SCIENTIFIC AMERICAN

BRUEGEL'S TECHNOLOGY

$1.50

March 1978

"It's pretty spooky rowing around inside a supertanker in a rubber boat."

"That's me in the back of the rubber boat," says Chief Mate George McShea. "We're not rowing on oil, but on a saltwater lake in the belly of a supertanker. And it's one aspect of life on a supertanker I can't get used to.

"Sure, a supertanker is three football fields long, and twelve stories high. But after a while those are just numbers.

A lake in a cave

"This tank inspection is something else. After the tanks are cleaned and the residual oil retrieved and stored, they're flooded with seawater, and inspection crews go inside, checking structural integrity.

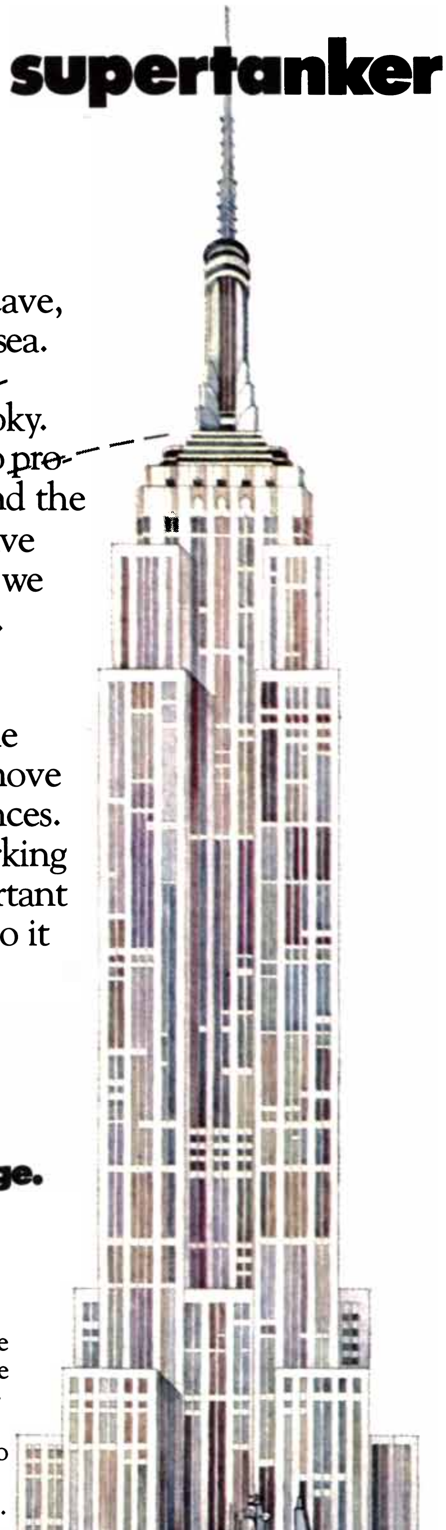
"It's like a lake in a cave, but it's inside a ship at sea. It's a big job, and sometimes it gets pretty spooky. But it has to be done, to protect the ship, the oil and the ocean. Gulf tankers have a fine record; we do all we can to keep it that way.

Most efficient

"Supertankers are the most efficient way to move so much oil such distances. It's a real challenge working on them. But it's important work, and I think we do it well."

Gulf

Gulf people: meeting the challenge.



Supertankers—some almost as long as the Empire State Building is tall—are the most efficient way to ship so much oil such great distances.

Gulf Oil Corporation

Established 1845 SCIENTIFIC AMERICAN March 1978 Volume 238 Number 3

ARTICLES

DEPARTMENTS

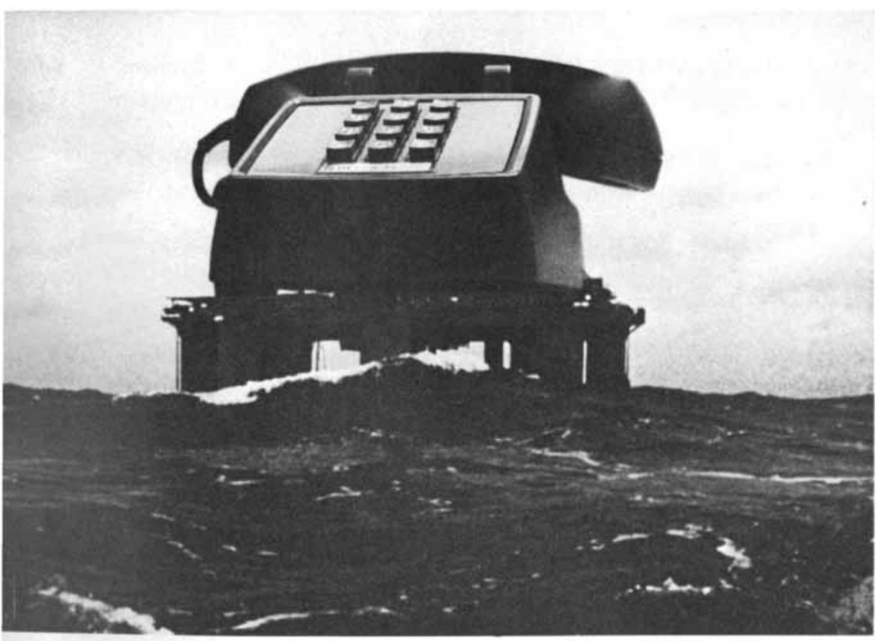
PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $18 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $22 PER YEAR, ALL OTHER COUNTRIES.

A report on a matter of public interest:

How the Bell System is pumping more service out of less oil.

In 1973, when the OPEC oil embargo went into effect, the Bell System committed itself to reducing its energy consumption. That commitment has been fulfilled. In four years, the Bell System has saved the energy equivalent of almost 24 million barrels of oil and over 415 million dollars in energy costs—savings that help hold down the cost of your telephone bills.

Today, the Bell System is actually using 11 per cent less energy than it did in 1973, even though the number of telephones in service has risen over 16 per cent and the volume of business has increased 33 per cent.



Here's how we are combining common sense with uncommon technology in four basic areas to achieve Zero Energy Growth.

Telephones are reconditioned three times.

The Bell System's energy needs begin with the power and fuel necessary to design and manufacture the basic communications components—cables, wire and equipment such as your home telephone.

In general, it takes much more energy to manufacture such items from scratch than it does to recycle them. Because Bell System equipment is designed by Bell Labs to be reliable, repairable *and* recyclable, extensive energy cutbacks have been realized through 40 years of recycling and reuse. New, more energy-efficient processes are constantly being devised by Western Electric, some of which entail modifying original designs for even greater materials and power savings.

Since 1974, the Bell conservation program has saved the energy equivalent of over three million barrels of oil by recycling metals. Also, more than six million equivalent barrels of oil have been saved through the reuse of equipment. The average telephone, for example, is reconditioned three times before it is unrepairable or obsolete.

New technology does more with less.

Another area in which the Bell System is effecting energy savings is in power for switching and transmission equipment. Constantly, new energy-saving technology is being added to the system. *Item:* Over two billion power-saving transistors, diodes and integrated circuits have been put into use. *Item:* Light Emitting Diodes (LEDs) are replacing incandescent bulbs in switchboards and telephones, saving over 90 per cent of the previously required power. *Item:* A new microprocessor called MAC-8 is less than one-tenth the size of a postage stamp yet contains the equivalent of over 7,000 transistors. The MAC-8 can execute several hundred electronic "thinking" functions, yet it will operate on only one-tenth of a watt of power.

Smaller vehicles power giant fleet.

Twenty-two per cent of Bell's energy requirements are in fuel for its fleet of over 170,000 vehicles, the largest privately owned and operated motor fleet in the world. Here, a number of commonsense procedures have been adopted: engines are carefully tuned for peak efficiency, smaller and more fuel-efficient vehicles are being used, and shuttle services have been set up between some company locations. In addition, New York Telephone Company is experimenting with nonpolluting, energy-saving electric-powered trucks. Due to these and other efforts, the Bell System in 1976 used over five per cent less motor fuel than in 1973.

Even employees' body heat is used.

Heating, lighting and air conditioning of Bell System's 25,000 buildings account for 45 per cent of its energy needs. Broad economies have been achieved simply by removing thousands of unnecessary lights; by lowering temperature settings; by cutting back on hot water temperatures; and by heating or cooling unoccupied areas only to the extent required for equipment operations.

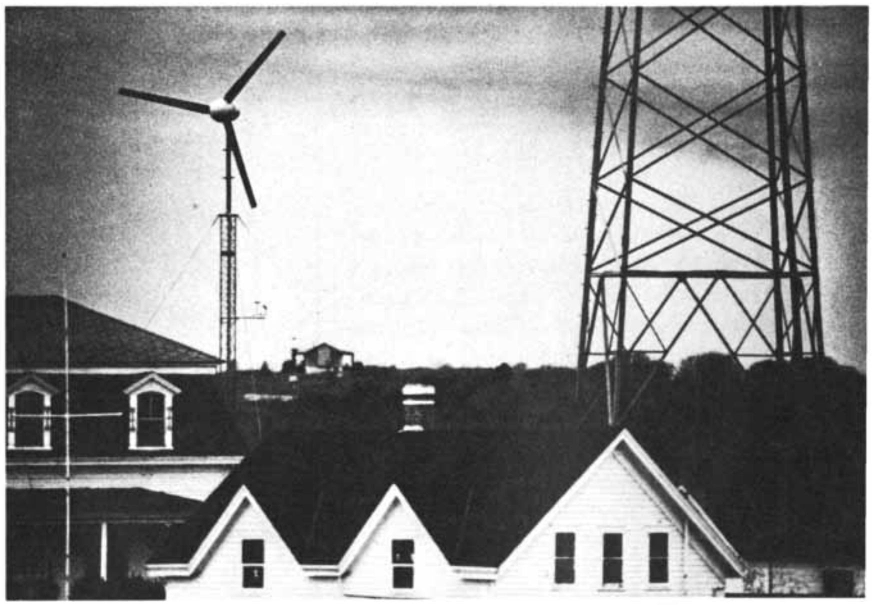
Moving beyond the obvious conservation measures, the Bell System created a building energy management program to redesign and retrofit existing buildings to improve their energy efficiency. Two examples of other power-saving programs at Bell facilities:

• On windy Block Island, Rhode Island, the New England Telephone Company began operating a wind dynamo in September, 1976. It can produce up to 15 kilowatts of electricity to power a central office and microwave radio terminal. Excess power from the windmill is fed back to the power company.

• In AT&T's new Basking Ridge, New Jersey, facility, an innovative computerized system heats about 1½ million feet of office space by recovering excess heat from the building environment—lights, equipment and the body heat of employees. It is estimated that the system uses 25 per cent less energy than conventional heating/cooling systems.

Bell trials of solar heating and cooling are providing valuable data which should lead to more widespread use of alternate energy systems.

Windmill helps power central phone office and microwave radio terminal (tower at right) on Block Island.



Today, throughout the Bell System, our commitment to energy conservation is more than a goal; it is an ongoing reality. And in looking to the future, we anticipate that in 1982 we will still be using no more energy than was used in 1973. *Keeping your phone system the best in the world.*

Bell System

THE COVER

The photograph on the cover shows a detail of a painting, the *Tower of Babel,* by the 16th-century Flemish artist Pieter Bruegel the Elder (see "Pieter Bruegel the Elder as a Guide to 16th-Century Technology," by H. Arthur Klein, page 134). The detail is a crane that was capable of lifting quite heavy objects even though its motive power was several men operating a treadmill. Independent evidence indicates that such cranes were employed in the 16th century for construction and also for loading and unloading ships. As one can see in a fuller reproduction of Bruegel's painting that appears on page 135, this crane is mounted on a ramp nearly halfway up the tower and is employed to lift materials, such as the slab of stone in the detail, that have been delivered to the site by ship. Bruegel's painting is in the Kunsthistorisches Museum in Vienna.

THE ILLUSTRATIONS

Cover painting courtesy of the Kunsthistorisches Museum, Vienna

Page	Source
25–28	Ilil Arbel
32	*Harvest of the Palm.* © 1977, Harvard University Press
33	*Submersibles and Their Use in Oceanography and Ocean Engineering.* © 1977, Elsevier Scientific Publishing Company
36	*Arachnida.* © 1977, Academic Press
43–49	Alan D. Iselin
51	Stanford Linear Accelerator Center
52–57	Gabor Kiss
59–60	Roman Kister, Bell-Northern Research
61–64	Adolph E. Brotman
77–86	National Aeronautics and Space Administration
87	Ilil Arbel
88	National Aeronautics and Space Administration
92	Robert L. Talerico, U.S. Forest Service
94–97	Tom Prentiss
98	Robert L. Talerico, U.S. Forest Service
99–100	Tom Prentiss
105	John N. Telford, Cornell University
106–116	Bunji Tagawa
125–131	Ilil Arbel
135	Kunsthistorisches Museum, Vienna
136	Capodimonte Museum, Naples
137	Groeningemuseum, Bruges (*top*); Devonshire Collection, Chatsworth, reproduced by permission of the Trustees of the Chatsworth Settlement (*bottom*)
138–140	Mr. and Mrs. Jacob I. Zeitlin
142–153	Michael Goodman

On Photographing the Invisible

To the naked eye, it was a Swedish 80-ore postage stamp. A rarity, and very valuable.

The camera, however, told quite another story. The stamp was a counterfeit.

Faint traces of tampering that were hidden to the naked eye were revealed by the camera. Someone, somewhere, had ingeniously altered the stamp by chemically removing a surprint. The stamp was worthless.

To the naked eye (left), the stamp was genuine. To the camera (right), it was a counterfeit. Note the faint, dark traces of tampering now revealed in the upper section.

What manner of exotic camera was this that could "see" the invisible?

The lens: one of the 20 in the Hasselblad arsenal, the 105mm Zeiss UV-Sonnar f4.3. Designed for photography within the ultraviolet portion of the electromagnetic spectrum, its costly quartz elements can detect radiations that are unseeable by the human eye.

It has peered at objects in outer space, examined forgeries, laid bare the secrets of counterfeit money. Not a lens for everyone, obviously, but an indication of just how awesomely comprehensive the Hasselblad System is.

The camera: an otherwise perfectly standard Hasselblad 500C/M, normally fitted with an 80mm Zeiss Planar f2.8 multi-coated lens.

This is the basic model that allows you to tap into the vast Hasselblad System. It is one of the most bewilderingly versatile cameras the world has ever known. Yet so marvelously simple to operate that it often plays the part of the family snapshot camera.

A True System.

The Hasselblad System is a prodigious array of 4 cameras, 20 lenses, 8 viewfinders, 9 film magazines, and over 300 other accessories. Choose the right pieces, and your 500C/M would be equipped for sports, aerial, architectural, and fashion photography.

And portrait, landscape, medical, underwater, and news photography.

And wildlife, laboratory, industrial, and child photography.

And you would always have the right film in the camera at the right time. You can shift from color to black-and-white and back again to color—and resume shooting at precisely the right frame—by popping in the protective dark slide and switching film backs.

The Camera with Nine Backs.

There is a small button on the film back of every Hasselblad 500C/M. Slide it sideways with your thumb and the back will come away in your hand.

The standard back holds 12 exposures. Each frame of film is 2¼ inches square, *almost four times the area of a 35mm frame.* (See box, below right, for actual size.)

This is only the beginning. There are eight other backs available: Backs that let you change to a 6 x 4.5cm format...or a 4.5 x 4.5cm superslide format for showing in any 35mm projector. Backs that give you a choice of 1, 12, 16, 24, 70, or 500 exposures. A back that is a sheet-film adapter. Even two backs for Polaroid film, so you can check composition, lighting, and exposure ahead of time.

You begin to realize why eight out of ten top commercial photographers surveyed name Hasselblad as the medium-format camera used in their work.

Retained Value vs. Obsolescence.

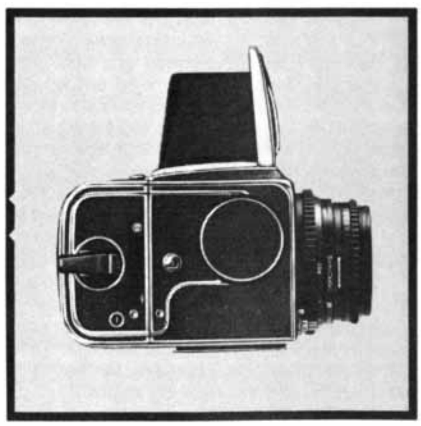
In an age when machines spew out cameras in the tens and hundreds of thousands, when flashy new models thrust last year's marvels into early obsolescence, Hasselblad goes its own way.

Planned obsolescence is taboo at Hasselblad. All but two of the accessories for the 500C/M will fit every Hasselblad made since 1957 (except the Super Wide C)...and will fit every future Hasselblad.

The greater part of a year is spent on building each camera, much of it crafted by hand. And fully one quarter of the work force devotes its time to nothing but quality control.

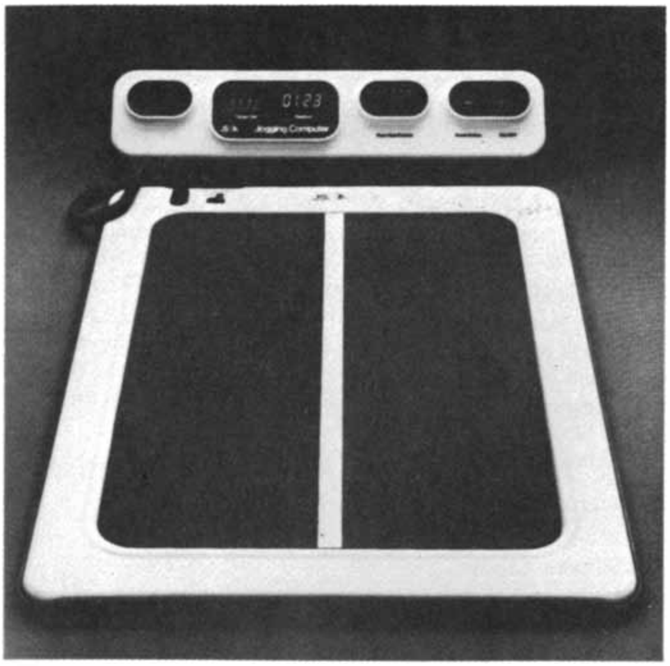
Little wonder, then, that a pre-owned Hasselblad commands such a high price...if its owner can be persuaded to part with it at all.

HASSELBLAD



The Hasselblad 500C/M.

A lavish brochure is available free if you write: Braun North America, Dept. SA378, 55 Cambridge Parkway, Cambridge, Mass. 02142. Braun North America is a division of The Gillette Company and exclusive marketer of Hasselblad cameras in the U.S.



Jogging Computer

Make jogging fun in the privacy of your home with a new space-age indoor exercise system.

The JS&A Jogging Computer is a total system of physical fitness and conditioning.

It's a fact. You reach your physical peak at age 25 and your mental peak at age 40. From then on it's downhill. But it needn't be. A 50 year old who exercises regularly can be healthier and in better physical shape than the average 25 year old.

When you're physically healthy, you are alert and better able to handle stress. You are better motivated and just plain happier. Jogging can keep you in good physical shape.

THE ADVANTAGES OF JOGGING

Jogging as a regular exercise has gained in popularity because it does three things for you. 1) It improves the functioning of the heart, lungs, blood vessels and lymph glands. 2) It helps control your weight without resorting to starvation diets, and 3) It is one of the few safe, strenuous exercises that creates the exertion necessary for good physical conditioning.

A NEW JOGGING COMPUTER

There is now a new, fun way to jog. The new JS&A Computer is a solid-state system that lets you jog in place in the comfort of your own home. It's fun, easy to use and convenient.

You simply set the distance and pace you wish to run and press the start button. An audible beep tone sounds and you jog in place to its rhythm. Each stride is registered on a large LED readout in the control unit so you can see how far you've run.

You jog on a large pad with sensors which register each stride. The pad is designed to feel like grass or soft earth so you can run either barefooted or with gym shoes. The idea is to gradually increase your distance and speed each day to build up your endurance.

Getting yourself to start jogging is often the hardest step. That is why the JS&A Jogging Computer is an ideal system for both the beginner and the experienced jogger.

FOR THE BEGINNER

The first time you step on the Jogging Computer, you run at a pre-selected pace and distance for approximately five minutes. (A chart will show you which speed to select based on your sex and age.) You then take your pulse rate for one minute by touching your wrist. The pulse/rate chart determines the settings and distance you should run the next time you jog.

You could be in poor, average or good shape, and this simple five minute test will accurately tell you. Start the jogger at the distance indicated on the chart, and gradually build up a little each day. In just one week you'll notice the difference, feel great, have greater endurance, and you won't tire as easily. That is what's so nice about the system–how easily and quickly it puts you into better shape.

FOR THE EXPERIENCED

If you jog regularly, you know the many benefits of jogging. But you also know the disadvantages–all overcome by owning a Jogging Computer. For example:

Forget about the ritual You wake up early, drive to your favorite indoor track, change clothes, and you're ready to run. With the Jogging Computer, just step out of bed and start running. The time you save in preparing to jog can be substantial.

Forget about the boredom Running around a track can be quite boring. And if you count laps, how many times have you lost your count? With the Jogging Computer, you can forget about counting, as the unit does it automatically for you. You can concentrate on problems or take flights of fancy–all while you strenuously exercise.

Forget about the weather Even in summer, there are days when you can't jog outdoors. And in a daily exercise program, you must resort to the indoor track. Not so with the Jogging Computer. It's always there when you need it–portable and ready to operate.

Forget about jogger's heel If you've run on indoor tracks, you know the pain of jogger's heel caused by leaning in around those curves. Jogging in place is easier on your whole body and eliminates this common jogging problem.

BRING IT ANYWHERE

The Jogging Computer is powered by four "C" cell batteries and requires no AC power so it goes anywhere–on your patio, in the garage or basement, or at your office. The control unit can be propped up with its built-in easel or placed on a wall using the four foot expansion cord. It's portable, so after you've run a few miles, just turn it off and put it away. There's no large exercise device to take up space.

QUALITY THROUGHOUT

The JS&A Jogging Computer is all solid state, and the 17"x 22" pad was pre-tested to take years of constant, hard pounding under all conditions. Service should never be required, but if anything ever does go wrong, JS&A's service-by-mail center will have it repaired and back to you in a matter of days. Be assured that we stand solidly behind our product's quality, construction and design. JS&A is America's largest single source of space-age products. We've been in business over a decade–further assurance that your modest investment is well protected.

We suggest that you order the JS&A Jogging Computer and use it for 30 days. Jog each day when you get up in the morning or before dinner. Enjoy the thrill of feeling your endurance build. Experience the convenience and fun. See how much better you feel and how much sharper you think. Then after 30 days, measure your progress. If you don't find the JS&A indoor jogger a convenient and fun way to stay trim and healthy, then return your unit for a complete and full refund including the $3.00 charge for postage and handling. You can't lose.

Simply send your check for **$149.95** plus $3.00 postage and handling (Illinois residents add 5% sales tax) to the address below or call our toll-free number. By return mail, we will send you the complete jogging computer system with instructions, charts, personal score card and a one year limited warranty.

Start today on an organized physical fitness program using the latest in solid-state, space-age technology. Order your JS&A Jogging Computer at no obligation today.

JS&A® NATIONAL SALES GROUP

Dept. SA One JS&A Plaza
Northbrook, Ill. 60062 (312) 564-9000
Call TOLL-FREE 800 323-6400
In Illinois Call ...... (312) 498-6900

HATS OFF TO IBM

Printer Break-through

A new 10-digit display calculator with the world's first dual-element integrated printing head will revolutionize the printing calculator.

The full-featured $89.95 Canon P10-D with its one-year parts and labor limited warranty is the greatest printer value ever offered by JS&A.

Hats off to IBM. Their single-element typing system did away with typewriter keys and started a new technology.

The new Canon P10-D printing calculator starts another new technology. Their dual-element printing system does away with the standard printing head which required a separate disc for each column. The Canon has only two discs–one with digits and the other with symbols.

The P10-D head weighs only ½ ounce compared to 31 ounces in a typical printing head. Its motor weighs only nine ounces–again much less than the heavier conventional motors required to drive larger heads. The Canon motor is smaller, lighter and more efficient because it moves less weight.

THE MOST EFFICIENT SYSTEM

The printing head is controlled by an LSI (large scale integrated circuit). As you press a key, a pulse is generated from this circuit and sent to the motor which does two things: 1) positions the two discs to print the numbers or symbols and 2) glides the numeric disc across the ten column width of the paper.

Conventional printers print from metal discs through thick fabric ribbon onto paper. The Canon system prints directly on paper so each impression is sharp, clear and easy to read. The synthetic polymer disc is first inked by a special cartridge before it prints. Each ink cartridge is easily replaceable. The cartridge lasts for more than 15 rolls of paper at a cost of 17¢ per roll–far less than any other system.

PLAIN PAPER PLUS

Using standard paper tape is only one of several advantages that make the Canon a truly spectacular value. Here are some other exciting new features:

Dual Power Operate the Canon from either your AC outlet or its built-in rechargeable batteries. It's totally portable, yet it also makes a handsome desk calculator.

Dual Display Just flip a switch and the 10-digit large green fluorescent display can be used with or without the printer.

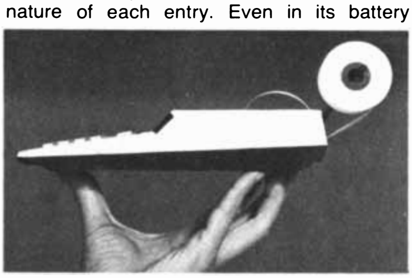
Space-Age Styling Compare the sleek appearance of the Canon with any other printer. It's small enough to fit in your briefcase and large enough to use as a space-saving desk unit. It measures only 1¾" x 4¼" x 8½", weighs only 24 ounces and the paper tucks into the body of the unit–perfect for travel.

Buffered Keyboard If you enter your problems faster than the printer can print them out, don't worry. The unit's memory stores your keystrokes and prints them out in rapid succession.

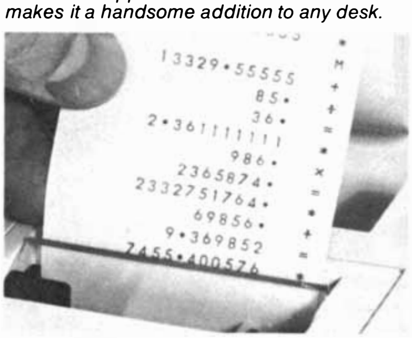
We have always looked at small printers as gimmicks–calculators that lack many important features. We were surprised with the Canon. It has features that far exceed most printers costing hundreds of dollars more.

The following is a list of those features: 10 digit capacity • full four-key memory • addition, subtraction, multiplication and division • percentage key • add-on and discount calculations • power and reciprocal calculations • repeat calculations • add-mode • switch for full-floating or second and third fixed decimal positions • round off or round down switch • paper tape advance.

There are other convenient features that make it perfect for people who spend hours at their calculators. There's a three-digit item counter that counts and prints out the number of entries while printing your total. The symbols on the right side of the tape tell you the nature of each entry. Even in its battery operated position, you could print out more than half a roll of tape before the unit signals you that its batteries are low.



The sleek appearance of the Canon P10-D makes it a handsome addition to any desk.



The direct-impression dual discs print cleaner and sharper on conventional paper tape.

A NEW WAY TO BUY

JS&A offers you a new way to buy your 10-digit Canon P10-D. First we give you the opportunity to use one for 30 days. Carry it in your briefcase. Put it on your desk and see how handy it becomes and how little space it takes up. Check the paper tape and see how clear and easy-to-read it is. Bring it home and let the whole family use it.

