. If, however, they are near, they may drift toward the region of high electric field. In this case, an electron created in the p-side may fall to the n-side, while a hole created on the n-side may fall to the p-side. In either case, these charges counteract the contact potential. Thus, the effect of light on a p-n junction is the lowering of the contact potential.



References

B. A. Bolto and D. E. Weiss, *Electronic Conduction in Polymers. II. The Electrochemical Reduction of Polypyrrole at Controlled Potential*, Australian Journal of Chemistry, **16(6)**, pp. 1076–1089, **1963**.

B. A. Bolto, R. McNeill, and D. E. Weiss, *Electronic Conduction in Polymers. III. Electronic Properties of Polypyrrole*, Australian Journal of Chemistry, **16(6)**, pp. 1090–1103, **1963**.

da Rosa, Aldo V., *Fundamentals of Electronics*, Optimization Software, Inc., **1989**.

Gordon, Roy G., Criteria for choosing transparent conductors, *Materials Research Society, MRS Bulletin*, August **2000**.

Greenham, N. C., X. Peng, and A. P. Alivisatos, *Charge separation and transport in conjugated-polymer/semiconductor-nanocrystal composites studied by photoluminescence quenching and photoconductivity Phys. Rev. B* **54**, p. 17628, **1996**.

Grätzel, Michael, Dye-sensitized solar cells, Journal of Photochemistry and Photobiology & Photochemistry Reviews, **4**, pp. 145–153, **2003**.

Gregg, Brian A., The essential interface; Studies in dye-sensitized solar cells, National Renewable Energy Laboratory, Golden, Colorado Part of Semiconductor Photochemistry and Photophysics, Editors: V. Ramamurthy and Rirk S. Schanze, CRC Press, **2003**.

H. Amano, N. Sawaki, I. Akasaki, Y. Toyoda, *Appl. Phys. Lett.* **48**, p. 353, **1986**.

Hamdy, Adel M., and D. E. Osborn, The potential for increasing the efficiency of solar cells in hybrid photovoltaic/thermal concentrating systems by using beam splitting, *Solar and Wind Technology* **7**, No. 23, pp. 147–153, **1990**.

Joannopoulos, John D., Steven G. Johnson, Joshua N. Winn, and Robert D. Meade, Photonic Crystals: *Molding the Flow of Light* (2nd edition), Princeton University Press, **2008**.

Koster, Lambert Jan Anton, Device physics of donor-acceptor/blend solar cells, Ph,D Thesis, University of Groningen, The Netherlands, **2007**.

Loferski, Joseph J., Theoretical considerations governing the choice of the optimum semiconductor for photovoltaic solar energy conversion, *J. Appl. Phys.* **27(7)**, pp. 777–784, July **1956**.

Ludman, Jacques E., Approximate bandwidth and diffraction efficiency in thick holograms, Rome Air Development Center, **1981**.

Ludman, Jacques E., Holographic solar concentrator, *Applied Optics* **212**, p. 3057, September 1, **1982**.

Ludman, Jacques E., Juanita Riccobono, Nadya Reinhand, Irina Semenova, José Martín, William Tai, Xiao-Li Li, and Geof Syphers, Holographic solar concentrator for terrestrial photovoltaics, *First WCPEC, Hawaii, IEEE.* December 5–9, **1994**.

Ludman, Jacques E., J. L. Sampson, R. A. Bradbury, J. G. Martín, J. R. Riccobono, G. Sliker, and E. Rallis, Photovoltaic systems based on spectrally selective holographic concentrators, SPIE, 1667 *Practical Holography VI*, p. 182, **1992**.

Luque, Antonio, and Antonio Martí, Increasing the efficiency of ideal solar cells by photon induced transitions at intermediate levels, *Phys. Rev. Lett.* **76(26)**, June 30, **1997**.

Mayer, Alex C., Shawn R. Scully, Brian E. Hardin, Michael W. Rowell, and Michael D. McGehee, REVIEW: Polymer-based solar cells, *Materials Today* **10(1)**, pp. 28–33, November **2007**.

Morrison, Orion, Michael Seal, Edward West, and William Connelly, Use of a thermophotovoltaic generator in a hybrid electric vehicle, Vehicle Research Institute, Western Washington University.

Petritsch, Klaus, Organic solar cell architectures, Ph. D. Thesis, Tech. Univ. Graz, July **2000**.

Prince, M. B., Silicon solar energy converters, *J. Appl. Phys.* **26(5)**, pp. 534–540, May **1955**.

Rappaport, Paul, The photovoltaic effect and its utilization, *RCA Rev.* **20**, pp. 373–397, September **1959**.

Rittner, Edmund S., An improved theory for Silicon *p-n* junctions in solar cells, *J. Energy 1*, p. 9, January **1977**.

Ross, B., and J. R. Madigan, Thermal generation of recombination centers in silicon, *Phys. Rev.* **108**, pp. 1428–1433, December 15, **1957**.

R. McNeill, R. Siudak, J. H. Wardlaw, and D. E. Weiss, Electronic Conduction in Polymers. I. The Chemical Structure of Polypyrrole, Australian Journal of Chemistry, **16(663)**, pp. 1056–1075, **1963**.

Schaller, R. D., and V. I. Klimov, High efficiency carrier multiplication in PbSe nanocrystals: Implications for solar energy conversion, *Phys. Rev. Lett.* **92**, p. 186601–4, **2004**.

Schaller, R. D., Vladimir M. Agranovich, and V. I. Klimov, High efficiency carrier multiplication through direct photogeneration of multi-excitations via virtual single-exciton states, *Nature Physics* **1**, December **2005**.

Shockley, William, and Han J. Queisser, Detailed balance limit of efficiency of *p*-*n* junction solar cells, *J. Appl. Phys.* **32(3)**, pp. 510–519, March **1961**.

S.Nakamura, *Jpn. J. Appl. Phys.* **30**, L1705, **1991**.

Wu, J., Wladek Walukiewicz, K. M. Yu, J. W., Ager III, E. E. Hallert, H. Lu, W. J. Schaff, Y. Saito, and Y. Nanishi, Unusual properties of the fundamental band-gap of InN, *Appl. Phys. Lett.* **80**, pp. 3967–3969, **2002**.

Wysocki, Joseph J., and Paul Rappaport, Effect of temperature on photovoltaic solar energy conversion, *J. Appl. Phys.* **31**, pp. 571–578, March **1960**.

Yu, K. M., W. Walukiewicz, J. Wu, W. Shan, J. W. Beeman, M. A. Scarpulla, O. D. Dubon, and P. Becla, Diluted II-VI oxide semiconductors with multiple band-gaps, *Phys. Rev. Lett.* **91(24)**, p. 4, December 12, **2003**.

PROBLEMS

14.1 What is the theoretical efficiency of cascaded photodiodes made of two semiconductors, one with a band-gap energy of 1 eV and the other with 2 eV when exposed to sunlight?

14.2 As any science fiction reader knows, there are many parallel universes, each one with different physical laws. In the parallel universe we are discussing here, the black body radiation at a given temperature, T, follows a simple law (P is the power density (W m^{-2}) and f is in Hz.):

$\frac{\partial P}{\partial f}$ is zero at $f = 0$,

$\frac{\partial P}{\partial f}$ grows linearly with f to a value of $1\,\mathrm{W\,m^{-2}\,THz^{-1}}$ at 500 THz.

From 500 THz it decreases linearly to zero at 1000 THz.

1. What is the total power density of the radiation?
2. What is the value of the band gap of the photodiode material that results in the maximum theoretical efficiency of the photodiode exposed to the above radiation?

14.3 Under circumstances in which there is substantial recombination of carriers in the transition region of a diode, the V-I characteristic becomes

$$I = I_\nu - I_R\left[\exp\left(\frac{qV}{2kT}\right) - 1\right] - I_0\left[\exp\left(\frac{qV}{kT}\right) - 1\right].$$

In solving this problem, use this more complicated equation rather than the equation given in the text, which is

$$I = I_\nu - I_0\left[\exp\left(\frac{qV}{kT}\right) - 1\right].$$

A silicon diode has 1 cm^2 of effective area. Its reverse saturation current, I_0, is 400 pA, and the current, I_R, is 4 μA. These values are for T = 300 K. Assume that this is the temperature at which the diode operates. Assume also 100% quantum efficiency—that is, that each photon with energy above 1.1 eV produces one electron-hole pair. Finally, assume no series resistance.

At onesun, the power density of light is 1000 W/m^2, and this corresponds to a flux of 2.25×10^{21} photons s^{-1} m^{-2} (counting only photons with energy above 1.1 eV).

1. What is the open-circuit voltage of this diode at one sun?
2. At what voltage does the diode deliver maximum power to the load?
3. What is the maximum power the diode delivers?
4. What load resistance draws maximum power from the diode?
5. What is the efficiency of the diode?

Now assume that a concentrator is used so that the diode will receive 1000 suns. This would cause the operating temperature to rise and would impair the efficiency. Assume, however, that an adequate cooling system is used so that the temperature remains at 300 K. Answer the five questions above using 100% concentrator efficiency.

In fact, the concentrator is only 50% efficient. Is there still some advantage in using this diode with the concentrator?

14.4 Assume that you are dealing with perfect black body radiation ($T = 6000$ K). We want to examine the theoretical limits of a photodiode. Assume no light losses by surface reflection and by parasitic absorption in the diode material. Consider silicon (band-gap energy, W_g, of 1.1 eV).

Clearly, photons with energy less than W_g will not interact with the diode because it is transparent to such radiation.

1. What percentage of the power of the black body radiation is associated with photons of less than 1.1. eV?
2. If all the energy of the remaining photons were transformed into electric energy, what would the efficiency of the photodiode be?
3. Would germanium ($W_g = 0.67$ eV) be more or less efficient?
4. Using Table 12.1 in Chapter 12, determine the percentage of the solar energy absorbed by silicon.
5. A photon with 1.1 eV will have just enough energy to produce one electron-hole pair, and, under ideal conditions, the resulting electron would be delivered to the load under 1.1 V of potential difference. On the other hand, a photon of, say, 2 eV, will create pairs with 0.9 eV excess energy. This excess will be in the form of kinetic energy and will rapidly be thermalized, and, again, only 1.1 eV will be available to the load. Thus, all photons with more than W_g will, at best, contribute only W_g units of energy to the load.
   Calculate what fraction of the black body radiation is available to a load connected to an ideal silicon photodiode.
6. The short-circuit current of a diode under a certain illumination level is 10^7 times the diode reverse saturation current. What is the relative efficiency of this diode compared with that of the ideal diode above?

14.5 Consider radiation with the normalized spectral power density distribution given by

$$\frac{\partial P}{\partial f} = 0 \quad \text{for} \quad f < f_1 \quad \text{and} \quad f > f_2,$$

$$\frac{\partial P}{\partial f} = 1 \quad \text{for} \quad f_1 < f < f_2,$$

where $f_1 = 100$ THz and $f_2 = 1000$ THz.

1. What is the theoretical efficiency of a photodiode having a band-gap energy of $W_g = hf_1$?
2. What band-gap energy maximizes the efficiency of the diode?
3. If the band-gap energy is $h \times 500$ THz, and if the material is totally transparent to radiation with photons of less energy than W_g, what fraction of the total radiation power goes through the diode and is available on its back side?
4. If behind the first diode, there is a second one with $W_g = h \times 100$ THz, what is the efficiency of the two cascaded diodes taken together?

14.6 Two photodiodes, each with an effective area of 10 cm^2, are exposed to bichromatic radiation having power densities of 500 W/m^2, in narrow bands one around 430 THz and the other, around 600 THz.

One diode has a band-gap energy of 1 eV, and the other has 2 eV. When the diode is reverse biased (in the dark), the saturation current is 10 nA.

The diodes operate at 300 K.

1. What are the short-circuit photo currents?
2. What is the open-circuit voltage of each diode?
3. What is the maximum theoretical efficiency of each diode?
4. What is the maximum power each diode can deliver to a load (assume no series resistance in the diodes)?

14.7 An ideal photodiode is made of a material with a band-gap energy of 2.35 eV. It operates at 300 K and is illuminated by monochromatic light with wavelength of 400 nm. What is its maximum efficiency?

14.8 What is the short-circuit current delivered by a 10 cm by 10 cm photocell (100% quantum efficiency, $R_s = 0$) illuminated by monochromatic light of 400-nm wavelength with a power density of 1000 W/m^2.

14.9 Under different illumination, the cell of Problem 14.8 delivers 5 A into a short circuit. The reverse saturation current is 100 pA. Disregard any internal resistance of the photodiode. What is the open-circuit voltage at 300 K?

14.10 The optical system of a solar photovoltaic system consists of a circular f:1.2[†] lens with a focal length, F, of 3 m.

When aimed directly at the sun (from the surface of planet Earth), what is the diameter, D_i, of the solar image formed in the focal plane? Assume a 90% efficiency of the optical system and a perfectly clear atmosphere at noon.

1. What is the concentration ratio, C?

[†] The f-number is, as in any photographic camera, the ratio of the focal length to the diameter of the lens.

2. What is the total light power, P, that falls on a photovoltaic cell exposed to the solar image?
3. What is the power density, p?
   Assume:
   3.1 no conduction losses;
   3.2 no convection losses;
   3.3 the silicon photovoltaic cell is circular and has a diameter equal to that of the solar image. The cell intercepts all the light of the image;
   3.4 the efficiency of the photocell is equal to 60% of the maximum theoretical value for a black body radiator at 5800 K.
   3.5 all the power the photocell generates is delivered to an external load;
   3.6 the effective heat emissivity, ϵ, is 0.4.
4. What is the power delivered to the load?
5. What is the temperature of the photocell?
6. What is the temperature of the photocell if the load is disconnected?
   If you do your calculations correctly, you will find that the concentrated sunlight will drive the cell to intolerably high temperatures. Silicon cells should operate at temperatures of 500 K or less.
7. How much heat must a coolant remove to keep the cell at 500 K when no electric power is being extracted?
   The coolant will exit the cell at 480 K and drives a steam engine that rejects the heat at 80 C and that realizes 60% of the Carnot efficiency.
8. How much power does the heat engine deliver?
9. What is the overall efficiency of the photocell cum steam engine system?

14.11 *Treat the photodiode of this problem as an ideal structure. Assume 100% quantum efficiency.*

A photodiode has an area of 1 by 1 cm and is illuminated by monochromatic light with a wavelength of 780 nm and with a power density of 1000 W/m^2. At 300 K, the open-circuit voltage is 0.683 V.

1. What is its reverse saturation current, I_0?
2. What is the load resistance that allows maximum power transfer?
3. What is the efficiency of this cell with the load above?

14.12 The power density of monochromatic laser light (586 nm) is to be monitored by a 1 × 1 mm silicon photodiode. The quantity observed is the short-circuit current generated by the silicon. Treat the diode as a perfect ideal device.

1. What current do you expect if the light level is $230\,\mathrm{W/m^2}$?
2. How does the temperature of the semiconductor affect this current? Of course, the temperature has to be lower than that which will destroy the device (some 150 C, for silicon).
3. Instead of being shorted out, the diode is now connected to a load compatible with maximum electric output. Estimate the load voltage.

14.13 A silicon photocell being tested measures 4 by 4 cm. Throughout the tests, the temperature of the device is kept at 300 K. Assume the cell has no significant series resistance. Assume 100% quantum efficiency. The band-gap energy of silicon is 1.1 eV.

Initially, the cell is kept in the dark. When a current of $100\,\mu$A is forced through it in the direction of good conduction, the voltage across the diode is 0.466 V.

Estimate the open-circuit voltage developed by the cell when exposed to bichromatic infrared radiation of 412 nm and 1300 nm wavelength. The power density at the shorter wavelength is $87\,\mathrm{W/m^2}$, while at the longer wavelength, it is $93\,\mathrm{W/m^2}$.

14.14 What is the ideal efficiency of a photocell made from a semiconducting material with a band-gap energy, $W_g = 2$ eV, when illuminated by radiation with the normalized spectral power distribution given as follows.

f	$\partial P/\partial f$
< 200 THz	0
200 to 300 THz	$0.01f - 2$
300 to 400 THz	$4{-}0.01f$
> 400 THz	0

In this table the frequency, f, is in THz.

Repeat for a semiconductor with $W_g = 1$ eV.

14.15 What is the theoretical efficiency of a photocell with a 2.5-V band-gap when exposed to $100\,\mathrm{W/m^2}$ solar radiation through a filter having the following transmittance characteristics:

Pass without attenuation all wavelengths between 600 and 1000 nm. Reject all else.

14.16 A photodiode is exposed to radiation of uniform spectral power density ($\frac{\partial P}{\partial f}$ = constant) covering the range from 300 to 500 THz. Outside this range there is no radiation. The total power density is $2000\,\mathrm{W/m^2}$.

1. Assuming 100% quantum efficiency, what is the short-circuit photocurrent of a diode having an active area of 1 by 1 cm?
2. When exposed to the radiation in Part 1 of this problem, the open-circuit voltage delivered by the diode is 0.498 V. A 1.0-V

voltage is applied to the diode (which is now in darkness) in the reverse conduction direction (i.e., in the direction in which it acts almost as an open circuit). What current circulates through the device? The temperature of the diode is 300 K.

14.17 The sun radiates (roughly) like a 6000-K black body. When the power density of such radiation is 1000 W/m^2—"one sun"—the total photon flux is 4.46×10^{21} photons per m^2 per second. Almost exactly half of these photons have energy equal or larger than 1.1 eV (the band-gap energy, W_g, of silicon).

Consider a small silicon photodiode with a 10 by 10 cm area. When 2 V of reversed bias is applied, the resulting current is 30 nA. This is, of course, the reverse saturation current, I_0.

When the photodiode is short-circuited and exposed to black body radiation with a power density of 1000 W/m^2, a short-circuit current, I_ν, circulates.

1. Assuming 100% quantum efficiency (each photon creates one electron-hole pair and all pairs are separated by the *p*-*n* junction of the diode), what is the value of this current?
2. What is the open-circuit voltage of the photodiode at 300 K under the above illumination?
3. Observe that the V-I characteristics of a photodiode are very steep at the high current end. In other words, the best operating current is only slightly less than that of the short-circuit current. This knowledge will facilitate answering this question:
   Under an illumination of 1000 W/m^2, at 300 K, what is the maximum power the photodiode can deliver to a load? What is the efficiency? Do this by trial and error and be satisfied with three significant figures in your answer. Consider an ideal diode with no internal resistance.
4. What is the load resistance that leads to maximum efficiency?
5. Now repeat the power, efficiency, and load resistance calculations for an illumination of 10,000 W/m^2.
6. What happens to the efficiency and the optimal load resistance when the power density of the illumination on a photodiode increases?

14.18 Everything else being the same, the efficiency of a photodiode rises

1. when the operating temperature rises.
2. when the operating temperature falls.
3. when the light power density rises.
4. when the light power density falls.

14.19 A photodiode with a band-gap energy of $W_g = 1.4$ eV is exposed to monochromatic radiation (500 THz) with a power density $P = 500$ w/m^2. The active area of the device is 10 by 10 cm.

Treat it as an ideal device in series with an internal resistance of $2\,m\Omega$.

All measurements were made at 298 K.

The open-circuit voltage is 0.555 V.

1. Estimate the short-circuit current.
2. How much power does the diode deliver to 200 mΩ load?
3. What is the efficiency of the device when feeding the 200 milliohm load?

14.20 *Suggestion: To solve this problem, use a spreadsheet and tabulate all the pertinent values for different hour angles (from sunrise to sunset). 5° intervals in α will be adequate.*

A flat array of silicon photodiodes is set up at 32° N. The array faces south and is mounted at an elevation angle that maximizes the year-long energy collection, assuming perfectly transparent air.

1. What is the elevation angle of the array?
2. On April 15, 2002, how does the insolation on the array vary throughout the day? Plot the insolation, P, versus the time of day, t, in hours.
3. What is the average insolation on the collector.
4. Assuming ideal silicon photodiodes with a reverse saturation current density of $10\,\text{nA/m}^2$, what is the average power delivered during the day (from sunrise to sunset) if a perfect load follower is used, that is, if the load is perfectly matched at all the different instantaneous values of insolation? What is the average overall efficiency?
5. Estimate the average power collected if the array is connected to a load whose resistance maximizes the efficiency at noon. In other words, the average power when no load-follower is used.

14.21 To simplify mathematical manipulation, we will postulate a very simple (and unrealistic) spectral power distribution:

$$\frac{\partial P}{\partial f} = \begin{cases} A & \text{for 300 THz} < f < \text{500 THz;} \\ 0, & \text{otherwise.} \end{cases}$$

1. If $A = 10^{-12}\,\text{W}\,\text{m}^{-2}\,\text{Hz}^{-1}$, what is the power density of the radiation?
2. Assuming 100% quantum efficiency, what is the short-circuit current density, J_ν, in an ideal photodiode having a band-gap energy smaller than the energy of 300 THz photon?
3. At 300 K, assuming a reverse saturation current density of $J_0 = 10^{-7}\,\text{A/m}^2$, what is the open-circuit voltage of the photocell?
4. At what load voltage, V_m, does this photocell deliver its maximum power output?
5. What is the current density delivered by the photocell when maximum power is being transferred to the load?

6. What is the efficiency of the photocell?
7. What is the load resistance under the above conditions?
8. Repeat all the above for a light power density of 2 W/m^2.
9. What would be the efficiency at these low light levels if the load resistance had the optimum value for 200 W/m^2?

14.22 It is hoped that high-efficiency cascaded photocells can be produced at a low cost. This consists of a sandwich of two cells of different band gaps. The bottom cell (the one with the smaller band gap) can be made using $\mathrm{CuIn_xGa_{1-x}Se_2}$, known as CIGS. This material has a band gap of about 1 eV and has been demonstrated as yielding cells with 15% efficiency.

The question here is what band gap of the top cell yields the largest efficiency for the combined cascaded cells. Assume radiation from a black body at 6000 K. Assume no losses; that is, consider only the theoretical efficiency.

14.23 The V-I characteristic of a photocell can be described by a rather complex mathematical formula, which can be handled with a computer but is too complicated for an in-class exam. To simplify handling, we are adopting, rather arbitrarily, a simplified characteristic consisting of two straight lines as shown in the figure above. The position of point C, of maximum output, varies with the I_ν/I_0 ratio. Empirically,

$$V_C = V_{OC}\left(0.7 + 0.0082 \ln \frac{I_\nu}{I_0}\right)$$

and

$$I_C = I_\nu\left(0.824 + 0.0065 \ln \frac{I_\nu}{I_0}\right).$$

Now consider silicon photodiodes operating at 298 K. These diodes form a panel, 1 m^2 in area, situated in Palo Alto (latitude 37.4° N, longitude 125° W). The panel faces true south and has an elevation of 35°. In practice, the panel would consist of many diodes in a series/parallel connection. In the problem here, assume that the panel has a single enormous photodiode.

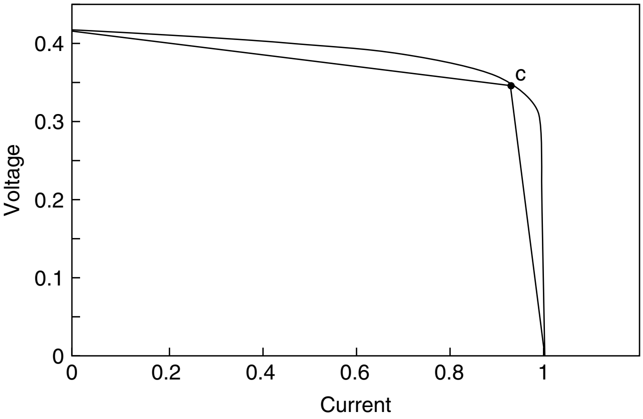
Calculate the insolation on the surface at a 1130 PST and at 1600 PST on October 27. Assume clear meteorological conditions.

Assume that the true solar time is equal to the PST.

1. Calculate the insolation on the collector at the two moments mentioned.
2. What are the short-circuit currents (I_ν) under the two illuminations? Consider the sun as a 6000-K black body.
3. When exposed to the higher of the two insolations, the open-circuit voltage of the photodiode is 0.44 V. What is the power

delivered to a load at 1130 and at 1600? The resistance of the load is, in each case, that which maximizes the power output for that case. What are the load resistances? What are the efficiencies?

4. Suppose that at 1600 the load resistance used was the same as that which optimized the 1130 output. What are the power in the load and the efficiency?
5. Let the load resistance be the same at both 1130 and 1600, but, unlike Question 4, not necessarily the resistance that optimizes the output at 1130. The idea is to operate the panel at slightly lower efficiency at 1130 and at somewhat higher efficiency than that of Question 4 at 1600 in the hope that the overall efficiency can be improved. What is the value of this common load resistance?



14.24 1. What is the ideal (theoretical) efficiency of a gallium phosphide photocell exposed to the radiation of a 6000-K black body? For your information: the corresponding efficiency for silicon is 43.8%.

2. What is the efficiency of an ideal silicon photocell when illuminated by monochromatic light with a frequency of 266 THz?
3. What is the efficiency of an ideal silicon photocell when illuminated by monochromatic light with a frequency of 541.6 THz?
4. A real silicon photocell measuring 10 by 10 cm is exposed to 6000-K black body radiation with a power density of 1000.0 W/m^2. The temperature of the cell is 310 K. The measured open-circuit voltage is 0.493 V. When short-circuited, the measured current is 3.900 A. The power that the cell delivers to a load depends, of course, on the exact resistance of this load. By properly adjusting the load, the power can be maximized. What is this maximum power?

14.25 Consider a solar cell made of semiconducting nanocrystals with a band-gap energy of $W_g = 0.67\,\text{eV}$. What is the theoretical efficiency when the solar cell is exposed to the radiation of a 6000-K black body? Assume that photons with less than $3.3\,W_g$ create each one single-electron/hole pair and that those with more than $3.3\,W_g$ create 2 electron/hole pairs each owing to impact ionization.

14.26 The Solar Power Satellites proposed by NASA would operate at 2.45 GHz. The power density of the beam at ionospheric heights (400 km) was to be $230\,\text{W/m}^2$. The collector on the ground was designed to use dipole antennas with individual rectifiers of the Schottky barrier type. These dipoles were dubbed **rectennas**.

The satellites would have been geostationary (they would be on a 24-hr equatorial orbit with zero inclination and zero eccentricity).

1. Calculate the orbital radius of the satellites.
2. Calculate the microwave power density on the ground at a point directly below the satellite (the subsatellite point). Assume no absorption of the radiation by the atmosphere.
3. The total power delivered to the load is 5 GW. The rectenna system has 70% efficiency. Assume uniform power density over the illuminated area. What is the area that the ground antenna farm must cover?
4. A dipole antenna abstracts energy from an area given in the text. How many dipoles must the antenna farm use?
5. Assuming (very unrealistically) that the only part of each rectenna that has any mass is the dipole itself, and assuming that the half-wave dipole is made of extremely thin aluminum wire, only 0.1 mm in diameter, what is the total mass of aluminum used in the antenna farm?
6. How many watts must each dipole deliver to the load?
7. If the impedance of the rectenna is 70 ohms, how many volts does each dipole deliver?

14.27 The Solar Power Satellite radiates 6 GW at 2.45 GHz. The transmitting antenna is mounted 10 km from the center of gravity of the satellite. What is the torque produced by the radiation?

14.28 Compare the amount of energy required to launch a mass, m, from the surface of the Earth to the energy necessary to launch the same mass from the surface of the moon. “Launch” here means placing the mass in question an infinite distance from the point of origin. Consult the *Handbook of Chemistry and Physics* (CRC) for the pertinent astronomical data.

14.29 Consider a simple spectral distribution:

$$\text{For } f < 300 \text{ THz},\ \partial P/\partial f = 0.$$

$$\text{For } f = 300 \text{ THz},\ \partial P/\partial f = A.$$

For $300 \leq f \leq 500$, $\partial P/\partial f = af$.

For $f > 500$, $\partial P/\partial f = 0$.

The total power density is $P = 1000$ W/m^2.

a. A silicon diode having 100% quantum efficiency is exposed to this radiation. What is the short-circuit current density delivered by the diode?

b. At 300 K, the reverse saturation current density, J_0, of the diode is 40 μA/m^2. What is the open-circuit voltage generated by the diode?

c. If the photodiode has an effective area of 10 by 10 cm, what load resistance will result in the largest possible output power?

14.30 1. Five identical photodiodes are connected in series and feed a single-cell water electrolyzer.

The whole system operates at 300 K, and the photodiode is exposed to a light power density that causes a photon flux of 2.5×10^{21} photons per second per square meter to interact with the device. Quantum efficiency is 100%.

For each photodiode, the reverse saturation current, I_0, is 0.4 μA, and the series resistance is 10 mΩ. The active area of the diode is 10 by 10 cm.

The electrolyzer can be represented by a 1.6-V voltage source, in series with a 100-mΩ resistance.

What is the hydrogen production rate in grams/day?

2. What photon flux is sufficient to just start the electrolysis?

14.31 What is the frequency at which a human body radiates the most heat energy per 1-Hz bandwidth?

14.32 The ideal photocells can exceed the black body spectrum efficiency if used in a configuration called a cascade. Thus the ideal efficiency of silicon cells exposed to a 6000-K black body radiation is 43.8%. If in cascade with an ideal cell with a 1.8 eV band gap, the efficiency is 56%. Clearly, using more cells the efficiency goes up further.

What is the efficiency of a cascade arrangement consisting of an infinite number of cells, the first having a band gap of 0 eV and each succeeding one having a band gap infinitesimally higher than that of the preceding cell? The cell with the highest band gap is on top (nearest the light source).

14.33 A high-precision photometer (300 K) equipped with a very narrow band-pass filter made the following measurements:

Light power density in a 1-MHz-wide band centered around 200 THz: 2.0×10^{-11} W/m^2

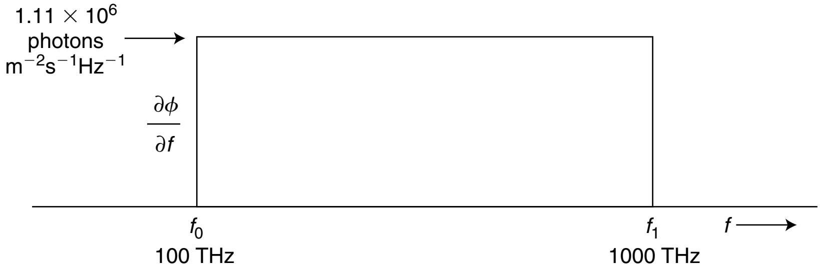
Light power density in a 1-MHz-wide band centered around 300 THz: 2.7×10^{-11} W/m^2

Assuming the radiation came from a black body, what is the temperature of the black body?

14.34 A silicon diode, operating at 300 K, is exposed to 6000-K black body radiation with a power density of 1000 W/m^2. Its efficiency is 20% when a load that maximizes power output is used. Estimate the open-circuit voltage delivered by the diode.

14.35 A GaAs photodiode, operating at 39 C, is exposed to 5500 K black body radiation with a power density of 675 W/m^2. The open-circuit voltage of the device is 0.46 V. What is the efficiency of the photodiode when delivering energy to a 2-milliohm m^2 load?

14.36



Consider radiation having the spectral distribution above. Observe that the plot is of photon flux versus frequency, not of power density versus frequency as usual. The flux is:

$$f < 100 \text{ THz}, \partial\phi/\partial f = 0,$$
$$100 < f < 1000 \text{ THz}, \ \partial\phi/\partial f = 1.11 \times 10^6 \,\text{photons}\,\text{m}^{-2}\,\text{s}^{-1}\,\text{Hz}^{-1},$$
$$f > 1000 \text{ THz}, \partial\phi/\partial f = 0,$$

1. What is the total flux of photons?
2. What is the total power density of the radiation (W/m^2)?
3. What is the ideal efficiency of a diode exposed to the above radiation if its band-gap energy is just slightly more than $10^{14}h$ joules? And if it is just a tad less than $10^{15}h$ joules?
4. What band-gap energy causes the ideal photodiode to attain maximum efficiency when illuminated by the radiation we are discussing? What is that efficiency?
5. If it were possible to split the spectrum into two regions—one extending from 100 THz to 500 THz and the other from 500 THz to 1000 THz—and if one were to use two independent photodiodes, one with a band-gap of $300\times10^{12}h$ joules, exposed to the lower of the two bands mentioned, and the other, with a band gap of $750\times\ 10^{12}h$ exposed to the higher of the two bands mentioned, what would the combined output and efficiency of the system be?