Then, within 30 days, decide. If the Canon is not perfect for you, return it for a prompt and courteous refund. And if you do return it, not only will we still consider you one of our good customers, but we will also refund your $2.50 postage, let you keep the paper tape, and thank you for giving us the opportunity of showing it to you. We couldn't be more positive about the quality and value of this incredible new product.

JS&A is America's largest single source of space-age products. We have been in business for over a decade–further assurance that your modest investment is well protected. Canon is one of the world's largest manufacturers of cameras and precision quality instruments and is highly respected as a quality manufacturer of electronic products.

The Canon costs only **$89.95** plus $2.50 for postage and handling and includes a free roll of tape, one ink cartridge, rechargeable batteries and a power cord/charger. It's an incredible value thanks to its new technology. To order, send your check to the address below (Illinois residents add 5% sales tax) or credit card buyers may call our toll-free number.

Space-age technology has produced another major product breakthrough. Order your Canon P10-D at no obligation today.

JS&A NATIONAL SALES GROUP

Dept.SA One JS&A Plaza
Northbrook, Ill. 60062 (312) 564-9000
Call TOLL-FREE 800 323-6400
In Illinois Call ...... (312) 498-6900

LETTERS

Sirs:

John U. Nef's article ["An Early Energy Crisis and Its Consequences," SCIENTIFIC AMERICAN, November, 1977] reveals interesting relations among the supply of fuel wood, coal mining and the origins of the Industrial Revolution in Britain. His analysis of the crisis in the supply of wood uses terms that perpetuate a misunderstanding of the extent of woodland in England in Tudor and Stuart times and of the nature of woodland management and its produce. Professor Nef states that this first energy crisis was a "crisis of deforestation," and that "although there was some reforestation during the 17th century, more and more forest was being cleared for farms and pastures."

Pollen analysis of British postglacial peats and sediments has clearly shown that the major period of forest clearance began about 3500 B.C. during the Neolithic. There has always been some doubt, however, about when this period ended and about the extent of the woodland that was left. The popular picture of medieval England is one of a vast expanse of mixed oak forest with occasional clearings for settlement and agriculture, and wisps of woodsmoke rising above the trees. Recent analyses of historical records of woodland management, in some cases dating back to Norman times (for example those by O. Rackham of the University of Cambridge and by me in southwestern England), indicate that woodland boundaries have changed little since the early Medieval period and were possibly established by Saxon times. It is significant that the Domesday Book makes scant reference to woodlands, except when occasionally indicating rights of pannage.

How can one reconcile this conclusion with Professor Nef's undoubtedly correct analysis of a shortage of fuel wood and other tree products culminating in an "energy crisis" beginning in Elizabethan times? I think the problem lies in the interpretation of words such as "forest," "felling" and "clear." Large areas of "forest" marked on British maps do not necessarily mean that those areas were completely wooded when maps were first being made. For example, there is no doubt that "Dartmoor Forest" was an open landscape before Roman times. Forest often means an area over which royalty exercised hunting rights and is independent of the concept of a wooded landscape.

Comparisons of forests in Britain with forests on the Continent are not realistic. Professor Nef quotes Biringuccio in *Pirotechnica* (1540): "Very great forests are found everywhere. . . ." They certainly were, and still are, on the Continent. Some of the Continental countries have as much as 40 percent of their land area wooded, whereas in Britain the figure is about 8 percent, and much of the woodland is the result of afforestation by the Forestry Commission in this century. The demands of farms and pastures in the 17th century to which Professor Nef refers are essentially enclosures of open land, not clearance of woodland. In addition a dearth of timber in urban environments cannot be equated with shortages of woodland.

Professor Nef writes: "The need for larger amounts of wood for construction and for heating, particularly for the smelting and refining of ores, called for a substantial increase in the felling of trees." Felling, however, does not necessarily mean deforestation. When the woodman cut down a tree, he expected it to be replaced naturally. The system of woodland management in medieval times (and in some areas up to about 1900) was "coppice" and "coppice with standards." When hardwood trees are cut down, they do not die but produce numerous shoots, the coppice, from the stool (the cut stump and the roots), many of which develop within 15 or 20 years into substantial stems. The medieval woodman in southwestern England called such shoots "sproutes" and "springes." These coppice stems could be harvested anytime from two to 20 years later depending on the produce required, which was known as "wood" and "underwood." They provided material for basket weaving, pea and bean sticks, bark for tanning, faggots for fuel wood and charcoal making, poles for fencing and the stirring of ore smelts. Scattered among the stools bearing coppice shoots were "standards," or trees, that provided "timber," mainly for buildings (and to a lesser extent for ships, whose influence on British woodlands has been exaggerated). Hence the royal proclamation of 1615, quoted by Professor Nef, distinguishing between "Wood and Timber." The material whose shortage the proclamation laments ("not only great and large in height and bulk, but hath also that toughness and heart") was not wood but timber.

Another misconception has arisen from the nature of a wood sale. It was the standing timber and wood that was sold, not the land. In such contracts the purchaser was required to clear the wood. That, however, meant not a total felling of an area but a requirement that the purchaser leave the resultant coppice stools and standards in a managed and undamaged condition. Wood-sale documents usually refer to sales of wood and underwood (coppice produce), but if they refer to standing timber, safeguards are written into the contract to maintain a statutory number of standard trees per acre.

There is, of course, a limit to the sus-

tainable yield of coppice with standards. Periodic heavy demands for timber and price inflation have been cited by Dr. Rackham in the case of the "great rebuild" of timber-framed buildings in East Anglia in the late 16th and early 17th centuries, and I have detected other local demands in Essex County associated with the addition of wood belfries, towers and shingled spires for churches in the 14th, 15th and 16th centuries.

Professor Nef's most interesting article refers to a crisis of fuel wood and timber demands of an increasing population from Elizabethan times onward. From his analysis it is clear that coal relieved the fuel-wood crisis with subsequent effects of the development of the Industrial Revolution. Periodic crises in the timber supply reoccurred, and to the present day they have only been relieved, as Professor Nef correctly states, by imports. Coal, in fact, saved coppice-with-standards management until one development of the Industrial Revolution, the mechanical saw (originally steam-driven), demanded straight and uniform timber that is better supplied by fast-growing conifers, which do not (with the exception of the redwood *Sequoia sempervirens*) coppice.

D. L. WIGSTON

Woodlands Research Group
Plymouth Polytechnic
Devon, England

Sirs:

I thoroughly enjoyed reading the article by G. Franklin Montgomery, "Product Technology and the Consumer" [SCIENTIFIC AMERICAN, December, 1977]. His group at the National Bureau of Standards does a commendable job that is bound to benefit the energy-conscious North American public.

I should like to suggest one relatively simple improvement to a residential appliance that has the potential of reducing energy consumption to a considerable degree. I am referring to the electric water heater that is very popular in certain parts of this continent. The standard version is equipped with an upper and a lower element, each of which is thermostatically controlled. Unfortunately the controls are inaccessible from the outside, and if you want to adjust either of them, you have to remove a panel and then search for the setscrew under a thick layer of insulation.

At present most consumers probably are unaware of the existence of these thermostats, and the few who do know about them are not likely to go to the trouble of changing the factory setting to suit their requirements. It seems to me that by means of a rather straightforward and inexpensive change in design, extensions to the two temperature controls could be devised that would be

THE CLASSIC GRAY'S ANATOMY 3 POUND, 1,257-PAGE COLLECTOR'S DELUXE EDITION

NOW ONLY $6.95!

AN EXTRAORDINARY PUBLISHING TRIUMPH—(ORIGINAL 1901 EDITION WORTH $200.00 OR MORE IF YOU CAN FIND ONE!)

For today's men and women truly interested in their own body and its functions. For students, artists and the medically curious—for everyone. The most fascinating book ever published!

IF you ever wanted to know how any part of your body really works, if you want to experience the same magic feeling a young medical student feels as the mysteries of the body are revealed for the first time, then this landmark edition of *Gray's Anatomy* is for you.

Considered one of the greatest texts of our time, *Gray's Anatomy* has been used by nearly every living physician and every physician has been exposed to it. But its appeal is not only to physicians, but to everyone, including students, artists, and collectors of classic books. Dr. John Crocco, Chief of Pulmonary Services, St. Vincent's Hospital and Medical Center of New York, writing in his new Introduction states:

"This stellar book represents the acme of anatomical description over the last century and will probably still be the premier text in anatomy over the next one hundred years."

Just what is Anatomy? Anatomy is the gateway to medicine and the queen of basic sciences. It is the parts and the whole of the human body and a prerequisite for entry into the field of human biology. In this, his masterpiece, Henry Gray unquestionably found the best method of imparting this knowledge to other minds.

To those interested, we say that this is an unprecedented opportunity to acquire *a magnificent classic* at the fantastically low price of only $6.95.

The book itself is divided into 16 main sections, with hundreds of major subdivisions.

The text includes:

The Skeleton. 181 pages covering all aspects of the 200 bones of the body.

The Articulations. 76 pages describing movements of joints, ligaments, cartilages, etc.

The Muscles. 158 pages covering every muscle from the tiniest in the eye to the largest in leg with every detail of their action and use.

The Blood-Vascular System. 167 pages describing circulation, arteries and veins. A veritable book within a book.

The Lymphatics. 14 pages on every lymph gland in the body from sex glands to neck and face glands.

The Nervous System. 171 pages on the brain, the extremities, the spinal cord, etc.

The Organs of Special Sense. 56 pages on the tongue, nose, eyes and ears.

The Organs of Digestion. 85 pages starting with the teeth and including the salivary glands, stomach, intestines, liver, etc.

The Organs of Voice and Respiration. 67 pages on the larynx, trachea, lungs, etc.

The Male Organs of Generation. 14 pages on the prostate glands, penis, testes, etc.

The Female Organs of Generation. 15 pages on the clitoris, vagina, ovaries, pregnancy, etc.

The Surgical Anatomy of Hernia. 20 pages including scrotal and femoral hernia.

General Anatomy or Histology. 73 pages on the animal cell, nutritive fluids, blood, the skin, secreting glands and much more.

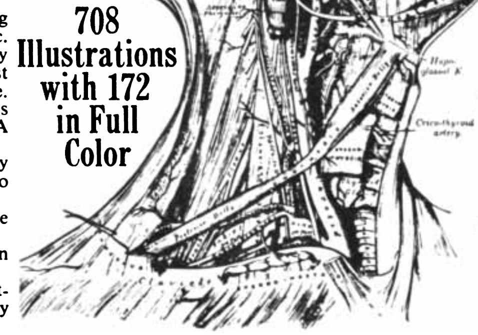
Embryology. 66 pages on the embryo, fetus, ovum, and development of a baby.

The above list of contents is by the limitation of space just a mere summation. The actual content of *Gray's Anatomy* is so massive that the table of contents in the book needs 16 entire pages with 1,932 separate category entries. And the index of this masterwork covers 41 pages with 8,541 separate listings.

Again, as Dr. Crocco states, "There have been many imitations. there have been anatomical narratives, dissection manuals and pictorial atlases with exquisite photographs. However, there is only one *Gray's Anatomy*. *Gray's* is the book all students turn to."

The medical contents alone makes this book invaluable, but the writing style makes it a true literary masterpiece. It combines both technical genius with a beautiful prose-like quality, usually associated with a great novel. The 708 illustrations, including 172 original full color plates have never been surpassed for their authenticity and clarity.

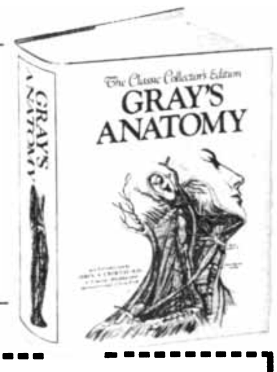
Originally published more than 100 years ago, this classic edition is now available complete and in a new format that is outstanding for both its visual beauty and the unprecedented low price of only $6.95.



708 Illustrations with 172 in Full Color

Fig. 283—Surgical anatomy of the arteries of the neck, showing the carotid and subclavian arteries.

Here is an extraordinary opportunity to acquire one of the greatest books of all time. To order, simply fill out form below and enclose check or money order. You'll be extremely pleased that you did.



CANBRIDGE BOOKS, Dept. 126
6 Commercial Street, Hicksville, N.Y. 11801

Gentlemen. Please send me ________ copies of the Classic Collector's Edition of Gray's Anatomy at the introductory prices shown below. I enclose check or money order for $________.

___Copies of Deluxe Heavy Laminated Soft Cover Edition @ only $6.95 each plus 75¢ for postage and handling.

___Copies of Special Hard Cover Library Edition @ only $8.95 each plus 75¢ postage and handling.

Name ________________

Address ________________

City ________________

State ____________ Zip ________

N.Y. Res. Add Applicable Sales Tax.

both conspicuous and easily accessible. In this way maximum water temperatures in the household tank could be maintained at the desired level and be altered to suit seasonal or other requirements, resulting in substantial energy savings.

T. R. SIMON

Laboratory Manager
Eaton's Product Research Bureau
Toronto

Sirs:

In reference to your review of Hugh Johnson's *The World Atlas of Wine* ["Books," SCIENTIFIC AMERICAN, January], many Europeans will be surprised to learn that "every vine in Europe today is grafted to an American rootstock." In Hungary, where the sandy soil in a large segment of the growing area is unsuitable to *Phylloxera*, only about 40 percent of the 2×10^5 hectares of vines are grafted (1971 figures). Romania has a considerable area in American vines that are ungrafted, and even France has some *Vitis vinifera* as well as hybrid "direct producers" that are not grafted. True, many grapes are grafted as a precaution against *Phylloxera* but certainly not "every vine."

STANLEY ASHMORE

Huntingtown, Md.

Scientific American, March, 1978; Vol. 328, No. 3. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the author's risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 35 cents.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name ______________________

New Address ______________________

____

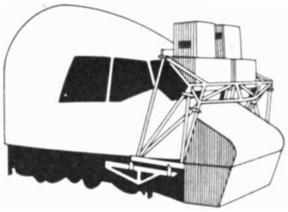
____

Old Address ______________________

____

____

**The subject is simulation:
a remarkable blending of science,
art–and magic–that makes what isn't
there at all seem very, very real.**



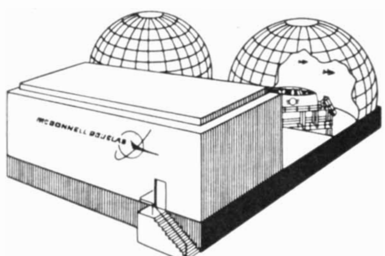
It sure *looks* like La Guardia to the pilot. Right down to the terminal lights and docking points.

The fact is, the "plane" is nowhere near New York. It's not even in the air.

The pilot is "flying" with a visual simulation system built by McDonnell Douglas for on-the-ground training of commercial and military pilots.

Called VITAL IV, it uses computers to draw sharp, true-to-life moving color images of destination cities, their approach, runway, and taxiway lights, their airport buildings and parking areas. It works for day, twilight, and night training. In clear weather, fog, even rainstorms (with or without lightning).

Today, airlines and military forces around the world use our VITAL systems for risk-free training that needs no fuel and costs a fraction of in-the-air training.



Landing at La Guardia is one thing. Air combat at 20,000 feet is quite another. For this, military pilots "fly" in MACS—our Manned Air Combat Simulator. Seated in a dome-enclosed, cockpit mockup, using real instruments and controls, pilots fly realistic combat against other pilots flying similar computer-controlled "hostiles." Pilots see, feel—and fly—as though in actual combat.

Not all simulation is visual, of course. In designing an aircraft, for instance, we accurately simulate the effects of a lifetime of heat and stress. We can measure the wear of 20 years of flying before we even start to build the plane.

And at MCAUTO,® our data service company, we work with another kind of simulation called linear programming. It helps business management find the most efficient uses for resources. For example, we completely "modeled" the production, transportation and sales for a large cement producer, helping management decide where, how, and when to undertake capital expansion.

Simulation—part science, part art, part magic. A very real, highly-developed technology at McDonnell Douglas. If you'd like to know more about how our simulation skills might work for you, just drop us a note on your letterhead.

Write: McDonnell Douglas, Box 14526, St. Louis, MO 63178.

MCDONNELL DOUGLAS

We bring technology to life.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MARCH, 1928: "Evolution, eugenics and the breeding of new varieties of crop plants and domestic animals—all these are affected by the work of Professor H. J. Muller of the University of Texas. Stated in three sentences, this is what Professor Muller's experiments signify: Evolutionary changes, or mutations, can be produced 150 times as fast by the use of X rays as they can by the ordinary processes of nature. This means that man may someday force the production of new and desirable plant and animal varieties far more rapidly than he has hitherto been able to get them. But X rays affect the human heredity cells too, and the reckless exposure of these cells to long and heavy doses of the rays is apt to inflict fearful penalties on our unborn grandchildren."

"A plate of quartz, cut from a clear and perfect crystal in the proper direction, has the curious property of expanding and contracting under the influence of a periodic electric field furnished by two thin metal plates in contact with the opposite surfaces of the plates and joined to the terminals of a high-frequency electric oscillator. A quartz plate one centimeter in thickness has a natural frequency of vibration of about 300,000 per second, and if we tune the electric oscillator to this frequency, resonance occurs, and the amplitude of the vibration of the quartz plate becomes very great. If a thread of glass is made to vibrate at this frequency and is pressed against a pine chip, the wood smokes and emits sparks, the thread rapidly burning its way through the chip, leaving a hole with charred and blackened edges. This is only one of the many remarkable effects obtained with high-frequency sound vibrations in a series of experiments carried out by R. W. Wood, professor of experimental physics at Johns Hopkins University, in collaboration with Mr. Alfred L. Loomis in his private laboratory located at Tuxedo Park, N.Y."

"Nitric acid produced by the oxidation of ammonia can now be made in any concentration, including fixed charges on plant, for at least $30 per ton less than it can be made from Chilean nitrate, even figured on the basis of two cents per pound. Fixed nitrogen in the form of ammonia is now obtainable for less than half the price of nitrogen and Chilean nitrate. This, together with oth-

er important developments in the past 10 years, points to early obsolescence of plants producing nitric acid by the old processes in America and Europe."

"An interesting set of circumstances surrounding the use of a trademark has recently been brought to light by the unsuccessful attempt of the B. F. Goodrich Company to prevent the Closgard Wardrobe Company from registering the term 'zipper' as a trademark for wardrobe bags. The well-known hookless fasteners are made by the Hookless Fastener Company and sold to various manufacturers. The Goodrich Company buys the fasteners and places them on boots and shoes made of rubber and fabric. In order to popularize this footwear the Goodrich Company coined the term 'zipper' and registered it as a trademark. Along comes the Closgard Company prepared to reap where it has not sown. It applies for registration of the name 'zipper' for wardrobe bags. The Hookless Fastener Company cannot protest; it makes the fastener but has nothing to do with the trademark. The Goodrich Company protests in vain, for although it originated the trademark and made it valuable, it has applied it only to footwear."

SCIENTIFIC AMERICAN

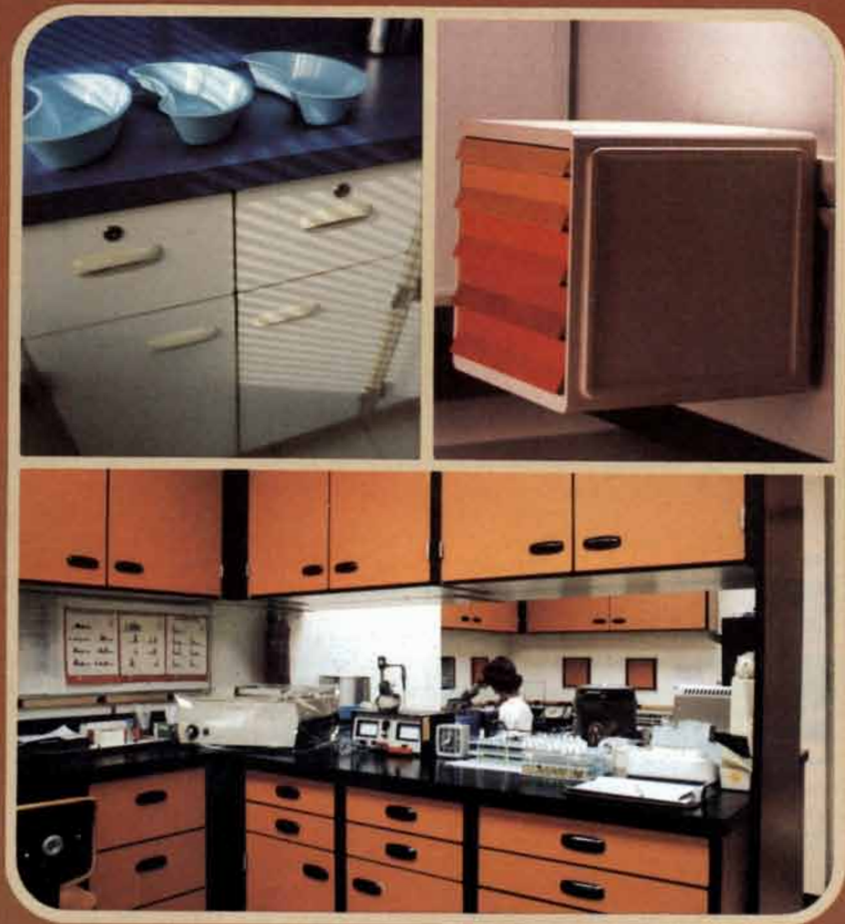
MARCH, 1878: "The new Otto horizontal gas engine closely resembles the ordinary horizontal steam engine. A cylinder receives a piston, which on starting draws in air and gas mingled in certain proportions. A flame brought in contact with this mixture produces its rapid combustion. The high temperature thus engendered results in considerable pressure, and the gases act by expansion upon the piston and drive it ahead; it in turn communicates its motion to connecting rod, crank, etc. There are four distinct phases of operation for every two turns of the flywheel, namely aspiration of air and gas, compression of the mixed gases, combustion and impulse given to the piston, and escape of the products of combustion."

"On being questioned concerning his telephone Prof. Thomas A. Edison said: 'I went to work before Professor Bell. Elisha Gray turned in at it, and got out the first machine. Bell's and mine came out about the same time. The machines are different. Bell's is called the magneto telephone and mine the carbon. Were it not for my hearing I would have discovered the telephone eight months before. While trying an experiment my deafness led me off on a wrong tack, and I was sloshing around on a false scent for months. But I have produced a good instrument. I have whispered into it here at Menlo Park and have been answered in a whisper by Henry Bentley in the Western Union office in Philadelphia.'"

"It has been discovered that several kinds of American vines withstand the ravages of the plant louse *Phylloxera*, notably the 'Clinton.' A vineyard at Montpellier, France, is now largely planted with stocks of the Clinton grape, on which the French Aramon vine is grafted. The grafting is necessary to obtain grapes whose wine has a satisfactory taste. A work recently published at Bordeaux describes the American vines best suited for stocks and specifies the French vines that can be grafted on them to advantage. For a while it was hoped that German stocks would have answered the purpose, but they are found unable to resist the insect pest."

"We have before us a record showing the number of lives lost in crossing the Atlantic during the past 37 years. In this period 56 fine steamers have been wrecked and in 29 instances lives were lost. Nine vessels were never heard from after leaving port; the number of lives that were thus blotted out aggregates 1,397. Of the remaining vessels four were burned, five were sunk by collision (two by colliding with icebergs), two foundered at sea and 34 were wrecked on various coasts. In only two cases of the entire total of steamers lost can the disaster be attributable to a breakdown of the machinery. The inference therefore is that the greatest loss of life is due not to lack of safe vessels but to failure in judgment."

"The supplement of the 35th annual report of the Register General of England presents some valuable and interesting statistics. The report singles out, in imagination, a generation of a million persons and traces its eventful journey from birth to the end of life. Of these more than a fourth die before they reach five years of age. During the next five years the tenure of life becomes surer, and between five and 10 years of age the number of deaths is less than a seventh part of that of the first five. Between 10 and 15 the average mortality is lower than that in any other period. From 15 to 20 the number of deaths increases again, particularly among women, many of whom fall prey to consumption and child-birth. In this period dangerous occupations also begin to affect the death rate. Fully eight times as many men as women die of violent deaths. Consumption is prevalent and fatal from 20 to 45, and is responsible for nearly half of the deaths. From 35 to 45 the effect of wear and tear begins to reveal itself, and many persons succumb to diseases of the important viscera. By 55 the imagined million has dwindled down to less than half. After this the death rate increases more rapidly until by degrees all disappear."



THE AUTHORS

ANDREW R. FLOWER ("World Oil Production") is a policy analyst with the British Petroleum Company specializing in long-term energy supply and demand. A native of Yorkshire, he received his bachelor's degree in mathematics at the University of Oxford in 1968 and his master's degree in statistics at the University of Birmingham the following year. He then joined British Petroleum as a statistician and spent three years studying the application of forecasting methods to company planning procedures. In 1972 he moved to the Corporate Planning Department as an economist responsible for forecasting energy demand in the European Economic Community; he was transferred to his present position with the Policy Review Unit in 1975. From 1975 to 1977 Flower was a member of the Workshop on Alternative Energy Strategies chaired by Carroll L. Wilson of the Massachusetts Institute of Technology; it was there that he developed many of the ideas that are discussed in his article.

MARTIN L. PERL and WILLIAM T. KIRK ("Heavy Leptons") are both on the staff of the Stanford Linear Accelerator Center (SLAC). Perl, professor of physics at Stanford University, obtained his bachelor's degree in chemical engineering at the Polytechnic Institute of Brooklyn in 1948 and worked as a chemical engineer for several years. His interests then turned to atomic physics and he went on to graduate school in that subject, receiving his Ph.D. from Columbia University in 1955. After joining the faculty of the University of Michigan he began to work in experimental elementary-particle physics and has remained in the field ever since. At SLAC Perl's work is centered on the physics of electron-positron colliding rings with the goal of better understanding the leptons, the simplest constituents of matter. Kirk, a technical writer, is assistant director of SLAC's research division and editor of the laboratory's monthly publication, *SLAC Beam Line*. He received his bachelor's degree in history at Cornell University in 1952 and, after four years in industry, joined the staff of the W. W. Hansen Laboratories of Physics at Stanford. His first assignment was to edit the "Project M" proposal that eventually led to the construction of SLAC and its two-mile-long electron accelerator in the early 1960's.

PETER P. LUFF ("The Electronic Telephone") is manager of the Electronic Subscriber Equipment Design group at Bell–Northern Research in Ottawa. Born in England, he was educated at Brighton Polytechnic, where he received his Ph.D. in applied physics in 1969. He then went to Canada and joined Bell–Northern Research to direct a research program on magnetic-bubble memories. After three years he was appointed manager of the Display Technology group, and in 1974 he began a feasibility exercise to develop "a cost-effective and system-compatible electronic telephone utilizing recent advances in bipolar integrated-circuit technology."

RAYMOND E. ARVIDSON, ALAN B. BINDER and KENNETH L. JONES ("The Surface of Mars") are geologists with a special interest in the planets. Arvidson, who is assistant professor of earth and planetary sciences at Washington University and head of the Viking Lander Imaging Team, received his Ph.D. in geoscience from Brown University in 1974. His work on the surface of Mars, he writes, has "awakened an interest in the desert regions of the earth and how they recede and expand, particularly in regions where the spread of deserts is economically devastating." Binder, a research scientist at the Institute of Geophysics at the University of Kiel in West Germany, obtained his Ph.D. from the University of Arizona. He is currently involved in analyzing Viking and lunar sample data and "defending the theory of the formation of the moon by fission from the earth." Jones, a research associate at Brown and a member of the Viking team, obtained his Ph.D. from Brown in 1974. He has recently founded a film production company specializing in scientific and technical documentaries.