6. What is the efficiency of an ideal diode with a band-gap energy of $250 \times 10^{12} h$ when exposed to the radiation under discussion? Then, assume that for each photon with energy equal or larger than $3 \times 250 \times 10^{12} h$ one additional electron becomes available to the output. What is the efficiency, now?

Chapter 15
Wind Energy

15.1 History

The use of wind energy dates back to ancient times when it was employed to propel sailboats. Extensive application of wind turbines seems to have originated in Persia where it was used for grinding wheat. The Arab conquest spread this technology throughout the Islamic world and China. In Europe, wind turbines made their appearance in the eleventh century and two centuries later became an important tool, especially in Holland.

The development of the American West was aided by wind-driven water pumps, cereal grinders, and sawmills. These wind machines drove their mechanical loads directly. Modern turbines generate electric energy.

The first significant wind turbine designed specifically for the generation of electricity was built by Charles Brush in Cleveland, Ohio. It operated for 12 years, from 1888 to 1900, supplying the needs of his mansion. Charles Brush was a mining engineer who made a fortune with the installation of arc lights to illuminate cities throughout the United States. His wind turbine was of the then familiar multivane type (it sported 144 blades) and, owing to its large solidity (see Section 15.10), rotated rather slowly and required gears and transmission belts to speed up the rotation by a factor of 50 so as to match the specifications of the electric generator.

The wind turbine itself had a diameter of 18.3 m, and its hub was mounted 16.8 m above ground.

The tower was mounted on a vertical metal pivot so that it could orient itself to face the wind. The whole contraption massed some 40 tons.

Owing to the intermittent nature of the wind, electric energy had to be stored—in this case in 400 storage cells.

Although the wind is free, the investment and maintenance of the plant caused the cost of electricity to be much higher than that produced by steam plants. Consequently, the operation was discontinued in 1900, and from then on the Brush mansion was supplied by the Cleveland utility.

In 1939, construction of a large wind generator was started in Vermont. This was the famous Smith-Putnam machine, erected on a hill called Grandpa's Knob. It was a propeller-type device with a rated power of 1.3 MW at a wind speed of 15 m/s. Rotor diameter was 53 m. The machine started operation in 1941, feeding energy synchronously directly into the power network. Owing to blade failure, in March 1945, operation was discontinued. It ought to be mentioned that the blade failure had been predicted, but during World War II there was no opportunity to redesign the propeller hub.

After World War II, the low cost of oil discouraged much of the alternate energy research, and wind turbines were no exception. The 1973 oil crisis reinvigorated interest in wind power, as attested by the rapid growth in federal funding. This led to the establishment of **wind farms** that at the time, were more successful in generating tax incentives than electric energy. The early machines used in such farms proved disappointing in performance and expensive to maintain. Nevertheless, the experience accumulated led to an approximately fivefold reduction in the cost of wind-generated electricity. In the beginning of 1980, the cost of 1 kWh was around 25 cents; in 1996, in some installations, it was down to 5 cents. To be sure, the determination of energy costs is at best an unreliable art. Depending on the assumptions made and the accounting models used, the costs may vary considerably. The calculated cost of the kWh depends on a number of factors including

1. The cost of investment—that is, the cost of the installed kW. This number was around \$1000/kW in 1997 and does not appear to have changed much since. In 2000, the largest wind power plant outside the United States was the 50-MW plant (84 turbines each with a 600-kW capacity) that the French company, Cabinet Germa, installed in Dakhla on the Atlantic coast of Morocco. The project cost \$60 million, or \$1200/kW.

   These investment costs are comparable with those of fossil-fueled and hydroelectric plants that operate with capacity factors of at least 50%, whereas wind-power plants operate with a factor of some 30%. The capacity factor compares the amount of energy produced over, say, a year with the amount that would be produced if the plant operated at full power 24 hours per day. The intermittent and variable nature of the wind is the cause of the low factors achieved. Thus, for a one-on-one comparison with fossil fuel electricity, the cost of wind power should be multiplied by $0.5/0.3 = 1.6$.

   The one-time investment cost must be translated to yearly costs by including the cost of borrowing the necessary funds. (See Section 1.12 in Chapter 1.) The cost of the kWh produced is extremely sensitive to the cost of the money (which may include interest, taxes, insurance, etc.). This is illustrated in Problem 15.17 at the end of this chapter.
2. Fuel costs, which are, of course, zero for wind-driven plants. The secular rise in fossil fuel cost confers a relative advantage to fuel-free energy processes such as wind, solar and hydroelectric.
3. Operating and maintenance costs.
4. Decommissioning costs.
5. Land costs.

Even though the real cost of wind-generated energy may be uncertain, what is certain is that it has come down dramatically during these last two decades.

The relative cheapness of oil in the 1990s resulted in another ebbing in development funds. However, beginning around the turn of the new millennium, the rate of installation of wind farms picked up in a most amazing manner as illustrated in Figure 15.1. The American interest in wind turbines mirrored the worldwide trend. At the end of 2007, over 94 GW (equivalent to 94 large nuclear plants) were in operation throughout the world, mostly in Germany (22.2 GW), the United States (16.8 GW), and Spain (15.5 GW). It is undeniable that wind energy is now an important player in the generation of electricity.

In addition to the growing economic attractiveness of wind energy, there are major ecological arguments for its use:

1. Wind-power plants emit absolutely no CO_2, by far the major pollutant when fuels (other than hydrogen or biomass) are burned.
2. The operation of wind turbines leaves behind no dangerous residues as do nuclear plants.
3. Decommissioning costs of wind turbines are much smaller than those of many other types of power plants, especially compared with those of nuclear generators.
4. Land occupied by wind farms can find other simultaneous uses such as in agriculture.

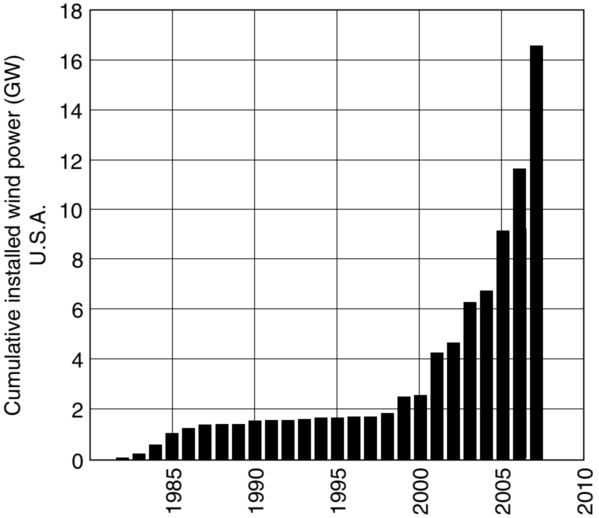


Figure 15.1 After a long period of sluggish growth, the rate of wind farm expansion in the United States has increased greatly beginning 2000.

On the other hand, some groups are opposed to wind turbines because of the danger they constitute to the birds that fly near the wind farms.

The optimal size of each individual wind turbine in a wind farm has been discussed. In the 1970s, government-sponsored research in both the United States and Germany favored large (several MW) machines, while private developers opted for much smaller ones. Large machines fitted in well with the ingrained habits of the power generation industry accustomed to the advantages of economy of scale. It could be argued, however, that such advantages might not apply to wind turbines. Consider an extremely oversimplified reasoning:

For a given wind regimen, the amount of energy that can be abstracted from the wind is proportional to the swept area of the turbine. The area swept out by a rotor with 100-m diameter is the same as that of 100 machines with 10-m diameter. The mass of the plant (in a first-order scaling) varies with the *cube* of the diameter. The aggregate mass of the 100 smaller machines is only 10% of the mass of the larger one. Hence, for the same amount of energy produced, the total equipment mass varies inversely with the diameter. Since costs tend to grow with mass, many small turbines ought to be more economical than one large one. This reasoning would suggest that the best solution is to use an infinite number of infinitely small turbines. Taken to this extreme, the conclusion is patently absurd.

Other factors play an important role in the economy (or diseconomy) of scale for wind turbines, complicating the situation to the point that a plausible model of how energy cost varies with turbine size becomes difficult to construct. Larger machines could arguably be more efficient and might simplify maintenance. They also would require less ancillary equipment (for instance, one single large transformer instead of many small ones) and possibly less land area. Small machines profit from mass production economies, from modularity (allowing an easy expansion of the capacity of a wind farm), and from a greater immunity to breakdown (the breakdown of a few turbines affects only a fraction of the total wind farm capacity). We must leave the verdict to practical experience. Early wind farms favored small turbines (say 100 kW). Progressively, larger machines have been used. In 2006, 3-MW machines were popular, and the trend toward larger power still seems to continue, in a way, vindicating the early government choice.

"Growian"[†] (100-m diameter, 3 MW) and the Boeing Mod-5B (98-m diameter, 3.2 MW) were among the largest early wind turbines.

[†] The German love for acronyms has given us words like *Stuka* and *Flak*. *Growian* stands for "Grosse Wind Energie Anlage."

15.2 Wind Machine Configurations

Several wind machine configurations have been proposed, including

1. drag-type turbines,
2. lift-type turbines (with vertical or horizontal axes),
3. Magnus effect wind plants, and
4. vortex wind plants.

Essentially all present-day wind turbines are of the lift type, and over 90% of these are of the horizontal axis type. Magnus effect and vortex plants have never played a serious practical role.

15.2.1 Drag-Type Wind Turbines

In a drag-type turbine, the wind exerts a force in the direction it is blowing. That is, it simply pushes on a surface as it does in a sailboat sailing before the wind. Clearly, the surface on which the wind impinges cannot move faster than the wind itself.

The ancient Persian wind turbine was a drag-type machine. Figure 15.2 is a sketch of such a mill seen from above. It consisted of a vertical axis to

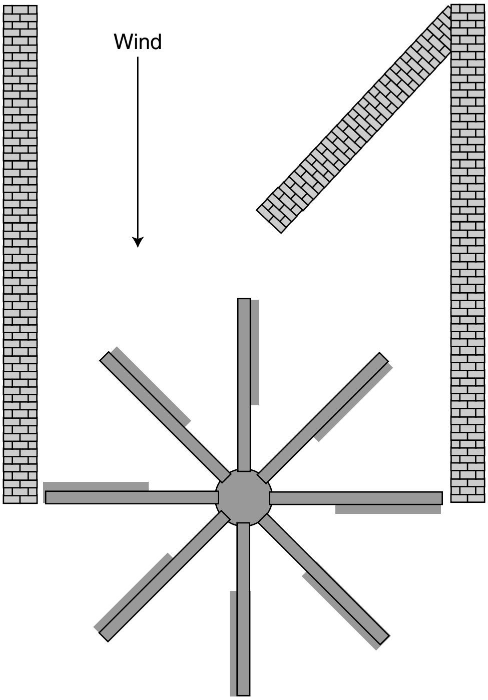


Figure 15.2 Top view of an ancient Persian wind turbine.

which horizontal radial arms were attached. Near the extremities of these arms, a vertical curtain was installed, and this was the surface on which the wind exerted its useful force. Two walls channeled the wind, forcing it to blow on only one side of the device, thus creating a torque. Notice that one wall forms a funnel concentrating the collected wind.

The bucket wind turbine, sketched in Figure 15.3, is another vertical axis drag-type device. It rotates because the convex surface offers less wind drag than the concave one. This device can be cheaply built by amateurs using an oil barrel cut along its vertical axis. It operates inefficiently.

Improved performance can be obtained by staggering the buckets as shown in Figure 15.4 so that a gap is left between them. The air is

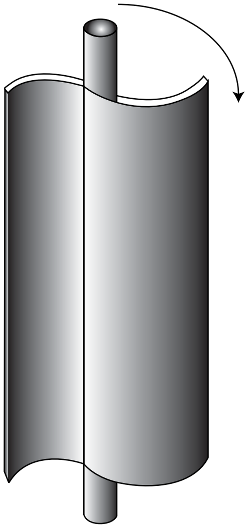


Figure 15.3 A 2-bucket wind turbine.

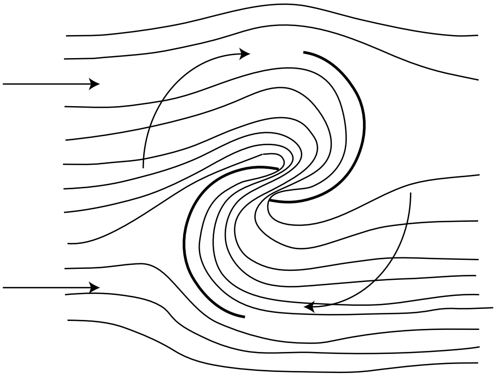


Figure 15.4 Air flow in a Savonius rotor.

accelerated as it passes the gap, reducing the front drag of the convex bucket. It is then blown on the reverse side of the bucket, aiding in the creation of torque. This type of device is called a **Savonius** rotor and actually uses a certain amount of lift (in addition) to drag.

Savonius turbines cannot compete in efficiency with pure lift-type machines, but they are easy to build and find application as sensors in anemometers and as starters for vertical axis lift-type machines.

15.2.2 Lift-Type Wind Turbines

In a lift-type machine, the wind generates a force perpendicular to the direction in which it is blowing. The familiar propeller wind turbines are of the horizontal axis lift-type. All lift-type turbines are analogous to sailboats sailing cross wind. The sailboat (or the blade of the turbine) can move substantially faster than the wind itself. Figure 15.5 shows such turbines.

Notice that the propeller-driven shaft that delivers the collected energy is high above ground level. This usually forces one of two solutions: either the electric generator is placed on top of the tower next to the propeller, or a long shaft, with associated gears, is used to bring the power to a ground-level generator. The first solution, although requiring reinforced towers, is preferred because of the cost and difficulties of transmitting large mechanical power over long shafts. Mounting the generator on top of the tower increases the mass of that part of the system that has to swivel around when the wind changes direction.

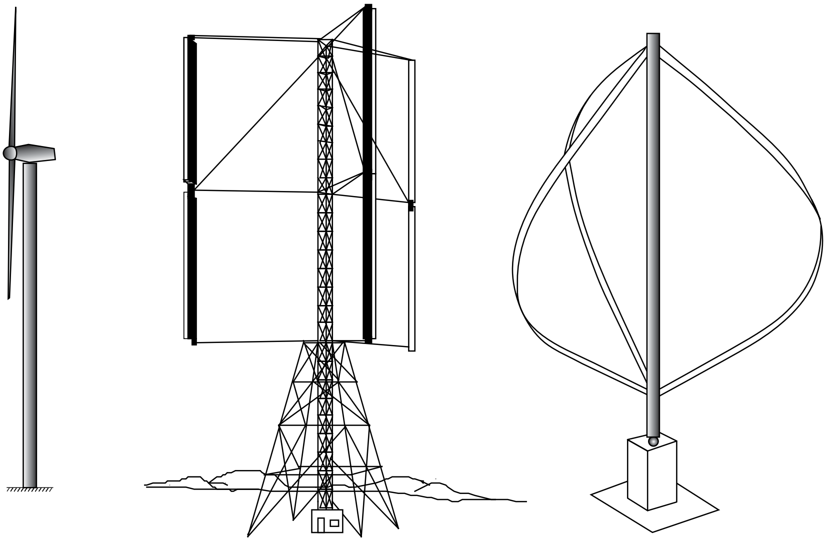


Figure 15.5 From left to right: a horizontal axis (propeller) type turbine, and two vertical axis machines—a Gyromill and a Darrieus.

Some wind turbines have the propeller upstream from the generator and some downstream. It has been found that the upstream placement reduces the noise produced by the machine.

A propeller wind turbine that employs a ground-level generator but avoids the use of a long shaft is the suction-type wind turbine. It resembles a conventional wind turbine but the rotating blades act as a centrifugal pump. The blades are hollow and have a perforation at the tip so that air is expelled by centrifugal action, creating a partial vacuum near the hub. A long pipe connects the hub to an auxiliary turbine located at ground level. The inrushing air drives this turbine. The system does not seem promising enough to justify further development.

One wind turbine configuration not only allows placing the generator on the ground but also avoids the necessity of reorienting the machine every time the wind changes direction: it is the vertical axis lift-type wind turbine. The one illustrated in the center of Figure 15.5 is a design that was proposed by McDonnell-Douglas and was called Gyromill. It would have been capable of generating 120 kW, but it was never commercialized.

One obvious disadvantage of the Gyromill is the centrifugal force that causes the wings to bend outward, placing considerable stress on them. An elegant way to avoid centrifugal stresses is to form the wings in the shape assumed by a rotating rope loosely attached to the top and bottom of the rotating shaft. This leads to the familiar “egg beater” shape and, of course, causes the wing to work only in tension.

The shape of such a rotating rope is called a **troposkein** and closely resembles a **catenary**. There is, however, a difference. The catenary is the shape taken up by a perfectly flexible cord of uniform density hanging freely from two fixed points. Each unit length of the cord is subject to the same (gravitational) force. In the case of the troposkein, the force acting on each section of the cord depends on the distance of the section from the axis of rotation.

The troposkein wing (right-hand drawing of Figure 15.5) was first suggested by a French engineer called Darrieus after whom this type of wind turbine is named.

15.2.3 Magnus Effect Wind Machines

Magnus effect machines have been proposed but look unpromising. This effect, discussed in Section 15.14, is the one responsible for, among other things, the “curve” in baseball.

When a pitcher throws a curve, he causes the ball to spin, creating an asymmetry: one side of the ball moves faster with respect to the air than the other and, consequently, generates the “lift” that modifies the trajectory of the ball. An identical effect occurs when a vertical spinning cylinder is exposed to the wind. The resulting force, normal to the wind direction, has been employed to move sailboats and wind machines.

15.2.4 Vortex Wind Machines

Finally, it is possible to abstract energy from the wind by making it enter tangentially through a vertical slit into a vertical hollow cylinder. As a result, the air inside is forced to gyrate, and the resulting centrifugal force causes a radial pressure gradient to appear. The center of this air column, being at lower than atmospheric pressure, sucks outside air through openings at the bottom of the cylinder. The inrushing air drives a turbine coupled to a generator. The spinning air exits through the open top of the cylinder, forming a vortex continuously swept away by the wind. This type of machine has been proposed by Gruman.

15.3 Measuring the Wind

Later in this chapter, we will show that the power, P_D, a wind turbine delivers is proportional to the cube of the wind velocity:

$$P_D = \frac{16}{27}\frac{1}{2}\rho v^3 A\eta, \tag{15.1}$$

where $\frac{1}{2}\rho v^3$ is the **power density** in the wind, $\frac{16}{27}\frac{1}{2}\rho v^3$ is the **available power density** from the wind, A is the **swept area**, and η is the **efficiency** of the wind turbine.

The mean power output from the wind turbine over a period from 0 to T is proportional to the cube of the **mean cubic wind velocity**, $<v>$:

$$<v> \equiv \left(\frac{1}{T}\int_0^T v^3 dt\right)^{1/3}. \tag{15.2}$$

Anemometers—instruments that measure or record wind velocity—can be used in wind surveys. Anemometric records have to be converted to eolergometric data—that is, data on wind power density. The mean cubic velocity, $<v>$, must be calculated from velocity measurements taken at frequent intervals.

The usual anemometric averages, $\overline{v}$ (the arithmetical averages of v), are not particularly suitable for siting wind turbines. Consider a wind that blows constantly at a speed of 10 m/s (average speed, $\overline{v} = 10$ m/s). It carries an amount of energy proportional to $v^3 = 1000$. A wind that blows at 50 m/s 20% of the time and remains calm the rest of the time also has a $\overline{v}$ of 10 m/s, yet the energy it carries is proportional to $0.2 \times 50^3 = 25{,}000$ or 25 times more than in the previous case. In the first case, $<v> = 10$ m/s, while in the second, $<v> = 29.2$ m/s.

The quantity, $<v>$, can be measured directly by dedicated instruments but is more conveniently derived from anemometers equipped with adequate electronics to process $\overline{v}$ into $<v>$ and store the data for later use.

Eolergometric surveys are complicated by the variability of the wind energy density from point to point (as a function of local topography) and by the necessity of obtaining vertical wind energy profiles. It is important that surveys be conducted over a long period of time—one year at least—so as to collect information on the seasonal behavior.

Values of $\overline{v}$ are easier to obtain than those of $<v>$ and, consequently, there is the temptation to guess the $<v>$ from the $\overline{v}$ values. However, the ratio, $<v>/\overline{v}$ is a function of the temporal statistics of the wind velocity and is strongly site dependent. For perfectly steady winds, this ratio will, of course, be 1. For the extreme case of the example in one of the preceding paragraphs, it is 2.92. When, for lack of better information, one assumes that the wind speed distribution follows the Rayleigh rule, then the ratio is 1.24.

If one could predict the exact behavior of the wind, one would be able to design a wind turbine optimally matching the local conditions. Unfortunately, the wind is notoriously fickle, varying substantially throughout a day, from season to season and even from year to year. This means that even if one has precise data about the wind collected over a full year, there is no guarantee that the next year will be identical.[†] Nevertheless, for planning wind farms, a year's worth of detailed data is useful. Failing this, it is possible to use statistical information to make a somewhat educated guess about wind behavior.

If all one knows about a site is its average wind velocity, $\overline{v}$, one can make the assumption that the wind obeyed a Rayleigh distribution (see, for example, Figure 15.6), that is, that the probability, $p(v)$, of the wind having a velocity, v, is

$$p(v) = \frac{v}{\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{v}{\sigma}\right)^2\right], \tag{15.3}$$

where σ is the **mode** of the distribution, that is, the value at which the probability density function (pdf) peaks. Although σ is not the mean value (in this case, $\overline{v}$), there is a relation between the average wind velocity and the mode of the Rayleigh pdf:

$$\sigma^2 = \frac{2}{\pi}\overline{v}^2. \tag{15.4}$$

Example

A turbine is to be installed at a site in which the average wind speed is 9.6 m/s. Plot the probable distribution of the wind speed throughout the year. What is the probability of having a 12-m/s wind?

(*Continues*)

[†] Portnyagin et al. (2006) point out that very strong year-to-year wind variations makes the estimation of long-term wind behavior a difficult task.

(Continued)

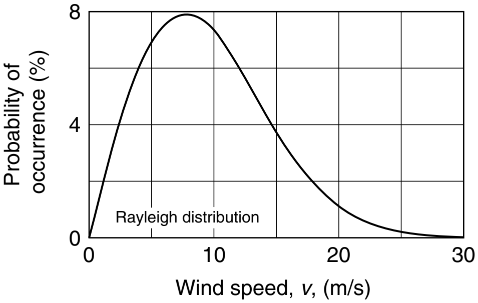


Figure 15.6 When only the average wind velocity is known, the best guess one can make is that the wind obeys a Rayleigh distribution.

Simply plot $p(v)$ as a function of v, using

$$p(v) = \frac{\pi}{2}\left(\frac{v}{9.6^2}\right)\exp\left[-\frac{\pi}{4}\left(\frac{v}{9.6}\right)^2\right]. \tag{15.5}$$

From the tabulation used to plot the graph, we find that there is a 6% probability of having a 12-m/s wind. However, this is not a very useful piece of information. A better question would be, “how many hours per year does the wind blow faster than 12 m/s?”

To answer this question, we must use the formula for the **cumulative Rayleigh function**,

$$F(v) = \exp\left[-\frac{\pi}{4}\left(\frac{v}{9.6}\right)^2\right] = 0.29, \tag{15.6}$$

where $F(v)$ is the probability of experiencing a wind larger than or equal to v; that is, there is a 29% probability of the wind being faster than or equal to 12 m/s. Since there are 8760 hours in a year, such a wind should blow $0.29 \times 8760 = 2568$ hours in a year.

If the cut-in wind velocity (the velocity at which the turbine starts to generate electricity) is 3.5 m/s, we can find, using Equation 15.6, that the turbine will produce a useful output 90% of the time.

Under the simplistic assumption of the Rayleigh distribution, the expected value of the cube of the cubic mean velocity is nearly twice the cube of the average velocity,

$$<v>^3 \approx 1.9(\overline{v})^3 \quad \text{(assuming a Rayleigh distribution).} \tag{15.7}$$

The Rayleigh statistic, being a single-parameter function, is but a poor representation of the wind behavior. It is somewhat better to match the observed wind data to a **two-parameter Weibull distribution function** in which the data for each site are characterized by an adjustable **shape factor**, k, in addition to a **scale factor**, c. Indeed, the shape of the Weibull distribution function (see Equation 15.8) depends on k, as illustrated in Figure 15.7:

$$p(v) = \frac{k}{c}\left(\frac{v}{c}\right)^{k-1} \exp\left[-\left(\frac{v}{c}\right)^{k}\right]. \tag{15.8}$$

The corresponding cumulative Weibull function is

$$F(v) = \exp\left[-\left(\frac{v}{c}\right)^{k}\right]. \tag{15.9}$$

Here, again, $F(v)$ is the probability that the wind velocity is equal to or larger than a chosen value, v.

It should be noticed that the scale factor, c, is *very roughly* equal to the average wind speed, $\overline{v}$. More accurately, the ratio, $c/\overline{v}$, is given by the approximation

$$\frac{c}{\overline{v}} = \left(0.568 + \frac{0.433}{k}\right)^{-\frac{1}{k}}, \tag{15.10}$$

which depends only weakly on k. In the plots of Figure 15.7, we used $c = 10$, in all cases. For the Rayleigh case,

$$c^2 = 1.275\overline{v}^2. \tag{15.11}$$

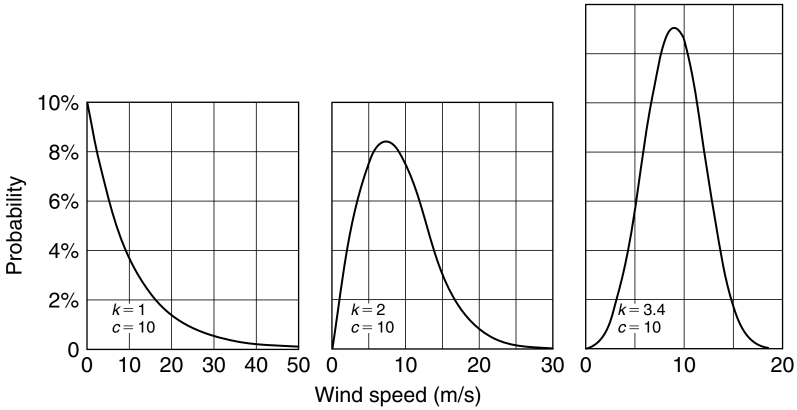


Figure 15.7 The Weibull distribution function can assume different shapes depending on the parameter, k. For example, when $k = 1$, it reduces to an exponential distribution function (left); when $k = 2$, it becomes the Rayleigh function (center), and, when $k = 3.4$, it mimics the normal distribution (right).

Weibull parameters are listed for a number of sites. They refer to statistics taken at a given reference height. If the proposed turbine is to be installed at a different height, corrections to the listed values of k and of c can be made. Appropriate formulas can be found, among other places, in Eggleston and Stoddard (1987).

When no information on the variation of wind velocity with height above ground is available, one can use the scaling formula

$$v(h) = v(h_0)\left(\frac{h}{h_0}\right)^{1/7}. \tag{15.12}$$

15.4 Availability of Wind Energy

In Chapter 1, we saw that 30% of the 173,000 TW of solar radiation incident on Earth is reflected back into space as the planetary albedo. Of the 121,000 TW that reach the surface, 3% (3600 TW) are converted into wind energy, and 35% of this is dissipated in the lower 1 km of the atmosphere. This corresponds to 1200 TW. Since humanity at present uses only some 15 TW, it would appear that wind energy alone would be ample to satisfy all of our needs.

This kind of estimate can lead to extremely overoptimistic expectations. For one thing, it is difficult to imagine wind turbines covering all the ocean expanses. If we restrict wind turbines to the total of land areas, we would be talking about 400 TW. Again, it would be impossible to cover all the land area. Say that we would be willing to go as far as 10% of it, which is more than the percentage of land area dedicated to agriculture. We are now down to 40 TW. But owing to the cubic dependence of power on wind speed, it is easy to see that much of this wind energy is associated with destructive hurricane-like winds, which actually generate no energy, since any reasonable wind turbine must shut itself down under such conditions.

The difficulty is that wind energy is very dilute. At a 10-m height, the wind-power density may be some 300 W/m^2 at good sites and, at 50 m height, it can reach some 700 W/m^2, as it does in the San Gregorio Pass in California and in Livingston, Montana. Notice that these are values of the power in the wind, not of the *available* wind-power densities, which are only 59% (16/27) of the quoted values. (See Subsection 15.6.5). In addition, one has to remember that only a fraction of the available power can actually be harnessed.

Currently, the only widely available datum about the wind is the mean wind speed, and this carries limited information about the power that can be generated. A system of site classification has been adopted, which is little more than an indication of what the average wind speed is. The value of this average velocity is transformed into an expected value of the cubic mean velocity under the assumption that the distribution can be represented by a Rayleigh function. In that case, there is a fixed relationship between the

Table 15.1 Wind Power Classes

Wind power class	Wind power density (W/m^2)	Mean wind speed (m/s)
1	200	5.6
2	300	6.4
3	400	7.1
4	500	7.6
5	600	8.1
6	800	8.9
7	2000	12.1

cube of the average velocity, $(\overline{v})^3$ and the cube of the mean cubic velocity, $<v>^3$. (See Equation 15.7.)

The power density in the wind is calculated from $<v>^3$ and is used to determine the "class" of the site,

Table 15.1 is for a hub height of 50 m.

Any sizable plant requires a large collecting area, which means many turbines spread over a large area of land.[†]

Winds tend to be extremely variable so that wind plants must be associated with energy storage facilities (usually in small individual installations) or must debit into the power distribution network (in large wind farms).

One can take advantage of the lack of correlation of wind behavior at different sites by interconnecting different wind farms so as to reduce the fluctuations in the amount of power generated.

In some areas of the world where the wind is quite constant in both speed and direction, wind turbines could operate with greater efficiency. All along the northeastern coast of Brazil, trade winds blow almost uninterruptedly from a northeasterly direction with steady speeds of some 13 knots (about 7 m/s), which corresponds to 220 W/m^2 of wind power density or 130 W/m^2 of *available* power density. This constant wind direction would allow the construction of fixed wind concentrators capable of substantially increasing the wind capture area.

15.5 Wind Turbine Characteristics

To minimize the cost of the produced electricity, the rated power of the generator must bear an appropriate relationship to the swept area of the wind turbine. The rated power of a generator is the maximum power it can

[†] Of course, land used for wind farms may still be available for agriculture.

deliver under steady conditions. Usually, the power actually generated is substantially lower than the rated. The ratio of the rated generator power to the swept area of the turbine is called **specific rated capacity** or **rotor loading**. Modern machines use rotor loadings of 300 to 500 W/m^2.

The ideal wind turbine would be tailored to the wind conditions of each site. With steady unvarying winds ($<v>/\overline{v} = 1$), the rated power of the generator should be the same as that of the turbine. However, when the ratio is larger than 1, which is invariably the case, the most economical combination has to be determined by considering many different factors. See, for example, Problem 15.3.