JAMES R. GOSZ, RICHARD T. HOLMES, GENE E. LIKENS and F. HERBERT BORMANN ("The Flow of Energy in a Forest Ecosystem") are biologists specializing in the ecology and management of forests. Gosz, associate professor of biology at the University of New Mexico, received his Ph.D. in forest science from the University of Idaho in 1968. His research has included numerous studies of the impact of human activities on natural systems. Holmes, professor of biology at Dartmouth College, obtained his Ph.D. in zoology from the University of California at Berkeley in 1964. His area of interest is the structure and function of animal communities. Likens, professor of ecology and systematics at Cornell University, received his Ph.D. in zoology from the University of Wisconsin in 1962 and was a member of the Dartmouth faculty for six years before he moved to Cornell in 1969. His research is focused on the cycling of nutrients and minerals in lake and forest ecosystems. Bormann, professor of forest ecology at Yale University's School of Forestry and Environmental Studies, obtained his Ph.D. in plant ecology from Duke University in 1952. A member of the National Academy of Sciences since 1973, he is currently on the Man in the Biosphere Directorate (for tropical forests) of the U.S. Department of State.

PETER C. HINKLE and RICHARD E. McCARTY ("How Cells Make ATP") are on the faculty of Cornell University. Hinkle, associate professor of molecular and cell biology, received his bachelor's degree at Harvard College and his Ph.D. from New York University in 1967. He then spent a postdoctoral year investigating biological energy transformation with Peter Mitchell at the Glynn Research Laboratories in England. His present work is focused on how various substances, particularly the sugar glucose, are transported across cell membranes. McCarty, professor of molecular and cell biology, did his undergraduate work at Cornell and went on to obtain his Ph.D. from Johns Hopkins University in 1964. In addition to investigating the membranes of chloroplasts, he teaches an introductory biochemistry course and lectures in more advanced courses.

RONALD L. GRAHAM ("The Combinatorial Mathematics of Scheduling") is head of the Discrete Mathematics Department at Bell Laboratories. He received his bachelor's degree in physics at the University of Alaska and his Ph.D. in mathematics from the University of California at Berkeley in 1962. His present work involves combinatorics, number theory, algorithms and the application of these disciplines to real-world problems. Of his nonacademic pursuits Graham writes: "For several years during my youth I was a professional trampoline performer with the Bouncing Baers (I still coach and perform now and then). I am also past president of the International Juggler's Association, a worldwide organization of some 500 professional and amateur jugglers."

H. ARTHUR KLEIN ("Pieter Bruegel the Elder as a Guide to 16th-Century Technology") is a writer based in Malibu, Calif. He has authored, coauthored or edited more than 20 books, mostly on science and the fine arts, of which the most recent is a 740-page work titled *The World of Measurements* (Simon and Schuster, 1974). He has written two books on Bruegel: *Graphic Worlds of Peter Bruegel the Elder* (Dover, 1963) and, with his wife Mina, *Peter Bruegel the Elder, Artist of Abundance* (Macmillan, 1968). Klein also wrote and produced the sound film *Bruegel's Seven Deadly Sins,* which is derived entirely from the artist's series of "demonic and didactic" engravings.

MATHEMATICAL GAMES

Count Dracula, Alice, Portia and many others consider various twists of logic

by Martin Gardner

"I now introduce Professor Smullyan, who will prove to you that either he doesn't exist or you don't exist, but you won't know which."

—MELVIN FITTING, introducing Raymond Smullyan to an undergraduate mathematics club

Raymond M. Smullyan's *What Is the Name of This Book?*, scheduled to be published in May by Prentice-Hall, Inc., is the most original, most profound and most humorous collection of problems in recreational logic ever written. It contains more than 200 brand-new puzzles, all concocted by the ingenious author and interspersed with mathematical jokes, lively anecdotes and mind-bending paradoxes. The book culminates with a remarkable series of story problems that lead the reader into the core of the late Kurt Gödel's revolutionary work on undecidability.

Who is Smullyan? He was born in 1919 in New York, studied philosophy under Rudolf Carnap at the University of Chicago and received his Ph.D. in mathematics at Princeton University. He is now professor of mathematics at Lehman College of the City University of New York and a member of the faculty of both the mathematics and the philosophy departments of the City University Graduate Center. Among experts on logic, recursion theory, proof theory and artificial intelligence he is best known as the author of two elegant little treatises: *First-Order Logic* (Springer-Verlag, 1968) and *Theory of Formal Systems* (Princeton University Press, 1961). His article in *The Encyclopedia of Philosophy* on Georg Cantor's famous continuum problem is a marvel of lucid compression. Last year Harper & Row published his first nontechnical work, *The Tao Is Silent,* one of the best introductions to Taoism I have seen.

Smullyan's main hobbies are music (he is an accomplished classical pianist), magic (in his youth he was a part-time professional) and chess. Several of his brilliant chess problems have appeared in this department. He has recently finished work on two collections of problems, *Chess Mysteries of Sherlock Holmes* and *Chess Mysteries of the Arabian Knights,* in which each problem is embedded in an appropriate pastiche. Add an enticing literary style, a large sense of humor and a Carrollian love of paradox and you have the flavor of *What Is the Name of This Book?*

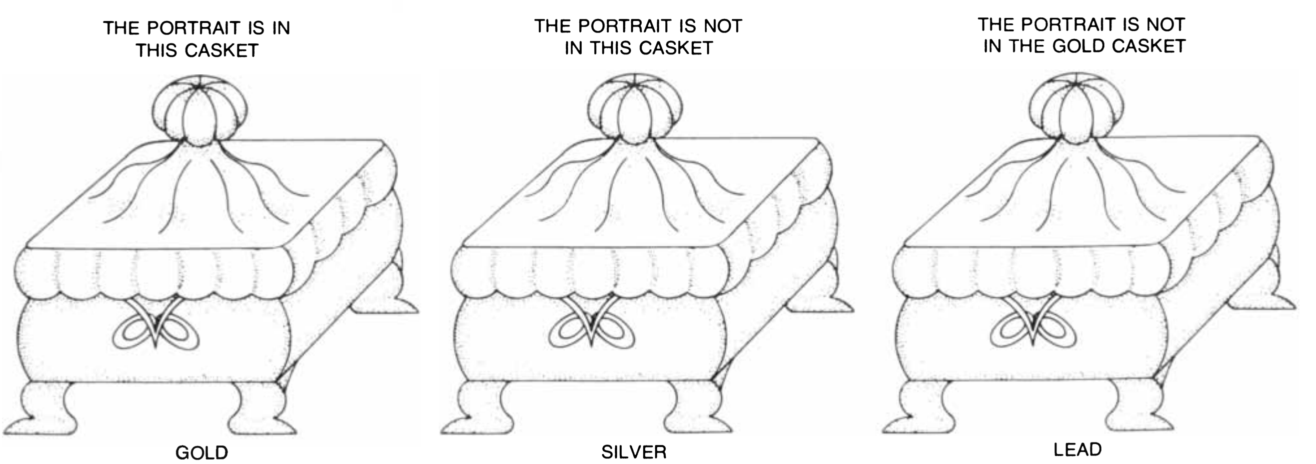
The book opens with a true story that introduces one of Smullyan's central themes. On April Fool's Day when Smullyan was six his older brother Émile told him he was going to be fooled as he had never been fooled before. All day long Smullyan waited for the prank, and he lay awake that night still waiting for it. Finally Émile revealed the joke: Raymond had expected to be fooled and so, by not doing anything, Émile had fooled him.

"I recall lying in bed long after the lights were turned out," Smullyan writes, "wondering whether or not I had really been fooled. On the one hand, if I wasn't fooled, then I did not get what I expected, hence I was fooled.... But with equal reason it can be said that if I was fooled, then I *did* get what I expected, so then, in what sense was I fooled?.... I shall not answer this puzzle now; we shall return to it in one form or another several times in the course of this book."

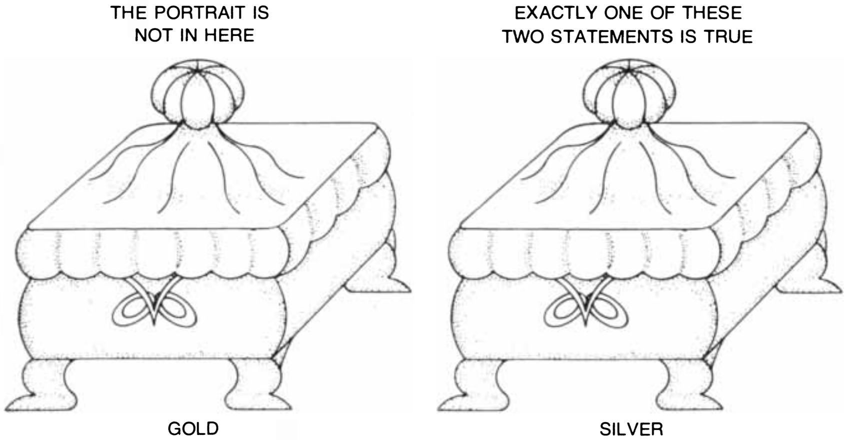
After an introductory section on some classic brainteasers, many with amusing new twists, Smullyan introduces three types of people who will participate in most of the problems that follow: "knights," who always tell the truth; "knaves," who always lie, and "normals," who sometimes tell the truth and sometimes lie.

It is amazing how much can be deduced from just a few lines of dialogue with these individuals. For example, on an island inhabited only by knights and knaves Smullyan comes on two men resting under a tree. He asks one of them: "Are either of you knights?" When the man—call him *A*—responds, Smullyan instantly knows the answer to his question. Is *A* a knight or a knave? What is the other man?

It seems as though there is not enough information to solve this problem, but the key lies in the fact that *A*'s answer enabled Smullyan to find the solution. If the answer had been yes, Smullyan would have learned nothing. (If *A* is a knight, one or both of the men could be knights, and if *A* is a knave, both of them could be knaves.) Therefore *A* must have said no. Now, if *A* were a knight, he would have had to say yes, but since he



Portia's first casket test from **What Is the Name of This Book?**



Where is Portia's portrait?

said no he must be a knave. Because he is a knave his no is false; therefore at least one knight must be present. Hence the other man is a knight.

Lewis Carroll's Alice soon enters the book. We find her wandering around in the Forest of Forgetfulness, where she is unable to remember the day of the week (see Chapter 3 of *Through the Looking-Glass*). In the forest she meets the Lion and the Unicorn. The Lion lies on Monday, Tuesday and Wednesday and the Unicorn lies on Thursday, Friday and Saturday. At all other times both animals tell the truth. "Yesterday was one of my lying days," says the Lion. "Yesterday was one of my lying days too," says the Unicorn. Alice, who is as smart as Smullyan, is able to deduce the day of the week. What is it? Next month I shall give the solution of this puzzle and the puzzles that follow.

Enter more *Looking-Glass* characters: the Tweedle brothers, the White King, Humpty Dumpty and the Jabberwock. Tweedledee behaves like the Unicorn, Tweedledum like the Lion. After Alice has solved a number of problems based on conversation with the Tweedles, Humpty Dumpty discloses a well-kept *Looking-Glass* secret: There is a third brother, identical in appearance with Dum and Dee, named Tweedledoo. Doo always lies. Alice is now dreadfully upset because all her previous deductions may be wrong. On the other hand, Humpty may be lying, and Tweedledoo may not exist. Four accounts are given of what happens next, and the reader is asked to deduce which one is correct and whether or not Tweedledoo exists.

The scene then shifts to Shakespeare's *The Merchant of Venice* and that famous puzzle occasion when Portia presents her suitor with three caskets, gold, silver and lead, each of which bears a different inscription. Only one casket contains Portia's portrait, and if the suitor guesses it correctly, Portia will marry him. (It is not generally realized, by the way, that Portia gives her suitor a whopping hint by singing, "Tell me where is fancy bred, / Or in the heart or in the head?" The last words of both lines rhyme with lead, the correct choice.)

By varying the inscriptions on the caskets Smullyan creates a series of remarkable problems that lead the reader closer and closer to Gödel's discovery. The first problem is shown in the illustration on the preceding page. Portia, who never lies, explains to her suitor that at most one inscription is true. Which casket should he choose?

Smullyan improvises clever variations on the Portia theme. In some problems each casket bears two inscriptions. We also learn that there are two casket makers: Bellini, who always puts true inscriptions on his caskets, and Cellini, who always puts on false ones.

Consider another casket problem, shown in the illustration above. The gold casket bears the inscription "The portrait is not in here" and the silver casket bears the inscription "Exactly one of these two statements is true." The two inscriptions present a logical dilemma of enormous importance in the history of modern semantics. The suitor reasons: If the silver statement is true, then the gold one is false. If the silver statement is false, then the inscriptions are either both true or both false. They cannot both be true if the silver statement is false, and so they are both false. In either case the gold statement is false; therefore the gold casket must contain the portrait. The suitor triumphantly opens the gold casket but finds to his horror that it is empty. The portrait is in the other casket. What is wrong with his reasoning?

The error he makes is in assuming that the statement on the silver casket is either true or false. This problem involves us in the modern concept of metalanguages. It is only permissible to discuss the truth values of a particular language in a larger language, or metalanguage, that contains the first language as a subset of its terms. When a language refers to its own truth values, the result is often a logical contradiction. Without metastatements about the truth or falsity of the casket inscriptions or information about how their truth values are related the inscriptions can be meaningless.

Inspector Leslie Craig of Scotland Yard now strides into the book, and Smullyan gives us a variety of mysteries from the inspector's files that can be solved by careful logical deduction. The first one is the simplest:

"An enormous amount of loot had been stolen from a store. The criminal (or criminals) took the heist away in a car. Three well-known criminals, *A, B* and *C,* were brought to Scotland Yard for questioning. The following facts were ascertained:

"(1) No one other than *A, B* and *C* was involved in the robbery.

"(2) *C* never pulls a job without using *A* (and possibly others) as an accomplice.

"(3) *B* does not know how to drive.

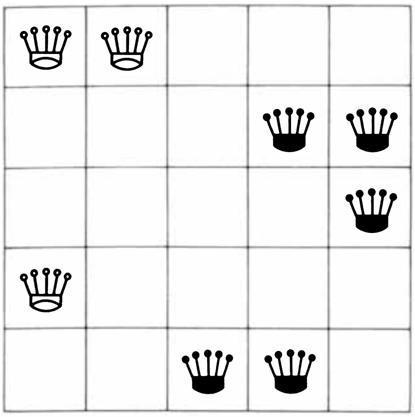
"Is *A* innocent or guilty?"

In the pages that follow Smullyan is concerned with such practical tasks as how to avoid werewolves, how to choose a bride, how to defend yourself in court and how to marry a king's daughter. For instance, suppose you want to convince a prospective bride who is unusually fond of knaves that you are a rich knave. (You must be rich or poor.) Can it be done with a single sentence? Yes. You have only to say, "I am a poor knave." The girl knows at once that you cannot be a knight because a knight would not lie and say he is a poor knave. Since you are a knave your statement must be false; therefore you are a rich knave. Suppose the girl is attracted only to knights. What sentence will convince her you are a rich knight?

The next section introduces logic puzzles based on the conditional statement of the form "If *P* is true, then *Q* is true." The two statements are connected by the relation of implication, an understanding of which is absolutely indispensable in understanding propositional calculus. Smullyan plays with the familiar paradoxes of implication and then gives 18 ingenious puzzles no reader can think through without acquiring a firm grasp on the logical principles involved.

The next setting is the Island of Baal, the only place on the earth where someone knows the answer to the superultimate metaphysical question: "Why is there anything at all?" The island is inhabited only by knights and knaves. After a series of encounters with the local folk Smullyan proves that the Island of Baal cannot exist.

The proof of nonexistence does not apply to the next island Smullyan visits: the Island of Zombies. Here there is no easy way to distinguish the zombies, who always lie, from the human beings, who always tell the truth. Life is further complicated by the fact that all yes-or-no questions are answered with either "bal" or "da," but we do not know which



Solution to last month's chess problem

means yes and which means no. Suppose you ask a native whether "bal" means yes and he answers "bal." You cannot determine what "bal" means, but can you tell whether the speaker is a human being or a zombie? Is it possible to find out what "bal" means in just one yes-or-no question?

Transylvania is equally confusing. Here the human beings (truthtellers) do not look different from the vampires (liars) and half of the inhabitants are insane. The insane believe all true propositions are false and all false propositions are true. Thus there are four types of Transylvanians: sane human beings, insane human beings, sane vampires and insane vampires. Anything a sane human being says, of course, is true and everything an insane human being says is false. Conversely, anything a sane vampire says is false and anything an insane vampire says is true. Luckily all questions are answered in English. How can you determine in one yes-or-no question whether a Transylvanian is a vampire? How can you find out in one yes-or-no question whether he is sane?

Smullyan, eager to know whether Dracula is alive or dead, puts the question to various Transylvanians. The reader is asked to deduce the answer from the dialogue. This section culminates in a grand ball at Count Dracula's castle, where the complexities are compounded by the fact that all questions are answered with "bal" and "da," as on the Island of Zombies. As a result there are three variables to worry about: Is the speaker sane, is he human and what does "bal" mean? Smullyan eventually discovers that Dracula is alive but insane.

One chapter of the book is titled "How to Prove Anything." After examining a sophistry in Plato's dialogue *Euthydemus* in which one speaker proves that another speaker's father is a dog, Smullyan discusses a number of curious devices by which one can seemingly prove that anything—God, Satan, unicorns, Santa Claus and so on—exists. One of the devices is derived from the traditional ontological argument for God. Some of the others are variants of a subtle method discovered by the mathematical logician J. Barkley Rosser.

For example, consider the following sentence: "If this sentence is true, then Santa Claus exists." Smullyan writes: "If the sentence is true, then surely Santa Claus exists (because if the sentence is true then it must also be true that if the sentence is true then Santa Claus exists, from which it follows that Santa Claus exists), hence what the sentence says is the case, so the sentence is true. Hence the sentence is true and if the sentence is true then Santa Claus exists. From this it follows that Santa Claus exists." The argument is unsound, but without an understanding of the role of metalanguages it is not easy to explain exactly why.

The penultimate chapter introduces the familiar "liar paradox" ("This statement is false") and its many disguises and variants. Smullyan presents some of the deepest paradoxes of logic and set theory in a way that makes them clearer than they have ever been made before. For instance, here is his explanation of the famous paradox, known as Richard's paradox, that is the basis for Gödel's undecidability proof.

A mathematician has a book called *The Book of Sets*. On each page he lists or describes a set of counting numbers. The pages are numbered consecutively. Can we describe a set of positive integers that cannot be listed in the book?

We can. If a number *n* belongs to the set listed on page *n*, call it an extraordinary number. If *n* does not belong to the set on page *n*, call it an ordinary number. Now consider the set of all ordinary numbers. Assume that this set is listed on a certain page. The number of the page cannot be ordinary because if it were, that number would be on the page and consequently it would be extraordinary. On the other hand, it cannot be extraordinary because in that case it would have to appear on the page and we assumed that the page lists only ordinary numbers. This contradiction forces us to abandon the assumption that the set of ordinary numbers can be listed. Therefore there is a set of positive integers that cannot be listed in the book.

Now we are ready for Smullyan's climactic chapter on Gödel's discovery; this final chapter is the best introduction I know to that great watershed in the history of the study of the foundations of mathematics. Since the days of Leibniz mathematicians have dreamed that someday all mathematics would be united in one vast system in which every statement that could be formulated could be proved true or false. Leibniz even extended the dream to philosophical disputes. "If controversies were to arise," he wrote, "there would be no more need of disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down to their slates, and to say to each other (with a friend to witness, if they liked): Let us calculate."

This dream was shattered forever by Gödel's paper of 1931. In the paper the 25-year-old Gödel showed that the deductive system of Alfred North Whitehead and Bertrand Russell's *Principia Mathematica,* as well as related systems such as standard set theory, contains undecidable statements, that is, statements that are true but cannot be proved true within the system. More precisely, Gödel showed that if a system like that of *Principia Mathematica* satisfies certain reasonable conditions such as consistency (freedom from contradiction), then it allows the formation of sentences that are undecidable. At the same time he showed that if such a system is consistent, there is no way to prove the consistency within the system.

These results apply to any deductive system rich enough to contain arithmetic. Even in ordinary arithmetic there are statements that are true but unprovable. (Very simple systems, such as arithmetic without multiplication, are free from any undecidable statements.) Moreover, it is not possible to prove the consistency of arithmetic within arithmetic.

One can, of course, enlarge arithmetic by adding new axioms so that the enlarged system allows proof of any formerly undecidable statement. Alas, the situation is as hopeless as it was before. It can be shown by Gödel's same arguments that the enlarged system will contain new undecidable statements and that the consistency of the enlarged system cannot be proved within it. This construction of increasingly large systems can go on forever, but it will never reach a level where undecidable statements can be banished from a system or a consistency proof can be devised within the system.

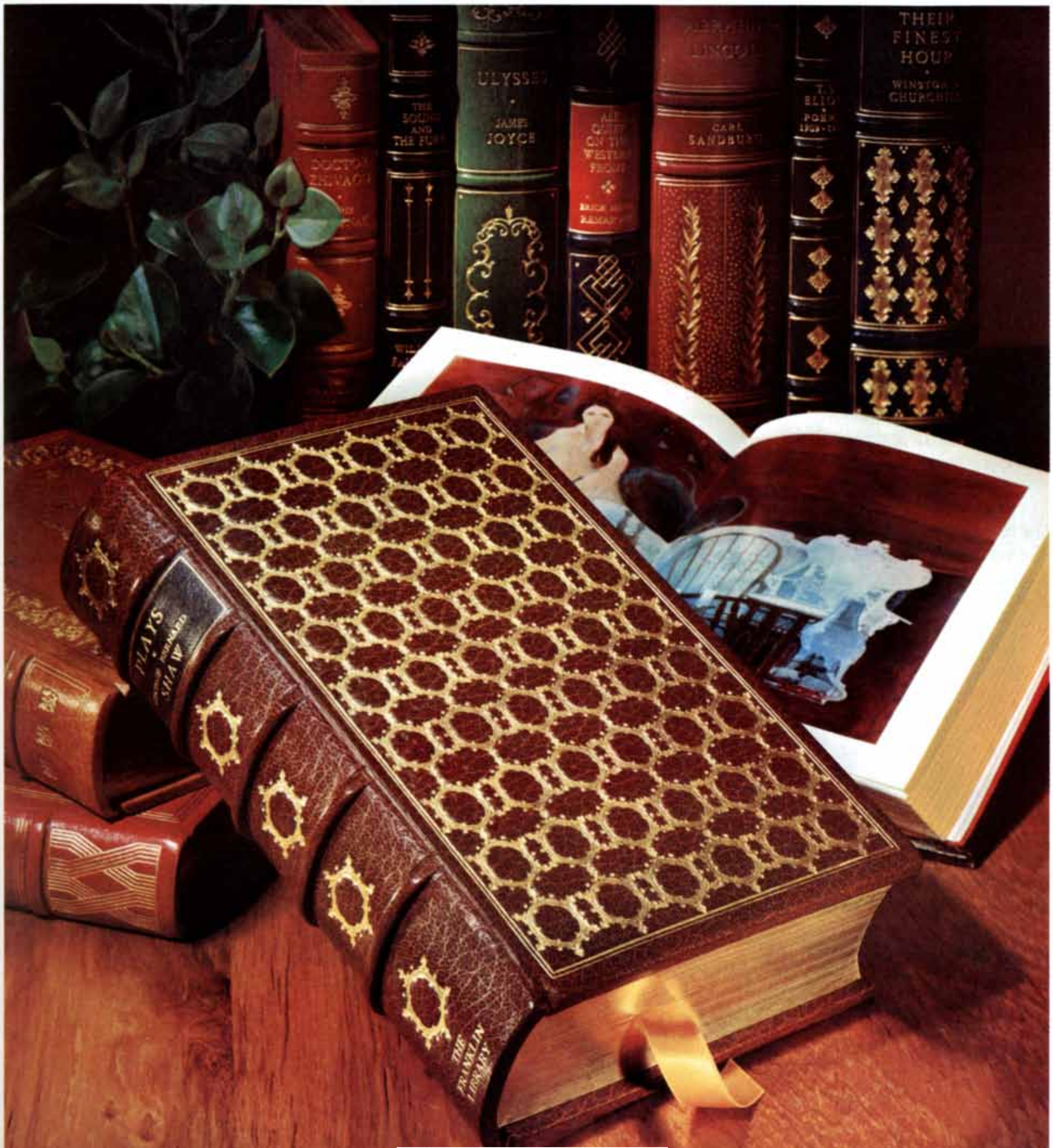
There is a famous unsolved problem in arithmetic called Goldbach's conjecture. It asserts that every even number greater than 2 is the sum of two primes. No one has proved it or found a counterexample. It is possible that Goldbach's conjecture is Gödel-undecidable. If that is so, it means the conjecture is true but unprovable within arithmetic. It is true because if it were false, a counterexample would exist, and that would make the conjecture decidable.

An even more disturbing aspect of this situation is the lack of any constructive way of showing that number theorists will not someday find an arithmetical proof that Goldbach's conjecture is true and also a proof that it is false! Mathematicians hope and believe it will never happen to any arithmetical theorem because if it does, it would reduce arithmetic and all higher mathematics to a shambles. (It is easy to show that if a deductive system contains even one contradiction, then it is possible to prove any statement whatsoever in the sys-

The Franklin Library presents

THE GREATEST BOOKS OF THE 20th CENTURY

A complete library of this century's greatest masterpieces, chosen by a distinguished international Board of Advisors. Each book fully bound in leather and ornamented in 22 karat gold.



Available by direct subscription only.
Subscription deadline: March 31, 1978.

THE TWENTIETH CENTURY has produced more great books than all previous centuries combined.

It has been the century of the novel—an art form largely perfected in our time. It has been an age of great drama and poetry. And it has been a time of unprecedented growth in knowledge—clearly explained in books by the foremost scientists, philosophers, biographers and historians of our day.

And now, from this vast treasury of major works, one of the most distinguished advisory boards ever assembled will select those which stand out as the greatest books of the twentieth century. Fifty towering masterpieces. The essential volumes of our time. *The very best writing of the modern world.*

Furthermore, this collection of *The Greatest Books of the Twentieth Century* will be published in a single, very special edition. Each volume will be set apart from the ordinary by individual design, full leather binding and rich ornamentation in 22 karat gold. And each will be privately printed and bound for subscribers by The Franklin Library, the world's leading publisher of limited edition books in fine leather bindings.

A treasury of the century's masterpieces

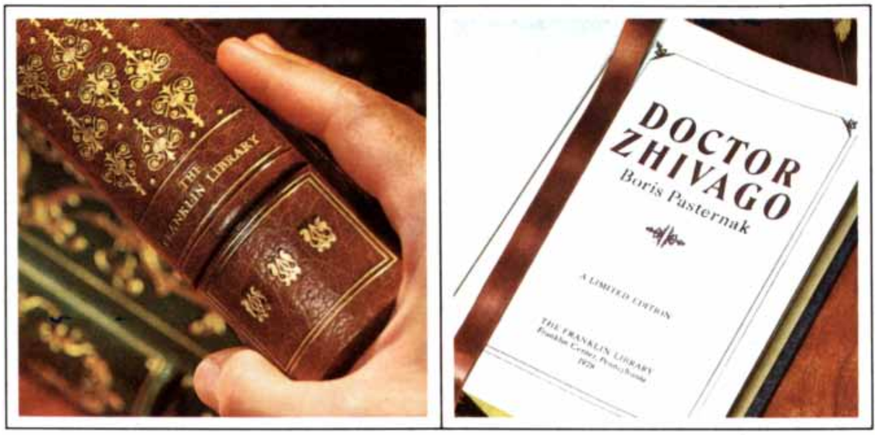
A major portion of the collection will be devoted to the most compelling novels of the twentieth century: *The Grapes of Wrath* by John Steinbeck. *Doctor Zhivago* by Boris Pasternak. *Look Homeward, Angel* by Thomas Wolfe. *The Magic Mountain* by Thomas Mann. *The Trial* by Franz Kafka.

In addition, there will be collections of *stories* by the master storytellers of our time: Ernest Hemingway. James Joyce. Jorge Luis Borges. *Drama* by the century's leading playwrights: Bernard Shaw. Anton Chekhov. Eugene O'Neill. *Poetry* by the modern masters: W. H. Auden, Robert Frost, William Butler Yeats. *Biographies* and autobiographies that recount the fascinating lives of great international figures. And brilliant accounts of the *history, philosophy* and *science* of our time.

Leather-bound volumes—individually designed and illustrated

All fifty volumes will be newly designed for this private edition. And each will be different from every other—in design, format and overall dimensions.

Each volume will be fully bound in leather. And leathers of different grains and colors will be used to create a harmonious yet varied appearance throughout the collection. The covers and the spine of each book will all be richly ornamented with 22 karat gold. The spine of every book will have raised horizontal "hubs," and the page edges will be *gilded with 22 karat gold*—the time-honored manner of enriching fine volumes.



Each volume will be beautifully crafted of leather, gold and moiré fabric.

Each volume will be printed on specially milled bookpapers and the endleaves will be of lustrous moiré fabric.

Original works of art will be commissioned to illustrate many of the books and will appear *only* in these leather-bound volumes. Famous works by outstanding artists of the past will grace other books in this collection. This tasteful use of illustration will add dramatically to the pleasure of reading and owning these magnificent books.

A single edition for subscribers only

This exclusive edition of *The Greatest Books of the Twentieth Century* will be privately printed and bound *solely* for those who subscribe to the complete collection. None of these leather-bound volumes will be made available in bookstores, nor will any ever be reissued.

The subscription rolls are now open. There is a firm limit of one subscription per person, and subscriptions will be accepted only until March 31, 1978. The edition will then be closed and no further applications *ever* accepted. Thus, the edition will be *permanently limited.*

The Franklin Library gratefully acknowledges the assistance of the following scholars, authors and literary historians in selecting *The Greatest Books of the Twentieth Century.*

DR. CARLOS BAKER, Princeton University.
MALCOLM COWLEY, author and critic.
DR. RICHARD ELLMANN, Oxford University.
ROBERT STUART FITZGERALD, Harvard University.
DR. ARTHUR H. F. HENKEL, University of Heidelberg.
JOHN HERSEY, recipient of the Pulitzer Prize.
J. FRANK KERMODE, Cambridge University.
KATHERINE ANNE PORTER, winner of the Pulitzer Prize.
DR. WALLACE STEGNER, Stanford University.
EUDORA WELTY, Pulitzer Prize winner.

Convenient acquisition plan

As a subscriber, you will be able to acquire this important collection on a truly systematic basis—at the convenient rate of one volume per month.

You will be billed for your books on the same monthly basis, at just $39 for each leather-bound volume. Your subscription application contains *a written guarantee* that this favorable issue price will be maintained for the entire collection.

To enter your subscription, complete and mail the attached application. No payment is required now. However, all applications must be postmarked by the subscription deadline of March 31, 1978, to be eligible for acceptance.

SUBSCRIPTION APPLICATION

THE GREATEST BOOKS OF THE 20th CENTURY

Valid only if postmarked by March 31, 1978.
Limit: One subscription per person.

The Franklin Library
Franklin Center, Pennsylvania 19091

Please accept my subscription for *The Greatest Books of the Twentieth Century*, consisting of fifty volumes to be privately printed and bound in genuine leather expressly for me. The books will be issued to me at the rate of one volume per month, and the issue price of $39.* for each leather-bound book will be guaranteed to me for the entire series. In addition, I will have the right to cancel my subscription at any time upon thirty days' written notice.

I need send no money now. I will be billed for each book, individually, in advance of its shipment.

**Plus my state sales tax and $1.75 per book for shipping and handling.*

Signature ____________________
ALL APPLICATIONS ARE SUBJECT TO ACCEPTANCE.

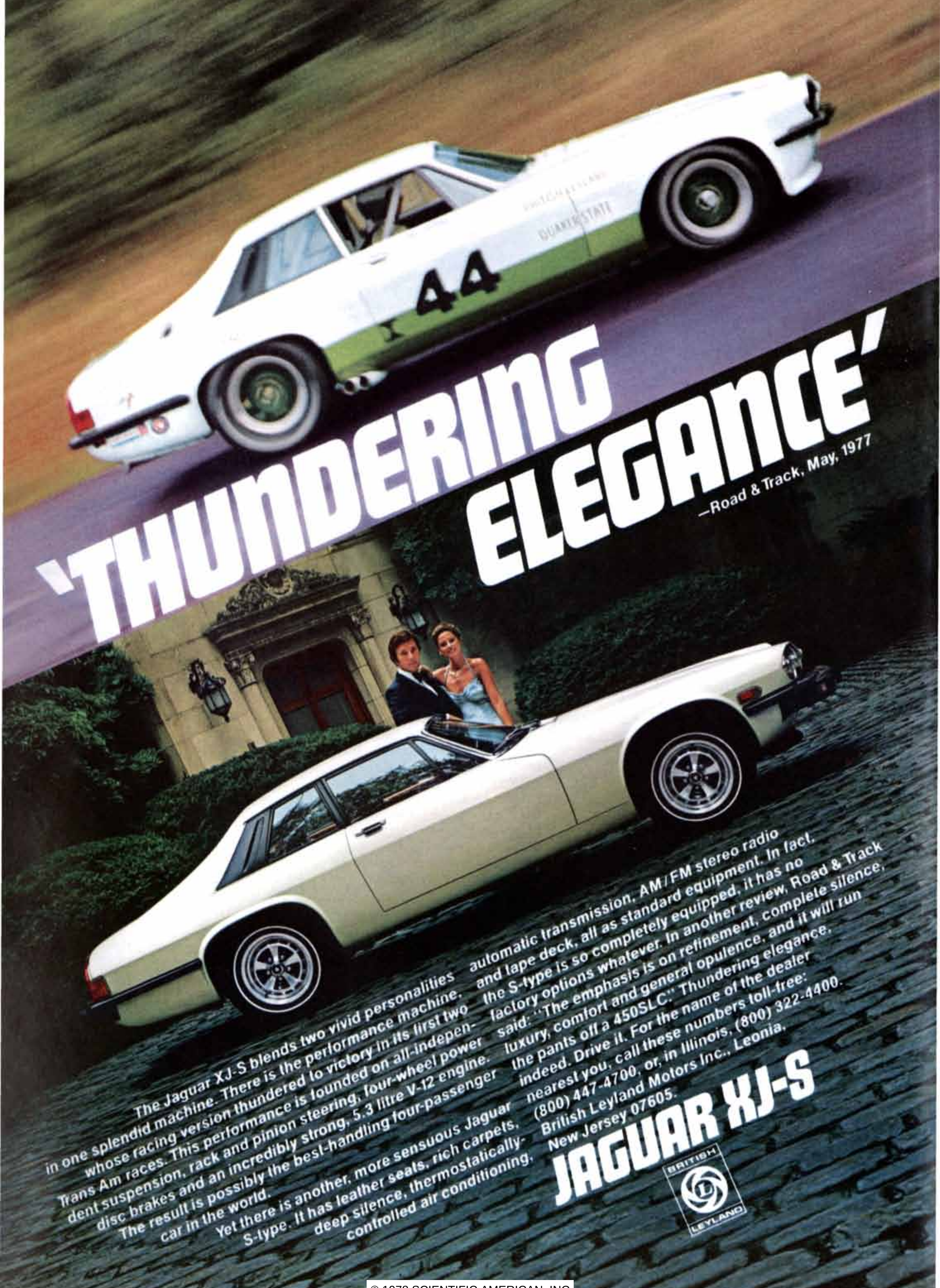
Mr.
Mrs.
Miss ____________________
PLEASE PRINT CLEARLY

Address ____________________

City ____________________

State ____________ Zip ____________

212


44
'THUNDERING ELEGANCE'
—Road & Track, May, 1977
The Jaguar XJ-S blends two vivid personalities in one splendid machine. There is the performance machine, whose racing version thundered to victory in its first two Trans Am races. This performance is founded on all-independent suspension, rack and pinion steering, four-wheel power disc brakes and an incredibly strong, 5.3 litre V-12 engine. The result is possibly the best-handling four-passenger car in the world.
Yet there is another, more sensuous Jaguar S-type. It has leather seats, rich carpets, deep silence, thermostatically-controlled air conditioning, automatic transmission, AM/FM stereo radio and tape deck, all as standard equipment. In fact, the S-type is so completely equipped, it has no factory options whatever. In another review, Road & Track said: "The emphasis is on refinement, complete silence, luxury, comfort and general opulence, and it will run the pants off a 450SLC." Thundering elegance, indeed. Drive it. For the name of the dealer nearest you, call these numbers toll-free: (800) 447-4700, or, in Illinois, (800) 322-4400. British Leyland Motors Inc., Leonia, New Jersey 07605.
JAGUAR XJ-S
BRITISH
LEYLAND

tem.) Platonist mathematicians who regard the axioms of arithmetic as being true and the rules of reasoning as being correct have no such worries because they believe no contradictions can arise. The purely constructive formalists, however, have no such guarantees.

Gödel's undecidable statements are undecidable only within a given system. In 1936 papers by A. M. Turing and Alonzo Church established the existence of problems that are undecidable in a deeper sense. They proved the existence of problems for which there is no finite algorithm, or step-by-step procedure, that will solve them. Examples of these absolute undecidables include the famous halting problem of Turing-machine theory, the color-domino problem in tiling theory, problems in John Horton Conway's game of life and many more. At no time in the future will it be possible in any logically consistent world to build a computer, however powerful, that by twiddling symbols will be able to solve such problems in a finite number of steps.

Since 1936 all kinds of ways of establishing Gödel's results and the related Church-Turing results have been devised, some of them simpler than Gödel's original method. Smullyan presents Gödel's proof in a delightful way by imagining a Gödel island inhabited only by knights and knaves. Knights who have "proved themselves" to be knights are called established knights. Knaves who have proved themselves to be knaves are called established knaves. The inhabitants have formed clubs for which the following conditions hold:

1. The set of all established knights forms a club.
2. The set of all established knaves forms a club.
3. Every club C has its complement: a club consisting of all people on the island who are not in club C.
4. Given any club, there is at least one inhabitant who professes to be a member of that club.

Smullyan is now able to show in just three paragraphs of simple, nontechnical argument that there is at least one unestablished knight on the island and at least one unestablished knave. If we regard the knights as being true sentences, the established knights as being provably true sentences, the knaves as being false sentences and the established knaves as being provably false sentences, the results of Smullyan's argument correspond to Gödel's results. In only three more sentences Smullyan establishes a related theorem of Alfred Tarski's, that neither the set of knaves nor the set of knights forms a club.

There is a well-known version of the old liar paradox where instead of a single sentence there are two sentences, one on each side of a card. One sentence states, "The sentence on the other side of this card is true." The other states, "The sentence on the other side of this card is false." Neither sentence refers to itself and yet the contradiction is obvious. Similarly, Smullyan imagines what he calls "doubly Gödelian islands" of knights and knaves that satisfy the following condition: Given any two clubs C_1 and C_2, there are inhabitants A and B such that A claims that B is a member of C_1 and B claims that A is a member of C_2. The study of such double islands is one of Smullyan's favorite hobbies. He discusses several of his discoveries about them and gives some new problems that have not yet been solved.

The book ends with a truly astonishing version of Gödel's construction of an unprovable sentence. Consider the following statement: "This sentence can never be proved." "If the sentence is false," writes Smullyan, "then it is false that it can never be proved, hence it *can* be proved, which means it must be true. So, if it is false, we have a contradiction, therefore it must be true.

"Now, I have just proved that the sentence is true. Since the sentence is true, then what it says is really the case, which means that it can never be proved. So how come I have just proved it?"

The fallacy, Smullyan explains, is that it has not been made clear just what is meant by provable. Consider a revised version of the sentence: "This sentence is not provable in system *S*." The paradox magically vanishes! "The interesting truth is that the above sentence must be a true sentence which is not provable in system *S*." (It is assumed, of course, that everything provable in *S* is true.) In this form the sentence is "a crude formulation of Gödel's sentence *X*, which can be looked at as asserting its own unprovability, not in an absolute sense, but only within the given system."

At this point Smullyan suddenly remembers that he has not yet answered the question "What is the name of this book?" The name is.... But see the book's last page for the answer.

Last month's problem was to place five queens of one color and three of another color on a five-by-five chessboard so that no queen of one color attacks a queen of another color. The unique solution (excluding rotations and reflections) is shown in the illustration on the opposite page.

Many readers sent corrections and new data for last November's column on Ramsey graph theory. The bounds now established for $R(4,5)$ are 25 through 28, and for $R(5,5)$ they are 42 through 55. The lower bound of 42 was obtained by Yi-Hsin Chen for his recent master's thesis at the University of Toronto. Gary Lorden corrected my statement that Léopold Sauvé had proved that any extremal graph can be two-colored so that all monochromatic triangles are the same color. A. W. Goodman had shown earlier that the statement is true for complete graphs with an even number of points. Sauvé showed it is false for graphs with an odd number of points, with the sole exception of K_7, found by Paul Erdös. Lorden's paper on "Blue-Empty Chromatic Graphs" (*The American Mathematical Monthly*, Vol. 69, No. 2, pages 114–120; February, 1962) gives easy methods for constructing the colorings. Thomas Hoover pointed out how an adjacency matrix can be used to find such colorings.

Robert E. Greenwood and Frank Harary have provided the early history of the famous problem of proving that among any six people there are at least three mutual acquaintances or three mutual strangers. It was making the mathematical-folklore rounds as a graph problem at least as early as 1950, and at Harary's suggestion it was included in the William Lowell Putnam Mathematical Competition of 1953.

In December I asked readers for words longer than "festoons" that are spelled with the first letters of the 10 digits. So many people wrote in that I cannot begin to list all their names. Steward Hartman, assisted by Kenneth Bell, sent the most complete compilation of words ranging from five letters (for example "often") through 12. Many readers hit on the eight-letter words "footnote," "fossette" and "nonsense," each of which can be lengthened to nine letters by making it plural. Eleven-letter words include "oftennesses," "tensenesses" and "sostenentes." ("Sostenente" means "sustained" in music.) Edward Kulkosky, Raymond J. Lovett and Stephen C. Koehler were the first to hit the jackpot with the 13-letter "noneffeteness," which can be stretched to 15 by adding "es." In the category of plausible made-up words, David Elwell proposed "nontennesseeness." Add "es" and the length is 18.

Alan Cassel of the Massachusetts Institute of Technology was the first to report that Dr. Matrix made a careless error when he reported on a large consecutive-digit prime. The correct number is 1234567891234567891234567891. Using a new and fast Monte Carlo test for primality, Cassel made the remarkable discovery that if the cycle 123456789 is repeated seven times and then followed by 1234567, the result is a pseudoprime with odds of about a trillion to one that it is a genuine prime. Perhaps some reader can prove that it is indeed a prime.

Stanley J. Benkoski wrote to say that my theory that cockroaches might trigger earthquakes by psychokinesis had gained credibility in California. "Governor Brown has asked that each state employee be released for one half hour per day to assist in eradication of cockroaches. Officials estimate that cockroaches will be extinct in California by July 13, 1978, and then earthquakes will be relegated to historical curiosities."

BOOKS

William Beebe and his successor aquanauts, and a people that lives on sweet palm juice

by Philip Morrison

NATURAL MAN: THE LIFE OF WILLIAM BEEBE, by Robert Henry Welker. Indiana University Press ($11.50). SUBMERSIBLES AND THEIR USE IN OCEANOGRAPHY AND OCEAN ENGINEERING, edited by Richard A. Geyer. Elsevier Scientific Publishing Company ($49). DIVING MEDICINE, edited by Richard H. Strauss, M.D. Grune & Stratton ($23). One morning in April, 1925, a naturalist of the New York Zoological Society took a few steps down a ladder into the waters of Darwin Bay in the Galápagos. In his log we read that he "found it most exciting experience. Trite but true to say it opens a new world." William Beebe was no quiet specialist but a literary figure of social prominence, a friend and colleague of Theodore Roosevelt's, an author with hundreds of articles and a dozen books to his name. That same year Putnam published a volume of felicitous essays, *Jungle Days.* The vivid field reports from British Guiana are still remembered by many a reader for their graceful language, engaged intelligence, affection for the natural world and informed and catholic perception. "The vast tangle of a fallen jungle tree" lived in Beebe's pages; the reader 50 years ago could learn of thrips and sloths, and of the chain of jungle life represented by the protozoon found in a frog eaten by a fish caught by a snake that had been taken by a bird found dead by Beebe. The theme was broader than the bright plumage and high adventures of Beebe's earlier books; at the height of his powers he wrote of ecology and evolution.

In *Natural Man,* a biography of a prolific writer of personal narrative who left no autobiography and no open papers, the man is sought behind the work. Beebe turns out to be less open, more veiled, than we had thought; he nonetheless remains the yea-sayer and the child of light, a comfortable Fifth Avenue Thoreau whose critique of his century indicted only one terror, war. Perhaps the most novel thread picked out by the acute biographer, who is at Case Western Reserve University, is Beebe's somewhat ambiguous feminism. His scientific field companions were often women (including his first wife, who was rather unfairly treated in Beebe's corpus), and "a long history of the liberation of women from restrictive conventions" can be traced in the photographs and the text. Beebe was "at least . . . a participant," if not clearly an advocate of the change, from his first wife's plea in Edwardian days for a split riding skirt to the easy costume and central responsibility of another young woman who is seen on deck in the 1930's, monitoring the essential telephone line from the bathysphere below.

The books stand on the shelves and in some memories today. But the new world Beebe saw in Darwin Bay opened a path. He was the "first famous and highly trained natural scientist to take to helmet diving" as fieldwork. First he collected and observed in shallow tropical waters (down to 50 feet) in a weighted helmet with an air hose, off Haiti and then off Bermuda. Within two years he dreamed of the depths. No living human being had ever returned from much below 500 feet of water, and yet life was there in plenty: the deep-sea dredges brought up strange creatures. Beebe had to visit that world, although it would tax his expressive powers once he saw it: dark and alien, untouchable, only a cramped view seen through two six-inch ports of fused quartz.

In 1926 *The New York Times* made known Beebe's vague project for a new vehicle "like a laundry boiler hanging from the mother ship by a steel cable . . . equipped with neither lights nor telephone wires." A young New England engineer named Otis Barton read the paper. He had already thought through the plans for a tethered steel ball that would be genuinely practical for the depths. Once he was able to meet Beebe ("as unapproachable as an Indian potentate") Barton unrolled his impressive blueprints and, even better, explained how an inheritance would enable him to cover the cost of construction. In time the 5,000-pound cast steel bathysphere entered the darkness; Beebe went into the sea with Barton 32 times, reaching a depth of 3,028 feet in the summer of 1934 off Bermuda. Auguste Piccard's stratospheric gondola and the bathysphere hung beside each other on display in Chicago, a popular success in that wretched year of the Great Depression. (This reviewer, then a half-hungry teenage hitchhiker, wandered those fairgrounds for weeks in delight.)

Other men and other vessels have since plumbed the depths. The balloonist Piccard developed his own postwar vessels: deep-sea free balloons with gasoline-filled floats, steel-shot ballast and a cast-steel gondola not unlike the bathysphere. The bathysphere is a frightening craft by today's engineering appraisal. There is total dependence on the umbilical cable, no recourse to emergency buoyancy. Any motion is "even more frightening" to contemplate. The passenger being towed below calls up for evasive action: "'A large coral head is dead ahead' would be the observer's calm remark."

Once the third Piccard submersible, *Trieste II,* refitted under U.S. Navy ownership, reached the ocean bottom in the Marianas Trench in 1960, depth had been mastered. A speculative boom in manned submersibles began. The American aerospace engineers and their opulent firms, confident in their outer-space triumphs and with an eye on Navy resources to come, built a "very large national base for deep-submergence engineering," with plenty of powerful one-of-a-kind vehicles. Then "one by one the first generation of submersibles was retired, scrapped or put into storage. There were few survivors."

The insider's symposium on submersibles edited by Richard A. Geyer devotes half a dozen knowing chapters to an appraisal of the various craft in operation today, manned and unmanned, free or tethered, in worldwide service for 14 nations. Some 150 submersibles of all types are listed, and just what they do in the present era of careful study of costs and benefits is detailed. Their major application is in the burial and inspection of undersea pipelines and cables, primarily in the oil fields of the North Sea, whose depth and cold are often beyond divers' limits. The harvesting off Hawaii of fresh pink and black coral of jewelry grade at a depth of 300 meters was the next most important activity until it was stopped by a fatal accident when *Star II* was mislaunched from its transport. (Two divers trying to free the vessel exceeded safe air-breathing limits in the fall of 1975.) Ten more papers review case histories in geology, ocean waste disposal, archaeology, the study of reefs and fishing grounds, the laying of underwater pipelines, rescue and retrieval and other subjects.

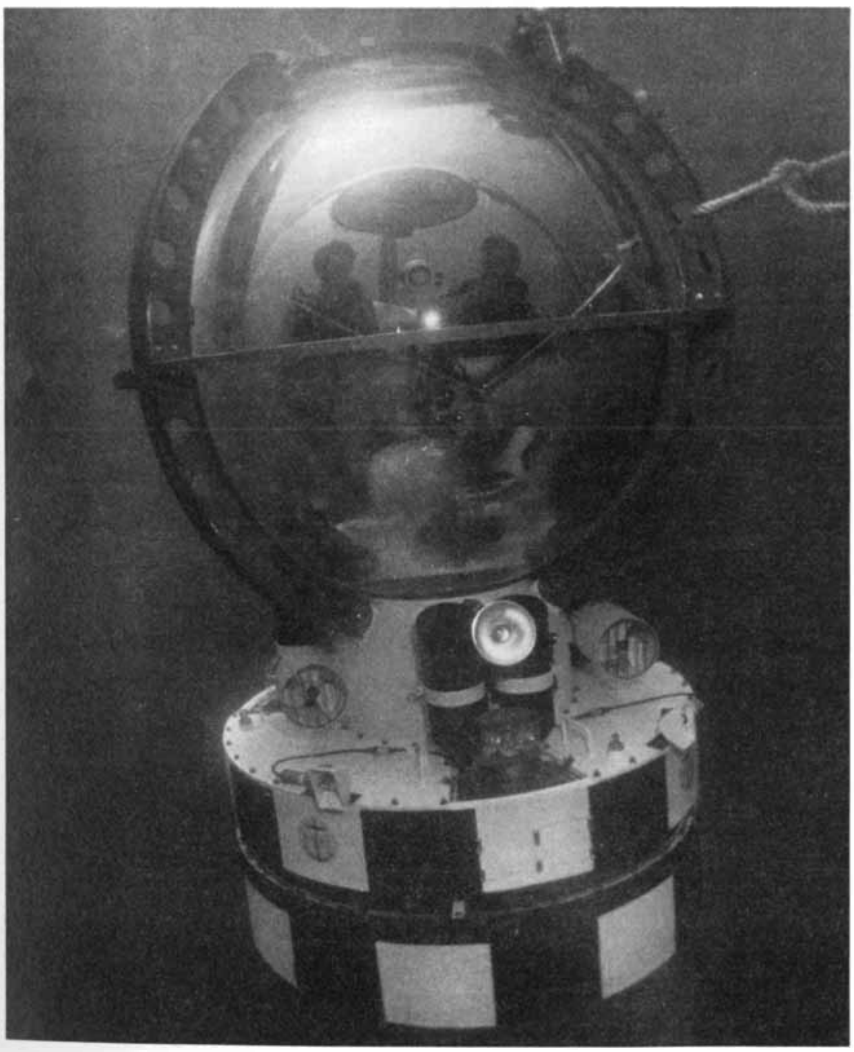
Three of these modern submersibles will be characterized to stand for the lot. Take *Curv III* (cable-controlled under-


You can tell a lot about an individual by what he pours into his glass.
INPORTED
OLD BUSHMILLS
IRISH WHISKEY
A BLEND
100% IRISH WHISKIES
1608
BUSHMILLS
Bushmills.
The world's oldest whiskey.
What individuals have poured into their glass since 1608.
The "Yachtsman" glass created for the Bushmills Collection by Henry Halem
A blend of 100% Irish Whiskies. 86 Proof. Bottled in Ireland. The Jos. Garneau Co., New York, N.Y. © 1974

water recovery vehicle). She (or it) resembles a sled with a cargo of hardware topped by a big flat box. She is no beauty but does handsomely, taking two television cameras and a film camera, a strobe unit and other lights and good sonar down 2,000 meters or more. There is a ton of payload tools, all mobile by air, and an electrohydraulic manipulator. *Curv I* replaced the manned *Alvin* after that brave craft became entangled with a lost thermonuclear bomb for an anxious moment under the seas off Spain in 1966. *Curv III* was flown all the way from San Diego to Cork in 1973 in time to find and attach a line to *Pisces III* and rescue her crew of two, sunk nearly 500 meters down in bad weather.

The Pisces class consists of about a dozen manned submersibles that have been operable since 1969. They are a commercial success and are particularly busy in the North Sea. Carrying keen-eyed crews helped by television, sonar, scoops and drills (a payload of a ton as against a dead weight of 11 tons), they are just what is needed to follow a costly pipeline across the bottom. Or take *RCV-225,* another commercial product from San Diego. Winch it over the side on a kind of dumbwaiter, a launcher held to your ship by a strong armored cable. Once down, the vehicle flies out of its launcher on a light, flexible tether, carrying lights, a low-light-level television camera and some instrumentation. About the size of a medicine ball, the drone carries sharp eyes down to 2,000 meters. Four motors propel it, keeping depth and heading automatically if desired. On shipboard up above are the controls and a careful watch on the television screen.

These and many other craft are discussed for the trade in this book, which is a little repetitious but nonetheless satisfying. How Beebe would have welcomed a trip in *Nemo* or its kin, free-swimming submersibles with a completely transparent sphere enclosing the two-man crew! These vessels offer a true panorama down to the stress limit of the cast acrylic balls, perhaps as much as 900 meters down.



Nemo, *from* **Submersibles and Their Use in Oceanography and Ocean Engineering**

Surprisingly enough the free divers, in the footsteps of Jacques-Yves Cousteau and his wartime partner Émile Gagnan, are not far behind. They are undertaking commercial contracts for the support of oil-recovery operations down to 1,500 feet; demonstrations ashore in a Marseilles test tank have been conducted at 600 meters of pressure. The main technique, described in *Diving Medicine,* is saturation diving. The workers are compressed in a tank, which is lowered to the working depth. When work is finished for the day, they return to their tank, which is hoisted out of the water, and in it they continue to be maintained at the working-depth pressure. Days may be spent on such a schedule. The support tank is a real dwelling and thus is relatively complex. Decompression also takes days, but the dangerous bubbles that can come with decompression do not form—much—because the gas-fluid system stays close to equilibrium.