The GE 3.6 MW wind turbine is of the horizontal-axis (propeller) type. The rotor has a diameter of 104 m, and the swept area is 8,495 m^2. The performance of this wind turbine is shown in Figure 15.8. Note that with wind speeds below 3.5 m/s, the propeller does not rotate; the power output rises rapidly as the wind speed increases and rated power output is achieved when the wind speed is 14 m/s, and then remains constant up to 27 m/s. Above this speed, the machine shuts itself down for safety reasons. Rotor speed is variable—8.5 to 15.3 rpm. Power control is achieved by active blade pitch adjustment. At high wind speeds, the turbine extracts but a small fraction of the available energy. The GE 3.6 generated 3.6 MW of electricity with any wind in the 14 m/s to 27 m/s range, notwithstanding there being seven times more energy at the higher speed. At even higher wind speeds, the machine delivers no energy at all, just when there is a largest amount of power in the wind.

Typically, the power delivered by a given wind plant depends on the wind velocity in a manner similar to the one displayed in Figure 15.8.

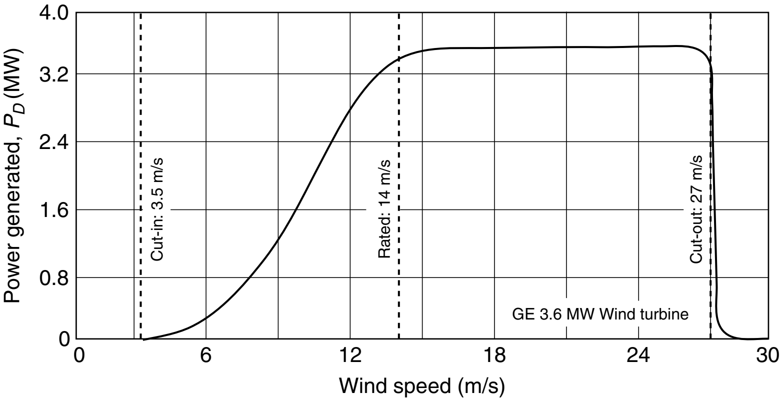


Figure 15.8 Power output of the GE 3.6 MW wind turbine as a function of wind velocity.

Wind turbines frequently deliver their energy to a utility-operated net and must do so with alternating current of the correct frequency. There are two general solutions to the synchronization problem:

1. Maintaining the rotation of the turbine at a constant rate (by changing the blade pitch, for instance).
2. Allowing the turbine to rotate at the speed dictated by load and wind velocity. In this case, dc is generated and electronically "inverted" to ac. Such **variable-speed** machines are somewhat more expensive but are more efficient and have a longer life.

15.6 Principles of Aerodynamics

The symbol, P, in this chapter stands for both *power* and *power density*—that is, power per unit area, depending on the context. The lower case, p, is reserved for *pressure.*

The following subscripts are used:

$P_W = \frac{1}{2}\rho v^3$ "Power density in the wind." This is the amount of energy transported across a unit area in unit time.

$P_A = \frac{16}{27}\frac{1}{2}\rho v^3$ "Available power density." This is the theoretical maximum amount of power that can be extracted from the wind.

$P_D = \frac{16}{27}\frac{1}{2}\rho v^3\, A\eta$ "Power delivered." This is the power that a wind turbine delivers to its load.

15.6.1 Flux

The flux of a fluid is defined as the number of molecules that cross a unit area (normal to the flow) in unit time. It can be seen that if n is the concentration of the molecules (number per unit volume) and v is the bulk velocity of the flow, then the flux, ϕ, is

$$\phi = nv \quad \mathrm{m}^{-2}\mathrm{s}^{-1}. \tag{15.13}$$

Consequently, the total flow across an area, A, is

$$\Phi = \phi A \quad \mathrm{s}^{-1}. \tag{15.14}$$

15.6.2 Power in the Wind

If the mean mass of the gas molecules is m, then the mean energy of a molecule owing to its bulk drift (not owing to its thermal motion) is $\frac{1}{2}mv^2$. The amount of energy being transported across a unit area in unit time is the power density of the wind:

$$P_W = \frac{1}{2}mv^2\phi = \frac{1}{2}mnv^3 = \frac{1}{2}\rho v^3 \,\mathrm{W\,m^{-2}}. \tag{15.15}$$

Notice that the power density is proportional to the cube of the wind velocity. The quantity, ρ, is the gas density—that is, the mass per unit volume:

$$\rho = mn \quad \mathrm{kg\,m^{-3}}. \tag{15.16}$$

At the reference temperature and pressure (RTP),[†] the density of air is

$$\rho = \frac{0.2 \times 32 + 0.8 \times 28}{24.5} = 1.18 \approx 1.2\,\mathrm{kg\,m^{-3}}. \tag{15.17}$$

The numerator is the average molecular mass of air containing 20% O_2 and 80% N_2, by volume. The denominator is the number of cubic meters per kilomole at RTP:

From the perfect gas law, at RTP,

$$V = \frac{RT}{p} = \frac{8314 \times 298.3}{1.013 \times 10^5} = 24.5\,\mathrm{m^3}. \tag{15.18}$$

Owing to the variability of air pressure, there is, at this stage, no point in calculating the air density with more precision than two significant figures.

15.6.3 Dynamic Pressure

Since $1\,\mathrm{m^3}$ of gas contains n molecules and each molecule carries $\frac{1}{2}mv^2$ joules of energy owing to its bulk motion, the total energy density—that is, the total energy per unit volume—is

$$W_d = \frac{1}{2}nmv^2 = \frac{1}{2}\rho v^2 \,\mathrm{J\,m^{-3}} \text{ or } \mathrm{N\,m^{-2}}. \tag{15.19}$$

Energy per unit volume has the dimensions of force per unit area—that is, of pressure. Thus, W_d is referred to as **dynamic pressure**.

[†] RTP (reference temperature and pressure) corresponds to 1 atmosphere and 25 Celsius or 298.3 kelvin, while STP (standard temperature and pressure) corresponds to 1 atmosphere and 0 Celsius.

15.6.4 Wind Pressure

Wind exerts pressure on any surface exposed to it. Consider the (unrealistic) flow pattern depicted in Figure 15.9. The assumption is that any molecule striking the surface is reflected and moves back against the wind without interfering with the incoming molecules.

Under such a simplistic assumption, each molecule transfers to the surface a momentum, $2mv$, because its velocity change is $2v$ (it impacted with a velocity, v, and was reflected with a velocity, $-v$). Since the flux is nv, the rate of momentum transfer per unit area, that is, the generated pressure, is $2mv \times nv = 2\rho v^2$. The assumption is valid only at very low gas concentrations, when, indeed, a molecule bouncing back may miss the incoming ones.

In a more realistic flow, the reflected molecules will disturb the incoming flow, which would then roughly resemble the pattern shown in Figure 15.10. This leads to a pressure smaller than that from the ideal flow case, a pressure that depends on the shape of the object. To treat this complicated problem, aerodynamicists assume that the real pressure is equal to the dynamic pressure multiplied by an experimentally determined correction factor, C_D, called the **drag coefficient**:

$$p = \frac{1}{2}\rho v^2 C_D. \tag{15.20}$$

The drag coefficient depends on the shape of the object and, to a certain extent, on its size and on the flow velocity. This means, of course, that

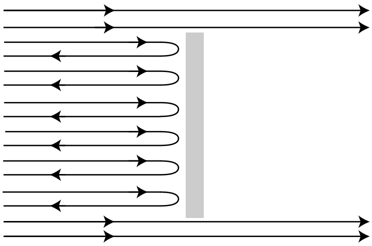


Figure 15.9 A simplistic flow pattern.

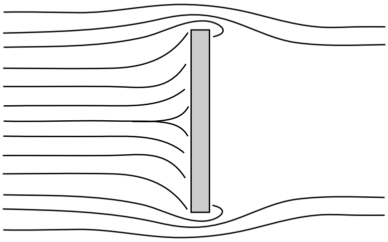


Figure 15.10 A more realistic flow pattern.

the pressure exerted by the wind on a surface is not strictly proportional to v^2 as suggested by Equation 15.20.

The drag coefficient of a large flat plate at low subsonic velocities is usually taken as $C_D = 1.28$.

15.6.5 Available Power (Betz Limit)

Electrical engineers are familiar with the concept of available power. If a source (see Figure 15.11) has an open-circuit voltage, V, and an internal resistance, R_s, the maximum power it can deliver to a load is $V^2/4R_s$. This occurs when $R_L = R_s$.

The same question arises when power is to be extracted from the wind. If the surface that interacts with the wind is stationary, it extracts no power because there is no motion. If the surface is allowed to drift freely downwind without any resistance, then it again will extract no power because the wind will exert no force on it. Clearly, there must be a velocity such that maximum power is extracted from the wind.

The power density, P, extracted from the wind is the product of the pressure, p, on the surface and the velocity, w, with which the surface drifts downwind. The wind pressure is

$$p = \frac{1}{2} p C_D (v - w)^2; \qquad (15.21)$$

hence,

$$P = pw = \frac{1}{2} \rho C_D (v - w)^2 w. \qquad (15.22)$$

Setting $\partial P/\partial w$ to zero, an extremum of P is found. This is a maximum and occurs for $w = v/3$ independently of the value of C_D. Thus,

$$P_{max} = \frac{2}{27} \rho C_D v^3 \,\mathrm{W\,m^{-2}}. \qquad (15.23)$$

The ratio of maximum extractable power to power in the wind is

$$\frac{P_{max}}{P_w} = \frac{\frac{2}{27}\rho C_D v^3}{\frac{1}{2}\rho v^3} = \frac{4}{27} C_D. \qquad (15.24)$$

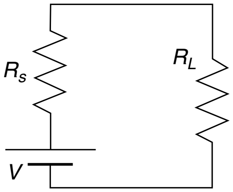


Figure 15.11 An electric source and its load.

The largest possible value of C_D is that predicted by our simplistic formula, which by stating that $p = 2\rho v^2$, implies that $C_D = 4$. Thus, at best, it is possible to extract 16/27 or 59.3% of the "power in the wind." This is the **available power density** from the wind:

$$P_A = \frac{16}{27}\frac{1}{2}\rho v^3. \tag{15.25}$$

The factor 16/27 is known as the **Betz limit** or **Betz efficiency**, η_{Betz}, and was discussed in a 1919 book by the German physicist, Albert Betz.[†]

The Betz limit applies to any type of wind-driven machine, and not only to drag types as in the previous derivation. We will consider here, as did Betz, a horizontal axis turbine (rotor) whose blades define a swept area—a vertical disk normal to the flow of the (horizontal) wind. Upwind, a certain distance from this disk, wind flows undisturbed with a velocity, v_1; sufficiently downwind, in the slip stream, the air flow is substantially slower—it has a velocity, v_3, because the rotors constitute a retarding obstacle to the airflow. We will reproduce here the derivation of the Betz limit following the steps in Betz's book (but using slightly different symbology closer to the one adopted in the current volume).

Betz starts with the "reasonable" assumptions that, at the disk, the flow velocity, v_2, is the average of the upstream and downstream velocities,

$$v_2 = \frac{v_1 + v_3}{2}. \tag{15.26}$$

Though reasonable, this is not obvious and requires demonstration. This is the **Rankine–Froude theorem** and is discussed in the Betz book and later, also in this chapter.

There can be no change in the velocity as the air flows across the disk (otherwise, air would either pile up or leave a vacuum). Thus, slightly ahead and slightly behind the disk, the velocity is the same (v_2).

The rate at which mass streams across the swept area, A, is

$$\dot{m} = \rho A v_2 = \rho A \frac{v_1 + v_3}{2} \text{ kg/s}. \tag{15.27}$$

Upstream, a mass, m, moving with a velocity, v_1, carries an energy $\frac{1}{2}mv_1^2$, and, if the rate of mass transport is $\dot{m}$, it carries a power $\frac{1}{2}\dot{m}v_1^2$.

[†] In the history of science, there a several instances in which a given discovery is named after a scientist who was clearly not the first to discuss it. The empirical rule that tells us the relative distance from the sun of the different planets in our system is known as Bode's law, notwithstanding the fact that Johann Daniel Titus announced it in 1766, whereas Johann Elert Bode simply popularized the notion six year later. A much more extreme case is that of the heliocentric theory, attributed to Copernicus (born 1473) but mentioned 18 centuries earlier by Aristarchus of Samos (born 310 BC). It appears that Betz's limit was first derived by the Englishman Frederick William Lanchester some five years earlier.

Far downstream, the power is $\frac{1}{2}\dot{m}v_3^2$. Therefore, the disk removed from the wind a power

$$P = \frac{1}{2}\dot{m}(v_1^2 - v_3^2) = \frac{1}{4}\rho A(v_1 + v_3)(v_1^2 - v_3^2). \tag{15.28}$$

Compare this extracted power with the power, $P_W A$, in the undisturbed wind,

$$\frac{P}{P_W} = \frac{\frac{1}{4}\rho A(v_1 + v_3)(v_1^2 - v_3^2)}{\frac{1}{2}\rho A v_1^3} = \frac{1}{2}\left[1 + \frac{v_3}{v_1}\right]\left[1 - \left(\frac{v_3}{v_1}\right)^2\right]. \tag{15.29}$$

This ratio is a function of v_3/v_1 and reaches a maximum when $v_3/v_1 = 1/3$; that is, when the far downstream velocity is one-third that of the undisturbed wind velocity.

The maximum power abstracted from the wind is, from Equation 15.29,

$$\frac{P_{max}}{P_W} = \frac{1}{2}\left[1 + \frac{1}{3}\right]\left[1 - \left(\frac{1}{3}\right)^2\right] = \frac{16}{27}. \tag{15.30}$$

which is the Betz limit.

15.6.5.1 The Rankine–Froude† theorem

As the wind blows through the swept area of a wind turbine, it exerts a drag force, F, on the machine, a force proportional to the pressure difference, Δp, that the wind builds up. In fact, although the wind turbine as a whole is surrounded by air at atmospheric pressure, p_1, the pressure is elevated (p_{2_1}) on the upstream side of the disk and reduced (p_{2_2}), on the downwind side, as suggested in Figure 15.12. The mentioned drag force is

$$F = ((p_{2_1}) - (p_{2_2}))A = \Delta p A, \tag{15.31}$$

where A, as before, is the swept area.

On the other hand, the wind velocity is continuous across the disk and has a value, v_2, which, owing to the drag exerted by the disk, is less than the upstream value, v_1, and more than the far slipstream value, v_3, as argued previously. We want to determine the relationship of v_2 to v_1 and to v_3. To accomplish this, we will resort to Bernoulli's principle. Far upstream the pressure is p_1 (atmospheric) and the velocity is v_1, while at the disk the pressure is p_{2_1} and the velocity is v_2:

$$p_1 + \frac{1}{2}\rho v_1^2 = p_{2_1} + \frac{1}{2}\rho v_2^2. \tag{15.32}$$

†Pronounced *Frood.* The theorem is named after the English hydrodynamicist, William Froude.

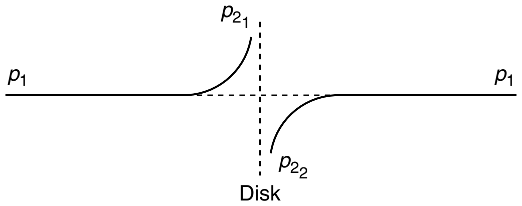


Figure 15.12 Although the static pressure around a wind turbine is mostly that of the undisturbed atmosphere, the presence of the rotor causes the pressure to rise upstream and fall downstream so that a Δp is established.

On the downstream side of the disk,

$$p_1 + \frac{1}{2}\rho v_3^2 = p_{2_2} + \frac{1}{2}\rho v_{2_1}^2. \tag{15.33}$$

The pressure difference, $\Delta p = p_1 - p_2$ (subtracting Equation 15.33 from Equation 15.32), is

$$\Delta p = \frac{1}{2}\rho(v_1^2 - v_3^2), \tag{15.34}$$

and the force on the disk is

$$F = \frac{1}{2}\rho A(v_1^2 - v_3^2). \tag{15.35}$$

Next, we must express F as a function of the air-flow velocity, v_2, across the swept area.

The mass flow rate upstream is $\rho A v_1$ kg/sec, whereas that at the far slipstream is $\rho A v_3$ kg/sec. This means that $\rho A(v_1 - v_3)$ kg/sec interacted with the disk where its flow velocity is v_2 m/sec. Thus,

$$F = \rho A(v_1 - v_3)v_2 \,\text{kg/sec}. \tag{15.36}$$

Comparing Equation 15.36 to Equation 15.35,

$$\rho A(v_1 - v_3)v_2 = \frac{1}{2}\rho A(v_1^2 - v_3^2), \tag{15.37}$$

from which

$$v_2 = \frac{v_1 + v_3}{2}. \tag{15.38}$$

Recapitulating, when a wind turbine extracts maximum power from a wind of velocity, v_1, the velocity at the swept area is $2v_1/3$ and at the far slipstream is $v_1/3$.

$$v_2 = 2\frac{v_1}{3} \tag{15.39}$$

$$v_3 = \frac{v_1}{3}, \tag{15.40}$$

and the power extracted is 16/27 of the power in the wind.

15.6.6 Efficiency of a Wind Turbine

The efficiency of a wind turbine is the ratio of the power, P_D, delivered to the load, to some reference power. There is a certain amount of arbitrariness in this definition. Some authors choose the "power in the wind" ($\frac{1}{2}\rho v^3$) as reference. Alternatively, one can use as reference the available power, $P_A = \frac{16}{27}\frac{1}{2}\rho v^3$. When talking about the efficiency of a wind turbine, it is important to specify clearly what the reference power is.

In a perfectly lossless wind machine, the power that can be generated, $P_{D_{ideal}}$, is, as we saw,

$$P_{D_{ideal}} = \frac{16}{27}\frac{1}{2}\rho v^3 A, \tag{15.41}$$

and the corresponding efficiency (using P_W as reference) is

$$\eta = \frac{P_D}{P_W} = \frac{16}{27} \equiv \eta_{Betz}. \tag{15.42}$$

Here, we can draw an analogy to the case of heat engines. Independently of the type of engine, the efficiency of a heat engine is limited by the Carnot ratio, so that the realized efficiency, η, can be written as

$$\eta = \frac{T_H - T_C}{T_H}\eta^* = \eta_{Carnot}\,\eta^*. \tag{15.43}$$

In a similar fashion, the efficiency of a wind turbine of whatever type is always limited by the Betz factor and can be written as

$$\eta = \frac{16}{27}\frac{1}{2}\rho v_1^3 \eta^* = \eta_{Betz}\,\eta^*, \tag{15.44}$$

where η^*[†] represents the performance of the particular turbine relative to the performance of the **ideal Betz turbine**. It goes without saying that no actual turbine has the Betz efficiency (just as no actual heat engine has the Carnot efficiency). In a well-designed wind turbine, the efficiency, η^*, can reach 0.7 but most often is between 0.4 and 0.6.

[†] The quantities we refer to as efficiencies, (because they are ratios of performances) are called power coefficient by many authors and are represented by the letter "c". We avoid this symbology because "c" is also used to represent the chord of the airfoil and the coefficient of lift or of drag.

15.6.6.1 Solidity

In a propeller-type turbine, the rotating blades describe a circular area (the **swept area**, A_v), which determines the power in the wind being intercepted. However, the total area occupied by the blades themselves (NA_p, where N is the number of blades and A_p is the area of each blade) is much smaller than A_v. The ratio between these two areas is called the **solidity**,

$$S = \frac{NA_p}{A_v}. \tag{15.45}$$

In a vertical axis turbine, an entirely equivalent solidity is defined.

The force the wind exerts on the blades, and, consequently, the torque on the propeller shaft, is obviously proportional to solidity. Yet, over a reasonable range of solidity, the efficiency of the turbine is independent of this parameter—the power, for a given wind velocity, depends only on the swept area. This means that increasing the solidity increases the torque but correspondingly decreases the rate of rotation: high-solidity turbines have great torque but rotate slowly.

In a low-solidity turbine, the disturbance created in the wind flow by the passage of a blade has been swept away by the moving air by the time the next blade comes around. In high-solidity machines, there can be considerable interference between the blades, reducing their individual efficiency and, thus, roughly compensating the increase in torque. This results in an approximate compensation that tends to ensure the mentioned independence of overall generated power on the value of solidity. Economic considerations dictate the choice of a few slender blades rather than a large number of wide ones.

15.6.6.2 Wake Rotation

When we derived the Betz limit, we assumed that all the air motion in a propeller-type turbine was axial, that is, perpendicular to the swept area. This is true as far as the upstream flow (the flow of air before interaction with the propeller) is concerned. However, the propeller extracts energy from the wind by developing a torque. Obeying Newton's law that calls for a reaction for each action, the propeller must induce a corresponding torque on the air flow, causing the wake to rotate. This swirling motion involves energy which is not captured by the propeller shaft—it constitutes loss.

Since the generated power is equal to the product of the torque, Υ, times the angular velocity, ω,

$$P_g = \Upsilon\omega, \tag{15.46}$$

for a given power, the larger the angular velocity, the smaller the torque and, consequently, the smaller the losses associated with wake rotation.

In the preceding subsubsection, we argued that the lower the solidity (up to a point), the lower the torque and, therefore, the lower the wake rotation loss.

Because, the angular velocity increases with the increasing **tip speed ratio**,

$$\lambda \equiv \frac{\omega R}{v_1}, \tag{15.47}$$

it follows that the larger λ, the smaller the wake rotation losses, P_{wake}.

The magnitude of wake rotation losses can be estimated theoretically,[†] but because the derivation is somewhat complicated, we will limit ourselves to quoting the results. When wake rotation losses are included, the power a wind turbine generates is given by

$$P_D = \frac{1}{2}\rho v_1^3 A_p \eta_{Schmitz}, \tag{15.48}$$

where $\eta_{Schmitz}$ replaces η_{Betz}.

To calculate η_{wake}, we define

$$\psi \equiv \arctan\left(\frac{1}{\lambda}\frac{R}{r}\right). \tag{15.49}$$

λ is the tip speed ratio, and r/R is our normalized variable of integration: the distance from the hub divided by the rotor blade length. r/R has a range from 0 to 1.

$$\eta_{wake} = 4\lambda \int\limits_0^1 \left(\frac{r}{R}\right)^2 \frac{\sin^3\left(\frac{2}{3}\psi\right)}{\sin^2(\psi)} d\left(\frac{r}{R}\right). \tag{15.50}$$

It is best to evaluate this integral numerically. The results are plotted in Figure 15.13 as a function of the tip speed ratio. For large values of the tip speed ratio, the wake efficiency approaches the Betz efficiency it replaced. In other words, the wake rotation losses become small when the λ is large.

15.6.6.3 Other losses

The Betz efficiency is entirely independent of the type of turbine; the wake rotation loss applies only to propeller-type machines, and, in these, it depends on the tip speed ratio.

[†] In 1935, H. Glauert, studied the effect of wake rotation in airplane propellers, and in 1956, G. Schmitz, did the same for wind turbines.

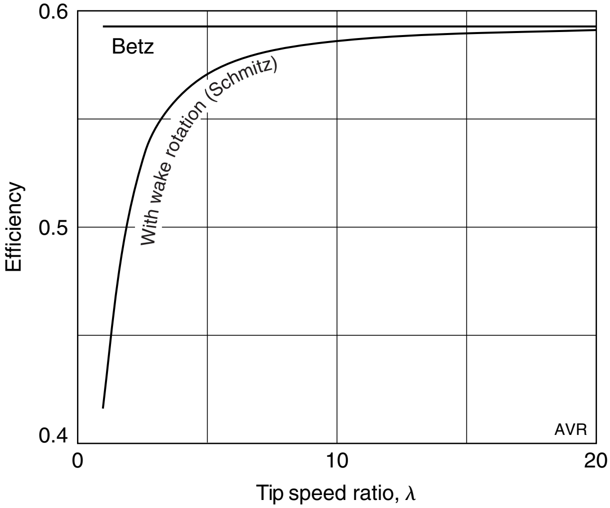


Figure 15.13 Wake efficiency is low in slow rotating turbines such as the traditional vane type. Modern electricity-generating turbines, which typically use $\lambda \approx 7$, approach the Betz efficiency.

Just as in any other device, wind turbines are plagued by a number of additonal losses including **wing tip losses** that, somewhat counterintuitively, tend to decrease when the number of blades—hence, the number of tips—increases. See Section 15.9, for an explanation. In addition, there are losses associated with the gear train, the generator, the bearings, and so on.

15.7 Airfoils

Airplane wings, helicopter rotors, empennage surfaces, and propeller blades are examples of aerodynamic surfaces (**airfoils**) that must generate a great deal of lift with a correspondingly small drag. The performance of an airfoil depends greatly on the shape of its cross section.

Figure 15.14 shows a section through an airfoil. The line (A.A′) represents the trace of an arbitrary **reference plane**. The region above this plane differs from the one below it—the airfoil is **asymmetric**. In **symmetric** airfoils, the reference plane is the plane of symmetry, and the region above it is a mirror image of that below.

When air flows relative to the airfoil along the x-axis in the figure, a force is exerted on the foil. Such force is usually decomposed into a **lift** component (normal to the velocity) and a **drag** component (parallel to the velocity). The corresponding pressures are indicated by the vectors, p_L and p_D, in the figure. The angle between the wind direction and the reference line is called the **angle of attack**, α.

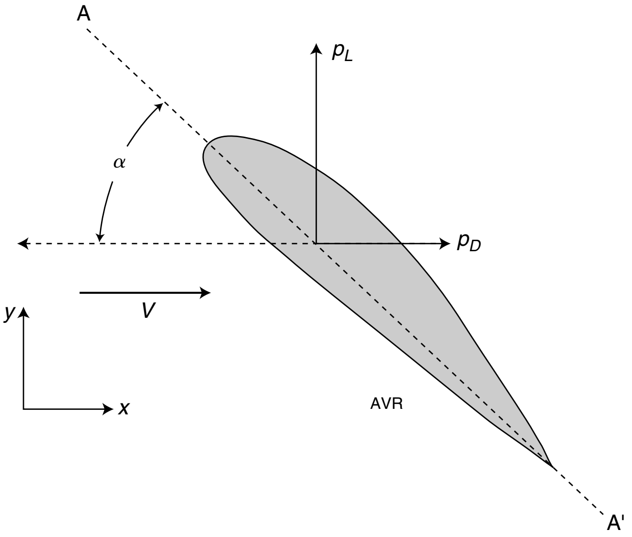


Figure 15.14 Pressure on an airfoil.

For each shape of the airfoil, p_L and p_D are determined experimentally in wind tunnels under specified conditions. The observed pressures are related to the dynamic pressure, $\frac{1}{2}\rho v^2$, by proportionality constants, C_L and C_D, called, respectively, the **lift coefficient** and the **drag coefficient**:

$$p_L = \frac{1}{2}\rho v^2 C_L \text{ and} \tag{15.51}$$

$$p_D = \frac{1}{2}\rho v^2 C_D. \tag{15.52}$$

These coefficients are functions of the angle of attack[†] and can be found in tabulations, many of which were prepared by National Advisory Committee for Aeronautics (NACA) (the forerunner of NASA) in the United States and by the University of Göttingen in Germany. An example of such a tabulation is found in Section 5.10, "Wind Turbine Analysis," while a plot of the lift coefficient of the Göttingen 420 airfoil can be seen in Figure 15.15. Since, airplanes presumably, move only forward, the tabulations are usually made for only a small range of angles of attack near zero. However, for some airfoils, data are available for all 360° of α as in Figure 15.16, which shows the dependence of C_L on α for the airfoil known as Göttingen-420.

[†] The coefficients also depend on the Reynolds number, R, and the aspect ratio, Λ, though more weakly, as we are going to show.

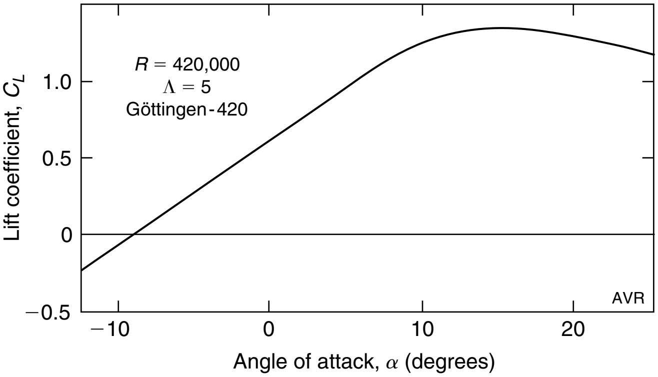


Figure 15.15 The lift coefficient of the Göttingen-420 airfoil.

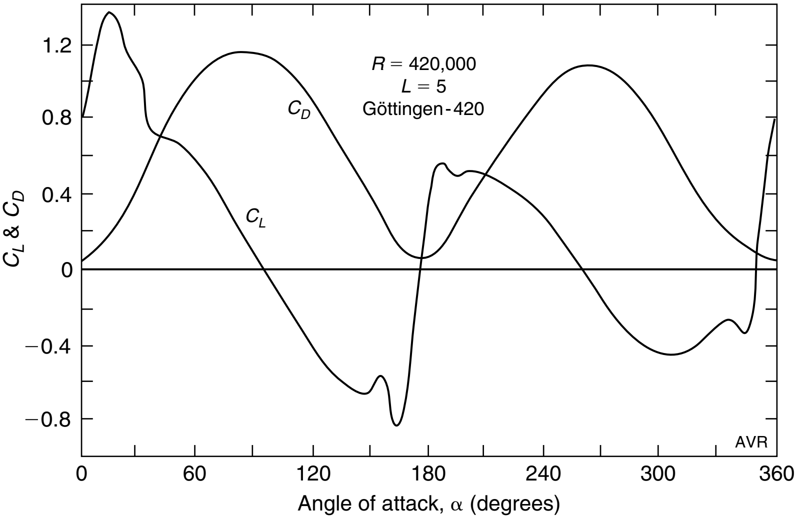


Figure 15.16 The lift and drag coefficients of the Göttingen-420 airfoil.

Notice that the airfoil under discussion generates lift even with negative angles of attack (as long as they remain small). When there is a positive angle of attack, one can intuitively understand the creation of lift even for a flat surface. After all, the air flow is hitting the surface from below, and the drag it exerts has a lift component. However, when the angle of attack is zero or slightly negative, the lift must be due to more complicated mechanisms. Observations show that the air pressure immediately above the airfoil is smaller than the pressure immediately below it, and this is the obvious cause of the lift. The problem is to explain how such a pressure difference comes about.

Further observations also show that

1. the air flow tends to follow the curvature of the top of the airfoil instead of simply being deflected away from it; and
2. the air-flow velocity is substantially increased, lowering the pressure, according to Bernoulli's principle.

We need some explanation for these observations.

The Coanda[†] Effect is the cause of the air flow's tendency to follow the shape of the airfoil. It is extremely easy to demonstrate this phenomenon. Open a faucet and allow a thin stream of water to fall from it. Now take a curved surface—a common drinking glass will do—and let the stream hit the side of the glass at a glancing angle. The water will run along the side and then, making a sharp turn, will flow along the bottom instead of simply falling vertically down. The water tends to follow the glass surface just as the air tends to follow the airfoil surface.

The bulging part of the airfoil restricts the air flow (as in the strangulation in a venturi), causing an acceleration of the flow. The transit time of air molecules along the path over the airfoil from leading to trailing edge is not the same as that for the flow under the airfoil. Such synchronism, frequently invoked in explanations of wing lift, does, in fact, not occur.

In the airfoil of Figure 15.16, the lift is linearly related to α up to some 10°. At higher angles the airfoil **stalls**—that is, a further increase in α actually reduces the lift. Near zero angle of attack, this airfoil develops a lift over 16 times larger than its drag.

Because the lift and drag coefficients are not strictly independent of the air velocity or the dimensions of the wing, it is impossible to scale any experimental results exactly. However, the data are valid for different sizes and speeds as long as the **Reynolds number** is preserved.

15.8 Reynolds Number

The size of most airplanes exceeds, by far, that of available wind tunnels, causing engineers to perform their tests and measurements on reduced scale models. The dimensions of such models are an accurate constant fraction of the original. One thing[††], however, can usually not be scaled down proportionally: the size of the molecules in air. For this reason, measurement on models may not be converted with precision to expected forces on the real plane. In fact, the forces observed on an object moving in a fluid are

[†] Henri-Marie Coanda, Romanian scientist (1885–1972), described this effect in 1930.