At depth one cannot safely breathe nitrogen. Above 10 atmospheres of pressure that gas induces nitrogen narcosis, the "rapture of the depths." The effect is biophysical in nature; molecular nitrogen is quite inert in biochemistry. The dissolved nitrogen appears to somehow cause an expansion of the key lipid membranes of the nerve synapses, thereby inducing the intoxication, a partial anesthesia. Helium does not dissolve as much, and it causes no narcosis at moderate pressures; divers have long breathed a helium-oxygen mixture. Increase the pressure, however, and finally even helium induces a related but contrary neurological syndrome, affecting motor function more than mental abilities: tremor and convulsion appear. There is evidence that lipid membranes are simply compressed under helium pressure so that the thinned layers then excite more easily. (In tadpoles such pressure can reverse anesthesia.) Mix nitrogen with helium and their two effects mutually compensate; provide as well necessary oxygen partial pressure and down the diver can safely go. The trimix (helium, nitrogen and oxygen), or perhaps some even more subtle atmosphere, is the gateway to greater depth.

What is the limit? We do not know. Presumably irreversible changes are eventually induced in proteins or other biopolymers. The energy-density increase of simple pressure amounts to about that of one degree Celsius of temperature for every few hundred meters of depth. On this naïve basis one might expect that human beings could indeed dive a long way, although a free dive kilometers down must not be very different from having a high fever.

This readable and even exciting volume is a practical guide for physicians written by a score of experienced diving physiologists and physicians. It is also the best kind of guide, with a background of theory and wide-ranging practical scope, including cases and comment. The jargon is pretty dense, but the general scientific reader should be able to find his way, and divers with any technical bent will be wonderfully served. Where else can we learn of the terrible ciguatera, a specific biointoxication? Ingest some uncertain kind of tropical algae, often by eating a perfectly edible fish that has itself chosen wrong, and you get it. There are no antidotes, and one in about eight cases is fatal. What the philosopher must wonder at is that the sufferer's sense of temperature is reversed: a hot object feels cold and a cold object feels hot! By the way, schools will find this volume a fountain of illustrations and problems in simple but deep physics, chemistry and human biology.

HARVEST OF THE PALM: ECOLOGICAL CHANGE IN EASTERN INDONESIA, by James J. Fox. Harvard University Press ($15). "It is not unusual on Roti to come upon an entire family dining regally from the contents of a couple of lontar [palm leaf] buckets, one containing five or six liters of sweet palm juice, the other brimfull of salty, slimy, slightly sour strands of green seaweed." These people drink more meals than they eat; that palm-sugar solution is their sucrose-rich staple.

The lontar fan-leaf palm (*Borassus sundaicus* Beccari) is a massive single stem two or three feet in diameter that soars straight up to a height of 90 feet, bearing high a crown of great leaves. It blossoms twice a year for a generation; at each blossoming the sweet sap can be made, by the right cut across the bud, to flow copiously for a couple of months. The yield can be astonishing: a good tree provides the skilled tapper with 300 or 400 liters of sweet juice per year containing a measured 15 to 17 percent of sucrose. Half a dozen such trees yield the calories of an acre of wetland rice. In season the juice is drunk directly, but it can be stored by cooking to a dark thick syrup or even to hard sugar.

On this extraordinary base (the trees grow wild almost everywhere on Roti) the rocky, poorly watered island supports a population density that here and there reaches the high values typical of wetland rice regions. Like all well-fed peoples the 75,000 islanders (a small cluster of islands nearby share the culture) are not limited to their staple. They grow some rice and maize and also fish, and a main protein source is the pig. A materials-flow sheet would be a delight (what do the pigs eat besides syrup?), but the book does not report any quantitative nutritional data.

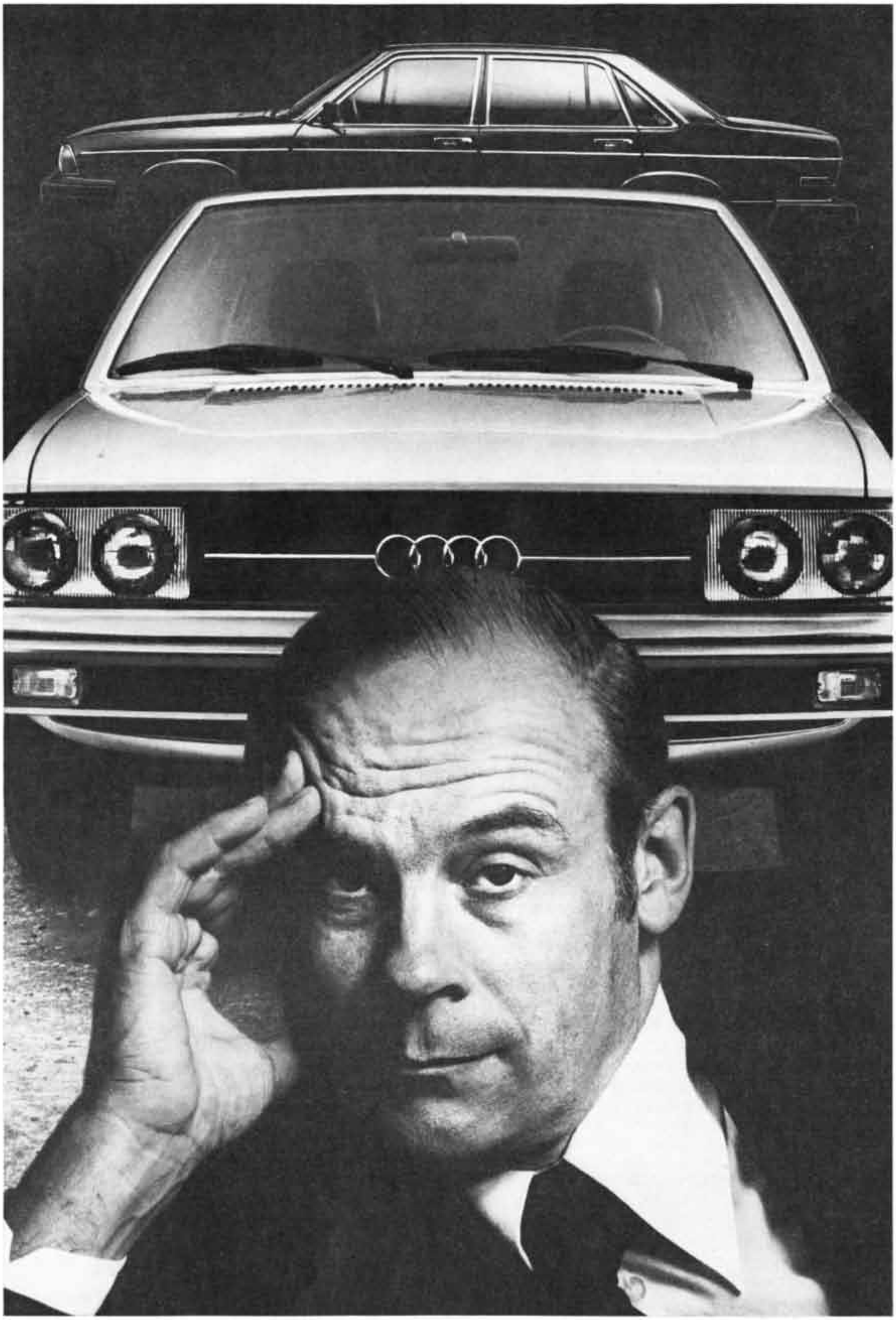
A Rotinese in the crown of a lontar palm, from **Harvest of the Palm**

The lontar has its relatives (often of the same species) in lands all the way from the two shores of South India through Burma, Thailand and Indochina and across Indonesia. Almost everywhere else the trees do not give sweet juice, only a raffish fermented product, soured or alcoholic or strongly flavored with lime or some other substance to inhibit swift fermentation as the juice collects in the heat of the day. On Roti the islanders harvest sweet juice without additives by collecting it twice a day, in the early morning and in the afternoon, and by scrupulously putting it only in clean, fresh receptacles.

It seems that the Rotinese came to their island long before they were a palm-tapping people. Initially they tapped the palm casually to eke out a grain grower's living. Eventually they perfected the elegant process of today. The men tap trees for a living, having given up the hardscrabble swidden grain farming of neighboring—and poorer—islands. There is a poem about a clan founder: "Tapping lontar twice a day / Comes from Dou-Danga. . . ." With Dou-Danga they entered the sweet life.

There can be no doubt that is the kind of life it is. The Rotinese are proud and confident; they are prosperous and are even spreading their complex subsistence culture to other islands in direct competition with more familiar husbandry, some of it introduced by the Dutch. Theirs is no relict culture, either. They have long distilled gin and practiced law; Rotinese are found throughout Indonesian life, from pillars of the establishment to leading Communists in exile. They adopted the Malay language long ago as a literary and unifying tongue, the better to deal with the colonial power. (Malay in a modern form is now the official language of the Republic.) Their schools are Malay and so is their Bible, but these litigious, talkative islanders stoutly maintain their many ministates and dialect differences in the mother tongue.

Fox is at pains to make it clear that this is in no way the automatic work of the generous palm, even accepting the



genius of Dou-Danga. The islanders' geography made them not worth economic invasion, yet they were assaulted for strategic advantage. They learned early that Christianity, education and having neither plantation lands nor much taxable crop all made for a workable accommodation with their rulers. Half of the volume discusses such matters, far from sugar and syrup, and we are promised an entire volume on the political tradition of their state. The author submits plenty of near-parallels to prove his case. The palm offered opportunity, and the Rotinese seized it over a period of centuries. Their neighbors on the island of Savu made a different world of the same strands.

The rajas and a tapper on his way up the tree alike wear the elegant textiles of these islands, yarn-dyed woven figures of world-famous intricacy. One can hope for more volumes, on both the history and the material culture—and the public health—of these people who live by strong, sweet draughts.

ARACHNIDA, by Theodore Savory. Second edition. Academic Press ($21.15). Our own class of animals, the mammals, may number some 15,000 or 20,000 species, counting extinct forms, and its major subdivisions, the orders, span such diversity as bats, marine animals and primates. The arachnids (it was Lamarck who grouped them first; Linnaeus "had given inadequate attention to the invertebrates") are a class too, with three times as many known forms, and the list is growing fast. The orders are indeed diverse. The smallest mite is about a tenth of a millimeter long, and a big scorpion of the Carboniferous measured 36 centimeters. The range is less great between the shrew and the blue whale!

Arachnida, by a distinguished British arachnologist who is a writer of grace and a classical scholar as well, achieves a comprehensive and even encyclopedic quality while remaining brief, engaging and personal. Part of the author's success is that he treats the best-known and most-studied forms with a light touch but adds detail and meaning to his account of the unusual and unfamiliar. Part is simply his gift for the meaningful. The zoological nomenclature is here and there rather dense (like most students of diversity, Savory spends a good deal of space on careful discriminations of form), but the book is both readable and authoritative.

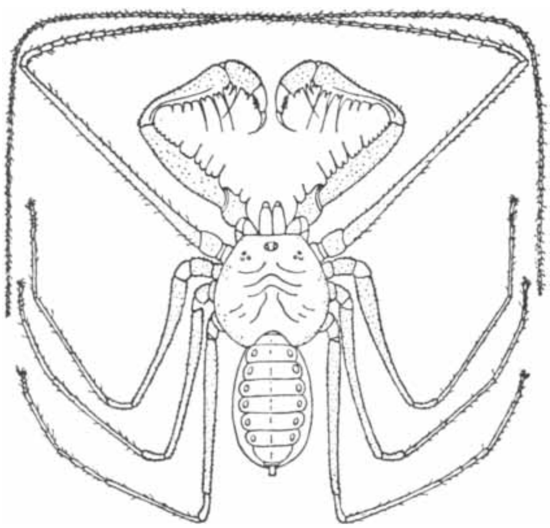
It is also wide-ranging: Savory writes both of spiders and scorpions and of their human students. He includes chapters on the arachnids' economic importance, history, collection and preservation and medical interest and even covers linguistic arachnology. Webs, courtship, arachnids in amber and unreasoning arachnophobia earn enlightening little essays. A hundred pages survey the class as a biological whole: organs, growth and development, habits, dispersal, evolution and taxonomy. Somewhat more space is devoted to a discussion of the 17 orders one by one, including very brief accounts of orders known only from the fossil record.

It appears that the scorpions are the oldest; perhaps they were the first of all animals to venture on dry land, probably from intertidal life. Their remarkable courtship, a promenade à deux, surely memorializes this daring change; the sperm are borne on a complex gelatinous structure attached to the ground. They are thus kept from drying out, and the ritual ensures quick transfer. If you seek scorpions now among the dunes of the California desert, you can hardly improve on the prowess of S. C. Williams. That doughty hunter used an ultraviolet lamp and located 2,000 within four hours by their greenish-yellow fluorescence, noting that "at one time seventeen specimens were seen glowing in the light of one lamp." The fluorescent substance can be detected in solution in museum preservatives, but it has not yet been characterized chemically.

Take the order Ricinulei, eyeless tropical forms looking rather like smooth-coated spiders. From the year 1838 to the 1930's the order was well known but excessively rare; "not only every species but actually every specimen captured could readily be traced." In 1933 a few hundred specimens were found in the Cameroons; between 1967 and 1973 the number of species known has been doubled to about 40 and the number of individuals taken has risen into the thousands. One Mexican cave yielded 1,035 specimens of a single species. They were recognized as having existed in the Carboniferous—the time of arachnid explosion—before any living form was found, and they remained incredibly rare for more than a century, features that "make them the most romantic order of the Arachnida."

Here are the formidable wind scorpions and the acetic-acid-secreting vinegarroons, the bizarre harvestmen ("daddy longlegs"), the ticks and the mites and all the rest. The study of ticks and mites, the order Acari, has the greatest direct economic importance; the animals are ubiquitous, intensely specialized and generally tiny and well hidden. Acarology is booming, with wild talk of half a million species.

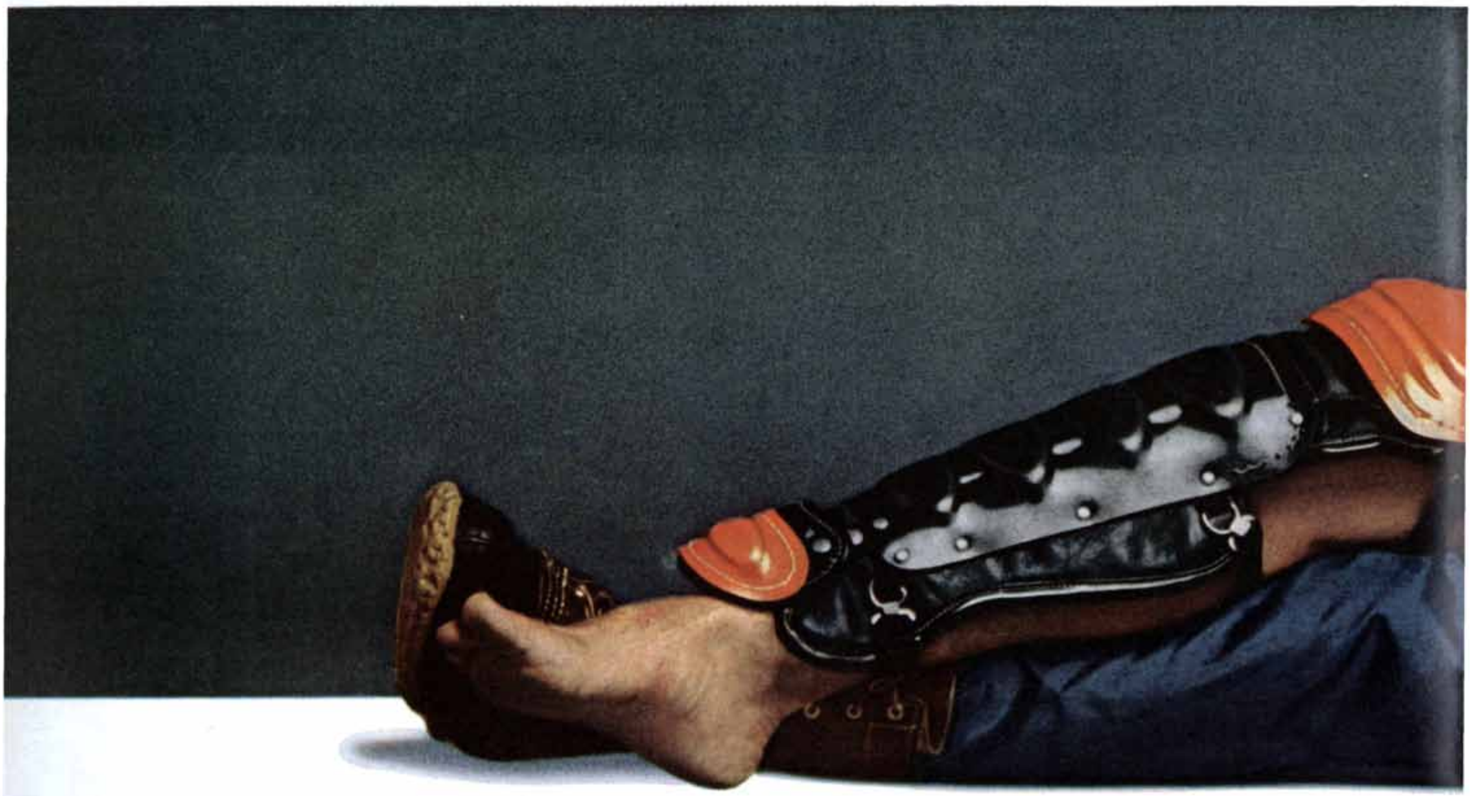
The most widespread affliction in humans "caused" by a spider was tarantism, common first in Italy and then in Spain from the 15th to the 19th century. It was a real disorder, its only cure was dancing and it reappeared in a patient summer after summer. We can exculpate the bite of the tarantula. Those spiders still abound around Taranto and people still dance the graceful tarantella, but the disorder has gone. An anonymous editor of a work in 1795 published the explanation Savory favors: "The women durst no longer act a frantic part in the character of Bacchantes. Unwilling to give up so darling an amusement, they devised other pretenses. Accident may have led them to the discovery of the tarantula; and upon the strength of its poison, the Puglian dames still enjoy their old dance, though time has effaced



Stegophrynus dammermani, *a species of spider of the order Amblypygi, from* **Arachnida**


Seagram's
Crown Royal
Fine De Luxe
Blended Canadian Whisky
Joseph E. Seagram & Sons Limited
WATERLOO ONTARIO CANADA
80 PROOF 750 ML (25.4 FL. OZ.)
Crown Royal
by
Seagram

DISCOVERED BY



A plastic foam that helps protect you 17 different ways

A Uniroyal plastic foam called Ensolite® has changed the shape of the old-fashioned life jacket. But that's only part of what Ensolite does. It helps protect people in places as different as football fields and mountaintops.

The first time you notice Ensolite may be the next time you buy a life jacket.

Life preservers used to be stiff and uncomfortable—a lot of people wouldn't wear them for love or money. Obviously, life jackets can't save lives when they're so uncomfortable that they're left in the bottom of the boat.

According to the U.S. Coast Guard, life preservers that are more comfortable are more likely to be *worn* than stiff, bulky types.

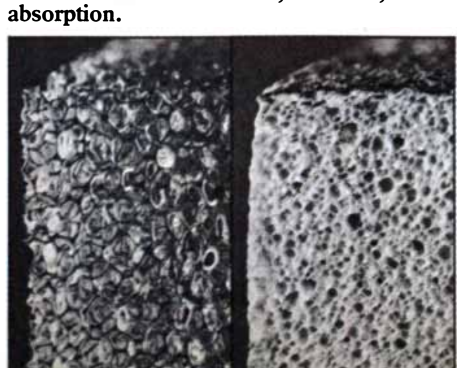
Buoyant Ensolite life preservers are compact, lightweight, soft, and drape easily around your body. (In fact, most manufacturers of quality life jackets now use Ensolite.)

But these are only some of Ensolite's remarkable properties.

Used in hundreds of products

Ensolite is also a shock-absorbent padding, used in everything from football

Open cell foam vs. Ensolite. Left: cross section of ordinary urethane foam. Open cells absorb water, readily collapse upon impact. Right: cross section of Ensolite. Closed cells with solid walls shut out air and water. Ideal for insulation, flotation, shock absorption.

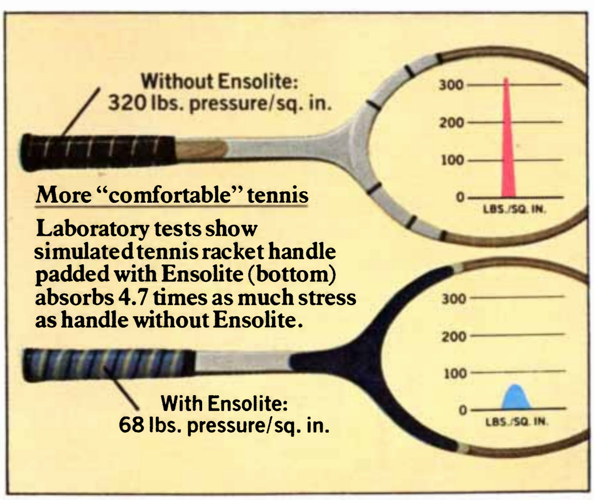


UNIROYAL:

Some of the ways shock-absorbing, buoyant, insulating Ensolite can protect you:
iners for baseball and football helmets, wrestlers' masks, life jackets, elbow and hip pads, shin guards, karate gloves, insulated boots, flotation suits.

helmets to operating-table mats. It cushions the falls of wrestlers, ballet dancers, and even U.S. astronauts—who were protected by Ensolite mats while practicing the maneuvers they'd make in weightless space.

Ensolite is also used to insulate cars, airplanes, and other vehicles against heat and cold. Mountain climbers have even *slept* on ½″ sheets of Ensolite in temperatures plunging to −30°F.



More "comfortable" tennis

Laboratory tests show simulated tennis racket handle padded with Ensolite (bottom) absorbs 4.7 times as much stress as handle without Ensolite.

But helping protect people isn't all that Ensolite does.

It cuts noise and vibration in air-conditioning units. Keeps oil-containment booms afloat at sea. Keeps condensation from forming in the missile tubes and hull of Polaris submarines. (It also reduced the weight of the sub by 22 tons.)

Forming plastic with gas

Ensolite evolved out of Uniroyal's expertise in commercial blowing agents—the chemicals used to turn plastic and rubber substances into spongy foams.

By working with highly sophisticated Uniroyal blowing agents, our chemists created a formidable range of different Ensolite foams—for everything from featherweight insulation to the bumpers used in docking supertankers.

Over 1,400 Uniroyal discoveries

Ensolite is only one of Uniroyal's discoveries. We've been granted more than 1,400 patents in the past 17 years. We have over 1,500 scientists, engineers, and technicians in many fields working to develop new Uniroyal products.

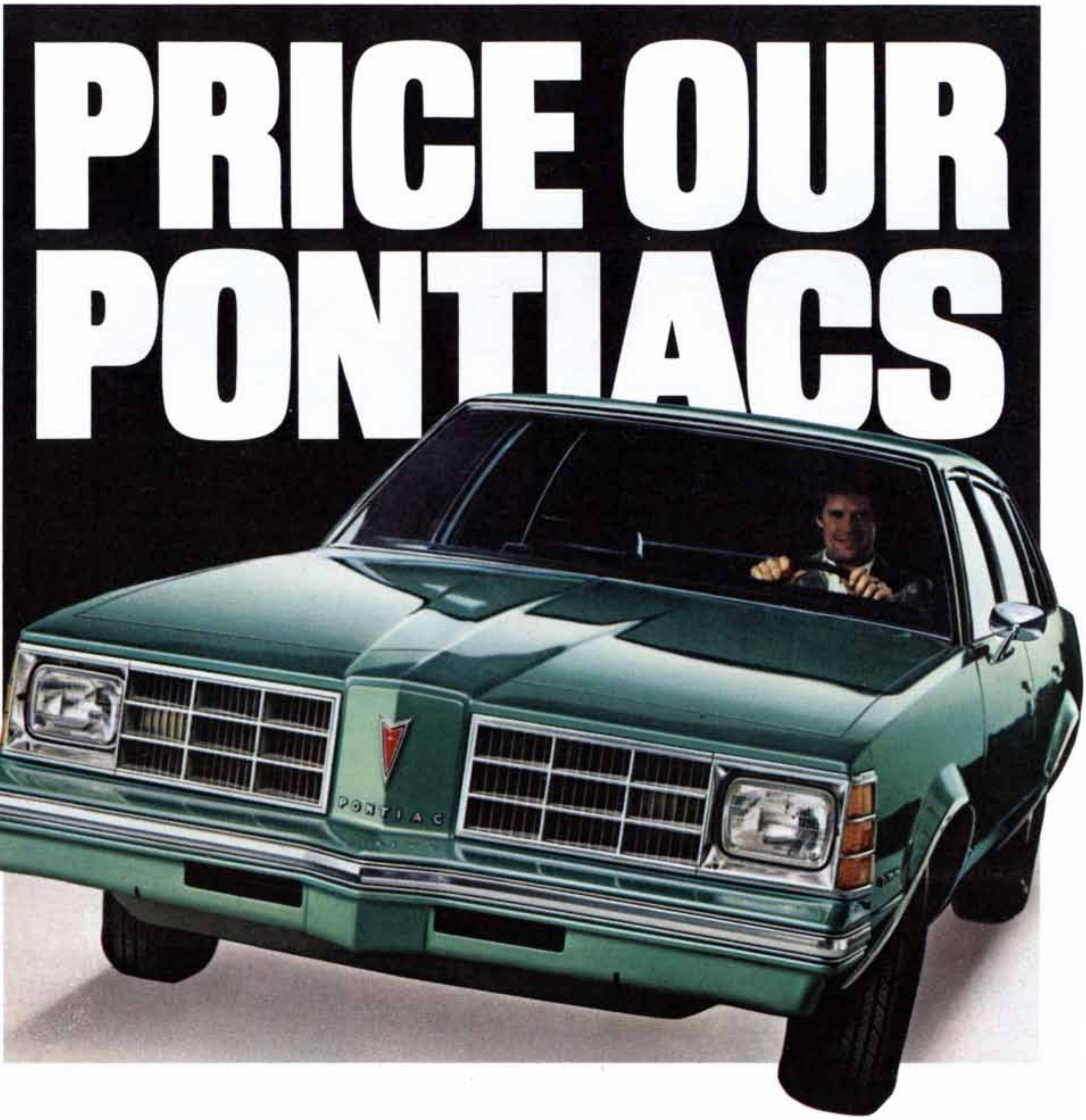
Uniroyal discoveries have paid off in substantial sales growth. We're a company with $1.6 billion in assets these days. But we like to think our biggest asset isn't even on the balance sheet.

Our urge to discover.

For more information on the many applications of Uniroyal Ensolite as an insulating, shock-absorbing, and flotation material, write to John Kroske, Uniroyal, Inc., 312 North Hill Street, Mishawaka, Indiana 46544. Or telephone (219) 255-2181.

UNIROYAL

chemical, rubber and plastic products worldwide



If You're Serious About Getting A Lot For Your Money.

$4481 is LeMans' manufacturer's suggested retail price including dealer preparation. Taxes, title, license, destination charges additional. Priced higher in California. And here's just a sample of all you get: More head, leg and luggage room than last year • New door and body seals to help seal out noise • Rich full-width cloth seats • Thick one-piece, cut-pile carpeting • Fully upholstered door panels • Deluxe cushion steering wheel • Concealed windshield wipers • New color-accented, soft-bumper design • New crosshatch grille • Radial Tuned Suspension and much, much more. Price LeMans. Price all our Pontiacs at your Pontiac dealer.

GM
MARK OF EXCELLENCE

1978 Pontiac's best year yet!

PONTIAC LEMANS $4481.

the memory of its ancient name and institution."

There is only one photograph. The many technical line drawings are curiously denatured, often showing mere fragments of the beasts. Although visually ordinary, this is a remarkable text, demanding to be supplemented by a museum collection or at least by the Little Golden Guide with its fine paintings.

COPPER: ITS GEOLOGY AND ECONOMICS, by Robert Bowen and Ananda Gunatilaka. John Wiley & Sons ($49.50). The red metal has been much sought after by our kind for a long time. Around Lake Superior there are rich veins worked 5,000 years ago, and "extensive ancient workings" are known near Ndola in the copper belt of Zambia. The biggest systematic producer documented in the early world, in about the time of Moses, was Egypt; world production has grown from that base by a factor of about 200,000. As a trace element copper is not rare. The upper kilometer of the earth's crust holds a mean content of some 50 parts of copper per million that has somehow migrated to the crust from the deep interior. There is copper enough for our industry for geological eras, but its recovery is economically daunting.

What makes a copper mine is concentration. The ancient miners looked for showy minerals with a copper content of 15 percent. The grade has steadily declined; it was below 8 percent in Europe by the time of the Renaissance, and today most copper is won from low-grade ores, the U.S. average grade being about .65 percent. Huge open pits, such as the biggest U.S. mine at Bingham in Utah and the world's largest producer at Chuquicamata in the red Chilean desert, are exploited with large-scale rock-moving techniques.

The pit at Bingham was once a narrow canyon. Since the years just before World War I it has been enlarged into an oval hole three kilometers long, two and a half kilometers wide and more than three-quarters of a kilometer deep. Every working day the bulldozers push some 200,000 tons of rock into the huge ore trucks. Such dilute ore bodies, each containing many cubic kilometers of rock, lie at various depths along a belt in western America from Alaska to Antarctica. Their copper is spread (mostly as grains of the double sulfide of iron and copper) through the mass of fine-grained intrusive rocks that supply the name: porphyry copper ores.

Going deeper for the porphyries, we shall win copper enough for a good many decades at the present rate of use, say 10 million tons per year. In 1976 a porphyry ore body in western Panama was announced that is thought to be the largest reserve in the world, reportedly three billion tons at .6 percent copper content. After the porphyries there will be the nodules. The ocean floor at great depth is covered with small dark nodules of layered iron and manganese oxide. Their total volume is immense, and already several international consortiums are planning to raise them by the millions of tons in the 1980's. The hope is for a major addition to the cobalt supply and an important amount of manganese, but the nodules can supply only 1 percent or so of world copper by 1985 (if we accept a United Nations forecast); their copper content can be recovered on a large scale only in the next century.

This volume is an up-to-date and expert review by two economic geologists from the School of Mines of the University of Zambia at Lusaka. About three-quarters of the book, with a few color photomicrographs and many line drawings, is a rather technical summary of the nature and origins of the principal copper deposits of the world; the rest is a well-informed, if openly opinionated, nontechnical survey of the economics, pricing, uses and control of copper mining and investment. The authors are keen partisans of the market system but offer a fair-minded country-by-country review of copper nationalization in recent times.

The geology is detailed, uncertain and complex for a general reader. A few points stand out tentatively among the controversies. It is fairly clear from the location and age of the American porphyry ores that they were formed as the Pacific Ocean floor underthrust the American plate. Four or five kilometers deep under the flat cone of some new volcano the hot brines brought copper in solution to fill the shattered fabric of the fine-textured rock. (There is even a drawing of the entire model system.) A 10 percent leaching of the normal copper content out of 100 cubic kilometers of the rock body involved would account for the copper concentration of a typical deposit. The geochemistry is not yet evident, but it is discussed at length. There is little doubt that we are slowly learning how to seek the strange and intricate consequences of the drifting continents.

Copper mines are running on short time today. The price is too low: the reasons are big stocks, plenty of scrap and industrial recession. Aluminum has won the long-distance power lines. Yet the London Metal Exchange, which fixes most of the world prices even though its dealers handle only small amounts of copper, in 1973 mixed oil with copper; the price tripled at the time of the oil embargo. By late 1974 the curve had fallen almost to the previous floor, and it has recovered only weakly. In the strange chemistry of international trade copper is more volatile than petroleum.

Established 1845 SCIENTIFIC AMERICAN March 1978 Volume 238 Number 3

World Oil Production

There is only a finite amount of oil and there are limits to the rate at which it can be recovered. Sometime before the year 2000 the decreasing supply of it will fail to meet the increasing demand

by Andrew R. Flower

The supply of oil will fail to meet increasing demand before the year 2000. As oil production inevitably levels off and then falls, alternative fuels will have to meet the demand for energy, which will continue to grow in the face of even vigorous attempts at conservation. The oil-importing countries have perhaps as few as five years or perhaps as many as 20 in which to accomplish a transition from dependence on oil to greater reliance on other fossil fuels, on nuclear energy and eventually on renewable energy sources. Because large investments and long lead times are required for developing new energy resources the effort must begin immediately.

These are among the major conclusions of a two-year international study, the Workshop on Alternative Energy Strategies. Some 70 people recruited from business and industry, government and the universities in 15 major non-Communist oil-importing countries came together under the direction of Carroll L. Wilson of the Massachusetts Institute of Technology to study world energy supply and demand, to identify potential problems in satisfying energy requirements and to consider strategies that might solve the problems. (The "world" we considered was the world outside the Communist areas. The U.S.S.R. and China are major producers and consumers of energy, but statistics on their reserves, production, consumption and trade are rough estimates at best and their trade with non-Communist countries does not loom large.) The workshop participants came from the U.S., Canada, Mexico, Venezuela, the United Kingdom, France, West Germany, the Netherlands, Denmark, Sweden, Norway, Finland, Italy, Iran and Japan.

The workshop developed projections of energy supply and demand under various assumptions, notably concerning economic growth and the price of energy. Two basic scenarios were formulated, one assuming a high economic growth rate and high energy prices and the other a low growth rate and constant prices. The high-growth, high-price scenario assumed that economic growth in the non-Communist world would average 5.2 percent per year between 1976 and 1985 and 4 percent between 1985 and 2000, and that energy prices would remain constant (in real terms) until 1985 and increase 50 percent by 2000. The low-growth, constant-price scenario assumed that economic growth would average 3.4 percent per year until 1985 and then 2.8 percent until 2000, and that real energy prices would remain constant during the entire period.

Within that framework energy-demand forecasts were prepared for the various regions for each fuel and were aggregated to project the total energy demand in 1985 and 2000. In the case of oil, which today meets more than half of the energy needs of industrial societies, the projections suggest that energy-conservation efforts will have some effect, so that the growth in demand will be less than the 6 percent average annual rate between 1960 and 1972. Given high economic growth and high oil prices, the demand for oil was postulated to grow at the rate of 3.6 percent per year until 1985 and at 2.6 percent per year between 1985 and 2000. With low growth and constant prices the projected rates of growth in oil demand were 2.5 percent and then 1.8 percent. These rates established two demand curves for oil. The next task was to consider how long the supply of oil is likely to satisfy those demands. (Although the workshop was concerned primarily with energy prospects up to 2000, the central importance of oil prompted us to consider the prospects for oil reserves and production until the year 2025; projections beyond 2000 are necessarily highly speculative.) In this article I shall tell how the workshop projected estimates of future oil supply and reached its conclusions on the relation of supply and demand.

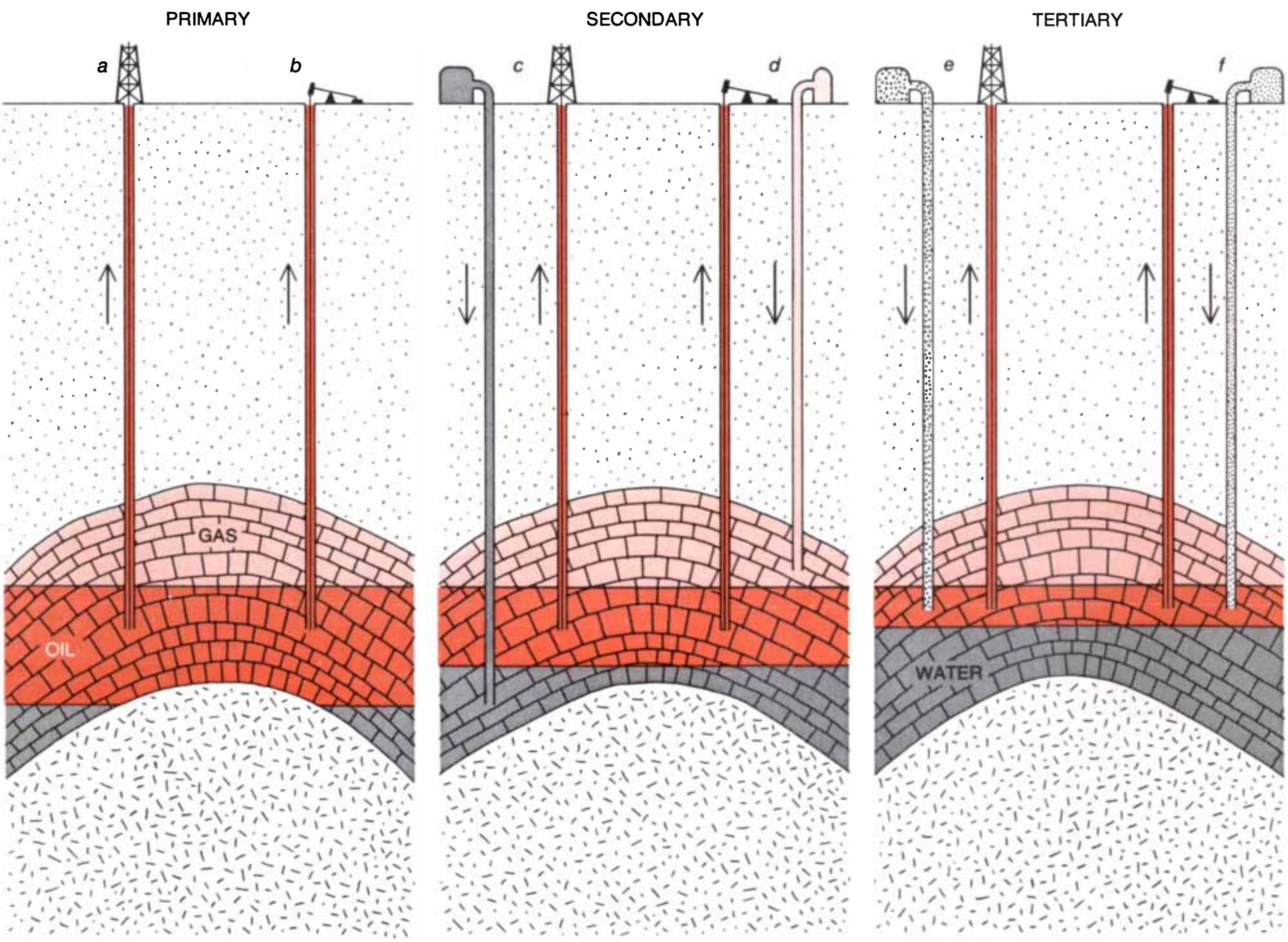
It is a popular misconception that oil is found in vast underground pools that need only to be pumped dry. Actually oil is found in small spaces in porous rock, somewhat like water in a sponge. It seeps slowly through the porous material until it is trapped by impervious rock, often as a layer between water and gas. The oil is usually under pressure, and a well releases that pressure; the oil flows into the well, where it either rises to the surface (if the pressure in the field is sufficient) or is pumped out. Oil recovery under natural pressure is considered "primary" production, and in the U.S. today such production recovers on the average about 25 percent of the oil in a reservoir. "Secondary" recovery is achieved, in fields where it is feasible, by pumping water or gas into the reservoir to produce more pressure on the oil. "Tertiary" recovery, not yet widely applied, can sometimes be achieved if the viscosity of the oil is lowered so that it flows more easily, either by heating the oil (by injecting steam, for example) or by injecting chemicals into the reservoir. Secondary and tertiary methods have increased recovery in the U.S. from about 25 percent of the oil in place in the 1940's to about 32 percent now. Most of the world's oil is produced from conventional fields by these methods and by the recovery of natural-gas liquids as a by-product of natural gas. The workshop considered only those sources, excluding oil from shale, tar sands and the liquefaction of coal.

The major factors that determine potential oil supply are the amount of proven reserves (oil recoverable from known reserves at current prices and with current technology) and the rate at which those reserves can be exploited. Each field has a potential production rate that depends on the size of the field, its geology and its installed facilities. In general it is impossible to produce more than 10 percent of the recoverable reserves in a field in any one year without reducing the amount of oil that can ultimately be recovered. In some fields the rate may be higher and in others it may be lower, but overall a reserves-to-production (R/P) ratio of 10 to 1 is probably the minimum ratio that is feasible for the world's oil reserves. To apply a 10-to-1 R/P ratio globally, however, would be to imply that all known oil fields can simultaneously produce at the maximum rate. The fact is that some fields will always be under development and will not yet be producing oil even though they contribute to the proven-reserves total. A more justifiable estimate of the maximum rate of production from worldwide proven reserves is provided by an R/P ratio of 15 to 1.

What one needs in order to project a basic oil-supply curve is therefore a yearly figure for the remaining proven reserves. That is obtained by adding to the proven reserves of a base year the gross additions to reserves and then subtracting the amount of oil that has been withdrawn from reserves: the cumulative actual production. The maximum potential production from those reserves is determined by dividing the reserves figure by 15. As long as oil demand is less than potential production, actual production by definition equals demand; when demand exceeds potential production, demand is necessarily limited to potential production, which then becomes actual production. The important question is: At what point does an R/P ratio of 15 bring potential production below projected demand? To establish that date we need (having estimated the demand) to determine the proven reserves in a base year and the probable annual rate of gross additions to reserves.

Estimates of proven recoverable reserves are uncertain, and they change from year to year. The workshop analysis was based on figures published at the end of 1975 by *Oil and Gas Journal*, compiled from estimates made by governments and oil companies. Remaining proven reserves were then put at 555 billion barrels, some 80 percent of it in the member nations of the Organization of Petroleum Exporting Countries (OPEC), mostly in the Middle East. Production to date, on the other hand, has been primarily in North America. In all some 846 billion barrels had been discovered in the non-Communist world by 1975, about a third of which had been consumed.

One guide to the rate of discovery in the future is the rate at which proven reserves have been increased in the past. There are two ways to evaluate past discoveries. One can simply compare consecutive year-end published figures or one can backdate the changing and generally increasing estimates of the oil in each field to the year in which the field was discovered. For example, a field discovered in 1960 might initially be estimated to contain a billion barrels of recoverable reserves. Five years later information obtained as the field is devel-



OIL IS GENERALLY FOUND under pressure in porous rock, often in a geological structure such as the one shown here. Drilling a well into the oil-bearing stratum releases the pressure, and in primary production the oil flows to the surface (*a*) or is pumped out (*b*). When the natural pressure is too low to bring the oil to the well, secondary recovery may be achieved by pumping water (*c*) or gas (*d*) into the field to increase the pressure. In tertiary recovery the oil's viscosity is lowered by injecting steam, which heats it (*e*), or by injecting a chemical (*f*).

oped or as recovery techniques improve might increase the estimate to 1.2 billion barrels. Comparing year-end estimates attributes the extra .2 billion to 1965; the backdating method attributes the .2 billion to the 1960 discovery.

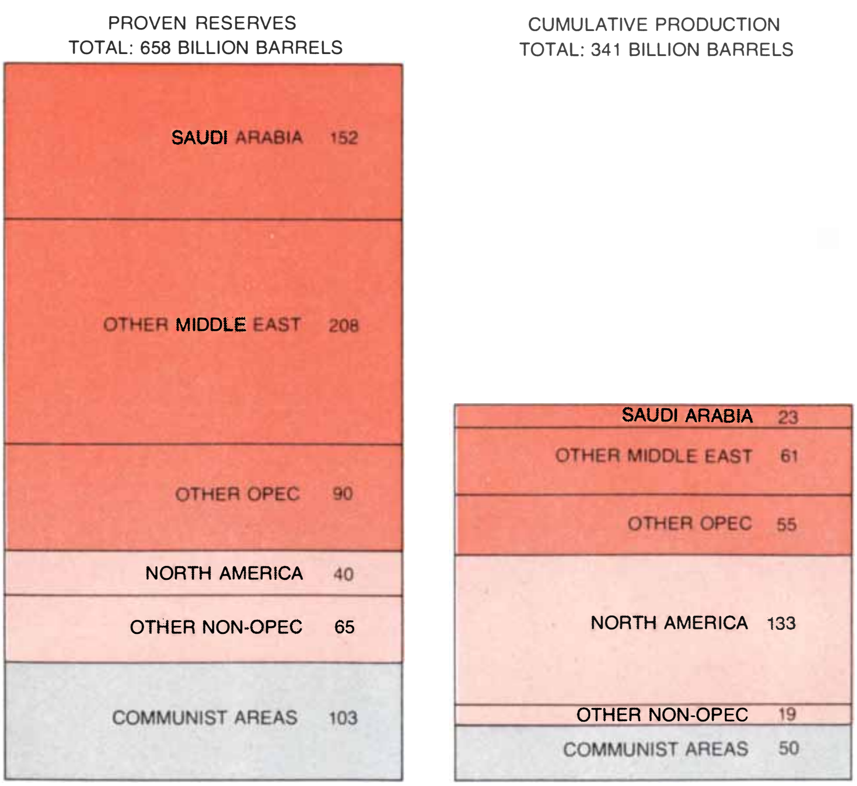
The year-to-year method emphasizes current changes in the rate of additions to reserves—new estimates for established fields as well as the impact of newly discovered fields [*see upper illustration on opposite page*]. It shows the five-year average (which smoothes out year-to-year fluctuations) rising from about 22 billion barrels per year between 1950 and 1965 to more than 50 billion between 1965 and 1970, when the magnitude of Middle East reserves in fields discovered before 1950 was being recognized. Since 1970 the yearly additions to reserves have been only about 25 billion per year. The backdating method, on the other hand, emphasizes that much of the late-1960's increase was in fields discovered before 1950 [*see lower illustration on opposite page*]. It flattens out the 1950-to-1970 rate to an average of about 18 billion barrels per year, with a drop to 15 billion after 1970.

In making estimates for the future it is necessary to consider two different components of gross additions to reserves: first, reserves in genuinely new discoveries and, second, additions to reserves in fields discovered before 1975, either through reassessment of what they contain or through enhanced ability to recover the oil.

The rate at which genuinely new reserves were added in the past would have been much lower without the discovery of the massive reserves in the Middle East, where about 60 percent of all the non-Communist world's known reserves have been found in a region that measures only some 800 by 500 miles. Might such a prolific oil-bearing region be found again? It is not very likely. Many of the remaining possible areas of the world have already been evaluated by seismic testing or exploratory wells and no evidence of a new Middle East has come to light. New discoveries probably will continue to be made in the Middle East itself, but there the rate at which new reserves are found has already fallen off.

In the Middle East and elsewhere it is in areas that are identified as being most likely to contain large accumulations of oil that exploratory wells will first be drilled, and the discovery rate will decline further as the search moves into less promising areas. At the same time the incentive for exploring will be reduced as the likelihood of finding large fields declines. Although some 30,000 oil fields have been discovered, about 75 percent of the remaining oil lies in a few very large fields, each of which holds more than 500 million barrels of recoverable reserves. Only 240 such fields have been discovered in the past 100 years; there cannot be many more. Yet in hostile environments, such as the North Sea, a field with 500 million recoverable barrels is about the smallest field that can justify the high cost of development. All in all it seems unlikely that genuinely new discoveries will continue to be made at even the rate of 15 billion barrels per year that has been achieved over the past five years.

What about additions to reserves in fields discovered before 1975? The estimation of recoverable reserves has become much more accurate than it was in the 1950's, so that the upward corrections of the past 25 years are not likely to be repeated on anything like the same scale. A more probable source of additions to reserves is the improvement of recovery techniques. Today the average world recovery rate is probably about 30 percent. Each 1 percent increase in the rate will therefore increase recoverable reserves by a thirtieth. Applying such a factor to the proven reserves of 555 billion barrels and to half of the cumulative production of 291 billion barrels (as an indication of reserves not previously considered recoverable) works out to an addition of about 25 billion barrels for each 1 percent increase in the recovery rate. John D. Moody, a petroleum consultant, and Robert E. Geiger of the Mobil Oil Corporation have suggested that the worldwide recovery rate will eventually average 40 percent, which would mean an addition of 250 billion barrels to reserves in currently known fields.



WORLD'S REMAINING OIL RESERVES as of the end of 1975 were estimated (*left*) by *Oil and Gas Journal.* Of the total of 658 billion barrels, 555 billion were in the world outside the Communist areas. Note that whereas most of the remaining recoverable reserves were in the Middle East, most of the production of oil up to 1975 (*right*) had been in North America.

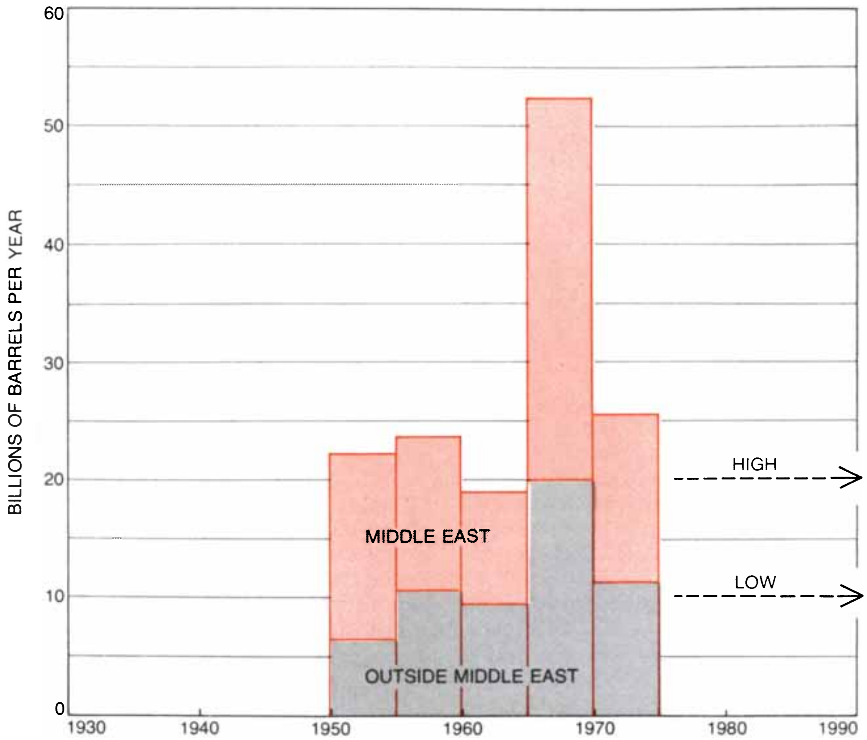
Taking all these factors into account, the workshop based its projections on two different gross-addition rates for the period from 1975 to 2000: 20 billion and 10 billion barrels per year. The higher rate assumes that new discoveries will decline gradually and that enhanced recovery will make a significant contribution, accounting for at least half of the gross additions by 2000. The lower rate assumes a more rapid decrease in the rate of new discoveries and a much smaller yield from enhanced recovery. We assumed that between 2000 and 2025 gross additions would decline gradually to four billion barrels per year (in the high-additions case) or three billion barrels (in the low-additions case) as total discoveries approach the inevitable ceiling: the total ultimately recoverable reserves, which we took to be about 1.6 trillion barrels in the non-Communist areas.

We were now in a position to develop oil-supply profiles for the two demand scenarios I outlined above, one assuming a high rate of economic growth and high energy prices and the other assuming a low growth rate and constant prices.

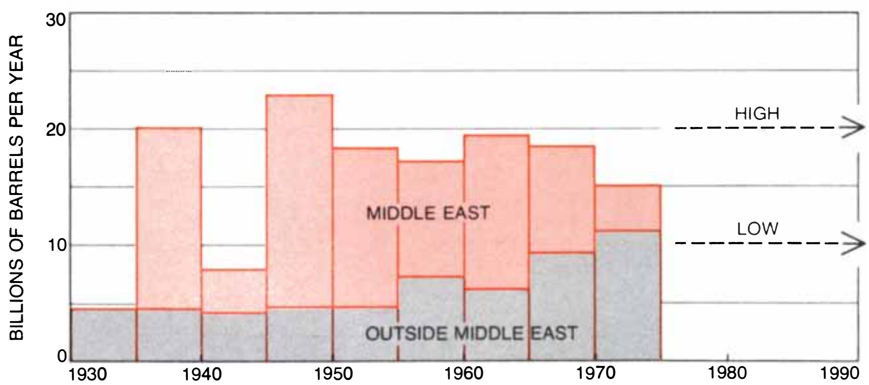
Given high growth and high prices—and also the higher rate of additions to reserves, which is more likely if oil prices are high—production first fails to meet demand in 1997. That is, applying the *R/P* ratio of 15 to 1 to the recoverable-reserves figure for 1997 yields a potential production in that year of 86 million barrels per day, whereas demand in that year is projected as being 87 million barrels per day. Thereafter, as reserves continue to decrease, the 15-to-1 *R/P* ratio reduces production to 80 million barrels per day in 2000 and about 30 million in 2025 [*see top illustration on page 47*]. If in this scenario we substitute the lower rate of additions to reserves (which is less likely but certainly possible), production ceases to meet demand in 1990, when it peaks at only 72 million barrels per day. Thereafter it declines to 50 million barrels per day in 2000 and about 20 million in 2025.

Given low growth, constant prices and the low rate of additions to reserves, oil production fails to meet demand in 1994, when production peaks at 66 million barrels per day compared with a projected demand in that year of 66.5 million barrels. Production then falls to 53 million and about 20 million barrels per day in 2000 and 2025 [*see bottom illustration on page 47*]. If, in the face of low growth and constant prices, additions to reserves proceed at the higher rate, production could meet demand until 2004, when production would peak at 81 million barrels per day.

The trouble with the oil-supply profiles I have just described is that they make a somewhat unrealistic assumption: that oil-producing countries will continue to increase their production to keep pace with demand, limited only by an *R/P* ratio of 15 to 1. In the real world some oil-producing countries outside OPEC may want to reduce their dependence on imported oil and may therefore produce oil at a rate higher than the rate set by a 15-to-1 ratio. More important, the countries that are net oil exporters (in particular the members of OPEC) may restrict production well before the 15-to-1 ratio is reached in an attempt to conserve their reserves. It is therefore necessary to modify the analysis: to look at production in the non-OPEC and the OPEC countries separately.



RATE OF OIL DISCOVERY in the past can be represented as is shown here: as five-year averages of the annual increase in proven reserves published each year by *Oil and Gas Journal*. Here both new discoveries and reassessments of reserves are credited to the year in which they were made. This method calls attention to the period of the late 1960's, when a major increase in known reserves resulted mainly from new estimates of vast reserves in Middle Eastern fields.