[††] Another characteristic that is not reproduced correctly in models is the rigidity. Real airplanes are much less rigid than most aerodynamic models. Fortunately, passengers, are not aware of how much a wing of a commercial airline flexes when flying through even moderate turbulence.

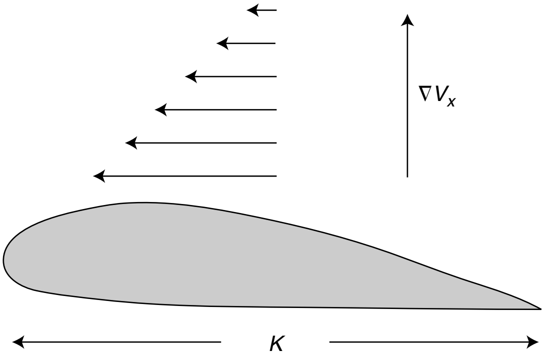


Figure 15.17 Wind shear over a wing.

not only the **dynamic** forces, F_d (proportional to $\frac{1}{2}\rho v^2$), but also include **viscous** forces, F_v.

When an airfoil moves through still air, molecules of gas in immediate contact with the surface are forced, through friction, to move with the velocity, v_x (assuming movement in the x-direction), of the foil, while those at a large distance do not move at all. A velocity gradient, ∇v_x, is established in the y-direction. See Figure 15.17.

The resulting velocity shear causes the viscous force, F_v, to appear. This force is proportional to the wing area, A, to the velocity gradient, $\partial v_x/\partial y$, and to the **coefficient of viscosity**, μ, a property of the medium through which the wing moves.

$$F_v = \mu \frac{\partial v_x}{\partial y} A. \tag{15.53}$$

For accurate scaling, it is necessary to preserve the ratio of the dynamic to the viscous forces. Define a quantity, R, called the **Reynolds number**, as

$$R \propto \frac{F_d}{F_v} = \frac{\frac{1}{2}\rho v_x^2 A}{\mu \partial v_x/\partial y A} \propto \frac{\rho}{\mu} \cdot \frac{v_x^2}{\partial v_x/\partial y}. \tag{15.54}$$

This is too complicated. Make two simplifying assumptions:

1. $\dfrac{\partial v_x}{\partial y}$ is independent of y,
2. The air is disturbed only to a distance, K, above the wing (where K is the **chord length**).

Under such circumstances,

$$\frac{\partial v_x}{\partial y} = \frac{v}{K}, \tag{15.55}$$

and

$$R = \frac{\rho v^2}{\mu v/K} = \frac{\rho}{\mu} v K. \tag{15.56}$$

The value of μ for air is 1.84×10^{-5} pascal second (same as kg m^{-1} s^{-1}), and, contrary to one's gut feeling, is independent of pressure and density (see the box at the end of this section). However, the ratio, μ/ρ, called the **kinematic viscosity**, increases when the pressure decreases. Fluids at low pressure exhibit great kinematic viscosity, and this explains why vacuum pumps need large-diameter pipes. For STP conditions, this ratio for air is 1/70,000 m^2/s because $\rho = 1.29\,\text{kg}\,\text{m}^{-3}$.

Since for any given angle of attack, the coefficients (C_L and C_D) are functions of R, measurements made with models cannot be extrapolated to life-size wings unless the Reynolds number is the same. Fortunately, the coefficients depend only weakly on R as illustrated in Figure 15.18, where C_D and $C_{L_{max}}$ are plotted versus R for the NACA 0012 symmetric airfoil. $C_{L_{max}}$ is the largest value that C_L can reach as a function of the angle of attack.

For first-order calculations, one can ignore the effects of a varying Reynolds number. In more precise calculations, the correct Reynolds number must be used especially because, as we shall see, the wing of a vertical axis wind turbine perceives a variable wind velocity throughout one cycle of its rotation.

In general, things tend to improve with larger Reynolds numbers (because the lift-over-drag ratio usually increases). This means that, on these grounds, larger wind turbines tend to be more efficient than smaller ones.

A 3-m chord wing moving at 360 km/h has a Reynolds number of

$$R = 7 \times 10^4 \times 100 \times 3 = 21 \times 10^6. \tag{15.57}$$

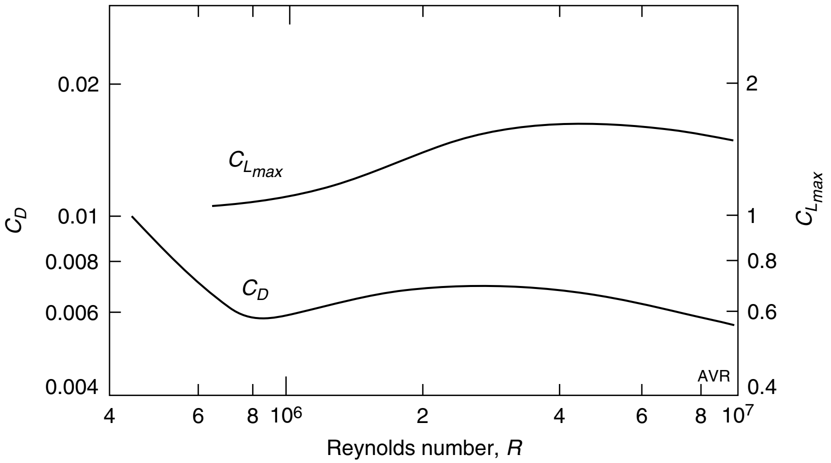


Figure 15.18 Effect of the Reynolds number on the lift and drag coefficients.

To measure the characteristics of this wing using a model with a 0.3-m chord under the same Reynolds number, one would need a wind velocity of 3600 m/s. However, the result would not be valid because the speed in question is supersonic. This explains why much of the wind tunnel data corresponds to modest Reynolds numbers. The Göttingen-420 data were measured at $R = 420{,}000$.

There is a class of wind tunnels, the NACA **variable-density tunnels**, in which R is increased by increasing the static air pressure. This raises ρ but does not affect μ. Thus, large Reynolds numbers can be achieved with small models at moderate wind speeds.

Why is μ independent of pressure or density? When a molecule of gas suffers a (isotropic) collision, the next collision will, statistically, occur at a distance one mean free path, ℓ, away. This locates the collision on the surface of a sphere with radius, ℓ, centered on the site of the previous collision. The projected area of this sphere is proportional to ℓ^2. One can, therefore, expect that

$$\mu \propto \nu n \ell^2, \tag{15.58}$$

where ν is the **collision frequency** and n is the **concentration** of the molecules. But ℓ is inversely proportional to the concentration, while the collision frequency is directly proportional to it. Thus,

$$\mu \propto n \times n \times \frac{1}{n^2}: \tag{15.59}$$

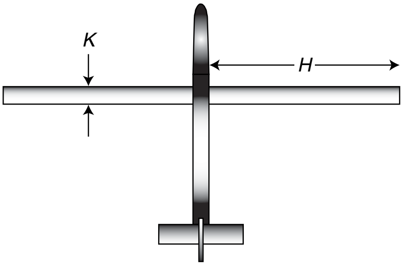
that is, μ is independent of n.

15.9 Aspect Ratio

In a rectangular wing, the ratio between the length, H, and the chord, K, is called the **aspect ratio**, $\mathcal{R}$.

$$\mathcal{R} \equiv \frac{H}{K} = \frac{H^2}{KH} = \frac{H^2}{A}, \tag{15.60}$$

where A is the area of the wing.



Hence, the aspect ratio can be defined as the ratio of the square of the wing length to the wing area. This definition must be used in the case of tapered (nonrectangular) wings, which have a variable chord length.

The drag experienced by a body moving through a fluid is caused by a number of different mechanisms. An ideal infinitely smooth and infinitely long wing would experience only a **pressure drag**. However, real wings are not perfectly smooth and the air tends to adhere to the surface, causing viscous shearing forces to appear and generating a **skin-friction drag**.

Wing lift results from the pressure being larger under the wing than over it. At the tip, there is a "short circuit" between the underside and the top and air circulates around it, forming a vortex. Energy is used to impart the circular motion to the air in the vortex, and this energy must come from the forward motion of the wing. It manifests itself as an additional drag force called the **induced drag**. The induced drag can be lessened by

1. increasing the number of wings (or the number of wing tips),
2. increasing the aspect ratio of the wing,
3. tapering the wing so that the chord is smaller near the tip, and
4. placing a vertical obstacle to the flow around the wing tip. Sometimes additional fuel tanks are mounted there.

Each wing tip generates its own vortex. Energetically, it is more economical to have many small vortices instead of one large one because the losses are proportional to the square of the vortex size. This is a consequence of the sum of the squares being smaller than the square of the sum. Biplanes (four wing tips) have less induced drag than monoplanes, everything else being the same. Soaring birds reduce the induced drag by spreading the feathers so that, near the end of the wing, there are many tips.

Obviously, the smaller the chord at the tip of the wing, the smaller the induced drag. In a rectangular wing of a given area, a larger length, H, results in a smaller chord, K. The wing has a larger aspect ratio and, consequently, a smaller induced drag.

Tapered wings have smaller wing-tip chords than rectangular wings of the same area and, again, have a smaller induced drag.

Gliders have long, slender wings to maximize the aspect ratio (minimizing the induced drag). In fast-moving airplanes, the **parasitic drag**[†] is dominant, making the induced drag unimportant: fast planes can tolerate small aspect ratios. Birds, on the other hand, have a variety of wing shapes, two of which can be seen in Figure 15.19.

When a wing is tested, the total drag measured includes the induced drag; hence it is customary to indicate the aspect ratio, Æ, of the test section when aerodynamic coefficients are tabulated.

[†] Parasitic drag is caused by parts of the machine that offer resistance to the flow of air but do not generate lift.

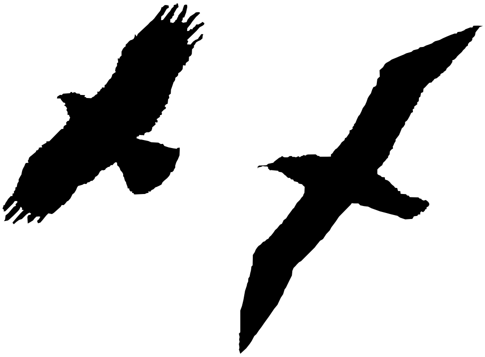


Figure 15.19 Nature's solution of the induced drag problem. Soaring birds (hawk, on left) have wings of moderate aspect ratio but with multiple wing tips, while petrels that skim over ocean waters far from shore, have pointed tips and wings with very large aspect ratio.

15.10 Wind Turbine Analysis

15.10.1 Horizontal Axis Turbines (propeller type)

Blades of large horizontal axis wind turbines in the twenty-first century are much more slender than earlier ones that were more paddle-like. The shape depicted in Figure 15.20 is approximately the one used on the Vestas V90 3 MW turbine. In the figure, the blade rotates clockwise with an angular velocity, ω, so that the leading edge of the airfoil is exposed to an induced wind of magnitude u, which is a function of r, the distance from the axis of rotation. We are going to consider a small section of the airfoil extending from r to $r + dr$. If the chord is K units long, the area of this section is $K\ dr$. An external horizontal wind, v_2, flows normal to the page, directly into the page. In other words, v_2 flows in the negative, z-direction.

In Figure 15.21, we have shown the section of airfoil seen from a different angle: the viewer is sighting the section along the y-axis (along the radius, r, of the blade), and the external wind blows up in the drawing (the direction of the vector $\vec{v}_2$). Observe that the magnitude of this wind is smaller than that of the undisturbed wind, v_1. It is the **retarded wind** discussed in the section on the Betz limit.

Repeated iterations will be necessary to arrive at an acceptable design of a wind turbine blade. To start, one has to know what power is to be extracted from the wind when it is at the chosen rated speed. As an example, let us assume we want a 3-MW output when the wind is 14 m/s. It is then easy to estimate the required swept area, A_v,

$$A_v = \frac{27}{16}\frac{P_D}{\frac{1}{2}\rho v_1^3 \eta}. \tag{15.61}$$

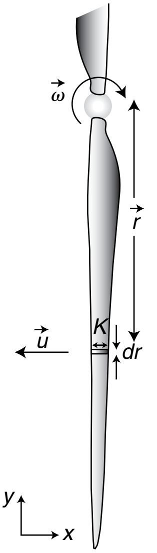


Figure 15.20 One blade of a horizontal axis wind turbine. This one rotates clockwise when driven by a wind blowing into the page.

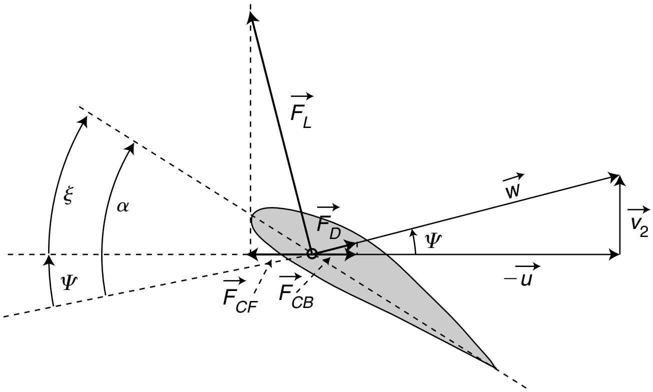


Figure 15.21 Aerodynamic forces on a section of a horizontal axis wind turbine blade.

A reasonable guess for the realizable efficiency, η, of a modern system is about 45%. Using $\rho = 1.2\,\text{kg/m}^3$ and the selected value of $v_1 = 14\,\text{m/s}$, we find that we need a swept area of $6833\,\text{m}^2$, which corresponds to a blade length of some 46.6 m and a rotor loading of $440\,\text{W/m}^2$.

Next, we should select the airfoil to be used. It does not need to be the same all along the blade. As a matter of fact, it may make sense to use thin airfoils near the tip of the blade where the air flow is fast, and thicker, sturdier, ones near the hub where the flow is relatively slow and the stresses are high. Let us assume that we opted for a NACA 4412 airfoil.[†] This is a thin (8%), high-lift airfoil. As in any airfoil, the lift/drag coefficient values depend on the Reynolds number. In the simple analysis that follows, we will conclude that the chord length, K, should be proportional to $1/r$, while the linear velocity of a given section of the blade varies with r. Hence the product, $u\ K$, which determines the Reynolds number, is the same for all sections of the blade. r is the distance, along the blade, from the hub. Larger, modern, wind turbines use rapidly moving, big blades and, therefore operate at relatively high Reynolds numbers. In Table 15.2, we tabulated the values of C_L and C_D for the NACA 4412 airfoil, measured with a Reynolds number of 9 million.

Another early decision is the choice of number of blades. Rotation becomes smoother with a larger number of blades, but the cost increases because blades are expensive. Today large turbines, almost invariably sport three blades. Thus, in our example, we will set $n = 3$.

The **tip speed ratio**, $\lambda = \omega R/v_1$, is one of the major parameters to be selected. Up to certain limit, the larger the λ, the better. However, too large a lambda will cause ωR to exceed the speed of sound. Actually, λ must be sufficiently low to avoid undue noise and undue stresses. Large tip speed ratios require good airfoils. One reason for repeated iterations is to home in on the value of λ that optimizes performance, given the choice of the airfoil. We will start with $\lambda = 8$. Among other things, high tip speed ratios lead to slender, low-mass blades that allow fast acceleration of the turbine when wind velocities change brusquely. Turbines that operate at constant speed (to generate alternating current at exactly the frequency of the power grid) cannot operate at constant λ, that is, at the optimum operating point. However, many modern turbines generate dc—which is then inverted to ac—and can, therefore, operate at the selected optimum tip speed ratio. To rapidly follow the fluctuations in wind velocity, it is advantageous to have slender, low inertia blades.

Different sections of the blade contribute different amounts to the power collected by the turbine because of the varying linear velocities and, possibly, because the airfoil and the setup angle may vary with the distance, r, from the hub. As a consequence, the blade is divided in a number of zones—between 10 and 20—and the contribution of each is calculated separately. Consider the ith zone. As the blade spins, the zone describes an annulus, which contributes a power, P_i, to the useful output of the turbine,

$$P_i = n\omega\Upsilon_i, \tag{15.62}$$

[†] If you want to play around with different airfoils, you may want to consult Abbott and Doenhoff (1949), which displays lift and drag data for a number of NACA airfoils. Though first published in 1949, this book is still being sold today.

Table 15.2 Characteristics of a NACA 4412 Airfoil
Reynolds number = 9 million

Angle of attack α (deg.)	**Lift coefficient C_L**	**Drag coefficient C_D**	C_L/C_D
−10	−0.66	0.0088	−75
−9	−0.55	0.0081	−68
−8	−0.44	0.0075	−59
−7	−0.34	0.0070	−48
−6	−0.23	0.0067	−35
−5	−0.12	0.0064	−19
−4	−0.02	0.0062	−3
−3	0.09	0.0062	15
−2	0.2	0.0062	32
−1	0.3	0.0061	49
0	0.41	0.0061	67
1	0.52	0.0062	84
2	0.63	0.0062	102
3	0.73	0.0063	116
4	0.84	0.0065	129
5	0.95	0.0072	132
6	1.05	0.0081	129
7	1.16	0.0092	126
8	1.27	0.0106	120
9	1.34	0.0116	115
10	1.43	0.0131	109
11	1.51	0.0146	104
12	1.59	0.0162	98
13	1.63	0.0171	96
14	1.67	0.0179	93
15	1.65	No data	
16	1.65	No data	
17	1.65	No data	
18	1.65	No data	
20	1.51	No data	

where n is the number of blades, ω is the angular velocity of the blade, and Υ_i is the torque generated by the ith zone. Refer to Figure 15.21:

$$\Upsilon_i = (F_{CF_i} - F_{CB_i})r_i. \tag{15.63}$$

F_{CF_i} is the x-projection of the lift vector, F_{L_i}, of the ith zone and is

$$F_{CF_i} = F_{L_i} \sin\psi_i, \tag{15.64}$$

while, F_{CB_i} is the x-projection of the drag vector, F_{D_i}, and is

$$F_{CB_i} = F_{D_i} \cos \psi_i. \tag{15.65}$$

The x-direction is the direction of motion of the blade.

$$\begin{aligned} \Upsilon_i &= (F_{L_i} \sin \psi_i - F_{D_i} \cos \psi_i) r_i = \frac{1}{2} \rho w_i^2 (C_{L_i} \sin \psi_i - C_{D_i} \cos \psi_i) A_{p_i} r_i \\ &\approx \frac{1}{2} \rho w_i^2 (C_{L_i} \sin \psi_i) A_{p_i} r_i = \frac{1}{2} \rho w_i C_{L_i} \frac{2}{3} v_1 K_i \Delta r \ r_i \end{aligned} \tag{15.66}$$

because it is hoped that $C_L \sin \psi >> C_D \cos \psi$; otherwise, the airfoil losses will be too large. The last step in Equation 15.66 results from the fact that the area, A_{p_i} of the airfoil section is $K_i \Delta r$, and from the realization (see Figure 15.21) that $w \sin \psi = 2/3 \, v_1$. Remember that $v_2 = 2/3 \, v_1$.

w is the wind speed seen by a given section of the blade. It is the resultant of the velocity, $u = r\omega$, owing to the rotation added vectorially to the velocity, v_2, due to the (retarded) external wind.

Combining Equations 15.62 with 15.66, we have

$$P_i = n \frac{1}{2} \rho w_i C_{L_i} \frac{2}{3} v_1 K_i \Delta r \ r_i \omega. \tag{15.67}$$

If this power is to be the maximum available power (under the Betz assumption), $P_{A_i} = \frac{16}{27} \frac{1}{2} \rho v_1^3 \, 2\pi r_i \Delta r$, where $2\pi r \Delta r$ is the swept area of the annulus,

$$n \frac{1}{2} \rho w_i C_{L_i} \frac{2}{3} v_1 K_i \Delta r \ r_i \omega = \frac{16}{27} \frac{1}{2} \rho v_1^3 \ 2\pi r_i \ \Delta r, \tag{15.68}$$

which yields a value of K_i of

$$K_i = \frac{16}{9} \frac{v_1^2 \pi}{n w_i C_{L_i} \omega}. \tag{15.69}$$

It is convenient to normalize the wind velocity by using, as a parameter, the tip speed ratio,

$$\lambda \equiv \frac{\omega R}{v_1} \quad \therefore \quad v_1 = \frac{\omega R}{\lambda}. \tag{15.70}$$

The tip speed ratio is the ratio between the linear speed of the tip of the rotor blade and the undisturbed wind velocity. The expression for the chord that extracts the Betz-limit power becomes

$$K_i = \left(\frac{16}{9} \frac{\pi \omega R^2}{n \lambda^2} \right) \frac{1}{C_{L_i} w_i}. \tag{15.71}$$

The quantity in parentheses in Equation 15.71 does not depend on the particular zone of the blade being considered. For the example we are dealing with, in which we selected $R = 46.6\,\text{m}$, $n = 3$, $\lambda = 8$, and $v_1 = 14\,\text{m/s}$—which yields $\omega = 2.40$ radians/sec—the equation reduces to

$$K_i = 151.6\frac{1}{C_{L_i} w_i}. \tag{15.72}$$

It is, therefore, easy to calculate the chord length for each zone. But things can be simplified further. $w_i = \sqrt{v_2^2 + \omega^2 r_i^2}$. For our example, $w_i = \sqrt{87.1 + 5.76 r_i^2}$. And, if we decide to use the same airfoil all along the blade with the same angle of attack, then C_L is independent of r. We could select an angle of attack that maximizes the C_L/C_D ratio, but this will lead to a relatively small coefficient of lift (0.95, for the chosen NACA 4412 airfoil). Perhaps it will be better to operate at a considerably higher α, say 13°, and reach a very large 1.63 while still having a respectable $C_L/C_D = 96$ and also, being several degrees away from stall. We will settle on $\mathrm{C}_L = 1.63$, and the expression for the chord length becomes a very simple

$$K_i \approx \frac{10}{\sqrt{1 + 0.0663 r_i^2}}. \tag{15.73}$$

We can get good insight into how the chord length depends on the distance from the hub by observing that the angle, ψ, decreases when the tip speed ratio, λ, increases, and when, r, increases. Thus, for wind turbines operating with large λ and with an airfoil that only starts at $r > 0.1R$, the rough approximation can be made that $w_i \approx \omega r$, which leads to

$$K \approx \frac{16}{9}\frac{\pi R^2}{n}\frac{1}{\lambda^2 C_L}\frac{1}{r}. \tag{15.74}$$

This equation shows that the chord of the blade would then be inversely proportional to the distance from the hub. It will also be inversely proportional to the square of the tip speed ratio and, of course, to the number, n, of blades, provided we use the same airfoil and the same angle of attack along the blade. Under these conditions, the area of each blade is

$$A_p = \frac{16}{9}\frac{\pi R^2}{n}\frac{1}{\lambda^2 C_L}\int_{r_0}^{R}\frac{1}{r}dr = \frac{16}{9}\frac{\pi R^2}{n}\frac{1}{\lambda^2 C_L}\ln\frac{R}{r_0}, \tag{15.75}$$

where r_0 is the distance from the hub at which the blade starts. For the not uncommon case in which $r_0 = 0.1R$, the equation reduces to

$$A_p = 4.1\frac{\pi R^2}{n}\frac{1}{\lambda^2 C_L}. \tag{15.76}$$

The ratio of the total area of the blades, nA_p, to the swept area, A_v, is called **solidity**, S:

$$S = \frac{nA_p}{A_v} = \frac{16}{9}\frac{1}{\lambda^2 C_L}\ln\frac{R}{r_0}. \tag{15.77}$$

In the example we are pursuing, the solidity is 3.7%. This is in the range of typical solidities of large modern turbines.

In order to ensure that we get the correct angle of attack selected for each zone (in our example we chose the same angle of 13° for all zones), we must vary the setup angle, ξ, accordingly. From Figure 15.21, we see that.

$$\xi = \alpha - \psi = \alpha - \arctan\left(\frac{2R}{3\lambda r}\right). \tag{15.78}$$

In constructing the blade, one must smoothly interpolate the shape and setup angle along the length of the blade.

The simplified approach we outlined above leads to a rough estimate of the turbine characteristics and can be considered as a starting point for an optimization study using a more sophisticated modeling program that includes at least some of the major loss mechanisms such as wake rotation and tip losses. An example of such a procedure can be found in the book by Manwell, McGowan, and Rogers (2002). A major consideration in turbine design is the fabricability of the blades, which may be an expensive and complicated task.

15.10.2 Vertical Axis Turbines

As an example of a vertical axis wind turbine analysis, we have chosen a Gyromill because, even though this type of turbine is not in use, its analysis is much simpler than that of the popular Darrieus type.[†] *Nevertheless, the analysis, brings out important conclusions that, with suitable modifications, are also applicable to other related devices.*

For those interested in an introductory analysis of propeller turbines, we recommend Duncan, Thom, and Young (1960) or Manwell, McGowan, and Rogers (2002).

Consider a McDonnell-Douglas vertical axis wind turbine. Consider also a right-handed orthogonal coordinate system with the z-axis coinciding with the vertical axis of the machine. The system is so oriented that the horizontal component of the wind velocity, v, is parallel to the x-axis.

We will assume that the airfoil, whose cross section lies in the x-y plane, has a chord that makes an angle, ξ, with the normal to the radius

[†] A small gyromill-like turbine called Windspire is being produced (2009) by Mariah Power.

vector. This is the **setup** angle chosen by the manufacturer of the wind turbine. It may be adjustable, and it may even vary during one rotation. In this analysis, it is taken as a constant.

The radius vector makes an angle, $\theta(t)$, with the x-axis. Figure 15.22 indicates the different angles and vectors involved in this derivation. The inset shows how the wind velocity, v, and the velocity, u (owing to the rotation), combine to produce an **induced** velocity, w, which is the actual air velocity perceived by the wing. Of course, u is the velocity the wing would perceive if there were no wind.

As seen in Figure 15.22, the angle of attack is $\alpha = \psi + \xi$.

$$\vec{\omega} = \vec{k}\omega, \tag{15.79}$$

$$\vec{r} = r(\vec{i}\cos\theta + \vec{j}\sin\theta), \tag{15.80}$$

$$\vec{u} = -\vec{\omega}\times\vec{r} = -\begin{pmatrix} \vec{i} & \vec{j} & \vec{k} \\ 0 & 0 & \omega \\ r\cos\theta & r\sin\theta & 0 \end{pmatrix} = u(\vec{i}\sin\theta - \vec{j}\cos\theta), \tag{15.81}$$

$$\vec{w} = \vec{u} + \vec{v} = \vec{i}(v + u\,\sin\theta) - \vec{j}\,u\,\cos\theta, \tag{15.82}$$

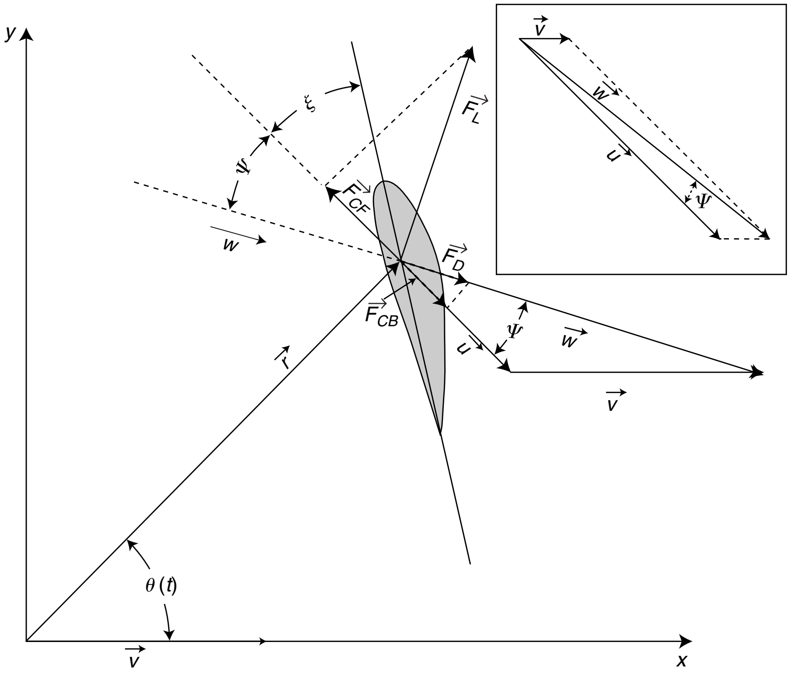


Figure 15.22 Angles and forces on a wing. Usually, $\vec{u}$ is much larger than $\vec{v}$, but for clarity in the drawing (but not in the inset), they were taken as approximately equal. This exaggerates the magnitude of the angle, ψ. The inset shows the wind vectors.

$$w = \sqrt{v^2 + u^2 \sin^2 \theta + 2uv \sin \theta + u^2 \cos^2 \theta}$$
$$= \sqrt{v^2 + u^2 + 2uv \sin \theta} \equiv \Gamma v, \tag{15.83}$$

where

$$\Gamma \equiv \sqrt{1 + \frac{u^2}{v^2} + 2\frac{u}{v} \sin \theta}, \tag{15.84}$$

$$\vec{u} \cdot \vec{w} = (u \sin \theta)(v + u \sin \theta) + u^2 \cos^2 \theta$$
$$= u^2 + uv \sin \theta = uw \cos \psi, \tag{15.85}$$

from which

$$\cos \psi = \frac{u^2 + uv \sin \theta}{uw} = \frac{u + v \sin \theta}{\sqrt{v^2 + u^2 + 2uv \sin \theta}} = \frac{u/v + \sin \theta}{\Gamma}. \tag{15.86}$$

For a given wind speed, v, and angular velocity, ω, of the wind turbine, the ratio, u/v, is constant. The quantity, Γ, and the angle, ψ, vary with the angular position of the wing; thus, they vary throughout the revolution. Consequently, the angle of attack also varies. Clearly, if there is no wind, α is constant. If there is a high u/v ratio (if the wind speed is much smaller than that of the rotating wing), then α varies only a little throughout the revolution.

Given a u/v ratio and a wind velocity, it is possible to calculate both α and w for any wing position, θ. For a given w and α, the wing will generate a lift

$$F_L = \frac{1}{2} \rho w^2 A_p C_L, \tag{15.87}$$

and a drag

$$F_D = \frac{1}{2} \rho w^2 A_p C_D. \tag{15.88}$$

In the preceding equations, A_p is the area of the wing and F stands for the force on the wing.

Note that $\vec{F}_L$ is normal to $\vec{w}$ in the x-y plane and that $\vec{F}_D$ is parallel to $\vec{w}$. The lift force, $\vec{F}_L$, has a component, $\vec{F}_{CF}$, normal to the radius vector, causing a forward torque. The drag force, $\vec{F}_D$, has a component, $\vec{F}_{CB}$, also normal to the radius vector, causing a retarding torque.