DISCOVERY RATE LOOKS DIFFERENT when only new discoveries are credited to the current year, with any revised estimates of the oil in known fields being backdated instead to the year in which each field was originally discovered. Backdating tends to smooth out the five-year fluctuations. In both this chart and the one at the top of the page the broken lines at the right indicate high and low assumptions concerning the future rate of additions to oil reserves.

To consider non-OPEC and OPEC production separately requires an estimate of how future additions to reserves will be divided between the two regions. There are no well-established geological estimates. We assumed that 50 percent of the additions will be in the non-OPEC countries and 50 percent in OPEC, which is to say 10 billion and five billion barrels per year for each region on the high-additions assumption and the low-additions assumption respectively. Much of the non-OPEC region is still unexplored and should yield a large proportion of the new discoveries; enhanced recovery will be more important within OPEC, where 80 percent of the known proven reserves are situated.

Each national team participating in the workshop estimated its country's oil production for the period from 1975 to 2000 and similar estimates were developed for the other non-OPEC oil-producing countries. For the period from 2000 to 2025 production in the non-OPEC countries was estimated by the methods outlined above, with the further assumptions that the *R/P* ratio would tend to decline toward 10 to 1 and that the rate of additions to reserves would also decline [*see upper illustration on page 48*].

Calculating future OPEC production is somewhat more complicated, because different countries have very different producing capabilities, economic interests and policies. What is clear is that to achieve the most optimistic oil-production profile would require OPEC production to reach more than 50 million barrels per day before 1995, which may not be feasible for technological reasons or for reasons of national policy. The currently installed "usable capacity" of the OPEC countries is about 37 million barrels per day, and in many countries the expansion of equipment and infrastructure for producing and transporting oil is unlikely to take place on the scale that would be required. The most telling influence on OPEC production, however, will probably be the insistence of particular countries on extending the life of their reserves by restricting production. Some OPEC governments have already suggested specific limits, which would hold OPEC production to a level some 4.8 million barrels per day below OPEC's usable capacity. Such restrictions are most likely to be imposed by countries that are "low absorbers" of oil revenues—countries, such as Saudi Arabia, Kuwait and Libya, that are earning more oil revenue than they can currently apply to domestic needs. On the other hand, Venezuela and Ecuador have already set a ceiling on production to prolong the flow of oil revenue, even though they are "high absorbers" of such income.

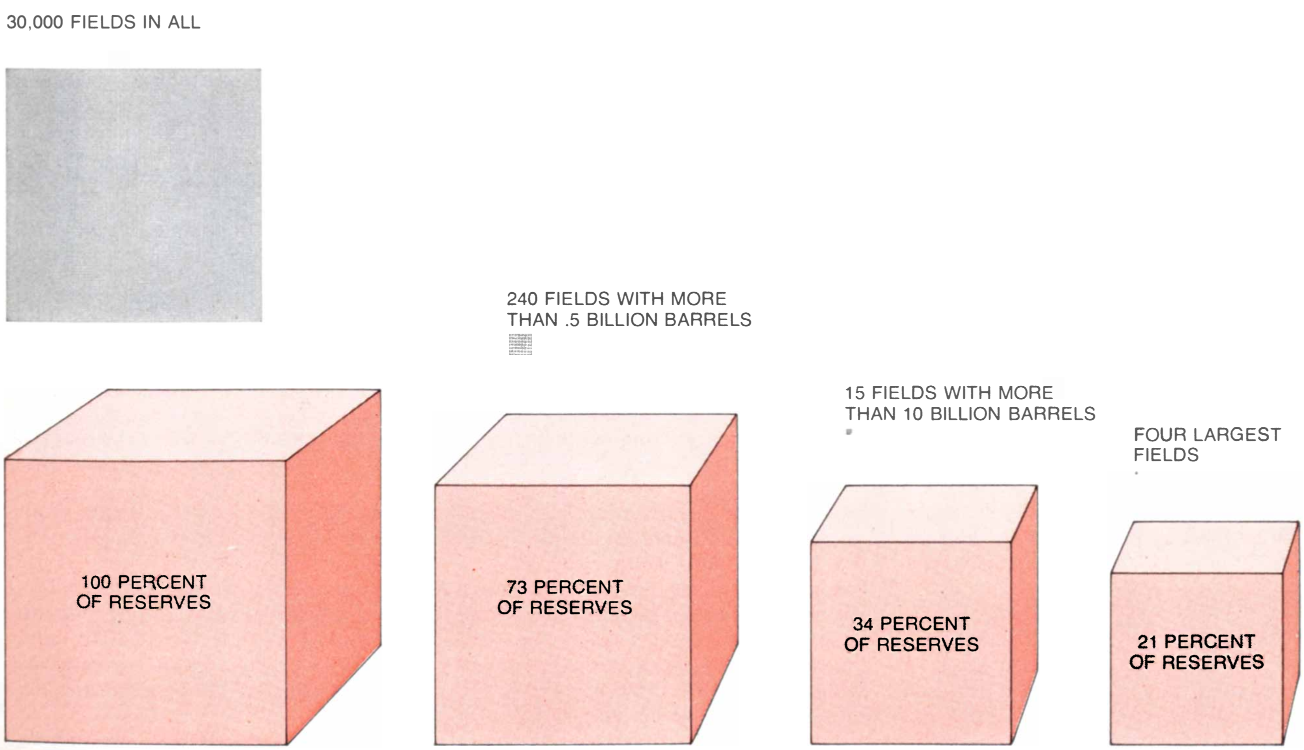
The government-imposed restrictions could have a considerable impact on potential production rates. For example, the OPEC countries of the Arabian Peninsula alone (Saudi Arabia, Kuwait, Qatar and the United Arab Emirates) have 58 percent of OPEC proven reserves. If an *R/P* ratio of 15 to 1 were the only constraint, they could today produce oil at the rate of 17.4 billion barrels per year, or 47 million barrels per day, which is about equal to the total non-Communist output. They already have in place a usable capacity of 16.8 million barrels per day. Yet in 1975 they actually produced only 11.3 million barrels per day. And if the restrictions they have suggested might be imposed are in fact imposed, then their production would be held at 12.8 million barrels per day in the future.

Taking account of such constraints as these, we made three different assumptions about ceilings on Arabian Peninsula production, ceilings whose effect would be felt before the 15-to-1 *R/P* ratio is reached. If government limits are set at or near present production levels, total Arabian Peninsula output would be held to about 13 million barrels per day. If Saudi Arabia allows its production to increase to about 15 million barrels per day (from some nine million barrels in 1977), the Arabian Peninsula's production would be held to 20 million barrels per day. If Saudi Arabia allows its production to rise to 20 million barrels per day, the output of the peninsula would be held to about 25 million barrels per day.

The OPEC countries outside the Arabian Peninsula have a usable capacity of 20.2 million barrels per day; restrictions proposed by three of these countries could limit non-Arabian OPEC production to .8 million barrels per day less than that capacity. Moreover, the reserves of some of these countries are already being reduced rather rapidly, so that any new additions are likely to enable them to maintain production at the present level rather than to increase it significantly. We therefore assumed an ultimate limit of 20 million barrels per day for OPEC production outside the Arabian Peninsula, a limit that takes into account both the physical constraints imposed by diminishing reserves and the formal restrictions that may be imposed by governments.

Adding the non-Arabian 20 million barrels per day to the three projected Arabian Peninsula limits (13, 20 and 25 million barrels per day) yields three assumed maximum figures for potential OPEC production: 33, 40 and 45 million barrels per day. In the revised analysis OPEC production is assumed to rise with demand only until it reaches one of these three limits, to stay at that level as long as the *R/P* ratio exceeds 15 and then to decline as remaining reserves decline [*see lower illustration on page 48*].

By combining the estimated produc-



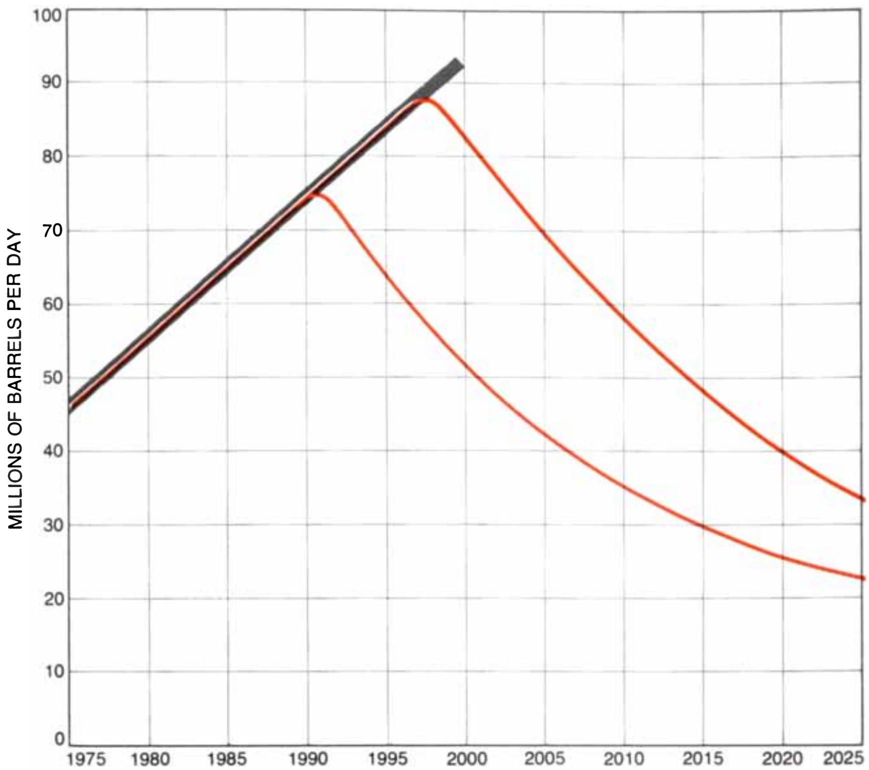
IMPORTANCE OF LARGE OIL FIELDS is indicated by this illustration, which shows that of the 30,000 fields that have been discovered only 240 contain more than half a billion barrels of recoverable reserves but that those 240 account for 73 percent of the non-Communist world's known reserves. Only 15 fields contain more than 10 billion barrels of reserves and those fields account for more than a third of the total. The four largest fields, one in Venezuela and three in the Arabian Peninsula, together contain a fifth of the known reserves.

tion of the non-OPEC countries with these various assumptions about constrained OPEC production we were able to develop a new and presumably more realistic set of oil-supply curves to set against projected demand [*see illustrations on page 49*]. The impact of the OPEC constraints is clear: production peaks earlier; the gap between supply and demand opens earlier. On the other hand, oil that is kept in the ground in the 1990's helps to maintain reserves and can be produced after 2000, making the decline in production somewhat less precipitous.

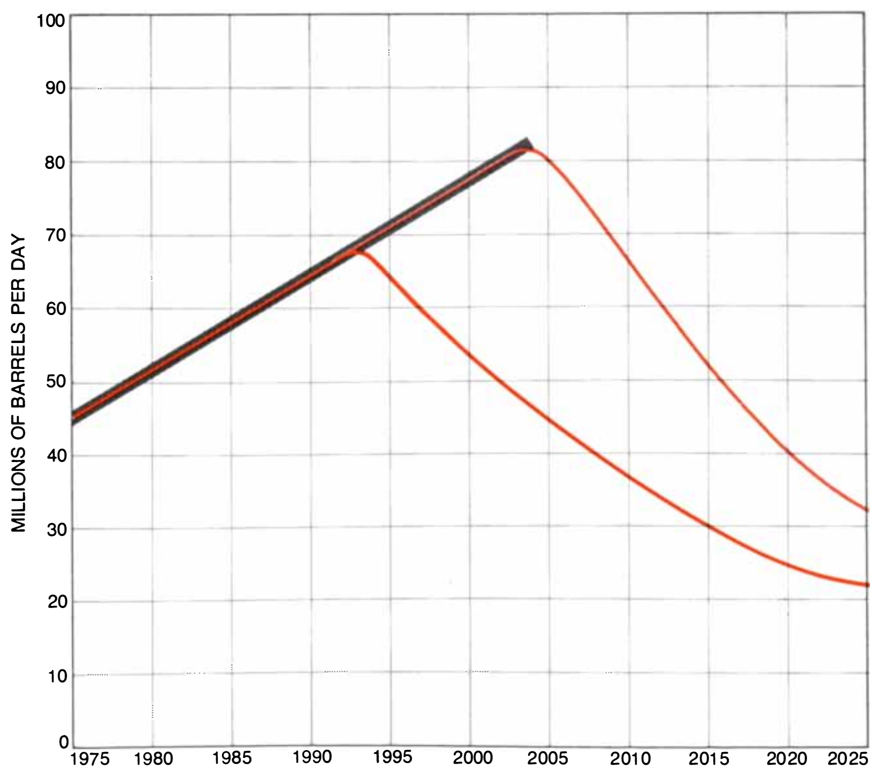
Given the high-growth, high-price scenario and the high rate of additions to reserves, for example, the 45-million-barrel OPEC limit moves the peak in the oil supply back to 1989 (compared with 1997 in the unlimited case) and down to 71 million barrels per day (compared with 86 million), but the supply can be maintained near 70 million barrels per day for some 15 years. With OPEC production limited to 33 million barrels per day, demand outruns supply as early as 1981, at only 55 million barrels per day; OPEC continues to produce at the 33-million-barrel level through 2025, however, so that total supply falls off only slowly as non-OPEC production declines.

With demand set by the low-growth scenario and with a low rate of additions to reserves, an OPEC production limit of 40 million barrels per day ordains a peak in supply five years earlier and five million barrels per day lower than in the unlimited case, but again the supply curve is flattened out. If OPEC production is limited to 33 million barrels per day, total supply is constrained to 55 million barrels per day in 1983 and falls off only with declining non-OPEC production until after 2013, when OPEC production finally drops below 33 million barrels.

Are these projections too pessimistic? Some observers might consider our assumptions about gross additions to reserves to be conservative. There have been optimistic estimates that the entire world's ultimately recoverable reserves may be as large as three trillion barrels, which means some 2.4 trillion barrels in the non-Communist areas instead of the 1.6 trillion we took as our central estimate. One basis for the optimistic forecasts is the expectation that new discoveries may be larger than we assumed, particularly in underdeveloped countries that have not yet been subjected to intensive exploration. The other basis is the expectation that major breakthroughs in techniques for secondary and tertiary production will make it possible to recover substantially more oil than we assumed. Neither of these possibilities can be ruled out. What, then, if one supposes that as much as 30 billion



OIL-SUPPLY PROFILE was developed by applying a reserves-to-production (*R/P*) ratio of 15 to 1 to projected figures for recoverable reserves. This is the profile according to the high-growth, high-energy-price scenario. Assuming a high rate of additions to reserves (which is more likely if oil prices are high), production (*dark colored curve*) fails to meet demand (*gray*) in 1997. Given a lower rate of additions, production (*light color*) fails to meet demand in 1990.



SIMILAR PROFILE for low-growth, constant-energy-price scenario shows production (*dark color*) failing to meet demand in 1994, given a low rate of additions to reserves (the more likely situation). If a high rate of additions is assumed, production (*light color*) would meet demand until 2004. Profiles on this page assume that production is not constrained by governments.

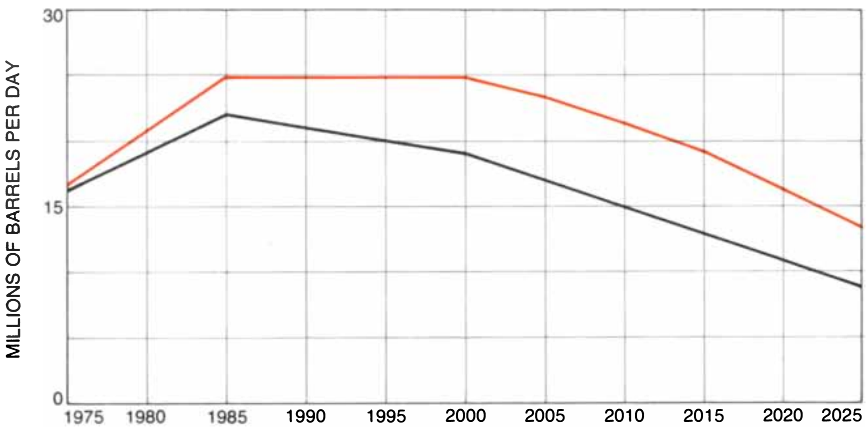
barrels per year will be added to proven reserves between 1975 and 2000 rather than 20 billion barrels per year, our high estimate?

If there were no constraints on OPEC production, the 30-billion rate would allow the oil supply to meet potential demand (in the high-growth case) until sometime between 2005 and 2010, when the limiting *R/P* ratio of 15 to 1 would be reached and production would peak and then decline. As I have shown, however, the level at which OPEC countries are willing to produce oil is more important than the theoretical limit imposed by the *R/P* ratio. The year in which an OPEC limit of, say, 45 million barrels per day is reached depends on the level of production in non-OPEC countries. Even with a discovery rate outside OPEC of 15 billion barrels per year (half the optimistic 30-billion estimate for the yearly addition to reserves) the OPEC limit would be reached, and would constrain total supply, soon after 1990. By 2000, with OPEC production limited to 45 million barrels per day, total production might be between 75 and 80 million barrels per day.

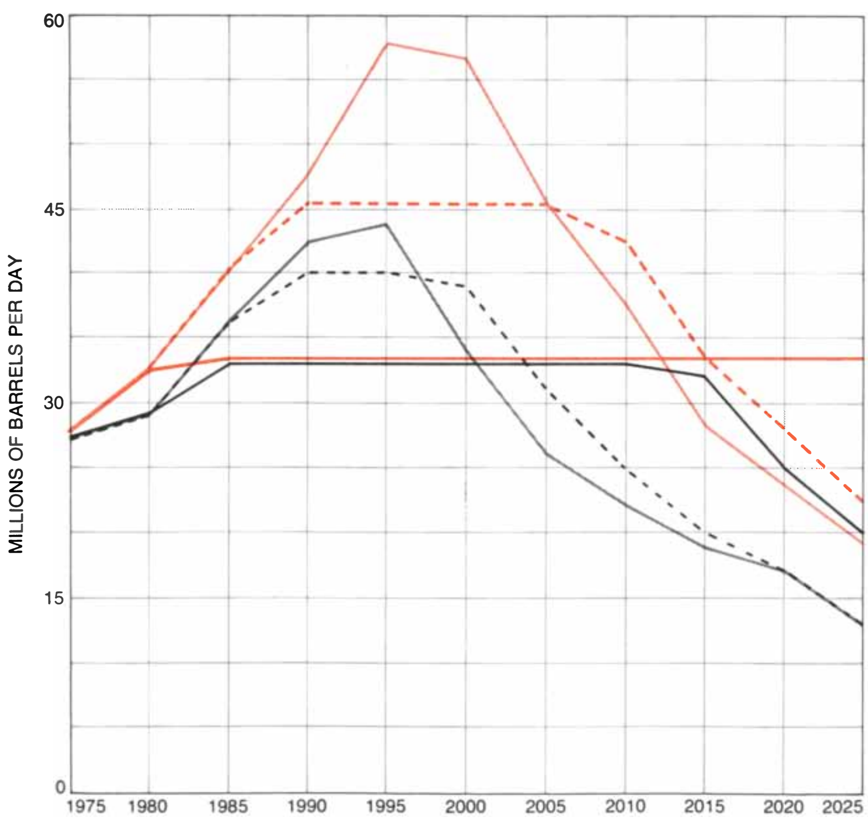
This would be less than 10 million barrels in excess of the figure for total production that we arrived at by assuming a discovery rate of 20 billion barrels per year. And it would still be some 15 million barrels per day below the demand projected for the high-growth scenario. In other words, adopting even the most optimistic estimates of additions to reserves delays the drop in production for only a few years and reduces the gap between supply and demand by only a few million barrels per day.

The future of the oil supply is uncertain, to say the least. The conclusion is clear, however, that the demand for oil in the year 2000 is unlikely to be satisfied by crude oil from conventional sources. Even in the absence of government limits on production, supply will meet demand in 2000 only in the case of the most optimistic assumption about additions to reserves. A more important constraint on production than the actual size of reserves is likely to be the level of production set by producing countries that have large reserves and a limited need for current oil revenue. Limits likely to be set by such members of OPEC will probably cause the oil supply to peak sometime around 1990 at the latest; lower limits could bring the date into the early 1980's.

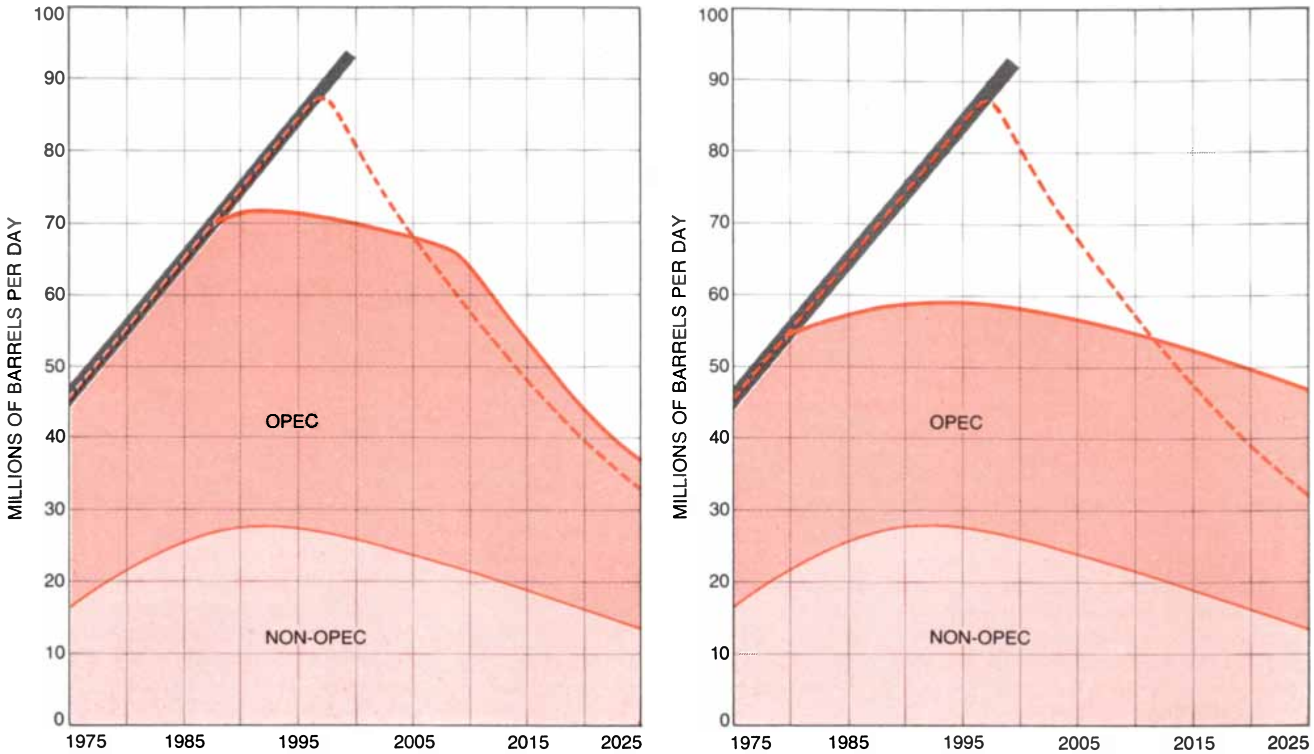
Although the end of the era of growth in oil supply is probably only 15 years away at the most, it may be followed by 10 years or so of fairly constant production, giving governments and consumers some time in which to make the adjustments that will be necessitated by a real decline in the oil supply. If oil is discovered, or recovery techniques are improved, on a scale greater than can now be foreseen, the effect would be only to delay for a few years—not to obviate—the necessary transition to other fuels. For governments and consumers to allow oil consumption to increase in the fond hope that more oil will somehow turn up is to run the risk that the complex interactions of geology, politics, economic growth and prices will instead dictate a drop in oil production even earlier than we have thought likely, thus increasing the difficulty of adjusting to a world in which oil is scarce.



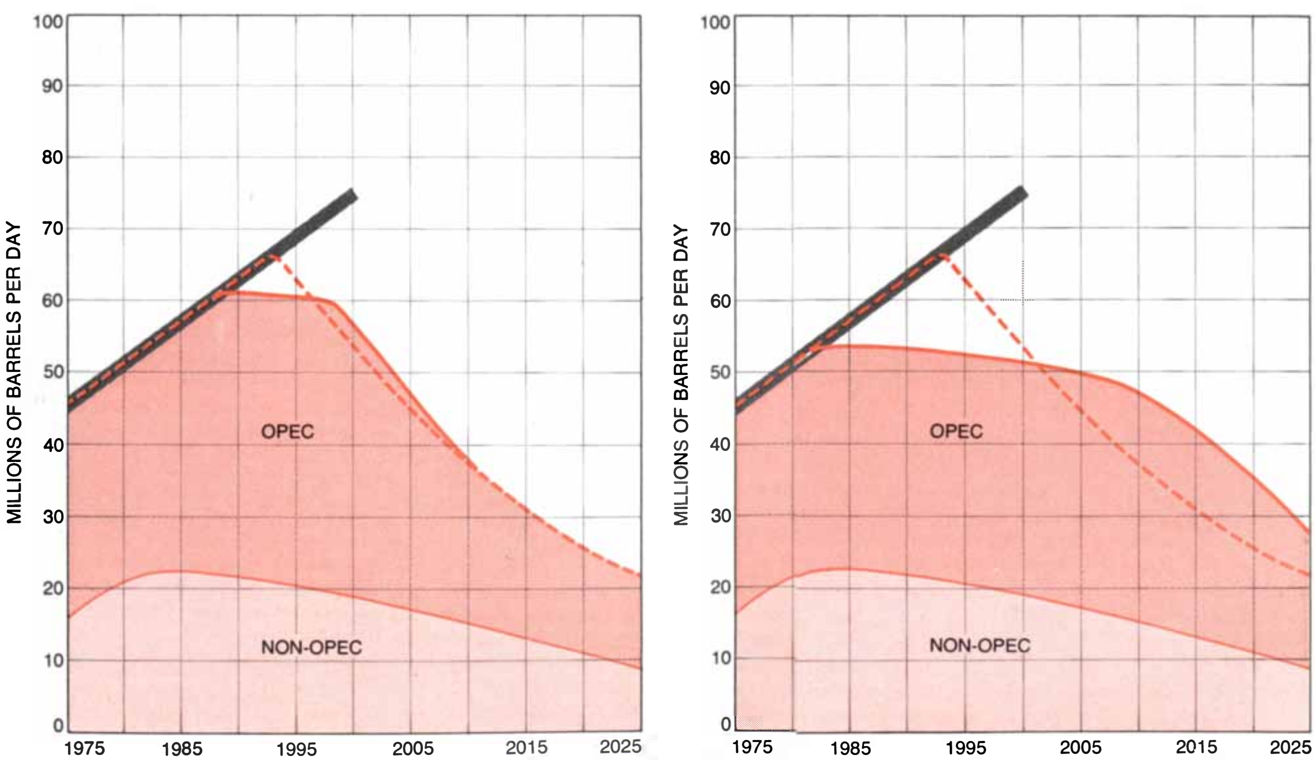
NON-OPEC PRODUCTION is separated from that of the Organization of Petroleum Exporting Countries (OPEC) in order to consider the effects of government limits. Curves show estimated non-OPEC production to 2000, and projections to 2025 assuming declining additions to reserves and declining *R/P* ratio, for high-growth (*color*) and low-growth (*black*) scenarios.