The resulting torque is

$$\Upsilon = r(F_{CF} - F_{CB}). \tag{15.89}$$

From Figure 15.22,

$$F_{CF} - F_{CB} = F_L \sin \psi - F_D \cos \psi = \frac{1}{2} \rho w^2 A_p (C_L \sin \psi - C_D \cos \psi). \tag{15.90}$$

Thus,

$$\begin{aligned}\Upsilon &= \frac{1}{2}\rho w^2 A_p r\,(C_L \sin\psi - C_D \cos\psi) \\ &= \frac{1}{2}\rho v^2 A_p r\left[\Gamma^2(C_L \sin\psi - C_D \cos\psi)\right]. \end{aligned} \tag{15.91}$$

The average torque taken over a complete revolution is

$$<\Upsilon> = \frac{1}{2\pi}\int_0^{2\pi} \Upsilon(\theta)d\theta. \tag{15.92}$$

In the expression for Υ, only the part in brackets is a function of θ. Let us define a quantity, D:

$$D \equiv \Gamma^2(C_L \sin\psi - C_D \cos\psi), \tag{15.93}$$

$$<D> = \frac{1}{2\pi}\int_0^{2\pi} D d\theta, \tag{15.94}$$

$$<\Upsilon> = \frac{1}{2}\rho v^2 A_p r <D>. \tag{15.95}$$

The power delivered by the turbine to its load is

$$P_D = \omega <\Upsilon> N. \tag{15.96}$$

Here, N is the number of wings on the wind turbine. The swept area is (see Figure 15.23)

$$A_v = 2rH, \tag{15.97}$$

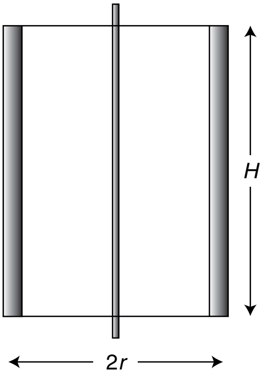


Figure 15.23 The aspect ratio of a wind turbine.

and the area of each wing is

$$A_p = KH, \tag{15.98}$$

where H is the (vertical) length of the wing and K is the chord (assumed uniform).

A **solidity**, S, is defined:

$$S = \frac{NA_p}{A_v} = N\frac{K}{2r}. \tag{15.99}$$

The **available** power from the wind is

$$P_A = \frac{16}{27}\frac{1}{2}\rho v^3 A_v, \tag{15.100}$$

$$\eta = \frac{P_D}{P_A} = \frac{\frac{1}{2}\rho v^2 N A_p r\omega <D>}{\frac{1}{2}\rho v^3 A_v \frac{16}{27}} = \frac{27}{16}\frac{u}{v} <D> S. \tag{15.101}$$

The efficiency formula derived above is correct only to first order. It ignores parasitic losses owing to friction and to the generation of vortices; it disregards the reduction in wind velocity caused by the wind turbine itself; it fails to take into account the interference of one wing blade on the next. In fact, Equation 15.101 predicts that with large enough solidities, the efficiency can exceed unity. We will discuss this question a little later.

Notice that the $u/v <D>$ product is a function of the parameter u/v.

$<D>$ must be obtained from numerical analysis, looking up values of C_L and C_D for the various α that appear during one revolution.

To gain an idea of the shape of the $u/v <D>$ versus u/v graph, consider the situation when $u = 0$. Clearly, $u/v = 0$ and since D cannot be infinite, $u/v <D>$ must also be 0.[†]

When $u \to \infty$, $w \to u$ and $\psi \to 0$. From Equation 15.93,

$$D = \Gamma^2(C_L \sin\psi - C_D \cos\psi) \to -\Gamma^2 C_D;$$

thus, for large values of u/v, $D < 0$ and, consequently, $u/v <D>< 0$.

[†] When $u=0$, Γ is unity (Equation 15.84) and, from Equation 15.86, $\cos\psi = \sin\theta$ and $\sin\psi = \cos\theta$. This makes

$$D = C_L \cos\theta - C_D \sin\theta, \tag{15.102}$$

and consequently, provided C_L and C_D are constant,

$$<D> = 0, \tag{15.103}$$

because the mean value of $\sin\theta$ and of $\cos\theta$ is zero. Since the torque is proportional to $<D>$, this type of wind turbine has no torque when stalled: it has zero starting torque and requires a special starting arrangement (such as a small Savonius on the same shaft). Actually, C_L and C_D do depend on Θ, and Equation 15.102 holds only approximately.

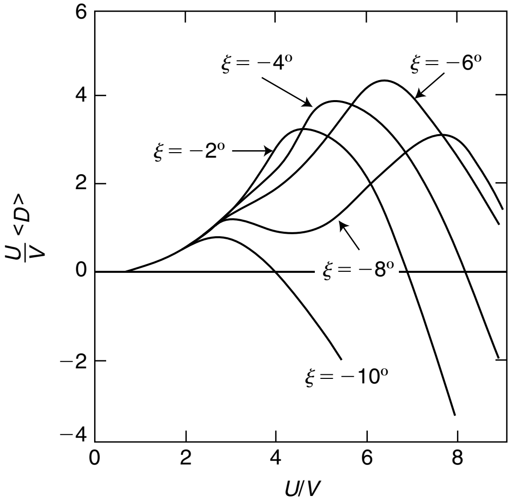


Figure 15.24 Performance of the Gö-420 airfoil in a vertical axis turbine.

This means that, at high rpm, the wind turbine has a negative torque and tends to slow down. One can therefore expect that the efficiency has a maximum at some value of u/v in the range $0 < u/v < \infty$.

As an example, we have computed u/v <D> for various values of u/v and a number of setup angles, ξ. The airfoil used was the Gö-420. The results are shown in Figure 15.24. It can be seen that the optimum setup angle (the one that leads to the highest $u/v < D >$ is $-6°$). Symmetric airfoils work best with $\xi = 0$.

For $\xi = -6°$, the airfoil reaches a u/v <D> of 4.38 (nondimensional) at a u/v of 6.5. Thus, in this particular case, the efficiency formula yields

$$\eta_{max} = 7.39\, S. \qquad (15.104)$$

Were one to believe formula 15.104, efficiencies greater than 1 could be reached by using solidities, S, larger than 0.135. Clearly, there must be some value of solidity above which the formula breaks down. In Figure 15.25, the efficiency of a wind turbine is plotted versus solidity. The linear dependence predicted by Equation 15.104 is represented by the dashed line with the 7.39 slope of our example. Using a more complicated aerodynamical model, Sandia obtained the results shown in the solid line. It can be seen that increasing the solidity beyond about 0.1 does not greatly affect the efficiency. One can distinguish two regions in the efficiency versus solidity curve: one in which, as predicted by our simple derivation, the efficiency is proportional to the solidity, and one in which the efficiency is (roughly) independent of the solidity.

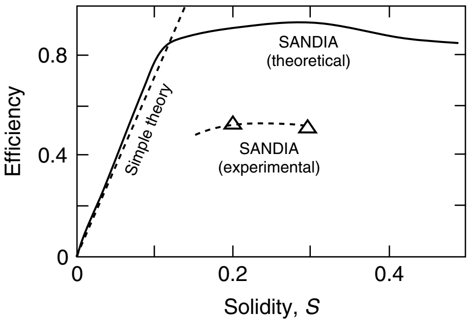


Figure 15.25 Effect of solidity on efficiency of a vertical axis wind turbine.

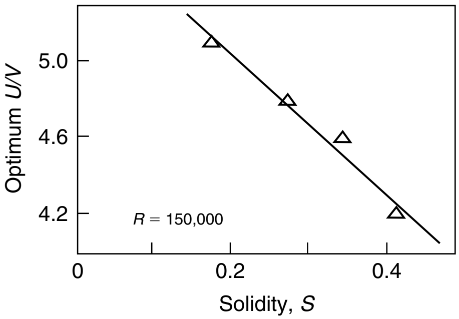


Figure 15.26 Dependence of the optimum u/v on solidity. Experimental data from the Sandia 2-m diameter Darrieus turbine.

Triangles in Figure 15.25 indicate values of efficiency measured by Sandia using small models of the wind turbines. Measured efficiencies are about half of the calculated ones. This discrepancy is discussed further on in this chapter.

The main reason for the behavior depicted in Figure 15.25 is that our simple theory failed to account for the interference of one wing with the next. The larger the solidity, the farther the disturbance trails behind the wing and the more serious the interference, thus counteracting the efficiency gain from a larger S. Consequently, the optimum u/v shrinks as S increases.

Figure 15.26 shows the experimentally determined effect of solidity on the optimum u/v. If the straight line were extrapolated, one would conclude that for $S = 1$, the optimum u/v would be about 0.7.

In the range of solidities that have only a small effect on the efficiency, increasing S results in a wind turbine that rotates more slowly (because of the smaller optimum u/v) and has more torque (because the efficiency—and consequently the power—is the same). Increasing S has the effect of "gearing down" the wind turbine. Since the cost of a wind turbine is roughly proportional to its mass, and hence to its solidity, one should prefer machines with S in the lower end of the range in which it does not affect the efficiency. That is why, for large machines, propellers are preferred to vanes. Nevertheless, for small wind turbines, the simplicity of vane construction may compensate for the larger amount of material required.

Although the equations derived above will help in the understanding of the basic wind turbine processes, they fall far short from yielding accurate performance predictions. Numerous refinements are needed:

1. Frictional losses in bearings must be taken into account.
2. The rotating wings create vortices that represent useless transformation of wind energy into whirling motion of the air. The effect of such vortices has to be considered.
3. The wind, having delivered part of its energy to the machine, must necessarily slow down. Thus, the average wind velocity seen by the blades is less than the free stream velocity and the power is correspondingly less than that predicted by the formulas.

   Single streamtube models are based on an average wind slowdown. These models ignore the nonuniformity of the wind velocity in the cross section of the wind turbine. By contrast, if the wind slowdown is considered in detail, then we have a **multiple streamtube** model.

   Figure 15.27 shows how the ratio of the streamtube to freestream velocity varies with position in the Sandia 2-m diameter

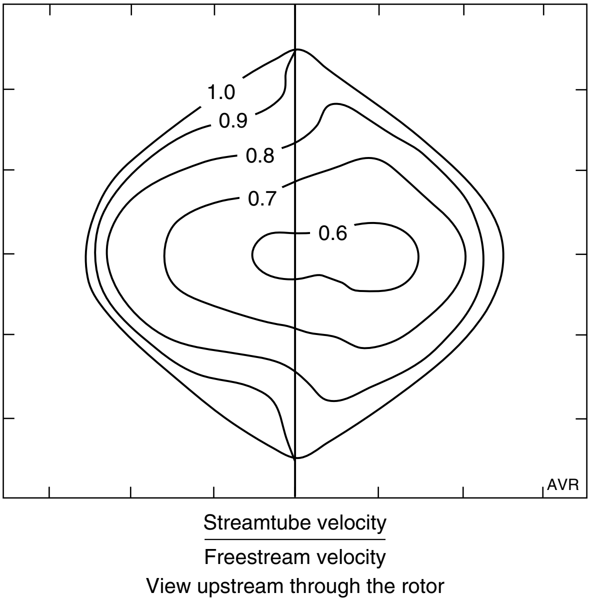


Figure 15.27 Streamtube velocity through the rotor of a Sandia 2-m Darrieus turbine.

Darrieus turbine. When the calculations are based on the multistreamtube model, the predicted performance approaches reality as can be seen in Figure 15.28 in which the measured efficiency (squares) is compared with the values predicted using both the single and the multistreamtube models. Even the multistreamtube model overestimates the performance.

4. The accuracy of the prediction is improved by using the appropriate Reynolds number, a quantity that actually varies throughout the revolution. Figure 15.29 shows the effect of the Reynolds number on the performance of the wind turbine. As expected, the measurements, made at Sandia show that the larger R, the better the performance.

Clearly, a refined wind turbine performance model is too complicated to be treated in this book.

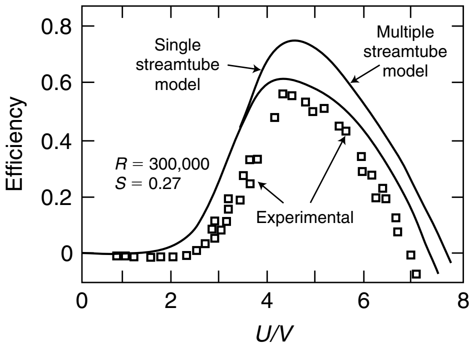


Figure 15.28 Calculated efficiencies compared with observed values. Data from Sandia.

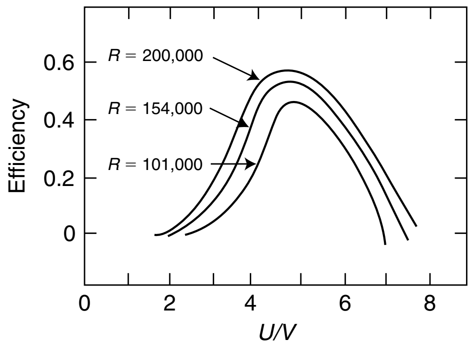


Figure 15.29 Influence of the Reynolds number on the performance. Data from Sandia.

15.10.2.1 Aspect Ratio (of a wind turbine)

In Figure 15.23 (in the previous section), the aspect ratio, AR_{turb}, of a Gyromill (McDonnell-Douglas vertical axis wind turbine) was defined:

$$AR_{turb} = \frac{H}{2r} = \frac{A_v}{4r^2} = \frac{H^2}{A_v}. \tag{15.105}$$

This is, of course, different from the aspect ratio of the wing, which is

$$AR_w = \frac{H}{K}, \tag{15.106}$$

where K is the chord length of the wing. Since the solidity is $S = NK/2r$, the wing aspect ratio can be rewritten as $AR_w = HN/2rS$. However, the aspect ratio of the wind turbine is $AR_{turb} = H/2r$; hence,

$$AR_{turb} = \frac{S}{N} AR_w. \tag{15.107}$$

For constant N and constant S (constant first-order efficiency), the aspect ratio of the wind turbine is proportional to that of the wing. Wingtip drag decreases with the increasing wing aspect ratio; consequently, wind turbine efficiencies actually tend to go up with increasing AR_{turb}.

From Equation 15.99, it can be seen that if both the swept area, A_v (power), and the solidity S (efficiency) are kept constant, then, to a first approximation, the mass of the wings remains constant because their area, A_p, is constant. However, larger aspect ratios require higher towers. The mass of the tower increases with height faster than linearly because higher towers must have a larger cross section. On the other hand, the struts that support the wings become smaller with increasing AR_{turb}, again nonlinearly.

Thus, we have a situation like the one depicted in Figure 15.30, which displays data for the McDonnell-Douglas Gyromill. There is a minimum in total mass of the wind turbine at a AR_{turb} of, roughly, 0.8.

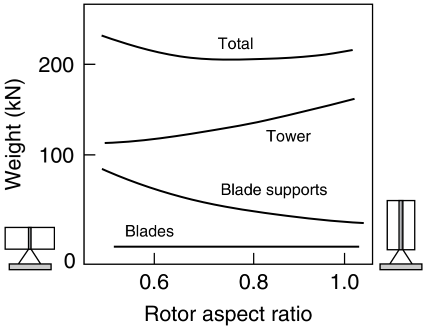


Figure 15.30 Influence of the aspect ratio on the mass of a wind turbine. *Source*: McDonnell-Douglas 120 kW Gyromill.

15.10.2.2 Centrifugal Force

Consider a section of a wing with mass M. Its weight is Mg (g is the acceleration of gravity). The centrifugal force acting on the section is

$$F_c = \frac{Mu^2}{r} = M\omega^2 r. \tag{15.108}$$

The ratio of centrifugal force to weight, w, is

$$\frac{F_c}{w} = \frac{\omega^2 r}{g}. \tag{15.109}$$

In a 10-m diameter wind turbine rotating at 100 rpm ($\omega = 10.5\,\text{rad/s}$), a section of the wing will experience a centrifugal force 55 times larger than its weight. This illustrates the need for careful design to minimize the centrifugal effects. For instance, what aspect ratio minimizes these effects?

Compare two wind turbines with the same swept area (see Figure 15.31). If they have the same solidity and use the same airfoil, they will deliver the same power when operated at the same u/v ratio. If they have different aspect ratios then they must have:

$$r_1 H_1 = r_2 H_2. \tag{15.110}$$

The centrifugal force per unit length is

$$F_c = \frac{Mu^2}{r}, \tag{15.111}$$

where M is the mass of the wing per unit length. The ratio of the centrifugal forces on the two wind turbines is

$$\frac{F_{c_1}}{F_{c_2}} = \frac{M_1 r_2}{M_2 r_1}. \tag{15.112}$$

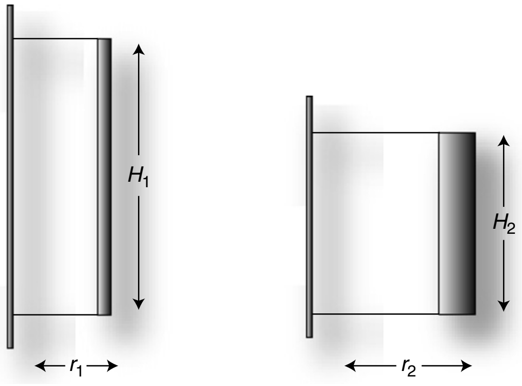


Figure 15.31 Two wind turbines with the same area but with different aspect ratios.

If the two wings have a solid cross section, then the masses (per unit length) are proportional to the square of the chord, K. In general,

$$M = bK^a, \tag{15.113}$$

where a is an exponent that depends on the type of construction. From this

$$K = \left(\frac{M}{b}\right)^{1/a}. \tag{15.114}$$

Since the wind turbines have the same solidity, $NK/2r$,

$$\frac{N_1}{2r_1}\left(\frac{M_1}{b}\right)^{1/a} = \frac{N_2}{2r_2}\left(\frac{M_2}{b}\right)^{1/a}, \tag{15.115}$$

$$\frac{M_1}{M_2} = \left(\frac{N_2 r_1}{N_1 r_2}\right)^{a}, \tag{15.116}$$

$$\frac{F_{c_1}}{F_{c_2}} = \left(\frac{N_2}{N_1}\right)^{a}\left(\frac{r_1}{r_2}\right)^{a-1}, \tag{15.117}$$

and, if the two wind turbines have the same number of wings,

$$\frac{F_{c_1}}{F_{c_2}} = \left(\frac{r_1}{r_2}\right)^{a-1}. \tag{15.118}$$

From the preceding formula, it becomes clear that for any $a > 1$, the larger the radius, the larger the centrifugal force. Since it is difficult to achieve low a's, this favors large-aspect ratios.

The centrifugal force in a Gyromill poses a difficult problem for the wind turbine designer. However, as explained in Subsection 15.2.2, this problem disappears in the case of the Darrieus (egg-beater) configuration.

15.10.2.3 Performance Calculation

Consider a vertical axis wind turbine with the characteristics given in Table 15.3.

Table 15.3 Wind Turbine Characteristics

Number of wings	$N = 3$
Height (length of wings)	$H = 16\,\text{m}$
Radius	$r = 10\,\text{m}$
Solidity	$S = 0.27$
Site	sea level
Wing performance	See Figure 15.32 and Table 15.4.

Assume that the performance data are valid for the actual Reynolds number. The wind turbine drives a load whose torque obeys the relationship

$$\Upsilon_L = 2000\,\omega. \tag{15.119}$$

The performance of the wind turbine is presented in the form of a torque versus angular velocity plot. A different plot must be constructed for each wind velocity of interest.

Take, for instance, $v = 10\,\text{m/s}$. The questions to be answered are:

1. What is the operating rpm?
2. What is the power delivered to the load?
3. What is the actual (average) Reynolds number?
4. What is the efficiency of the wind turbine?

Using the technique described in Section 15.7, a plot of efficiency versus u/v is constructed. Such a plot (similar to those in Figure 15.24) is shown in Figure 15.32. The corresponding numerical data appear in Table 15.4.

Now, let us construct the torque versus angular velocity plot:

$$\omega = \frac{u}{r} = \frac{u}{v} \times \frac{v}{r} = \frac{u}{v}. \tag{15.120}$$

Observe that in this particular example, the value of v/r is 1.

$$A_v = 2rH = 320\,\text{m}^2, \tag{15.121}$$

$$\rho = 1.29\,\text{kgm}^{-3}\ \text{(sea level)}, \tag{15.122}$$

$$P_D = \frac{16}{27}\frac{1}{2}\rho v^3 A_v \eta = 122{,}000\,\eta, \tag{15.123}$$

$$\Upsilon = \frac{P_D}{\omega} = 122{,}000\,\frac{\eta}{u/v}. \tag{15.124}$$

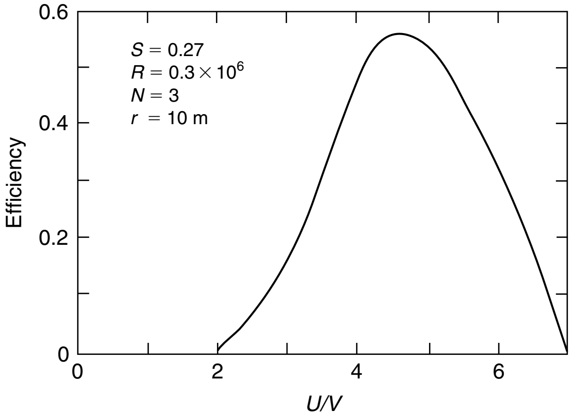


Figure 15.32 Efficiency versus u/v for the wing in the example.

Table 15.4 Efficiency versus u/v for the wing in the example

u/v	**Efficiency**
2.0	0.00
2.2	0.02
2.4	0.05
2.6	0.08
2.8	0.12
3.0	0.15
3.5	0.33
4.0	0.49
4.5	0.56
5.0	0.53
5.5	0.44
6.0	0.33
6.5	0.17
7.0	0.00

Notice that, Υ in Equation 15.124, is the total torque of the wind turbine, not the torque per wing, as before.

Using the performance table (Table 15.4), it is easy to calculate η for any u/v (or for any ω). The plot of Υ versus ω is shown in Figure 15.33. In the same figure, the Υ_L versus ω characteristic of the load is plotted. It can be seen that two different values of ω yield wind turbine torques that match that of the load. In point A, if there is a slight increase in wind velocity, the wind turbine will speed up, its torque will become larger, and the turbine will accelerate, increasing the torque even further. It is an unstable point. In point B, an increase in wind velocity will reduce the wind turbine torque, causing it to slow down back to its stable operating point.

At the stable point, B, the angular velocity in our example is 5.4 rad/s, equivalent to 0.86 rps or 52 rpm. The power generated will be

$$P = \Upsilon\omega = 2000\,\omega^2 = 58{,}300\,\mathrm{W}. \tag{15.125}$$

The efficiency that corresponds to $u/v = \omega = 5.4$ is 0.46.

The solidity is

$$S = \frac{NK}{2r} \quad \therefore \quad K = \frac{2rS}{N} = 1.8\,\mathrm{m}. \tag{15.126}$$

The Reynolds number is

$$R = 70{,}000\,wK \approx 70{,}000\,uK = 6.8 \times 10^6. \tag{15.127}$$

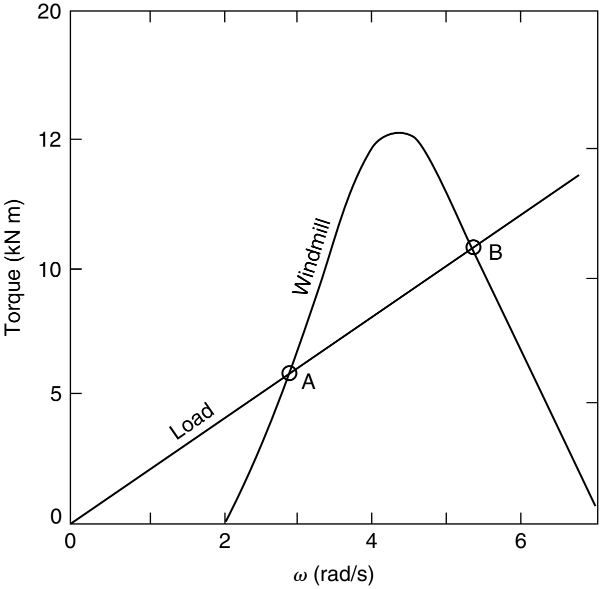


Figure 15.33 Operating point of the turbine.

15.11 Magnus Effect

In the introduction to this chapter, we made reference to the use of rotating cylinders to generate lift useful in driving wind turbines. The lift is caused by the **Magnus** effect, used by baseball pitchers when throwing a "curve."

Consider a rotating cylinder exposed to a stream of air.

Sufficiently far away from the cylinder, the air is undisturbed; that is, it moves with the velocity of the wind.

Immediately in contact with the cylinder, at point a (see Figure 15.34), the air flows from right to left in a direction opposite to that of the wind because the (rough) cylinder surface exerts friction on the air and accelerates it in the direction of its own motion. Its velocity is equal to the linear velocity of the cylinder surface. A gradient of velocity is established as illustrated in the figure.

On the opposite side, b, the cylinder impels the air from left to right, causing it to flow faster than the stream and establishing a velocity gradient. The average stream velocity is larger on the b-side than on the a-side. According to Bernoulli's law, the high-velocity side, b, experiences a smaller pressure than the low-velocity side, a—a force from a to b is exerted on the cylinder. This force, the result of the aerodynamic reaction on a

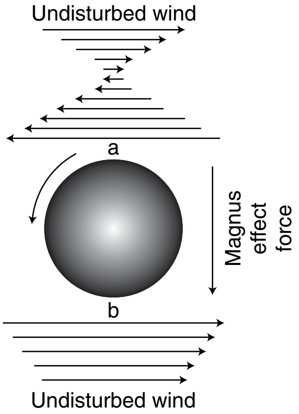


Figure 15.34 A rotating cylinder induces a velocity gradient in the wind from which a lateral force results.

rotating body, is called the **Magnus effect**. It is proportional to $\vec{\Omega} \times \vec{v}$, where $\vec{\Omega}$ is the angular velocity of the cylinder and $\vec{v}$ is the velocity of the wind.

References

Abbott, I. H., and A. E. von Doenhoff, *Theory of Wing Sections*, Dover, New York, **1949**.

Duncan, W. J., A. S. Thom, and A. D. Young, *An Elementary Treatise on the Mechanics of Fluids*, The English Language Book Society, **1960**.

Eggleston, D. M., and Stoddard, F. S., *Wind Turbine Engineering Design*, Van Nostrand Reinhold, New York, **1987**.

Gipe, Paul, *Wind Energy Comes of Age*, John Wiley, **1995**.

Glauert, H., "Airplane Propellers", *in* W. F Durand (ed.), *Aerodynamic Theory*, Division I, Volume 4, Berlin: Springer, **1935**.

Manwell, J. F., J. G. McGowan, and A. L. Rogers, *Wind Energy Explained*, John Wiley, West Sussex, **2002**.

Portnyagin, Yu. I., E. G. Merzlyakov, T. V. Solovjova, Ch. Jacobi, D. Kuerschner, A. Manson, C. Meek, A. N. Fahrutdinova, and A.Yu. Elkin, Long-term trends and year-to-year variability of the mid-latitude MLT winds, *J. Atmos. Solar-Terr. Phys.* 68(17), pp. 1890–1901. Dec. **2006**.

Schmitz, G. Theorie und Entwurf von Windrädern optimaler Leistung, *Wiss. Zeitschrift de Universität Rostock*, 5. Jahrgang **1955/1956**.

PROBLEMS

15.1 At a given sea-level location, the wind statistics, taken over a period of one year and measured at an anemometer height of 10 m above ground, are as follows:

Number of hours	Velocity (m/s)
90	25
600	20
1600	15
2200	10
2700	5
remaining time	calm

The velocity is constant in each range (to simplify the problem).

Although the wind varies with the 1/7th power of height, assume that the velocity the windmill sees is that at its center.

The windmill characteristics are:

Efficiency	70%
Cost	150 $/m^2$
Weight	100 kg/m^2

The areas mentioned above are the swept area of the windmill.

If h is the height of the tower and M is the mass of the windmill on top of the tower, then the cost, C_T, of the tower is

$$C_T = 0.05\, h\, M.$$

How big must the swept area of the windmill be so that the average delivered power is 10 kW? How big is the peak power delivered—that is, the rated power of the generator? Be sure to place the windmill at the most economic height. How tall will the tower be?

Neglecting operating costs and assuming an 18% yearly cost of the capital invested, what is the cost of the MWh?

If the windmill were installed in La Paz, Bolivia, a city located at an altitude of 4000 m, what would the average power be, assuming that the winds had the velocity given in the table above? The scale height of the atmosphere is 8000 m (i.e., the air pressure falls exponentially with height with a characteristic length of 8000 m).

15.2 A utilities company has a hydroelectric power plant equipped with generators totaling 1-GW capacity. The utilization factor used to be exactly 50%—that is, the plant used to deliver every year exactly half of the energy the generators could produce. In other words, the river that feeds the plant reservoir was able to sustain exactly the above amount of energy. During the wet season, the reservoir filled but never overflowed.

Assume that the plant head is an average of 80 m and that the plant (turbines and generators) has an efficiency of 97%.

What is the mean rate of flow of the river (in m^3/s)?

With the development of the region, the utilities company wants to increase the plant utilization factor to 51%, but, of course, there is not enough water for this. So, they decided to use windmills to pump water up from the level of the hydraulic turbine outlet to the reservoir (up 80 m).

A survey reveals that the wind regime is as given in the following table:

$<v>$	Θ
(m/s)	
5	0.15
7	0.45
10	0.30
12	0.10

$<v>$ is the mean cubic wind velocity, and Θ is the percentage of time during which a given value of $<v>$ is observed. The generator can be dimensioned to deliver full power when $<v> = 12$ m/s or when $<v>$ is smaller. If the generator is chosen so that it delivers its rated full power for, say, $<v>$ m/s, then a control mechanism will restrict the windmill to deliver this power even if $<v>$ exceeds the 10 m/s value.

Knowing that the cost of the windmill is \$10 per m^2 of swept area, that of the generator is \$0.05 per W of rated output, the efficiency of the windmill is 0.7, and that of the generator is 0.95, calculate the most economic-limiting wind velocity.

What is the swept area of the windmill that will allow increasing the plant factor to 51%? (The pumps are 95% efficient.) What is the cost of the MWh, assuming an annual cost of investment of 20% and neither maintenance nor operating costs. The windmills are at sea level.

15.3 A windmill is at a sea level where the wind has the statistics:

v (m/s)	% of time
0	30
3	30
9	30
12	8
15	2

The velocity in this table should be the mean cubic velocity of the wind. However, to simplify the problem, assume that the wind actually blows at a constant 3 m/s 30% of the time, a constant 9 m/s another 30% of the time, and so on. The windmill characteristics are:

Efficiency (including generator)	0.8
Windmill cost	200 \$/m^2 of swept area
Generator cost	200 \$/kW of rated power.

Rated power is the maximum continuous power that the generator is designed to deliver. The duty cycle is 1—that is, the windmill operates continuously (when there is wind) throughout the year. Consider only investment costs. These amount to 20% of the investment, per year.

The system can be designed so that the generator will deliver full (rated) power when the wind speed is 15 m/s. The design can be changed so that rated power is delivered when the wind speed is 12 m/s. In this latter case, if the wind exceeds 12 m/s, the windmill is shut down. It also can be designed for rated power at 9 m/s, and so on.

We want a windmill that delivers a maximum of 1 MW designed so that the cost of the electricity over a whole year is minimized. What is the required swept area? What is the cost of electricity?

15.4 *For this problem, you need a programmable calculator or a computer.*

Consider an airfoil for which $C_L = 0.15\alpha$, $C_D = 0.015 + 0.015|\alpha|$, for $-15° < \alpha < 15°$ (α is the angle of attack). A wing with this airfoil is used in a vertical axis windmill having a radius of 10 m. The setup angle is 0.

Tabulate and plot α as well as the quantity D as a function of θ, for $u/v = 6$. Use increments of θ of 30° (i.e., calculate 12 values).

Considerations of symmetry facilitate the work. Be careful with the correct signs of angles. It is easy to be trapped in a sign error. Find the mean value of $<D>$.

15.5 In the region of Aeolia, on the island of Anemos, the wind has a most peculiar behavior. At precisely 0600, there is a short interval with absolutely no wind. Local peasants set their digital watches by this lull. Wind velocity then builds up linearly with time, reaching exactly 8 m/s at 2200. It then decays, again linearly, to the morning lull.

A vertical axis windmill with 30-m-high wings was installed in that region. The aspect ratio of the machine is 0.8, and its overall efficiency is 0.5. This includes the efficiency of the generator.

What are the average and the peak power generated? Assuming a storage system with 100% turnaround efficiency, how much energy must be stored so that the system can deliver the average power continuously? During what hours of the day does the windmill charge the storage system? Notice that the load always gets energy from the storage system. This is to simplify the solution of the problem. In practice, it would be better if the windmill fed the load directly and only the excess energy were stored.

15.6 A vertical axis windmill with a rectangular swept area has an efficiency whose dependence on the u/v ratio (over the range of interest) is given, approximately, by

$$\eta = 0.5 - \frac{1}{18}\left(\frac{u}{v} - 5\right)^2 .$$

The swept area is 10 m^2, the aspect ratio is 0.8, and wind velocity is 40 km/h.

What is the maximum torque the windmill can deliver? What is the number of rotations per minute at this torque? What is the power delivered at this torque? What is the radius of the windmill, and what is the height of the wings? If the windmill drives a load whose torque is given by

$$\Upsilon_L = \frac{1200}{\omega}\ \text{Nm}$$

where ω is the angular velocity, what is the power delivered to the load? What are the rpm when this power is being delivered?

15.7 An engineering firm has been asked to make a preliminary study of the possibilities of economically generating electricity from the winds in northeastern Brazil. A quick and very rough estimate is required. Assume that the efficiency can be expressed by the ultrasimplified formula of Equation 15.48 in this chapter. The results will be grossly overoptimistic because we fail to take into account a large number of loss mechanisms, and we assume that the turbine will always operate at the best value of u/vD, which, for the airfoil in question is 4.38. Our first cut will lead to unrealistic results but will yield a ballpark idea of the quantities involved.

In the region under consideration, the trade winds blow with amazing constancy, at 14 knots. Assume that this means 14 knots at a 3-m anemometer height. The wind turbines to be employed are to have a rated power of 1 MWe (MWe = MW of electricity).

For the first cut at the problem we will make the following assumptions:

a. The configuration will be that of a Gyromill, using three wings.

b. Wind turbine efficiency is 80%.

c. The wings use the Göttingen-420 airfoil (see the drawing at the end of the problem). Use the simple efficiency formula derived in this chapter.

d. The wind turbine aspect ratio will be 0.8.

e. The wings will be hollow aluminum blades. Their mass will be 25% of the mass of a solid aluminum wing with the same external shape. Incidentally, aluminum is not a good choice for wind turbine wings because it is subject to fatigue. Composites are better.

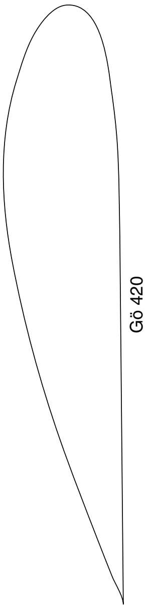
f. The total wind turbine masses three times the mass of the three wings taken together.

g. Estimated wind turbine cost is \$1.00/kg. Notice that the cost of aluminum was \$1200/kg in 1852, but the price is now down to \$0.40/kg.

h. The cost of investment is 12% per year.

Estimate:

1. The swept area.
2. The wing chord.
3. The wing mass.
4. The Reynolds number when the turbine is operating at its rated power. Assume that this occurs at the optimum u/v ratio.
5. The rpm at the optimum u/v ratio.
6. The torque under the above conditions.
7. The tension on each of the horizontal supporting beams (two beams per wing).



8. The investment cost per rated kW. Assuming no maintenance or operating cost, what is the cost of the generated kWh? Use a utilization factor of 25%.

15.8 A sailboat has a drag, F, given by

$$F = aW^2,$$

where F is in newtons, W is the velocity of the boat with respect to the water (in m/s), and $a = 80\,\text{kg/m}$.

The sail of the boat has a 10 m^2 area and a drag coefficient of 1.2 when sailing before the wind (i.e., with a tailwind).

Wind speed is 40 km/h.

When sailing before the wind, what is the velocity of the boat? How much power does the wind transfer to the boat? What fraction of the available wind power is abstracted by the sail?

15.9 A vertical axis windmill of the Gyromill configuration extracts (as useful power) 50% of the available wind power. The windmill has a rectangular swept area with a height, H, of 100 m and an aspect ratio of 0.8.

The lower tips of the wings are 10 m above ground. At this height, the wind velocity is 15 m/s. It is known that the wind increases in velocity with height according to the 1/7th power law.

Assuming that the windmill is at sea level, what power does it generate?

15.10 *Solve equations by trial and error. Use a computer.*

Can a wind-driven boat sail directly into the wind? Let's find out (forgetting second-order effects).

As a boat moves through the water with a velocity, w (relative to the water), a drag force, F_w, is developed. Let $F_w = 10w^2$.

The boat is equipped with a windmill having a swept area of $100\,\text{m}^2$ and an overall efficiency of 50%. The power generated by the windmill is used to drive a propeller, which operates with 80% efficiency and creates a propulsive force, F_p, that drives the boat.

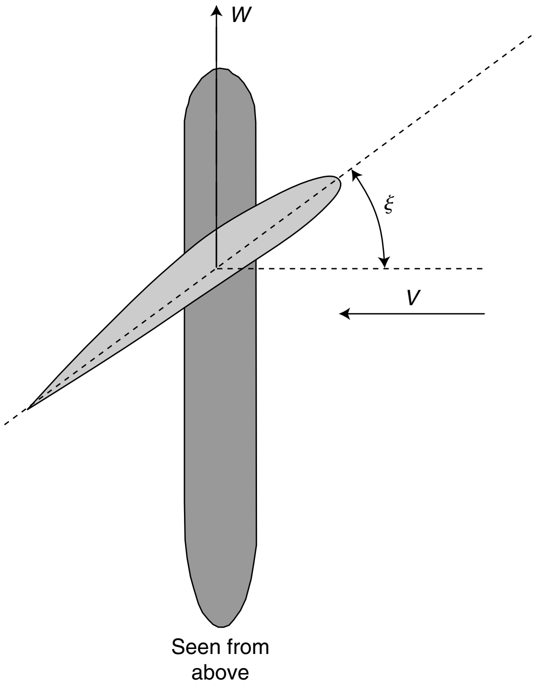
The windmill is oriented so that it always faces the *induced* wind, that is, the combination of v and w.

The wind exerts a force, F_{WM}, on the windmill. This force can be estimated by assuming a $C_D = 1.1$ and taking the swept area as the effective area facing the wind.

Wind velocity, v, is 10 m/s. What is the velocity, w_S, of the boat in the water? Plot w as a function of the angle, ϕ, between the direction of the wind and that of the boat. In a tailwind, $\phi = 0$ and in a headwind, $\phi = 180°$.

The boat has a large keel, so that the sideways drift caused by the "sail" effect of the windmill is negligible.

15.11 *Solve equations by trial and error. Use a computer.*



A vehicle is mounted on a rail so that it can move in a single direction only. The motion is opposed by a drag force, F_W,

$$F_W = 100W + 10W^2,$$

where W is the velocity of the vehicle along the rail. Notice that the drag force above does not include any aerodynamic effect of the "sail" that propels the vehicle. Any drag on the sail has to be considered in addition to the vehicle drag.

The wind that propels the vehicle is perpendicular to $\vec{W}$ and has a velocity, $v = 10$ m/s. It comes from the starboard side (the right side of the vehicle).

The "sail" is actually an airfoil with an area of 10.34 m^2. It is mounted vertically; that is, its chord is horizontal, and its length is vertical. The reference line of the airfoil makes an angle, ξ, with the normal to $\vec{W}$. (See the figure.)

The airfoil has the following characteristics:

$$C_L = 0.15\alpha,$$

$$C_D = 0.015 + 0.015|\alpha|.$$

α is the angle of attack expressed in degrees. The two formulas above are valid only for $-15° < \alpha < 15°$. If α exceeds 15°, the wing stalls and the lift falls to essentially zero.

Conditions are STP.

Calculate the velocity, W, with which the vehicle moves (after attaining a steady velocity) as a function of the setup angle, ξ. Plot both W and α as a function of ξ.

15.12 An electric generator, rated at 360 kW at 300 rpm and having 98.7% efficiency at any reasonable speed, produces power proportionally to the square of the number of rpm with which it is driven.

This generator is driven by a windmill that, under given wind conditions, has a torque of 18,000 Nm at 200 rpm but produces no torque at both 20 and 300 rpm. Assume that the torque varies linearly with the number of rpm between 20 and 200 and between 200 and 300 rpm.

What is the electric power generated?

15.13 A car is equipped with an electric motor capable of delivering a maximum of 10 kW of mechanical power to its wheels. It is on a horizontal surface. The rolling friction (owing mostly to tire deformation) is 50 N regardless of speed. There is no drag component proportional to the velocity. Frontal area is $2\,\text{m}^2$, and the aerodynamic drag coefficient is $C_D = 0.3$.

Assume calm air at 300 K and at 1 atmosphere. What is the cruising speed of the car at full power?

With the motor uncoupled and a tailwind of 70 km/h, what is the car's steady-state velocity. The drag coefficient is $C_D = 1$ when the wind blows from behind.

15.14 A building is 300 m tall and 50 m wide. Its C_D is 1. What is the force that the wind exerts on it if it blows at a speed of 10 m/s at a height of 5 m and if its velocity varies with $h^{1/7}$ (h being height above ground).

15.15 A windmill has a torque versus angular velocity characteristic that can be described by two straight lines passing through the points:

$$\begin{aligned} 50 \text{ rpm},&\ \text{torque} = 0.\\ 100 \text{ rpm},&\ \text{torque} = 1200\,\text{Nm}.\\ 300 \text{ rpm},&\ \text{torque} = 0. \end{aligned}$$

1. If the load absorbs power according to $P_{load} = 1000\omega$, what is the power taken up by this load? What is the torque of the windmill? What is the angular velocity, ω?
2. If you can adjust the torque characteristics of the load at will, what is the maximum power that you can extract from the windmill? What is the corresponding angular velocity?

15.16 A windmill with a swept area of 1000 m^2 operates with 56% efficiency under STP conditions.

At the location of the windmill, there is no wind between 1800 and 0600. At 0600, the wind starts up and its velocity increases linearly with time from zero to a value that causes the 24-h average velocity to be 20 m/s. At 1800, the wind stops abruptly.

What is the maximum energy the windmill can generate in one year?

15.17 U.S. Windpower operates the Altamont Wind Farm near Livermore, California, and reports a utilization factor of 15% for its 1990 operation. The utilization factor is the ratio of the energy generated in one year, compared with the maximum a plant would generate if operated constantly at the rated power. Thus, a wind turbine rated at 1 kW would produce an average of 150 W throughout the year. Considering the intermittent nature of the wind, a 15% utilization factor is good. Hydroelectric plants tend to operate with a 50% factor.

1. It is hoped that the cost of the kWh of electricity will be as low as 5 cents. Assuming that the operating costs per year amount to 15% of the total cost of the wind turbine and that the company has to repay the bank at a yearly rate of 12% of the capital borrowed, what cost of the wind turbine cannot be exceeded if the operation is to break even?

   For your information, the cost of a fossil fuel or a hydroelectric plant runs at about 1000 dollars per installed kW.

2. If, however, the wind turbine actually costs $1000 per installed kW, then, to break even, what is the yearly rate of repayment to the bank?

15.18 Many swimmers specialize in both freestyle (crawl) and butterfly. The two strokes use almost the same arm motion, but the crawl uses an alternate stroke, whereas the butterfly uses a simultaneous one. The kicks are different but are, essentially, equally effective. Invariably, swimmers go faster using the crawls.

From information obtained in this course, give a first-order explanation of why this is so.

15.19 *Solve equations by trial and error. Use a computer.*

An airfoil of uniform chord is mounted vertically on a rail so that it can move in a single direction only. It is as if you had cut off a wing of an airplane and stood it up with its longer dimension in the vertical. The situation is similar to the one depicted in the figure of Problem 15.11.

There is no friction in the motion of the airfoil. The only drag on the system is that produced by the drag coefficient of the airfoil.

The *setup angle* is the angle between the airfoil reference plane and the normal to the direction of motion. In other words, if the rail is north-south, the airfoil faces east when the setup angle is zero and faces north when the setup angle is 90°.

Consider the case when the setup angle is zero and the wind blows from the east (wind is abeam). Since the wind generates a lift, the airfoil will move forward. What speed does it attain?

Conditions are STP. The area of the airfoil is $10\,\mathrm{m}^2$. Wind speed is $10\,\mathrm{m\,s^{-1}}$.

The coefficients of lift and drag are given by

$$C_L = 0.6 + 0.066\alpha - 0.001\alpha^2 - 3.8 \times 10^{-5}\alpha^3.$$
$$C_D = 1.2 - 1.1\cos\alpha.$$

15.20 Consider a vertical rectangular (empty) area with an aspect ratio of 0.5 (taller than wide) facing a steady wind. The lower boundary of this area is 20 m above ground, and the higher is 200 m above ground. The wind velocity at 10 m is 20 m/s, and it varies with height according to the $h^{1/7}$ law.

Calculate:

1. the (linear) average velocity of the wind over this area.
2. the cubic mean velocity of the wind over this area.
3. the available wind power density over this area.
4. the mean dynamic pressure over this area.

Assume now that the area is solid and has a C_D of 1.5. Calculate:

5. the mean pressure over this area.
6. the torque exerted on the root of a vertical mast on which the area is mounted.

15.21 The retarding force, F_D, on a car can be represented by

$$F_D = a_0 + a_1 v + a_2 v^2.$$

To simplify the math, assume that $a_1 = 0$.

A new electric car is being tested by driving it on a perfectly horizontal road on a windless day. The test consists of driving the vehicle at constant speed and measuring the energy used up from the battery. Exactly 15 kWh of energy is used in each case.

When the car is driven at a constant 100 km/h, the distance covered is 200 km. When the speed is reduced to 60 km/h, the distance is 362.5 km.

If the effective frontal area of the car is 2.0 m^2, what is the coefficient of aerodynamic drag of the vehicle?

15.22 The army of Lower Slobovia needs an inexpensive platform for mounting a reconnaissance camcorder that can be hoisted to some height between 200 and 300 m. The proposed solution is a kite that consists of a Göttingen-420 airfoil with 10 m^2 of area tethered by means of a 300-m-long cable. To diminish the radar signature, the cable is a long monocrystal fiber having enormous tensile strength so that it is thin enough to be invisible, offers no resistance to airflow, and has negligible weight.

In the theater of operation, the wind speed is a steady 15 m/s at an anemometer height of 12 m, blowing from a 67.5° direction. It is known that this speed grows with height exactly according to the 1/7–power law.

The wing loading (i.e., the total weight of the kite per unit area) is 14.9 $\mathrm{kg\,m^{-2}}$.

The airfoil has the following characteristics:

$$C_L = 0.5 + 0.056\alpha,$$
$$C_D = 0.05 + 0.012|\alpha|,$$

where α is the attack angle in degrees.

The above values for the lift and drag coefficients are valid in the range $-10° < \alpha < 15°$.

The tethering mechanism is such that the airfoil operates with an angle of attack of 0.

The battlefield is essentially at sea level.

The questions are:

1. Assume that the kite is launched by somehow lifting it to an appropriate height above ground. What is the hovering height?
2. What modifications must be made so that the kite can be launched from ground (i.e., from the height of 12 m). (No fair changing the wing loading.) Qualitative suggestions with a minimum of calculation are acceptable.

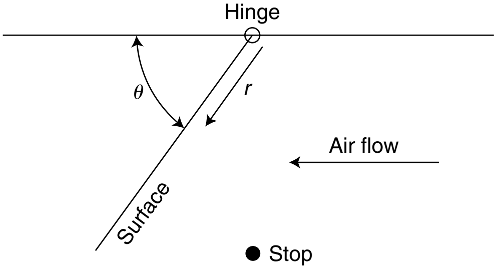
15.23 A car massing 1000 kg has an effective frontal area of 2 m^2. It is driven on a windless day on a flat, horizontal highway (sea level) at the steady speed of 110 km/h. When shifted to neutral, the car will, of course, decelerate, and in 6.7 seconds its speed is down to 100 km/h.

From these data, estimate (very roughly) the coefficient of aerodynamic drag of the car. What assumptions and/or simplifications did you have to make to reach such an estimate?

Is this estimate of C_D an upper or a lower limit? In other words, do you expect the real C_D to be larger or smaller than the one you estimated?

15.24 In early airplanes, airspeed indicators consisted of a surface exposed to the wind. The surface was attached to a hinge (see figure), and a spring (not shown) torqued the surface so that, in absence of air flow, it would hit a stop and thus assume a position with $\theta = 90°$.

The wind caused the surface to move, changing θ. This angle, seen by the pilot, was an indication of the air speed.



In the present problem, the surface has dimensions L (parallel to the axis of the hinge) by D (perpendicular to L). $L = 10\,\text{cm}$, $D = 10\,\text{cm}$.

The spring exerts a torque

$$\Upsilon_{spring} = \frac{0.1}{\sin\theta}\ \text{Nm}.$$

The coefficient of drag of the surface is 1.28. Air density is $1.29\,\text{kg/m}^3$. Calculate the angle, θ, for wind velocities, v, of 0, 10, 20, and 50 m/s.

15.25 An EV (electric vehicle) is tested on a horizontal road. The power, P, delivered by the motors is measured in each run, which consists of a 2-km stretch covered at constant ground velocity, v. Wind velocity, w, may be different in each run.

Here are the test results:

Run	Wind direction	Wind speed (m/s) W	Car speed (km/h) v	Power (kW) P
1	—	0	90	17.3
2	Headwind	10	90	26.6
3	Headwind	20	90	39.1
4	—	0	36	2.1
5	Tailwind	35	90	4.3

How much power is required to drive this car, at 72 km/h, into a 30-m/s headwind?

15.26 Here are some data you may need:

Quantity	Earth	Mars	Units
Radius	6.366×10^6	3.374×10^6	m
Density	5517	4577	kg/m^3
Surface air pressure	1.00	0.008	atmos.
Surface air temperature	298	190	K
Air composition	20% O_2, 80% N_2	100% CO_2	
Gravitational constant	6.672×10^{-11}		$\text{N m}^2\text{kg}^{-2}$

A parachute designed to deliver a 105-kg load to Mars is tested on Earth when the air temperature is 298 K, and the air pressure is 1.00 atmospheres. It is found that it hits the surface with a speed of 10 m/s.

Assume that the mass of the parachute itself is negligible. Assume the drag coefficient of the parachute is independent of the density, pressure, and temperature of the air.

If we want to have a similar parachute deliver the same load to Mars, what must its area be? Compare with the area of the test parachute used on Earth.

15.27 An EV experiences an aerodynamic drag of 320 N when operated at sea level (1 atmosphere) and 30 C.

What is the drag when operated at the same speed at La Paz, Bolivia (4000-m altitude, air pressure 0.6 atmospheres) and at a temperature of -15 C?

15.28 A trimaran is equipped with a mast on which a flat rigid surface has been installed to act as a sail. This surface is kept normal to the induced wind direction. The boat is 25 km from the shore, which is due north of it. A 36-km/h wind, v, blows from south to north. How long will it take to reach the shore if it sails straight downwind? Ignore any force the wind exerts on the boat except that on the sail.

The area, A, of the sail is $10\,\mathrm{m}^2$.

The coefficient of drag of a flat surface is $C_D = 1.28$.

The air density is $\rho = 1.2\,\mathrm{kg/m}^3$.

The water exerts a drag force on the trimaran given by

$$F_{water} = 0.5 \times W^2,$$

where W is the velocity of the boat relative to the water (there are no ocean currents).

15.29 Two identical wind turbines are operated at two locations with the following wind characteristics:

Location 1

Percent of time	Wind speed
50	10 m/s
30	20 m/s
20	25 m/s

Location 2

Percent of time	Wind speed
50	15 m/s
50	21 m/s

Which wind turbine generates more energy? What is the ratio of energy generated by the two wind turbines?

15.30 What is the air density of the planet in Problem 1.22 if the temperature is 450 C and the atmospheric pressure is 0.2 MPa?

15.31 One may wonder how an apparently weak effect (the reduction of pressure on top of an airfoil caused by the slightly faster flow of air) can lift an airplane.

Consider a Cessna 172 (a small 4-seater). It masses 1200 kg and has a total wing area of $14.5\,\mathrm{m}^2$. In horizontal flight at sea level, what is the ratio of the average air pressure under the wing to the pressure above the wing?

15.32 A car has the following characteristics:

$$\begin{aligned}\text{Mass, } m, &= 1200\,\text{kg}.\\ \text{Frontal area, } A, &= 2.2\,\text{m}^2.\\ \text{Coefficient of drag, } C_D &= 0.33.\end{aligned}$$

The experiment takes place under STP conditions.

When placed on a ramp with a $\theta = 1.7°$ angle, the car (gears in neutral, no brakes) will, of course, start moving and will accelerate to a speed of 1 m/s. This speed is maintained independently of the length of the ramp. In other words, it will reach a **terminal velocity** of 1 m/s.

When a steeper ramp is used ($\theta = 2.2°$), the terminal speed is 3 m/s.

Now place the car on a horizontal surface under no wind conditions. Accelerate the car to 111.60 km/h and set the gears to neutral. The car will coast and start decelerating. After a short time, Δt, the car will have reached the speed of 104.4 km/h.

What is the value of Δt?

15.33 The observed efficiency of a Gyromill-type wind turbine is

$$\begin{aligned}&\eta = 0, && \text{for } u/v \le 2,\\ &\eta = 0.280(u/v - 2), && \text{for } 2 \le u/v \le 5,\\ &\eta = -0.420u/v + 2.940, && \text{for } u/v > 5.\end{aligned}$$

The turbine has two blades or wings each 30 m long, and the radius of the device is 9 m.

When operating at sea level under a uniform 15-m/s wind, what power does it deliver to a load whose torque is 50,000 Nm independent of the rotational speed? What is the rotation rate of the turbine (in rpm)?

15.34 A standard basketball has a radius of 120 mm and a mass of 560 grams. Its coefficient of drag, C_D, is 0.3 (a wild guess), independent of air speed.

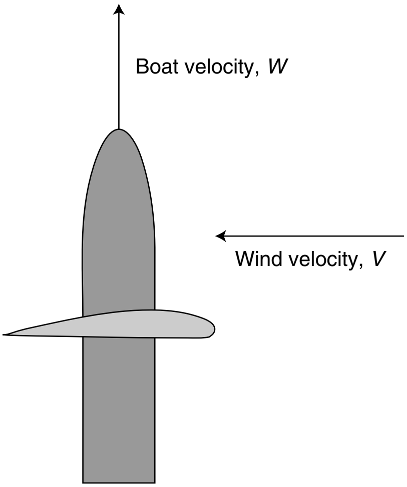
Such a ball is dropped from an airplane flying horizontally at 12-km altitude over the ocean. What is the velocity of the ball at the moment of impact on the water?

Make reasonable assumptions.

15.35 This is a terrible way to sail a ship but leads to a simple problem.

In the absence of wind relative to the boat, a boat's engine power, P_{Eng}, of 20,680 W is needed to maintain a speed of 15 knots (1 knot is 1852 meters per hour). The efficiency of the propeller is 80%. Assume that water drag is proportional to the water speed squared.

Under similar conditions, only 45 W are needed to make the boat move at 1 m/s.



This very boat is now equipped with a 10-m^2 airfoil, mounted vertically and oriented perpendicularly to the boat's axis. (See figure.)

The coefficient of lift of the airfoil is

$$C_L = (0.05\alpha + 0.5)$$

and is valid for $-10 < \alpha < 10$. In these two equations, α is in degrees.

The airfoil exerts a lift that, it is to be hoped, propels the boat due north when a 15 knot wind blows from the east.

What is the speed of the boat?

15.36 A sail plane (a motorless glider) is at a 500-m altitude and is allowed to glide down undisturbed. The atmosphere is perfectly still (no wind, no thermals [vertical wind]). Air temperature is 0 C, and air pressure is 1 atmosphere.

The wings have a 20-m^2 area, their lift coefficient is $C_L = 0.5$, and their drag coefficient is $C_D = 0.05$. Assume, to simplify the problem, that the rest of the sail plane (fuselage, empennage, etc.) produces no lift and no drag. The whole machine (with equipment and pilot) masses 600 kg.

Naturally, as the sail plane moves forward, it loses some altitude. The *glide ratio* is defined as the ratio of distance moved forward to the altitude lost.

1. What is the glide ratio of this sail plane?
2. What is the forward speed of the plane?
3. To keep the plane flying as described, a certain amount of power is required. Where does this power come from, and how much is it?

15.37 The drag force on a car can be expressed as a power series in v, the velocity of the car (assuming no external wind):

$$F_D = a_0 + a_1 v + a_2 v^2.$$

For simplicity, assume $a_0 = 0$.

A car drives 50 km on a horizontal road (at sea level) at a steady speed of 60 km/h. Careful measurements show that a total of 1.19×10^7 J were used. Next, the car drives another 50 km at a speed of 120 km/h and uses 3.10×10^7 J. The frontal area of the car is 2.0 m^2. What is the coefficient of drag, C_D, of the car?

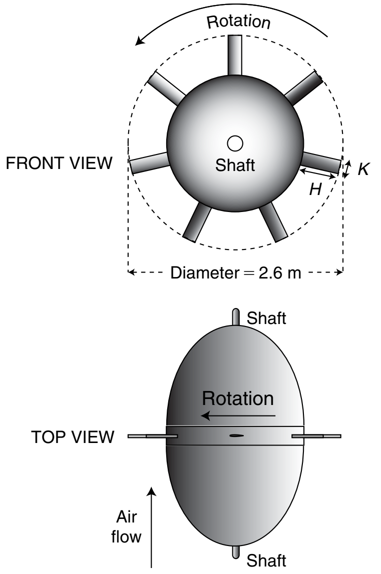
15.38

Percentage of time	**m/s**
10	Calm
20	5
40	10
30	15

The wind statistics (over a whole year) at a given site are as shown in the table.

When the wind has a speed of 15 m/s, the wind turbine delivers 750 kW. What is the number of kWh generated in a one-year period?

15.39



Consider a turbine consisting of seven blades each shaped as a NACA W1 symmetric airfoil with a 32-cm chord, K, and sticking out 52 cm, H, above a fairing. These blades have their reference plane aligned with the plane of rotation of the turbine. The diameter is 2.6 m. The device spins at 1050 rpm in such a direction that, if the turbine rotates in calm air, the angle of attack is zero.

To an acceptable approximation, the coefficients of the airfoil are

$$C_L = 0.08\alpha - 0.0001\alpha^3.$$

For $|\alpha| < 11°$:

$$C_D = 0.0062\exp(0.2|\alpha|).$$

For $11° < |\alpha| < 21°$:

$$C_D = -0.415 + 0.0564|\alpha| - 0.001|\alpha|^2.$$

In these formulas, α is in degrees.

The air stream that drives the turbine flows vertically up in the drawing (it flows parallel to the turbine shaft) and has a density three times that of the air at STP. It's velocity is 28.6 m/s.

How much power does the turbine deliver to the shaft?

15.40 Compare two different sites for wind farms: in one the wind blows at a steady 15 m/s, all the time; in the other the wind blows at a steady 20 m/s exactly half of the time during the other half, there is no wind. Both are at sea level.

a. In the two sites, what is the yearly amount of electrical energy generated by a propeller type of wind turbine having a rotor diameter of 50 m and an overall efficiency of 60%?

b. If the plants above, were installed in Cochabamba, 2558-m high in the Bolivian Andes, what would be their estimated yearly production of electricity? If you need more information, read the section on Boltzmann's equation in Chapter 2 of the book. Observe that the wind speed statistics at Cochabamba are exactly the same as at sea level and, in either case, refer to turbine height.

15.41 A wind turbine with 30-m radius and 60% efficiency is installed at a sea-level site in which the average wind regimen is given in the following table.

	Wind (m/s)	**% of time**
1	0	10
2	5	25
3	10	35
4	15	25
5	20	5

Assume that the wind velocity in each wind slot (1 though 5) is the cubic mean in that slot. The cost of the generator is \$0.25 per rated watt, and that of the turbine is \$900/m^2 of swept area. The yearly cost during the lifetime of the plant is 15% of the cost of the plant for both the generator and the wind turbine. Assume no other costs.

You can choose a generator of a rated power equal to the power the turbine delivers when the wind is 5 m/s. This means that if the wind speed exceeds 5 m/s, the plant will only generate the power corresponding to a 5-m/s wind. Another possibility is the choice of a generator rated at the power corresponding to 10 m/s.

You can also choose a generator rated at 15, or at 20 m/s.

You have four different choices. Clearly, the larger the generator, the more expensive the plant.

a. Calculate the cost of the kWh generated for each of the above choices. If you aim to produce the cheapest electricity, what size generator would you pick?

b. For the best choice of Item a, what is the total cost of the facility in dollars per rated kW?

c. For the best choice of Item a, what is the rotor loading of the turbine?

15.42 A matter accelerator is mounted at the edge of a mesa and launches a spherical projectile horizontally toward a level plain which is 100 m below the launching device. The problem is to determine how far the projectile goes before impacting the ground. First, assume that there is no air so that there is no aerodynamic drag. This will set the maximum value of the range (distance from launcher to impact point.) Next consider the case when there is air at STP, however, to simplify the problem, assume that the air drag influences only the horizontal component of the motion of the projectile, not the vertical component of the motion.

The necessary data to solve the problem include the following.

a. Initial velocity, v_0, is 720 km/hr.

b. The spherical projectile does not spin.

c. The projectile is hollow and is made of iron 2 cm thick.

d. The outer diameter of the projectile is 1 m.

e. The density of iron is 7874 kg/m^3.

f. Assume that the drag coefficient, C_D, is 1.0.

15.43 You are in charge of a sea-level wind farm that has bought a number of horizontal axis propeller-type wind turbines having the following specifications:

Cut-in wind speed:	3.5 m/s
Cut-out wind speed:	27 m/s
Rotor:	3-blade, 104-m diameter
Efficiency of the turbine/generator system (referred to the available wind power):	50%

A generator of selected power and the corresponding gear box can be matched to the turbine.

The turbine plus tower cost \$2.5 million. The generator cost depends, of course, on the chosen power: it is \$278 per kW of electric output. The cost of money is 12%/year. Operating cost per unit (1 turbine plus generator) is \$230,000/year. This includes insurance, taxes, and salaries.

All you know about the wind regime at the selected site is that it has traditionally been 12 m/s on average. This is obviously a very good site for wind power generation.

There will be times of the year when there will be enough wind to drive the generator to full power. Develop a formula that yields the total amount of energy (in kWh) delivered in one full year by the generator counting only the occasions when the generator is developing its full rated power, (P_g). Disregard the power generated at less than full rated power.

1. What is the cost of the electricity above, in mills/kWh[†] (use three significant figures, and, as instructed above, disregard the energy generated at less than full power) when the rated generator power is $P_g = 3.0\,\text{MW}$. Repeat for, $P_g = 3.6\,\text{MW}$.
2. Make a very rough guess: what is the torque the propeller exerts on the shaft that leads to the input of the gear box (just a ballpark figure). Assume the machine is operating at the rated wind speed, the one that just delivers full power. Make plausible assumptions. Consider the 3.6-MW case.
3. What would happen (qualitatively) to the torque if the turbine had only two blades instead of three, that is, if the solidity were decreased?
4. Which of the two turbines in the question above would presumably have less wake rotation loss?

15.44 An object masses 10 kg, has a frontal area of 0.3 m^2, and a drag coefficient of 1.1. This object is dropped from an airplane flying at 5000 m (initial vertical velocity is zero). The horizontal velocity plays no role—it can also be taken as zero. The atmospheric density (in

[†] A mill is one-tenth of a cent.

kg/m^3) is a function of height and is given by

$$\rho = 1.29 \exp\left(-\frac{h}{8000}\right).$$

The height, h, is in meters above sea level. Assume that the acceleration of gravity, g, is height independent. Describe quantitatively what happens. Calculate the maximum vertical velocity acquired by the object.

This will lead to a differential equation. Do not attempt to solve it analytically. Use a numerical solution with a time step of 1 second. Few iterations will be needed, and notwithstanding the very coarse time steps, will yield a good estimate of the velocity. Surprisingly, you will overestimate the correct velocity by less than 0.5%.

15.45 The GE 2.5-xl wind turbine has the following characteristics:

$$\begin{aligned} P_{rated} &= 2.5 \text{ MW}.\\ v_{cutin} &= 3.5 \text{ m/s}.\\ v_{rated} &= 12.5 \text{ m/s}.\\ v_{cutout} &= 25 \text{ m/s}. \end{aligned}$$

Three 50-m rotor blades.

What is the rotor loading when the wind velocity is 10 m/s? A new generator is used having a rated power equal to the power the turbine delivers when there is a 10 m/s wind.

15.46 In some parts of Southern California, the Santa Ana winds blow steadily at 100 km/h during 10% of the year. During the rest of the time, there is negligible wind. You want to adapt the GE 2.5-xl to these conditions. You keep the same generator but change the gear box and the rotor (still three blades). Assuming all efficiencies are still the same, what is the length of the rotor blade?

15.47 Assuming optimum adjustment, what is the wind velocity just behind the rotor disk of the turbine in Problem 15.46?

15.48 Estimate the force the wind exerts on the rotor disk of the turbine in Problem 15.46. We are not talking about the torque force that turns the rotor, but rather, the pushing force that tries to topple the tower.

Chapter 16
Ocean Engines

16.1 Introduction

In Chapter 4, we discussed the utilization of the thermal energy of the oceans. Other forms of ocean energy are also available, such as mechanical energy from waves, currents and tides, and chemical energy from salinity gradients. A summary of the oceanic energy resources is presented by Isaacs and Schmitt (1980).

16.2 Wave Energy

OTECs took advantage of the ocean acting as an immense collector and storer of solar radiation, thus delivering a steady flow of low-grade thermal energy. The ocean plays a similar role in relation to the wind energy, which is transformed into waves far steadier than the air currents that created them. Nevertheless, waves are neither steady enough nor concentrated enough to constitute a highly attractive energy source notwithstanding their large total power: about 2 TW can possibly be captured according to the World Energy Council.

16.2.1 About Ocean Waves[†]

The following specialized terminology and symbology is required for the discussion of ocean [surface] waves:

Duration: Length of time the wind blows.
Fetch: Distance over which the wind blows.
d: The **depth** of the water.
g: The **acceleration of gravity**, 9.81 m/s^2.
h: The **height** of the wave—the vertical distance between the through and the crest of a wave.
T: The **period**—the time interval between two successive wave crests at a fixed point.
v: The **phase velocity** of the wave—the ratio between the wavelength and the period.

[†] A much more extensive discussion of ocean waves can be found in *Introduction to Physical Oceanography*, Chapter 16, by Robert H. Stewart, Texas A&M University accessible at http://oceanworld.tamu.edu/resources/ocng_text book/.

v_g: The **group velocity** of the wave—the velocity of wave energy propagation.

λ: The **wavelength**—the horizontal distance between two successive wave crests, measured along the direction of propagation.

16.2.1.1 The Velocity of Ocean Waves

There is little net horizontal motion of water in a surface ocean wave. A floating object drifts in the direction of the wave with about 1% of the wave velocity. A given elementary cell of water will move in a vertical circle, surging forward near the crest of the wave but receding by an almost equal amount at the trough. Near the surface, the diameter of the circle is equal to the wave height. As the depth increases, the diameters diminish—the motion becomes negligible at depths much larger than one wavelength. Thus, in deep waters ($d \gg \lambda$) the wave does not interact with the sea bottom, and its behavior is independent of depth. The wave velocity is, then, a function of the wavelength,

$$\underline{\text{Deep water}}, (d \gg \lambda): \qquad v = \sqrt{g}\left(\frac{\lambda}{2\pi}\right)^{1/2}. \tag{16.1}$$

Any system in which the wave velocity depends on wavelength is called **dispersive**; hence the deep ocean is dispersive. However, when the water is sufficiently shallow, the circular motion of the water is perturbed by the seafloor and the wave loses some of its energy. When $\lambda \gg d$, the velocity no longer varies with λ (the system is no longer dispersive) but depends now on the depth, d.

$$\underline{\text{Shallow water}}, (\lambda \gg d): \qquad v = \sqrt{g}\ d^{1/2}. \tag{16.2}$$

For intermediate depths, there is a transitional behavior of the wave velocity. If the water is very shallow (at $d \approx \lambda/7$), the velocity of the crest of the wave is too fast compared to that of the trough and the wave breaks.

We saw that in shallow waters, the wave velocity diminishes as the depth is reduced. This is the reason waves tend to come in parallel to a beach. In waves coming in at a sharp angle, the part closer to the shore (presumably, in shallower water) slows down, allowing the more distant parts to catch up. If one constructs an undersea mound in the shape of a spherical segment, it will act as a lens focusing the waves into a small region.

In a sloping beach, the wave farther out moves faster than one nearer shore, so the wavelength diminishes as the waves come in. The period, however, is not affected.

All the preceding refers to the phase velocity of the wave. The group velocity—the velocity of energy propagation—differs from the phase velocity in a dispersive medium. Using the two limiting approximations (deep and shallow water), we find that

$$\underline{\text{ Deep water}},\ (d \ll \lambda): \qquad v_g = \frac{v}{2}. \tag{16.3}$$

$$\underline{\text{Shallow water}},\ (\lambda \gg d): \qquad v_g = v. \tag{16.4}$$

16.2.1.2 Wave Height

The height of the wave depends on both the fetch and the duration of the wind. Figure 16.1 shows the empirically determined relationship involved. The fetch data assume that the wind has been blowing for a long time and the duration data assume that the wind has been blowing over a sufficiently long fetch. Both sets of curves seem to approach some final asymptotic value for each wind velocity.

The largest storm[†] wave ever observed (1933) had a height of 34 m and a wavelength of 342 m. Its period was 14.8 s. This leads to a phase velocity of 23.1 m/s and a group velocity of about 11.5 m/s.

16.2.1.3 Energy and Power

The energy density of an ocean wave is (approximately)

$$W = g\rho\left(\frac{h}{2}\right)^2 \qquad \text{J/m}^2, \tag{16.5}$$

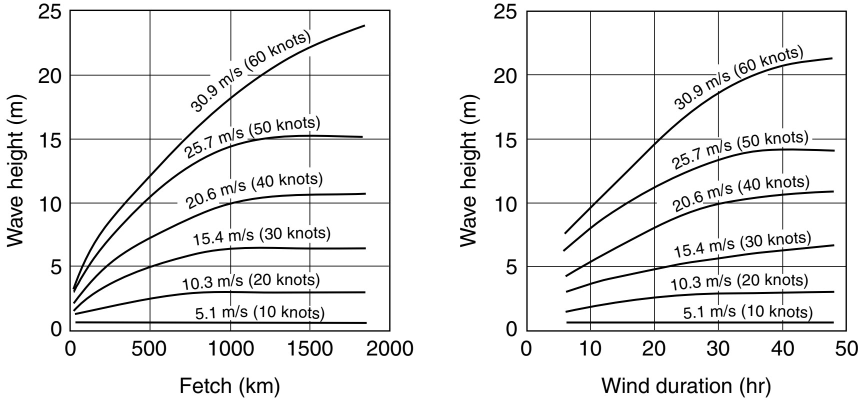


Figure 16.1 Wave height as a function of the fetch (left) and of the duration of the wind (right). A second-order regression was fitted through observed data points.

[†] Tsunamis can be much larger than storm waves. A tsunami in Lituya Bay, Alaska, on July 9, 1958 reached a height of 524 m.

where ρ is the density of the water (roughly 1000 kg/m^3). The power associated with a wave is

$$P = v_g W \qquad \text{W/m}. \tag{16.6}$$

Consider the power of the 1933 record wave (see above).

$$W = 9.8 \times 1000 \times \left(\frac{34}{2}\right)^2 = 2.8 \times 10^6 \qquad \text{J/m}^2. \tag{16.7}$$

Since the group velocity of this wave was 11.5 m/s, the power density was $11.5 \times 2.8 \times 10^6 = 32.5 \times 10^6$ W/m.

16.2.2 Wave Energy Converters

Perhaps the major difficulty with wave energy converters is the great range of powers in ocean waves. A moderate sea may have waves with a power of some 50 kW/m, while a big storm may generate powers of 10 MW/m. This 200:1 range is difficult of accommodate. Machines capable of economically using 50 kW/m seas tend to be too frail to resist large storms.

Renewable energy resources are frequently unsteady (variable wind velocities, fluctuating insolation, changing sea states, etc.) and have low power densities. A good wind farm site may operate with 400 W/m^2 rotor loading. Average insolation even in sunny Arizona barely exceeds some 250 W/m^2. Wave energy may be argued to have a more attractive mean power density. Indeed, one can find a number of sites throughout the world where ocean wave power density is above 50 kW/m. Assuming that a power plant uses, say, 10 m of land perpendicular to the shore, such sites would provide some 5 kW/m^2, an order of magnitude more than wind and sun. Clearly, this kind of comparison is only marginally meaningful but may still be used as a point in favor of the use of ocean waves.

Another advantage of ocean wave power systems is that they can be integrated into coastal structures, including sea walls and jetties, constructed to combat erosion caused by the action of the sea. Such integration may be economically interesting. However, wave energy converters do not necessarily have to be built on the shore line; they may also be placed offshore, even though this may create substantial mooring or anchoring problems.

16.2.2.1 Offshore Wave Energy Converters[†]

Offshore wave energy converters usually belong to one of four categories:

1. Heaving buoys.
2. Hinged contour.
3. Overtopping.

[†] An extensive review of offshore devices can be found in E21 EPRI Assessment.

4. Oscillating water column. This category of wave-energy converters can be used both offshore or on shoreline. We will discuss these machines when shoreline devices are examined later in this chapter.

Heaving Buoy Converters

Although the particular arrangement for the conversion of wave energy into mechanical energy, and then, presumably into electricity depicted in Figure 16.2 does not appear to be practical, it serves to illustrate the manner in which more complicated devices of this class operate. It is a wave-operated pump consisting of a buoy attached to a long vertical pipe equipped with a check valve at the bottom.

When a wave lifts the device, water in the pipe is accelerated upward because it cannot escape through the check valve. As the wave recedes, the pipe accelerates downward against the movement of the internal water, creating a considerable hydraulic pressure. The water can be allowed to escape through the top to drive a turbine. Simultaneously, more water enters through the check valve.

Let A be the cross-sectional area of the pipe, L its length, δ the density of water, and γ the peak acceleration of the pipe.

The mass of water in the pipe is $M = \delta AL$, and the resulting force is $F = \gamma\delta AL$. The corresponding pressure (when the pipe is accelerated vertically) is $p = \gamma\delta L$. If $L = 100\,\text{m}$ and $\gamma = 5\,\text{ms}^{-2}$, then the pressure is $0.5\,\text{MPa}$ (5 atmos).

The vertical oscillation of the water can act on floats that drive levers and pistons, but such an arrangement may be too complex and too vulnerable to stormy conditions. Simpler schemes may work better. Heaving buoys in several different versions seem, at this point, to be the most successful

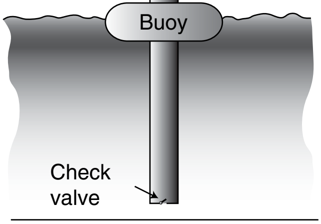


Figure 16.2 Wave-operated pump.

solution. The technology is still very young, but a few projects seem to be close to practical operation.

AquaBuOY, a subsidiary of Finavera Renewables, has developed a floating heaving buoy device based roughly on the principles described above. The vertically oscillating water column drives a piston, which in turn pumps water at a higher pressure into an accumulator acting as a low-pass filter. The accumulator discharges the water into a turbine for generation of electricity. A prototype was tested using a 6-m diameter buoy equipped with a pipe that dips 30 m down. (The system requires water depths of 50 m or more.) The prototype had a rated power of 250 kW at reasonable wave heights. The system uses slack mooring. AquaBuOY has three wave energy projects under construction, one in Figueira da Foz (near Coimbra, Portugal), one in Makah Bay, Washington, and one in Ucluelet, British Columbia, Canada, planned, respectively, for 2008, 2009, and 2010. The combined power of the three projects is 200 MW peak.

Seadog (Independent Natural Resources, Inc.) is a bottom-dwelling heaving buoy system. The 5.7-m diameter float pumps water into an elevated basin onshore. Discharge from the basin drives a conventional hydroturbine. The system requires a depth of 20 m.

AWS Ocean Energy, promises to deploy a 250-kW demonstrator (called Archimedes Waveswing) in 2009, with commercial rollout to follow by 2011.

Hinged Contour Converters

Both the AquaBuOY and the Seadog mentioned are **point absorbers** because they interact with only a small ocean area. To generate large amounts of energy, a multitude of these devices must be deployed, each with its own piston and power takeoff equipment. The solution developed by Pelamis Wave Power, formerly (prior to 2007) known as Ocean Power Delivery, Ltd., although also using a system of buoys, is capable of interacting with a much larger ocean area.

The 750-kW prototype called "Pelamis" (see Figure 16.3), installed at the European Marine Energy Centre in Orkney,[†] consisted of four tubular steel floats measuring 4.63 m in diameter attached to one another by hinges. The length was 150 m.

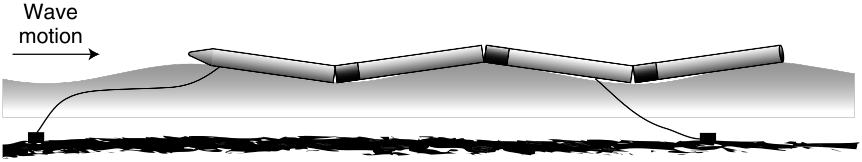


Figure 16.3 The Pelamis hinged contour converter.

[†] The Orkney Islands form an archipelago some 50 km north of the northernmost tip of Scotland.

The force the waves exert in moving each segment relative to its neighbors is captured by hydraulic rams that press a biodegradable fluid into accumulators, which, in turn, power a number of 125-kW generators.

The system was moored by a three-point configuration that allowed Pelamis to orient itself normal to the incoming wave front. The natural resonance of the system is automatically altered to match the frequency of the waves. However, when exposed to storm conditions, the system is detuned to minimize the stress on the mooring. Simulations and actual tests show that Pelamis appears to exhibit excellent storm survivability.

In 2008, three Pelamis P1 machines were installed and were in the process of being commissioned at the Aguçadoura ocean energy farm, 5 km offshore in northern Portugal.[†] These three units will deliver 2.25-MW peak (3 × 750) with an estimated plant factor of between 25% and 49%. The plant factor depends on the average yearly ocean behavior and will probably be narrowed down, as electricity production data are accumulated.

A slightly larger Pelamis-equipped wave farm is planned for the Orkney site where the original prototype was tested. It will consist of four units (3 MW, peak).

The P1 production machine differs slightly from the prototype. It is 140-m long and 3.5 m in diameter.

Overtopping Converters

Wave Dragon ApS (originally a Danish company, which has now moved its center of operation to Wales) is developing a fundamentally simple wave energy converter (Figure 16.4). It consists of a large floating basin or reservoir a few meters above sea level. Waves (concentrated by a pair of reflector arms) hit a ramp rising up and spilling into the reservoir. The water flows out back to the ocean driving a number of Kaplan turbines, thus, generating electricity. One advantage of this design is that the only moving parts (other than the flowing water) are turbines and generators.

The manufacturer feels confident that the Wave Dragon will withstand intense winds (owing to its low-in-the-water profile) and large waves that simply flow over the installation. A 57-m-wide prototype rated at 20 kW has been in operation since March 2003.

Observers are optimistic about the Wave Dragon's possibilities. Its one obvious disadvantage is the unfavorable mass-to-power ratio: the proposed 4-MW, 300-meter-wide machine masses over 30,000 tons!

[†] Portugal, currently (2008), gets 39% of its electricity from renewable sources and is vigorously expanding this area. Up to 2008, the 11-MW photovoltaic facility at Serpa in southeast Portugal was the world largest. It is now surpassed by the 16.1-MW Göttelborn facility in Germany. Solarpark Waldpolenz, also in Germany, is being expanded to 40 MW.

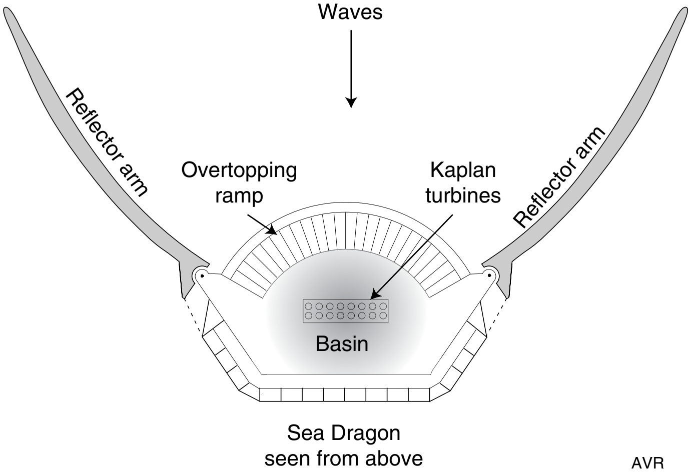


Figure 16.4 The Wave Dragon.

A 7-MW Wave Dragon was deployed offshore of Pembrokeshire[†] in 2008. The reflector arms intercept a 300-m wavefront and focus the wave on a 140-m-wide ramp. Currently, it is the largest wave energy converter in the world. The unit will remain in that area for three to five years collecting operational experience. Since the area is a "Special Area of Conservation," it is probable that the unit will eventually be moved farther out to sea.

16.2.2.2 Shoreline Wave Energy Converters

Shoreline wave energy converters usually belong to one of two categories: the first is the **tapered channel** type also known as **tapchan**, and the second is the **oscillating water column** (OWC). OWCs can be installed either onshore or in deep water **offshore sites**. The tapchan by its very nature can only be used as a shoreline device.

As discussed in Sub-subsection 16.2.1.1, underwater structures can focus a large wave front into a smaller aperture of a wave energy converter.

Tapered Channel System[††]

An ingenious wave energy system was developed by a Norwegian company appropriately called Norwave. The system was demonstrated in a fjord at

[†] The British Isles consist essentially of two islands, the largest being, not surprisingly, Great Britain. On the southwest there is a peninsula known as Wales, the westernmost part of which is Pembrokeshire.

[††] In 1877, Giovanni Schiaparelli, using a favorable opposition of Mars, drew a detailed map of the planet and described what he perceived as "canali." Unfortunately, this was translated into English as "canals" instead of "channels" and supported the view that Mars was inhabited. The expression "tapered channel" suffers from the opposite error. Almost invariably, we are referring to artificially built canals, not to natural channels.

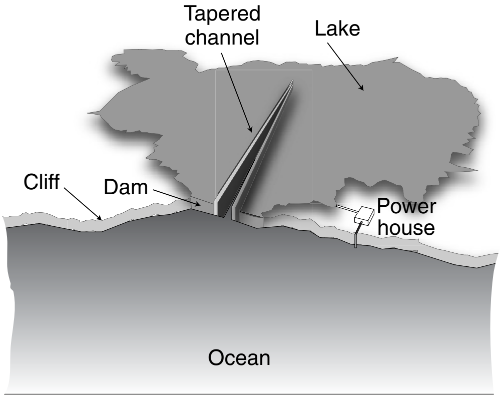


Figure 16.5 NORWAVE wave power plant.

Toftestallen in the vicinity of Bergen, Norway. The general arrangement is sketched out in Figure 16.5.

A narrow, tapered, concrete canal, whose walls rise well above sea level connects the ocean with a bay or lake. Waves entering the tapered canal move along it, increasing progressively in height (because of the gradual narrowing) until water spills over the canal walls. Thus, the waves pump water into the lake, causing its level to become substantially higher than that of the ocean. This height differential was used to drive a hydraulic turbine that generated electricity.

The only artifact that is directly exposed to the waves is the concrete canal, which can be designed to resist any reasonable storm, although in the Norwave prototype it apparently suffered severe damage that caused the discontinuation of the project in the early 1990s after being in intermittent operation since 1986. The actual generating equipment is a normal low-head hydroelectric power plant except that it must be built to resist the corrosive action of the salt water. The Norwave prototype was modest in size. It was equipped with a 350-kW generator fed by a reservoir with 5500 m^2 surface area and a water level 3 m above sea level. The opening to the sea was 60 m wide, but the concrete canal started with only a 3-m width and narrowed progressively to 20 cm at its far end. Its length was 80 m and its depth 7 m. The system was to operate with a 14-m^3/s pumping capacity. This corresponds to an available water power of slightly more than 400 kW, enough to drive the 350-kW generator.

After the Norwave program closed, interest in tapered channel wave energy plants seems to have waned. Perhaps this is because tapchans work only in selected sites where the wave regimen is steady and where tides are less than, say, 1 m. Very low tides are not too common, the exception being in the Mediterraneam, where tides are only a few centimeters and thus escaped the attention of Aristotle, who does not mention the phenomenon in writings.

Oscillating Water Column (OWC)—Wavegen System

Another promising scheme for using ocean wave energy was constructed by the Scottish company WAVEGEN and was commissioned in November 2000. It bears the whimsical name **LIMPET** (Land Installed Marine Powered Energy Transformer) and is not a commercial unit, but rather, a further research and development unit. It is a scale-up of a 75-kW prototype that operated from 1991 to 1999. LIMPET is a 500-kW plant installed on the island of Islay.[†] It is powered by the oscillating water column generated by wave action. A slanted concrete tubelike structure is open on one side to the ocean. Waves cause alternate compression and suction in the air column inside the tube. The resulting forward and backward air flow drives a Wells turbine, which has the unusual property of rotating in the same direction regardless of the direction of the air flow. (See Problem 15.39).

The Wells turbine has much lower efficiency than a normal, unidirectional flow turbine. Its efficiency is between 40% and an optimistic 70%, in part because it must use symmetrical airfoils rather than the usual highly asymmetrical ones. In addition, Wells turbines are not self-starting; they require that the generator be used as a motor for starting purposes. The alternative solution for extracting energy from a periodically reversing air stream would be a mechanically complicated rectification system using fast-acting valves. This would be expensive and would lead to costly maintenance.

A water inlet consists of three 6-by 6-m rectangular ducts whose air columns merge into a common turbine inlet. The system is depicted (in an idealized fashion) in Figure 16.6. The two Wells turbines measure 2.6 m in diameter and operate at between 700 and 1500 rpm. To accommodate this large speed variation and to facilitate interfacing with the local power grid, the output of the two generators is rectified and then inverted into alternating current of the proper frequency.

Among other advantages, the LIMPET represents a low visual intrusion and causes a minimal impact on fauna and flora. According to WAVEGEN, "ongoing projects include development of a floating device and a tunnelled cliff shoreline array in the Faroe Islands (ultimately up to 100 MW)."

WAVEGEN, now a subsidiary of Voith Siemens, a major manufacturer of hydraulic machinery, operates the 500-kW LIMPET plant in Islay, which is advertised as the first commercial-scale grid-connected wave power station in the world.

16.3 Tidal Energy

Tides can be utilized as energy sources through the currents they cause or through the associated variations in ocean level. The most effective way of

[†] Islay is one of the larger islands of a complex archipelago, known as the Hebrides, west of Glasgow and north of Northern Ireland.

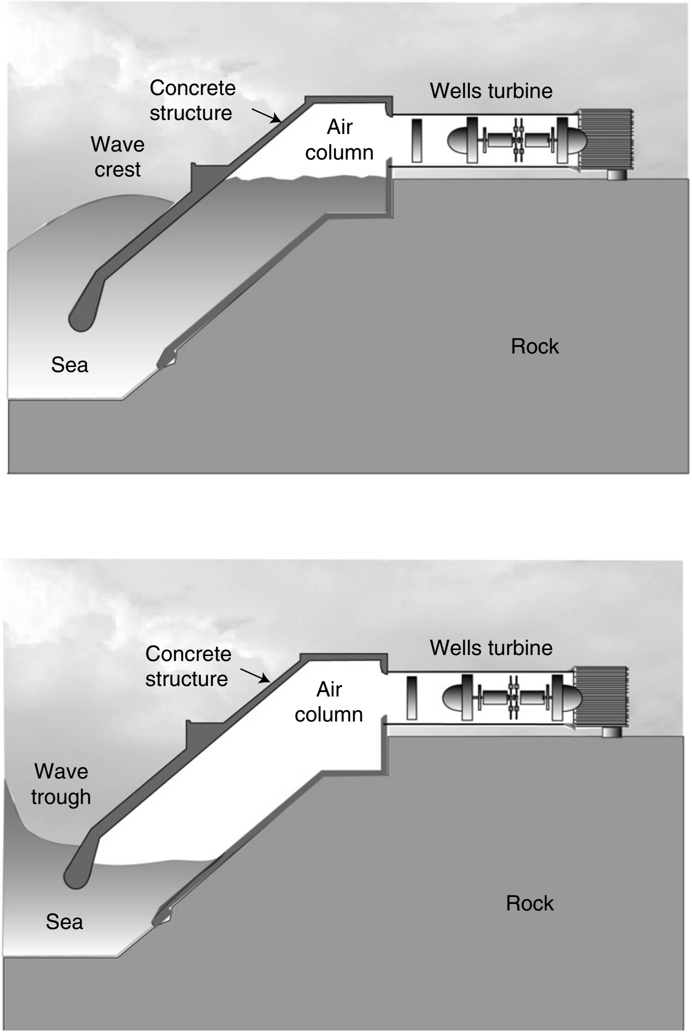


Figure 16.6 The crest of the wave (top picture) corresponds to maximum compression into the Wells turbine, while the trough (bottom picture) corresponds to maximum suction. Owing to the resonance of the water column in the concrete structure, the amplitude of the oscillations in the structure exceeds the amplitude of the waves. The turbine will generate electricity regardless of the direction of the air flow.

taking advantage of these vertical water displacements is through impoundment, and this is the only approach being considered seriously. Average tidal fluctuations in the open sea are small; however, local resonances can magnify the tidal effect considerably, as demonstrated dramatically by the high tides in Fundy Bay, especially at Cobequid Bay† where they reach a 16-m swing. The amount of water moved by the Fundy Bay tides equals the amount of water discharged by all the world's rivers: something like 100 cubic kilometers per day!

Impounding tides without "detuning" the system can be a problem, as has happened to some extent in the only major tidal power plant in existence—the one at the mouth of the Rance River, in France. It appears that the installation reduced the power flow of the estuary by 20%. Nevertheless, the system delivers an average of 160 MW (peak of 240 MW). It consists of 24 reversible turbines and a dam built across part of the estuary. At high tide, the water flows through the turbines into the estuary. At low tide, the water flows out, again through the turbines, generating power at both ebb and flood. Mean tides are 8.5 m, a large value for tides, but only a modest head for hydraulic turbines. These are of the Kaplan type and have a large diameter: 5.35 m. Near high and low tides, the turbines are used as pumps to accentuate the working head. The Rance River plant, located near Saint Malo, has been in operation since the 1970s. One of the difficulties it has experienced is corrosion owing to the use of salt water.

16.4 Energy from Currents

A recommended reading for those interested in this subject is the article by Fraenkel (2002).

Currents are an example of the enormous amount of energy stored in the oceans under unfavorably low-power densities. As with wind and direct solar, there is a problem with the intermittence of the available power. But whereas wind and sunlight can vary at random, ocean currents are more predictable, especially when caused by tides. The report "Marine Currents Energy Extraction" estimates that European waters can support at least 12 GW of ocean turbines able to generate an average of 5.5 GW. This is a load factor of 45%, which may be a bit optimistic. Other experts estimate a factor of between 35% and 40%. At any rate, it is much better than the 20%, or so, for land-based wind turbines. Offshore wind turbines may reach 30%. A 1.2-GW U.S.-operated nuclear power plant delivers an average of more than 1.08 GW—a load factor larger than 90%. European ocean turbines could replace about five large nuclear plants. *Caveat*: We are comparing an estimated performance of ocean turbines with actual performances of wind turbines and nuclear plants. There is always some degree of optimism in estimates.

†The Bay of Fundy is a substantial body of water between Nova Scotia and New Brunswick on the Canadian mainland. It has a general SSW to NNE orientation and ends in two "horns," the southernmost of which is Cobequid Bay.

Ocean currents are driven by numerous forces: winds, salinity gradients, thermal gradients, the rotation of Earth, and, mainly, by tides. Undisturbed tides are synchronized with the phases of the moon. If the moon had no phases—that is, if, for instance, one had an eternal full moon—then a high tide would come every 12 hours as a given part of the world rotated under the moon. There are two daily tides, one near the sublunar point and one on the opposite side of Earth. However, the moon orbits the Earth in 27.32 days and goes from full moon to new and back to full. On average, two successive full moons recur every 29.53 days. This is called a lunar month or a **lunation**. To understand the distinction between the lunar month and the length of the lunar orbital period, do Problem 12.28.

Meteorological and topographic influences make it impossible for simple theory to predict the exact amplitude or the exact frequency of the tides. In addition, tidal flow is often turbulent so that velocities cannot easily be translated into recoverable energy. Nonetheless, using the expressions developed in Chapter 15, we can estimate available power densities,

$$P_A = \frac{16}{27}\frac{1}{2}\rho v^3 A_v. \tag{16.8}$$

In a good site, water-flow velocities are around 3 m/s, but there are places with velocities as high as 5 or 6 m/s. The world's first underwater turbine servicing consumers on a regular basis uses a 2.5 m/s flow. It is operated by Hammerfest Stroem in northern Norway, way beyond the Arctic Circle. Compared with wind, these are modest velocities; however, the high density of water (1000 kg/m^3,[†] almost 800 times larger than that of air) causes these small velocities to lead to attractive power densities.

Some of the early proposals for extracting energy from currents seemed impractical. One example is the idea of employing a parachute-like device dragged along by the waters, unwinding a long cable that drives the shaft of a generator. When all the cable has been unwound (perhaps several kilometers), the parachute is collapsed and retrieved by running the generator as a motor, thus rewinding the cable. More practical are schemes that employ what are essentially underwater "wind" turbines.

16.4.1 Marine Current Turbine System

Marine Current Turbine, Ltd. (MCT) has a program for harnessing ocean currents by using windmill-like underwater turbines. MCT demonstrated a prototype (called Seaflow shown in Figure 16.7) rated at 300 kW and capable of generating an average power of 100 kW. The 11-m-diameter rotor can have its pitch adjusted to accommodate tides flowing from opposite directions. The turbine-driven induction generator has its output rectified

[†] Ocean water, being salty, has a somewhat higher density—1027 kg/m^3 being typical for surface water where, presumably, most turbines will operate.

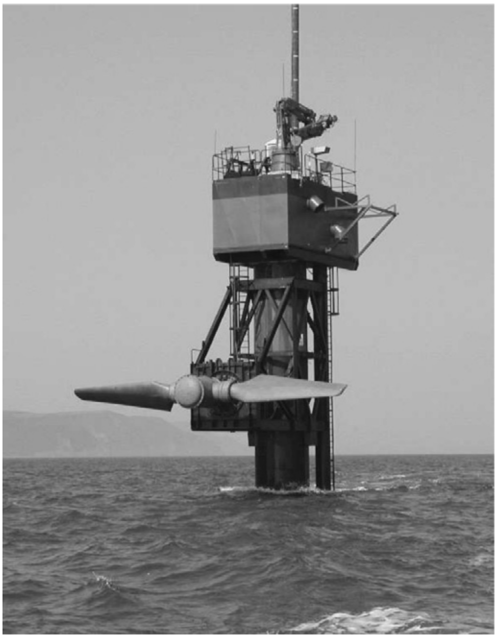


Figure 16.7 Prototype of a tidal turbine installed off the coast of Devon, England. The rotor/turbine assembly is shown in the raised position for maintenance. *Source*: Marine Current Turbine, Ltd.

and inverted to 50-Hz ac, which is dissipated in a resistive test load. The company's next project will be a 1-MW system connected to the grid.

Though similar to wind-driven turbines, there are major differences.

16.4.1.1 Horizontal Forces

The power generated by a wind turbine is

$$P_{wind} = \frac{16}{27}\frac{1}{2}\rho_{air}v_{air}^3 A_{wind}\eta_{wind}, \tag{16.9}$$

and that generated by a water turbine is

$$P_{water} = \frac{16}{27}\frac{1}{2}\rho_{water}v_{water}^3 A_{water}\eta_{water}. \tag{16.10}$$

Let us assume that both types of turbines have the same efficiency, η. If the two turbines generate the same amount of power, then

$$\rho_{air}v_{air}^3 A_{wind} = \rho_{water}v_{water}^3 A_{water} \tag{16.11}$$

$$\frac{A_{wind}}{A_{water}} = \frac{\rho_{water}}{\rho_{air}}\left(\frac{v_{water}}{v_{air}}\right)^3. \tag{16.12}$$

The horizontal forces on the turbines are proportional to $\rho v^2 A$. Hence,

$$\frac{F_{water}}{F_{air}} = \frac{\rho_{water} v_{water}^2 A_{water}}{\rho_{air} v_{air}^2 A_{wind}} = \frac{v_{air}}{v_{water}}. \qquad (16.13)$$

If, for example, the wind velocity is 15 m/s and that of the water is 3 m/s, the horizontal forces on the water turbine will be five times larger than that on a wind turbine with the same power output. The density, ρ, of the fluids plays no role in the above formula because, for a fixed power output and selected flow velocities, the ρA product must be constant. If the density of the fluid is increased, the area needed to produce the same power will be correspondingly decreased.

From the above, one should expect substantial horizontal forces in ocean devices. This requires heavy structure (seaflow masses 130 tons) and strong anchoring systems. Megawatt-sized ocean turbines may be subjected to over 100 tons of horizontal forces.

16.4.1.2 Anchoring Systems

The anchoring system of ocean turbines must withstand the extremely large horizontal forces mentioned earlier. In moderate depth, the turbine might simply be weighed down by an adequate ballast, but it appears that a single pile driven into the ocean floor is being preferred. Great depth would probably require floating platforms tethered to a sunken pile. One difficulty with this arrangement is that as the direction of the current reverses (as happens with tide-driven currents), the floating platform will change position, causing all sorts of difficulties, especially with the transmission line that carries the electric energy to the consumer.

Fortunately, builders of ocean oil rigs have accumulated a vast experience in constructing undersea structures and in anchoring them.

16.4.1.3 Corrosion and Biological Fouling

The ocean is a hostile environment requiring careful choice of materials and of passivation procedures as well precautions against biological fouling.

16.4.1.4 Cavitation

As we saw in Chapter 15, lift-type turbines operate by generating a pressure differential between opposite sides of an airfoil (or of a water foil, as it were). The pressure on the suction side may become low enough to cause water to boil, generating vapor bubbles. When reaching regions of higher pressure, these bubbles will implode, releasing energy and reaching high temperatures, notwithstanding the environment acting as an excellent heat sink.[†] Such

[†] Taleyarkhan et al. (2002) report the extremely high temperature of 10 million kelvins of imploding bubbles in deuterated acetone. The temperature is high enough to provoke the nuclear fusion of the deuterium. These results are being disputed by some scientists.

implosions are surprisingly damaging to the rotating blades of hydraulic machines and propellers. Severe pitting and vibration can result, and in the case of submarines, noise is generated reducing the stealthiness of the boat. This phenomenon is called **cavitation**. Clearly, increased depth (because of increased pressure) will retard cavitation, which, therefore, tends to occur at the upper part of the rotor sweep, where the static pressures are at a minimum. The tip speed of the rotor of an ocean current turbine must be kept low enough to avoid cavitation. The safe top speed varies with, among other factors, the depth as explained above. In shallow sites, cavitation may develop when the linear speed exceeds some 15 m/s. One can get a rough estimate of the rotor tip velocity, v_{tip}, that will cause cavitation in typical ocean conditions as a function of the depth, d,

$$V_{tip} \approx 7 + 0.31d - 0.0022d^2. \tag{16.14}$$

WAVEGEN reports having relatively minor difficulties with cavitation. The main problem it has experienced from cavitation is loss of efficiency.

16.4.1.5 Large Torque

Because of the large density of the medium, ocean current turbines operate at low rpm. For a given power, large torques are developed. Costly gearing is required to match the characteristics of most electric generators, which are usually designed for high-speed/low-torque conditions. In wind turbines, there is a modern trend toward the development of low-speed generators. If they prove practical, these will benefit ocean turbines as well.

When the tip speed of the rotor is limited to a fixed value (owing to cavitation), the torque delivered by the turbine is proportional to the generated power raised to 3/2; that is, the torque grows faster than the power. Thus, the larger the power, the worse the torque problem.

16.4.1.6 Maintenance

Maintenance contributes heavily to the cost of operation. Turbine design should be such as to minimize maintenance frequency. Sea flow has the capability of raising the rotor and gear box to make it easily accessible when repairs become necessary. Unattended operation and remote control and metering will probably be required for economic operation of the system.

16.4.1.7 Power Transmission

Ocean current turbines must be located at some distance from shore. This calls for expensive undersea transmission lines.

16.4.1.8 Turbine Farms

Ocean turbines of a given power are more compact than equivalent wind turbines; hence the ocean turbines can be more densely packed than the wind turbines. In addition, in most wind turbine farms, the location of individual units must take into account that the wind direction may be quite variable, and one turbine should not shade the next. The direction of ocean currents is reasonably constant, so that turbines can be set up in close proximity. The packing density of these devices can lead to up to 100 MW/km^2, one order of magnitude larger than their wind-driven cousins. Close packing has a number of advantages, including reduced cost in the power transmission system.

16.4.1.9 Ecology

Arguments can be made that sunken structures in the ocean can act as artificial reefs benefiting the local flora and fauna.

16.4.1.10 Modularity

The modularity of ocean current turbines can contribute significantly to the reduction of operation costs.

In 2008, Marine Current Turbines (MCTs) started operating SeaGen, the world's first commercial-scale tidal turbine, located in Northern Ireland's Strangford Lough. When fully operational, the installation, will generate 1.2 MW of power. This is comparable to the power delivered by a modern wind turbine.

In a few regions in the world, very strong tidal currents occur, with representative water flow rates of up to 6 m/s or 21 km/hr. The power density is then very large: 65 kW/m^2. It would take a wind of 440 km/hr to reach this same power density using air instead of water. With 20 m/s or 72 km/hr, a very high-wind velocity, the available power density for a wind turbine is only 2.8 kW/m^2. The problem is that places with large tidal currents created extreme mooring problems owing to the lateral forces exerted by the water. Although the average power generated is a function of the average flow speed, the anchoring system must resist the *peak* forces, a function of the *peak* velocity. At Saltstraumen, Norway, the peak velocity of the water flow exceeds 10 m/s.

Large whirlpools, caused by the peculiar topography of the region, are the source of these high tidal currents. They are called **maelstroms**, and when they have a substantial vertical velocity component, they are technically **vortexes**. Many have been made famous by (exaggerated) depictions in the literature. Among the better known whirlpools are

Moskstraumen, in the Lofoten islands off the coast of Norway.

Salstraumen, in Norway.

Corryvreckan, in Scotland. This site is being considered for testing tidal energy converters.

Old Sow between New Brunswick and Maine.

Naruto in Japan.

Garofalo in the strait of Messina between Sicily and Calabria, in Italy. This is almost certainly the Charybdis mentioned by Homer in the *Odyssey*.

16.5 Salination Energy

The highest power density available in the ocean is that associated with the salination energy of fresh river waters mixing with the sea. Indeed, the osmotic pressure of fresh water with respect to sea water is over 2 MPa. Thus, a flow of $1\,\text{m}^3\ \text{s}^{-1}$ will result in the release of energy at a rate of

$$P = p\dot{V} = 2 \times 10^6 \times 1 = 2\ \text{MW}. \tag{16.15}$$

To produce the same power per unit flow, a hydroelectric plant would require a head of 200 m. The estimate worldwide salination energy potential is 160 GW, the equivalent of some 160 commercial nuclear reactors. It is not insignificant.

Most proposals for salination engines are based on the use of semipermeable membranes. These can be employed in engines driven by the osmotic pressure between fresh and salt water (**pressure-retarded osmosis, PRO**), or in electrodialysis engines (**reverse electrodialysis, RED**), based on the electric potential difference established across a membrane that is permeable to cations and not to anions or vice versa.

Membranes are expensive and tend to have short lives. The Norwegian company, Statkraft, has for years been developing a commercial osmotic system, which is now ready for actual trials in Sunndalsø. It is a 1–2 MW pilot plant of the PRO type. The main breakthrough was the development of inexpensive long-lived (7–10 years) membranes probably based on a modified form of polyethylene, capable of 4 to 6 W/m^2. In the next section, we are going to see that osmotic pressures, p_O, can be very large. If brackish water is separated from fresh water by an osmotic membrane, the fresh water will tend to flow into the brackish side. If the pressure of the brackish water is higher than that of the fresh water, the flow will be diminished, that is, **retarded**. It will be (from Equation 16.26),

$$\dot{V} = k(p_O + p_F - p_S) \approx k(p_O - p_S) \quad \text{m}^3\text{/s per m}^2 \text{ of membrane,} \tag{16.16}$$

where k is the **permeability** of the membrane in m^3/s per m^2, p_O is the osmotic pressure, p_F is the pressure on the fresh water side, and p_S is the pressure on the salt water side. The power transferred is

$$P = \dot{V}p = k(p_O - p_S)p_S. \tag{16.17}$$

Maximum power is transferred when the operating pressure, p_S, is equal to half the osmotic pressure, p_O,

$$p_S = \frac{1}{2}p_0. \tag{16.18}$$

One simple realization of a PRO engine is shown in Figure 16.8. The central component is a **mass exchanger** that consists of a series of

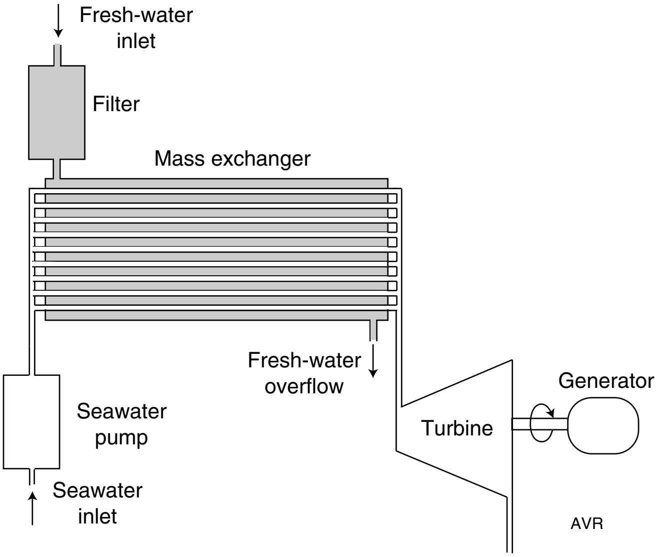


Figure 16.8 A pressure retarded osmosis engine, basically like the machine developed by Statkaft.

permeable osmotic tubes through which sea water is pumped under pressure. These tubes are immersed in fresh river water (after filtering out most of the suspended material). Osmotic pressure causes a flow of fresh water into the sea water, diluting it and substantially increasing its mass. The high-pressure water is depressurized through a turbine, producing useful work. Some work had been done by the seawater pump, which, of course handles a much smaller water volume. The net power generated is, to a first order, the difference between the turbine output and the pump input, and depends on the amount of water transferred from the fresh water to the seawater side of the mass exchanger.

Olsson, Wick, and Isaacs (1979) proposed an ingenious salination engine that does not use membranes. It works best between fresh water and brine (defined as a saturated sodium chloride solution) but, presumably, can be made to operate also with ocean water.

Salt water boils at a higher temperature than fresh water. In other words, salt water has a lower vapor pressure than fresh water. Figure 16.9 shows how the vapor pressure of fresh water depends on temperature. It also shows the difference between the vapor pressure of fresh water and that of brine.

The proposed engine is sketched out schematically in Figure 16.10 Compartment A contains brine, while compartment B contains fresh water.

If the liquids are at the same temperature, the vapor pressure in A is lower than that in B; hence, water vapor flows from B to A, driving the turbine in the interconnecting pipe. Evaporation of the fresh water

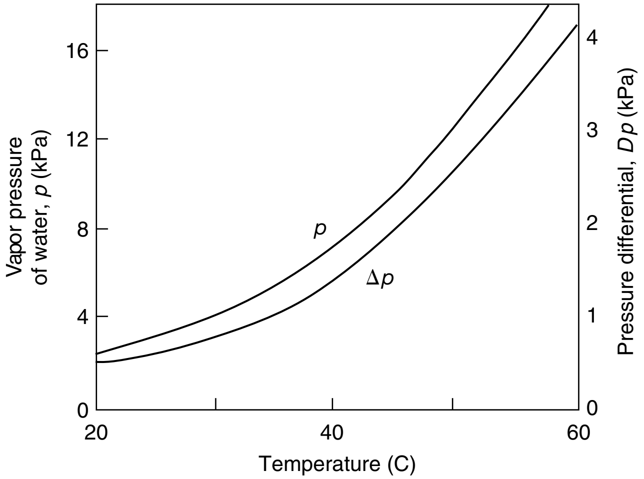


Figure 16.9 Vapor pressure of water and the difference between the pressure over fresh water and that over concentrated brine.

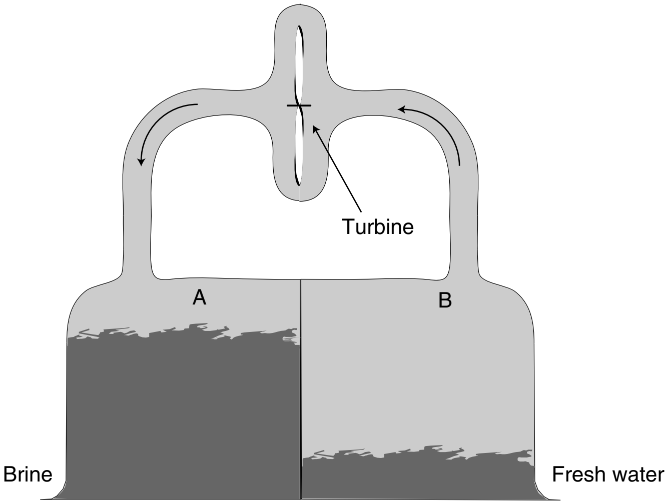


Figure 16.10 A membraneless salination engine.

cools it down, lowering the pressure, while dilution of the brine causes the temperature (and the pressure) of compartment A to rise. Soon the system will reach equilibrium, and the turbine will stop.

To avoid the equlibrium from occuring, one feeds the heat of condensation back from B into A. If all the heat is returned, there will be no temperature change, and, eventually, all the water will be transferred to the brine compartment. The heat transfer between the compartments can be accomplished by placing them side by side separated by a thin heat-conducting wall.

In practice, the brine would be replaced by ocean water, and rivers would supply the fresh water. Let us determine how much energy can be extracted when 1 kilomole of fresh water is transferred to the brine. Let p_{FR} be the pressure of the water vapor over the fresh water container, p_{BR}, the pressure in the brine container. Let V_{FR} be the volume of the water vapor that flows into the turbine from the fresh water side and V_{BR} be the volume that exits the turbine on the brine side, all in a given time period.

The water vapor expands through the turbine doing work:

$$W = \int_{p_{FR}}^{p_{BR}} pdV, \tag{16.19}$$

and since the expansion can be taken as adiabatic,

$$p_{FR}V_{FR}^{\gamma} = p_{BR}V_{BR}^{\gamma}. \tag{16.20}$$

The integral becomes

$$W = \frac{p_{FR}V_{FR}}{\gamma - 1}\left[1 - \left(\frac{p_{FR}}{p_{BR}}\right)^{\frac{1-\gamma}{\gamma}}\right]. \tag{16.21}$$

For 1 kilomole of any perfect gas, $pV = RT$. Assume the system operates at a constant temperature of 25 C (298 K), then the $p_{FR}V_{FR}$ product is $8314 \times 298 = 2.49 \times 10^6$ J.

For water, $\gamma = 1.29$.

The work done by 1 kilomole is then

$$W = 8.54 \times 10^6 \left[1 - \left(\frac{p_{FR}}{p_{BR}}\right)^{-0.225}\right]. \tag{16.22}$$

At 25 C, the fresh water vapor pressure is $p_{FR} = 3.1$ kPa, and the pressure difference between the fresh water and brine compartments is 0.59 kPa (see Figure 16.9). This means that $p_{BR} = 3.1 - 0.59 = 2.51$ kPa. Introducing these values into Equation 16.22, we find that each kilomole of water vapor that flows through the turbine generates 396 kJ of mechanical energy.

Let us compare the above energy with that generated by an OTEC operating between 25 C and 5 C, a ΔT of 20 K. Assume that half of this ΔT is applied across the turbine. In the Lockheed OTEC, the power extracted from the turbine is 90% of the Carnot power. So, let as take the turbine efficiency as equal to the Carnot efficiency, which in this case would be 0.033. If one-fourth of the ΔT is the warm water temperature drop, then the input thermal energy (per kmole of warm water) is 10 MJt and the mechanical output from the turbine is 330 J. This is in the same order as the 396 J/kmole calculated for the salination engine. The salination engine could theoretically achieve 100% efficiency. In a laboratory model, Olsson and coworkers (1979) demonstrated 40% efficiency.

One problem with this type of engine is the outgassing of the water necessary to ensure that the pressures in the compartment are due only to the water vapor, not to the presence on incondensable gases. If instead of concentrated brine more realistically ocean water is used, then the power output of the machine falls substantially.

16.6 Osmosis

Salination engines are being seriously proposed as a possible energy converter, while the osmotic engine described in this section will probably never be more than an academic curiosity. Its study is nevertheless an interesting intellectual exercise.

Osmosis is a (quantitatively) surprising phenomenon. If two solutions of different concentrations are separated by an **osmotic membrane** permeable to the solvent but not to the solute, then there is a net flow of the solvent from the more dilute to the more concentrated side. Such flow will persist until the pressure on the concentrated side is sufficiently large.

Osmotic pressures can be measured by means of a U tube with an osmotic membrane at the bottom as suggested in Figure 16.11. When equilibrium is reached—that is, when the pressures on the two sides of the membrane are the same—the column on the concentrated side is higher than that on the dilute side. The hydrostatic pressure of the brine must equal

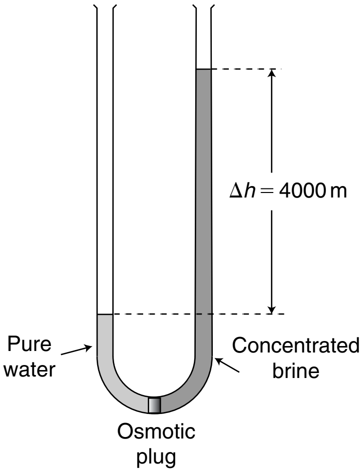


Figure 16.11 Apparatus for measuring osmotic pressure.

the sum of the hydrostatic pressure of the fresh water, plus the osmotic pressure. What is surprising is the magnitude of the osmotic pressure.

For concentrated NaCl solution at room temperature, the height difference between the two columns would be 4000 meters! The osmotic pressure is 400 atmospheres or 40 MPa.

Osmotic pressure depends on concentration and on temperature. Figure 16.12 shows the osmotic pressure of salt water versus fresh water as a function of salinity at two different temperatures.

To understand the operation of our osmotic engine, consider a cylindrical column of water with base area, A, and a depth, d. If the density of the water column is δ, then the mass is

$$M = A\delta d \tag{16.23}$$

and the weight is

$$W = gA\delta d, \tag{16.24}$$

where g is the acceleration of gravity. The pressure on the base is

$$p = g\delta d. \tag{16.25}$$

Consider now a pipe immersed vertically in the ocean with its top just above the surface and having, at its bottom, an osmotic plug, as depicted in Figure 16.13. Let the pipe be filled with fresh water up to a height, d_F. p_S is the pressure exerted by the salt water (on the outside) on the osmotic plug and p_F the opposing pressure of the fresh water. Besides the hydrostatic pressures, there is also an osmotic pressure, p_O, tending to force the fresh water into the ocean.

Thus, p_F and p_O act in the same direction and oppose p_S. In equilibrium,

$$p_S = p_F + p_O. \tag{16.26}$$

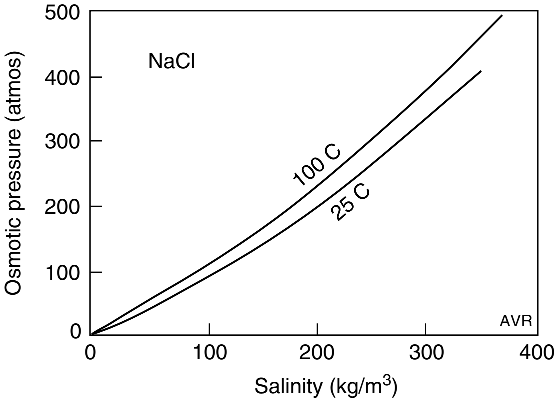


Figure 16.12 The osmotic pressure of saltwater at two temperatures.

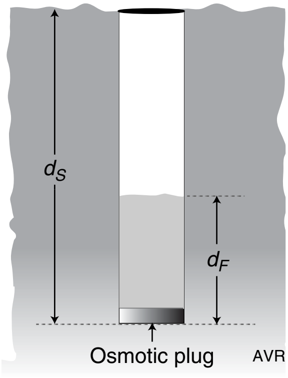


Figure 16.13 A pipe with an osmotic plug in the ocean.

The density of fresh water is 1000 kg/m^3, while that of ocean water is 1025 kg/m^3. Thus

$$1025gd_s = 1000gd_F - p_O, \tag{16.27}$$

from which

$$d_F = 1.025d_s - \frac{p_O}{1000\,g}. \tag{16.28}$$

The osmotic pressure of fresh water with respect to ocean water is 2.4 MPa. Taking $g = 10\,\text{m s}^{-2}$,

$$d_F = 1.025d_S - 240. \tag{16.29}$$

Down to a depth, d_S, of about 240 m, $d_F < 0$, that is, the osmotic pressure is sufficient to keep the saltwater from entering the pipe. If the pipe goes deeper, reverse osmosis takes place and fresh water from the salty ocean is forced into the pipe. When, for instance, the pipe goes down to 1000 m, the fresh water column will rise 785 m coming within 215 m from the surface.

At what depth will the water rise just to the ocean level? When that happens,

$$d_F = d_S \equiv d = 1.025d - 240, \tag{16.30}$$

or

$$d = 9600\,\text{m}. \tag{16.31}$$

If the pipe goes even deeper, fresh water will fountain out of its top.

A possible implementation of an osmotic engine is shown in Figure 16.14.

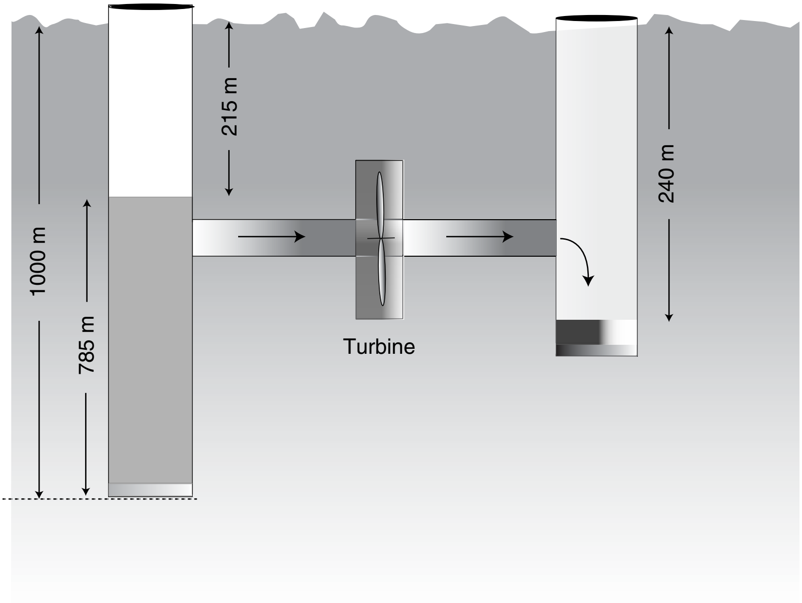


Figure 16.14 A possible configuration for an osmotic engine.

References

Fraenkel, Peter L., Power from marine currents, *Proc. Inst. of Mechanical Engineers* **216**, *Part A: Journal of Power and Energy*, **2002**.

Isaacs, John D., and W. R. Schmitt, Ocean energy: Forms and prospects, *Science 207*, p. 265, **1980**.

Olsson, M., G. L. Wick, and J. D. Isaacs, Salinity gradient power: utilizing vapor pressure difference, *Science 206*, p. 452, **1979**.

Taleyarkhan, R. P., C. D. West, J. S. Cho, R. T. Lahey, Jr., R. Nigmatulin, and R. C. Block, Evidence for nuclear emissions during acoustic cavitation, *Science* **295**, p. 1868, **2002**.

Further Reading

Duckers, L. J. (Coventry University, Coventry, UK), Wave energy: Crests and troughs, *Renewable Energy*, **5** (5–8), pp. 1444–1452, August **1994**.

E21 EPRI WP–004–US, Offshore Wave Energy Conversion Devices, June 16, **2004**

Krock, Hans-Jurgen, ed., Ocean energy recovery, Proceedings of the First International Conference, ICOER **1989**, Honolulu, Hawaii, November 28–30, 1989, held under the auspices of the Ocean Energy Committee of the Waterway, Port, Coastal, and Ocean Division of the American Society of Civil Engineers, cosponsored by the Pacific International Center for High Technology Research, School of Ocean and Earth Science and Technology of the University of Hawaii, New York, **1990**.

McCormick, Michael E., and Young C. Kim, eds., Utilization of Ocean Waves—wave to energy conversion, *Proceedings of the International Symposium*, Scripps Institution of Oceanography, La Jolla, California, June 16–17, 1986, sponsored by the Waterway, Port, Coastal, and Ocean Division of the American Society of Civil Engineers and the National Science Foundation, Symposium on Utilization of Ocean Waves (**1986**: Scripps Institution of Oceanography), New York, **1987**.

McCormick, Michael E., Compendium of international ocean energy activities, prepared under the auspices of the Committee on Ocean Energy of the Waterway, Port, Coastal, and Ocean Division of the American Society of Civil Engineers. New York, **1989**.

Mccormick, Michael E. *Ocean Wave Energy Conversion*, John Wiley, **1981**.

Practical Ocean Energy Management Systems, Inc.
<http://www.poemsinc.org/home.html>

Seymour, Richard J. ed., Ocean Energy Recovery: *The State of the Art. American Society of Civil Engineers*, New York, **1992**.

The exploitation of tidal marine currents (Non-nuclear energy Joule II project results), Report EUR 16683EN, DG Science, Research and Development, The European Commission Office for Official Publications, Luxemburg L-2920, **1996**.

PROBLEMS

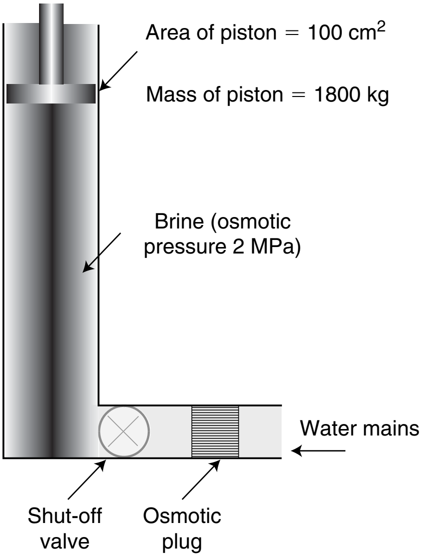
16.1 Reverse osmosis is a technique sometimes employed to extract fresh water from the sea.

Knowing that the osmotic pressure of fresh water with respect to sea water is 2.4 MPa, what is the minimum energy necessary to produce 1 cubic meter of fresh water?

The price of the equipment is \$200/kW and that of electricity is \$50/MWh. What is the cost of a cubic meter of fresh water if there is no operating cost and if the cost of the capital is 20% per year?

How realistic are your results? Point out reasons for having underestimated the cost.

16.2 Refer to the figure. Initially, the shutoff valve is closed, and the piston is at a height of 50 m. Pressure of the water in the mains is 3×10^5 Pa. How high will the piston rise when the valve is opened? Assume that the osmotic pressure of the brinewater system is 2 MPa independently of dilution.

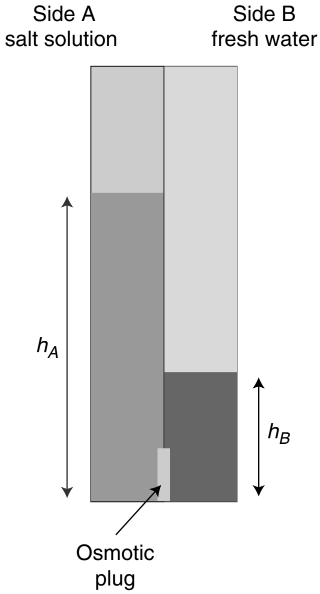


16.3 At 25 C, the osmotic pressure of a NaCl (versus fresh water) is given approximately by the empirical expression,

$$p = 0.485 + 0.673\sigma + 1.407 \times 10^{-3}\sigma^2,$$

where p is the pressure in atmospheres and σ is the salinity in kg/m^3.

Two tubes with a square internal cross section, A, of 8.86×8.86 mm are interconnected via an osmotic wall. In one side (Side A)



Problem 16.3

of the assembly, a NaCl solution containing 2 g of salt and measuring 10 ml is poured in. Initially, no liquid is poured into Side B. For simplicity assume that the volume of the solution is equal to the volume of the solvent (water).

How high is the salt water column in Side A? How high is the fresh water column in Side B?

How much distilled water has to be poured into Side B so that, after equilibrium has been reached, the fresh water column is 5 cm high? Again, for simplicity, assume the brine of any concentration has the same density as distilled water.

16.4 Imagine a very tall underwater tower on top of which there is a platform (which, of course, does not change its height above the seafloor). This arrangement, although impractically expensive, permits easy observation of the ocean waves. On a given day, the average height between the trough and the crest of a wave is measured as exactly 2.6 m, and the waves follow one another at a 8.2 secs interval. How far apart in space are the waves? Assume that the depth of the water is much larger than the wavelength of the waves.

16.5 Both the sun and the moon cause ocean tides on Earth.

1. Which tide is bigger? the solar tide or the lunar tide?
2. Calculate the gravitational accelerations at Earth caused by the sun and the moon.
3. If you did the calculations correctly, then you will have found that the gravitational field of the sun on Earth is vastly larger

than that of the moon. We must conclude that the tides are not proportional to the gravitational field.

What is the nature of the influence that causes the tides? Calculate the ratio of the lunar to the solar influences.

16.6 Assume that the orbits of Earth around the sun and that of the moon around the Earth are both circular. The Earth's orbit around the sun is completed in 3.1558157×10^7 seconds, and the moon's orbit around the Earth, in 2.36055×10^6 seconds or 27.32 days.

Remember that a full moon occurs when the sun, the Earth, and the moon are in the same plane—that is, their projection on the ecliptic lies in a straight line.

Calculate the number of days between consecutive full moons. Express this with four significant figures.

16.7 Owing to the possibility of cavitation, the tip speed, v_{tip}, of the rotor of an ocean current turbine must not exceed a certain speed, $v_{tip_{max}}$, which depends on, among other factors, the depth. Show that for a constant tip speed, the torque, Υ, is proportional to $P_g^{3/2}$, where P_g is the generated power.