OPEC PRODUCTION is projected on various assumptions about government-imposed limits. For the high-growth scenario with high additions to reserves three curves are shown: government limits constrain production at 33 million barrels per day (*solid colored line*), or at 45 million barrels (*broken colored line*), or production is not constrained except by the *R/P* ratio (*light colored line*). For the low-growth scenario and low additions to reserves the curves assume that production is held to 33 million barrels per day (*black line*), or to 40 million barrels (*broken black line*), or is not constrained (*gray*). The projections are speculative beyond 2000.



MORE REALISTIC PROFILES take into account government limits. These curves are for the high-growth, high-price scenario and a high rate of additions to reserves. Here non-OPEC production is shown combined with OPEC production that is limited to 45 million barrels per day (*left*) and to 33 million barrels per day (*right*) to give total production in the non-Communist areas of the world (*heavy colored line*). In both cases a comparison with the unconstrained total (*broken colored line*) shows that the imposition of government limits would cut supply below demand (*gray line*) sooner (as early as 1981) but that the subsequent fall in oil supply would be less precipitate.



REALISTIC PROFILE for the low-growth, constant-price scenario and a low rate of additions to reserves shows similar effects. With OPEC production limited to 40 million barrels per day (*left*) or to 33 million barrels per day (*right*), production fails to meet demand sooner than in the unconstrained case, but later production of reserves that have been conserved keeps supply fairly steady for some time.

Heavy Leptons

The class of elementary particles of matter that includes the electron and the muon has a new member: the tau. It may be the first in a sequence of charged heavy leptons

by Martin L. Perl and William T. Kirk

The small family of elementary particles known as the leptons (from the Greek for "light ones") can be distinguished from the two other major classes of subatomic matter by a number of criteria, among which mass is perhaps the least important. For example, the leptons differ from the generally heavier hadrons primarily by being insensitive to the strong nuclear force, the dominant short-range force that binds together the particles of the atomic nucleus. The leptons share this immunity to the strong force with the photon, the massless carrier of the electromagnetic force, which forms a one-member class of its own. Unlike the photon, however, both the leptons and the hadrons are capable of interacting by means of the weak nuclear force, the even shorter-range force responsible for the radioactive decay of nuclear particles.

One of the most striking things about the leptons is that there are so few of them. Here again they stand in contrast to the hadrons, which have proliferated notoriously in recent years, reaching the point where they now total several hundred distinct particles, arrayed in various subclasses. The most familiar of the hadrons are the two main constituents of the atomic nucleus, the proton and the neutron; one of the latest additions to this class is the psi, or *J*, particle, discovered in 1974. (Burton Richter of Stanford University and Samuel C. C. Ting of the Massachusetts Institute of Technology shared the 1976 Nobel prize for physics for their independent discovery of this particle.) The lepton family, on the other hand, has for some time been thought to consist of only four particles (together with their corresponding antiparticles): the electron, discovered in the form of cathode rays more than 80 years ago; the muon, first observed in cosmic-ray showers some 40 years ago, and two kinds of neutrino, one associated with the electron (called the electron neutrino) and the other associated with the muon (the muon neutrino). Neutrinos, long suspected on theoretical grounds, were detected for the first time some 20 years ago.

Another sense in which the leptons are special has emerged just recently. A considerable body of evidence points to the conclusion that hadrons are not elementary particles at all but rather are composite structures built up of the simpler constituents termed quarks. Prior to the discovery of the psi particle all the known hadrons could be accounted for by assuming that they represented various combinations of three different kinds of quark (labeled "up," "down" and "strange") and their corresponding antiquarks. The main significance of the discovery of the psi particle was that it provided compelling evidence for the existence of a fourth kind of quark, which had earlier been named the "charmed" quark. According to the revised quark picture, the psi particle is a hadron consisting of a charmed quark and a charmed antiquark.

There is no evidence that the leptons are anything but pointlike objects, and so it seems that they, unlike the hadrons, are truly elementary particles, in the traditional sense of being indivisible. The list of the already known particles that can still be considered elementary in this sense has accordingly become quite short: four kinds of quarks (and their antiquarks), four kinds of leptons (and their antileptons) and the photon (which is its own antiparticle). In this context the search for additional members of the lepton family has acquired new meaning, because such particles, if any exist, would also be counted among the few genuinely fundamental constituents of matter.

During the past few years a group of physicists led by one of us (Perl) has been conducting just such a search at the Stanford Linear Accelerator Center (SLAC) as part of a larger experimental program carried out jointly by research groups from SLAC and from the University of California's Lawrence Berkeley Laboratory. We have had at our disposal a potentially powerful tool for creating new leptons: the SPEAR electron-positron storage ring, a device in which counterrotating beams of matter (electrons) and antimatter (positrons) can be made to pass through each other, causing occasional high-energy collisions in which the original particles are annihilated and, out of the resulting flash of energy, new particles are created. With this system we have uncovered evidence of the existence of a fifth kind of lepton. The new particle, which is electrically charged (like the electron and the muon), is much heavier than any of the previously known leptons; indeed, it is heavier than some hadrons. We shall relate here the story of the discovery of the new heavy lepton and its antiparticle, which we have named the tau and the antitau.

How does one go about finding a new kind of elementary particle? It helps to follow a few guiding principles. First, have a clear idea of what you are looking for; then you will know when you have found it. Second, if the object must be made artificially, as we had to do in this case, then find a way to make it in copious quantities. Third, make sure that the new object has some distinguishing characteristics that will tell you when you have it.

To get an idea of what we were looking for we took our lead from the known properties of the electron and the muon. In other words, we decided to look for a particle that has an electric charge of either -1 or $+1$ and that is acted on by both the electromagnetic force and the weak force but not by the strong force. Two questions followed. First, in our search for a new charged lepton what mass should we look for? Second, bearing in mind that the electron lasts indefinitely but that the muon decays in about two millionths of a second, what should we expect the lifetime of the new particle to be?

The question of mass was a difficult one, because there was (and still is) no theory that accounts either for the observed masses of the muon and the electron or for the ratio of their masses (approximately 200 to 1). All that was known as of four years ago was that some experiments carried out at the ADONE electron-positron colliding-

ELECTRON-POSITRON STORAGE RING at the Stanford Linear Accelerator Center (SLAC) is seen from directly overhead in this aerial photograph of the SLAC experimental area. The buildings in the photograph are arrayed in a roughly fan-shaped pattern that radiates from the high-energy end of the two-mile linear particle accelerator, located just out of the picture to the right. The SPEAR colliding-beam storage ring, which was used to create the new heavy leptons described in this article, is the oval structure near the bottom. Electrons and positrons generated by the linear accelerator are injected into the SPEAR device through the two tangential arms. The two buildings athwart the straight sections of the ring house the interaction regions, where the two counterrotating beams are made to pass through each other, causing the matter-antimatter collisions that lead to the creation of new particles. The charged-particle detector shown in the diagram on page 53 is housed in the larger of the two interaction buildings. The storage ring itself is about 80 meters across.

PARTICLE	MASS	LIFETIME	ANTIPARTICLE
ELECTRON (e^-)	.51 MeV	STABLE	POSITRON (e^+)
ELECTRON NEUTRINO (ν_e)	0	STABLE	ELECTRON ANTINEUTRINO ($\bar{\nu}_e$)
MUON (μ^-)	106 MeV	2.2×10^{-6} SECOND	ANTIMUON (μ^+)
MUON NEUTRINO (ν_μ)	0	STABLE	MUON ANTINEUTRINO ($\bar{\nu}_\mu$)
TAU (τ^-)	1,800 TO 1,900 MeV	LESS THAN 5×10^{-12} SECOND	ANTITAU (τ^+)
TAU NEUTRINO (ν_τ)	LESS THAN 600 MeV (MAY BE 0)	NOT KNOWN (MAY BE STABLE)	TAU ANTINEUTRINO ($\bar{\nu}_\tau$)

LEPTON FAMILY has until recently been thought to consist of just four weakly interacting particles (and their corresponding antiparticles): the electron, the muon and two kinds of neutrino. The mass of the newly discovered charged heavy lepton, the tau, and its antiparticle, the antitau (given here in MeV, or millions of electron volts), is approximately 4,000 times the mass of the electron and 20 times the mass of the muon. The significance of this mass ratio is not known. Muons decay abruptly into electrons through the weak interaction. Taus decay even more quickly into electrons, muons or other particles, also through the weak interaction. Not all the theoretically possible decays of the tau and the antitau have been observed.

beam facility in Italy had set a lower limit of about 1,000 million electron volts (MeV) on the mass of any new charged lepton: an energy equivalent to roughly 10 times the mass of the muon and 2,000 times the mass of the electron. Our group at SLAC, however, had no idea whether the mass of a new charged lepton, if one existed, would be within reach of our experimental equipment, which could detect particles with a mass as high as 3,500 MeV.

The question of the lifetime of the hypothetical particle, it turned out, had a better theoretical answer: less than a hundred-billionth of a second for any charged lepton with a mass greater than 1,000 MeV. The lightest charged lepton, the electron, is stable simply because there is no lighter charged particle into which it can decay. The muon is unstable because it can decay into an electron. The muon-to-electron decay does not take place in the simplest conceivable way, however, which would be for the muon to change spontaneously into an electron and a photon through the electromagnetic interaction, even though more than enough energy would be released in the process to produce a photon. Instead the muon is observed to decay through the weak interaction into three particles: an electron, an electron antineutrino and a muon neutrino.

Physicists explain this odd behavior by invoking an empirical rule whose basic significance remains unclear. They ascribe to the electron and the muon separate intrinsic properties, which we shall refer to here as "electronlikeness" and "muonlikeness," and they postulate that each of these properties is exhibited by a group of four related particles. Thus the electron, the electron neutrino, the antielectron (or positron) and the electron antineutrino are all said to exhibit the property of electronlikeness (or antielectronlikeness in the case of the antiparticles), whereas the muon, the muon neutrino, the antimuon and the muon antineutrino all exhibit muonlikeness (or antimuonlikeness). The rule then states simply that in any particle interaction or decay process involving leptons the properties of electronlikeness and muonlikeness must be separately conserved.

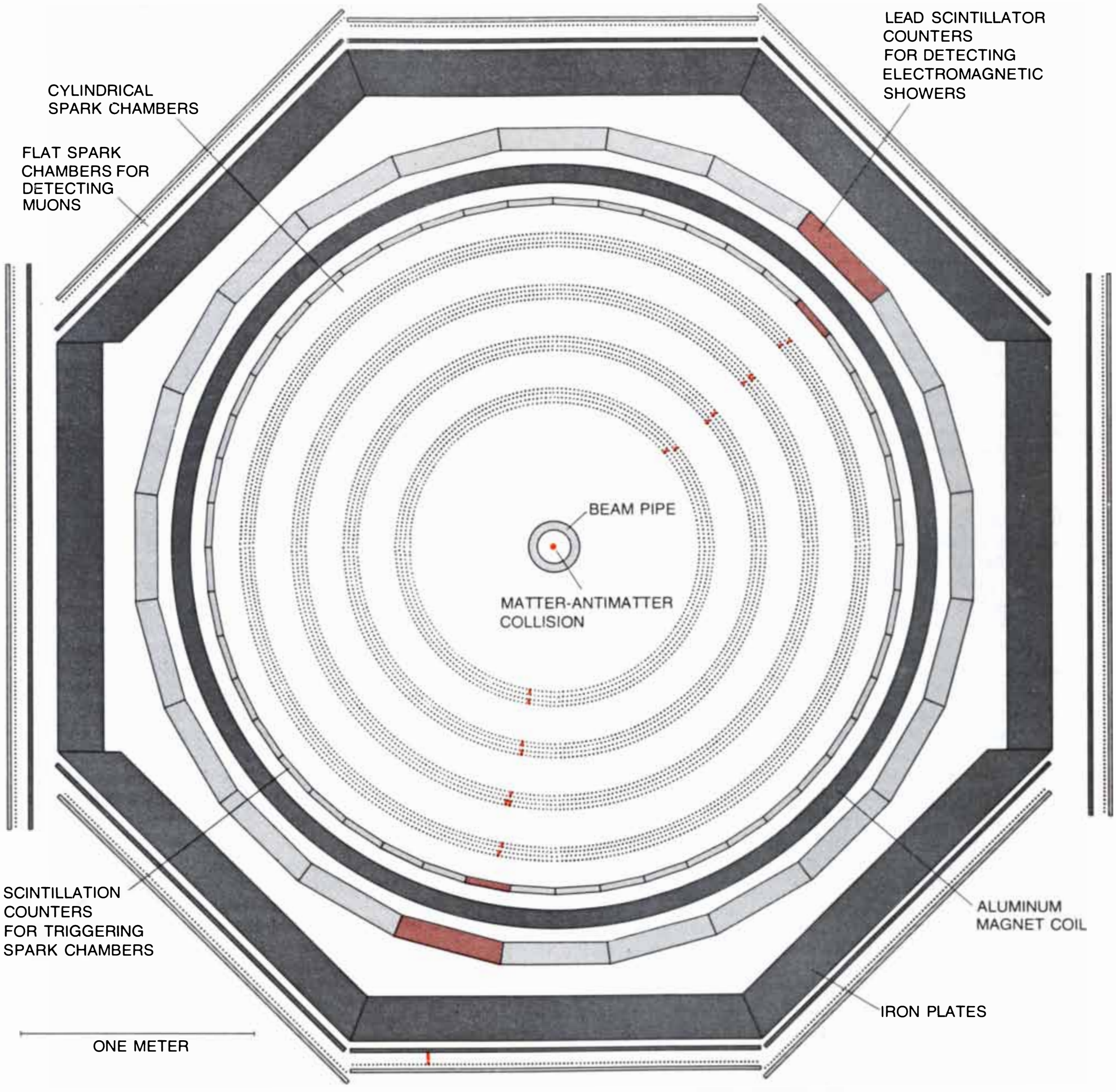
It follows that the simple electromagnetic decay of a muon into an electron and a photon is not possible because in the process the muonlikeness of the muon would have to change into the electronlikeness of the electron. The more complicated weak decay of the muon into an electron, an electron antineutrino and a muon neutrino does occur, because the muon neutrino preserves the muonlikeness of the muon, whereas the electronlikeness of the electron is exactly canceled by the antielectronlikeness of the electron antineutrino. (In actual practice particle physicists employ the terms electron lepton number and muon lepton number to denote the properties of electronlikeness and muonlikeness, and the rule in question is called the conservation of lepton number: in our view, however, these formal terms tend to obscure the fundamental mystery of the separate unique properties of the electron and the muon.)

In any event we had to decide on some comparable property for the hypothetical charged heavy lepton we were looking for. One possibility was to assign an electron lepton number (or electronlikeness) to the new lepton, thereby allowing it to decay electromagnetically into an electron and a photon. A more intriguing possibility was to assume that the new lepton-antilepton pair came with their own separate lepton number and their own associated neutrino-antineutrino pair. According to this view, the electron and the muon might be just the beginning of a sequence of charged leptons, each with its own unique "likeness." The general name adopted for these highly speculative particles at the time was sequential charged heavy leptons, with the "heavy" indicating that their masses would be greater than that of the electron or the muon. Soon after the first fragmentary evidence for such a particle was obtained, we began to use the symbol *U* (for "unknown" particle), but now that we have substantial evidence for its existence we call it tau, after the first letter of the Greek word τριτον, meaning "third." The name is meant to indicate that the tau is the third charged lepton in the sequence beginning with the electron and the muon.

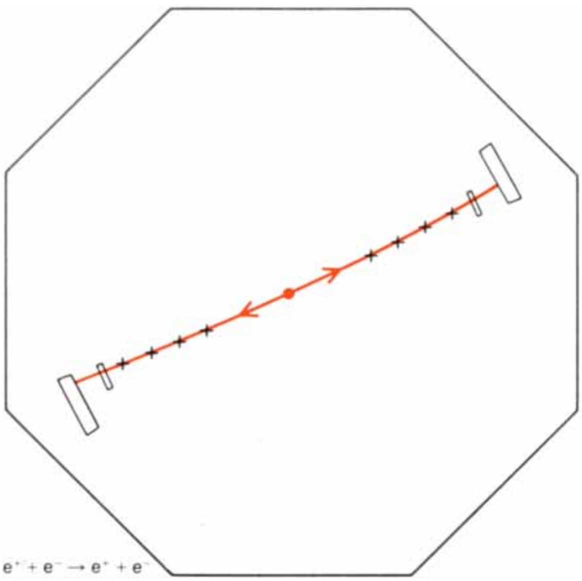
As we have noted, an effective search for a new elementary particle requires some idea of its properties, a method for producing the particle in sufficient quantities and a means of distinguishing the new particle from the particles already known. Experiments utilizing the SPEAR electron-positron storage ring at SLAC can satisfy both of the last two requirements. This machine, which began operating in 1972, consists of about 100 magnets in a ring-shaped array some 80 meters in diameter. The electron and positron beams for SPEAR, which are generated by the SLAC two-mile linear accelerator, are injected into the storage ring during a "filling time" of anywhere from 10 to 30 minutes. The beams circulate in a vacuum chamber that passes through the ring of magnets, and the deflection and focusing provided by the magnets hold the beams in stable orbits for periods of several hours. The circulating beams do not encounter each other except at two points on opposite sides of the ring, where they are made to pass through two interaction regions. In order to maximize the chance of collisions each beam contains about 100 billion particles in a single tightly packed "bunch" that is only a few centimeters long. Although each bunch makes a complete trip around the ring more than a million times a second, the chance that a particle in one beam will make a direct hit on a particle in the other beam is so remote that such a collision typically happens only once every few seconds.

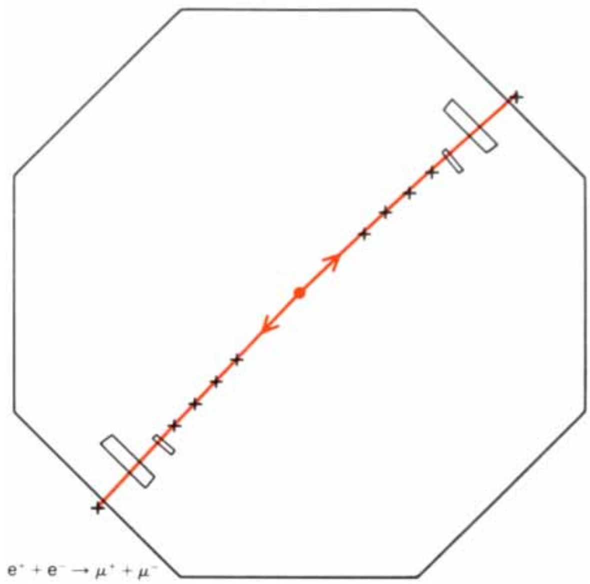
In order to study all the particles that might be produced in a single electron-positron annihilation two groups of physicists from SLAC joined forces several years ago with two groups from the Lawrence Berkeley Laboratory to build a general-purpose particle detector to surround one of the interaction regions of the SPEAR device. The detector has a cylindrical central section containing four concentric spark chambers immersed in a strong magnetic field. Surrounding the spark chambers is a system of scintillation counters for detecting charged particles. Whenever two or more charged particles are detected by the counters, the inner spark chambers are actuated and the paths of the charged particles are recorded with the aid of a computer on magnetic tape. The computer can then reconstruct the paths of the particles, yielding a diagrammatic "picture" of the event.

Outward from the scintillation counters is the aluminum magnet coil, followed by another cylindrical system of counters for detecting electromagnetic showers. These counters can distinguish electrons or positrons, which generate large showers, from hadrons or muons, which produce smaller showers. Outside all this hardware is an octagonal set of iron plates at least eight inches thick



GENERAL-PURPOSE PARTICLE DETECTOR surrounding one of the interaction regions of the SPEAR electron-positron storage ring is viewed here in cross section, with the trajectories of the colliding beams at right angles to the plane of the diagram and the collision point in the plane of the diagram at the center (*colored dot*). The cylindrical inner part of the detector contains four concentric spark chambers, each consisting of four layers of closely spaced wires (*black dots*). In the actual device there are some 100,000 such wires, strung about a millimeter apart. When the detector is in operation, a high voltage is applied across pairs of adjacent wire layers, and spark discharges (*colored dashes*) mark the ionization trails of charged particles passing through the inert gas that fills the chambers. Outside the spark chambers are two systems of charged-particle counters, separated by an aluminum magnet coil. Whenever the inner system of scintillation counters detects the passage of two or more charged particles, the cylindrical spark chambers are actuated and the paths of the particles are recorded. (The paths are bent slightly by the magnetic field; the amount of bending indicates the particle's momentum.) The outer system of lead scintillator counters registers the distinctive electromagnetic showers generated by different types of particles. Enclosing all this equipment is an octagonal set of thick iron plates. Of all the charged particles that can be produced in the electron-positron collisions at the center of the detector the only ones that can penetrate the iron and be detected in the outermost system of flat spark chambers are the muons. An interpretation of the "electron-muon" event shown in color here is given in the bottom illustration on page 55.





SIMPLIFIED DIAGRAMS of four possible outcomes of the mutual annihilation of an electron and a positron were drawn with the aid of a computer-assisted display system connected to the charged-particle detector at the Stanford colliding-beam facility. In representations of this kind electric discharges caused by the passage of charged particles through either the inner cylindrical spark chambers or the outer flat ones are symbolized by small black crosses. Only those counters that have detected particles are shown; the small rectangles are the inner "trigger" counters, and the larger, flatter rectangles are the outer electromagnetic-shower counters. The octagonal outline indicates the outer surface of the iron plates. The colored lines connecting the detection signals are bent into circular arcs by the strong magnetic field inside the magnet coil, but they are straight outside the coil. The first picture at the left shows a typical electron-positron event, in which the original electron-positron pair disappear and, out of the resulting energy, a new electron-positron pair emerge back to back, each with half of the total energy of the colliding particles. (On emerging from the collision point negatively charged particles curve right

and then more spark chambers. In general hadrons cannot penetrate the iron plates because they interact with the iron nuclei through the strong force. Electrons and positrons are also unable to penetrate the iron because they have already lost most of their energy through the production of large electromagnetic showers. Muons, however, do penetrate the iron and are detected in the outer spark chambers. Hence the detector can not only measure the direction and the momentum of the newly created particles but also separately identify hadrons, electrons and muons. It was its ability to distinguish these different kinds of particles that enabled us to find the tau-antitau pair.

The SPEAR ring can be adjusted to store beams of electrons and positrons with an energy of up to four billion electron volts (GeV) in each beam. Since the collisions that occur are between matter and antimatter, a possible outcome of such a collision is what is called an annihilation reaction. The reaction actually takes place in two steps. First the colliding electron and positron disappear, and in so doing they create the short-lived state of pure electromagnetic energy called a virtual photon. Then after an immeasurably brief time (10^{-25} second) the virtual photon rematerializes into any one of a very large number of possible combinations of new particles. Some of the possible outcomes are the re-creation of an electron-positron pair, the creation of a muon-antimuon pair, the creation of a hadron-antihadron pair (for instance a proton and an antiproton) or the creation of a large number of hadrons. What we proposed to do was to use the same method to produce a heavy lepton-antilepton pair.

If such pairs of new heavy leptons were actually being created in the electron-positron collisions at SPEAR, how would we be able to recognize them? Our earlier estimate of the tau's lifetime—less than a hundred-billionth of a second—meant that a newly created tau particle would travel less than a centimeter from its point of creation to the point where it decays, even if its velocity were roughly equal to that of light (30 billion centimeters per second). This path is too short to be directly detected by our experimental apparatus, and so we had to look for a way to recognize tau particles indirectly by identifying some distinctive pattern that appears when they decay.

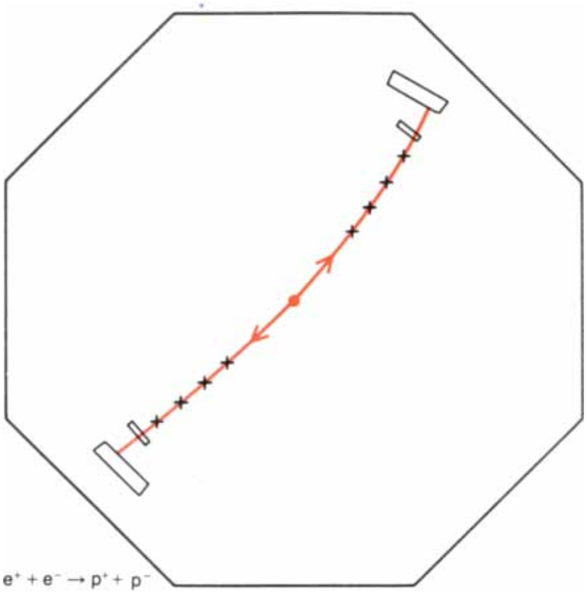
Our decision to attribute a unique lepton number, or "taulikeness," to the tau and its neutrinos meant that we did not expect the tau to decay electromagnetically into either an electron or a muon with the accompanying emission of a photon. Instead we assumed that the tau, like the muon, would decay through the weak interaction and that several different decay modes would be possible for the tau because of its large mass. We selected two of these possible weak-decay modes for special attention.

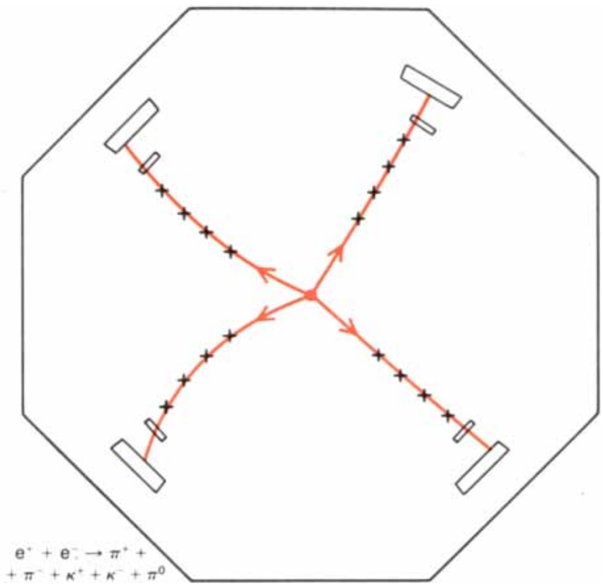
In the first decay mode the tau decays into a tau neutrino, a muon and a muon antineutrino, and in the inverse process the antitau decays into a tau antineutrino, an antimuon and a muon neutrino. In the second decay mode the tau decays into a tau neutrino, an electron and an electron antineutrino, and the antitau decays into a tau antineutrino, a positron and an electron neutrino.

These two decay modes were picked because they were expected to occur frequently. More important, the seemingly complex decay processes described above are really very simple when they are viewed in the context of the actual experiment. The reason is that all the neutrinos and antineutrinos produced in the decay processes are such elusive particles that the experimental apparatus simply does not "see" them at all. Thus what the apparatus actually records in each case is merely the tracks of two charged particles: either an electron and an antimuon, or a positron and a muon.

This final outcome has a distinctive experimental "signature" for two reasons. First, the appearance of an electron and a muon (or their antiparticles) in the final state seems to violate the principle of the conservation of lepton number; the rule is not really violated, however, because the undetected neutrinos make up the balance. Second, there is a lot of missing energy, but that of course is also accounted for by the unseen neutrinos.

We first began to find evidence of such electron-muon events in 1974. In a sample of 10,000 events of all types we identified 24 as electron-muon events. Al-



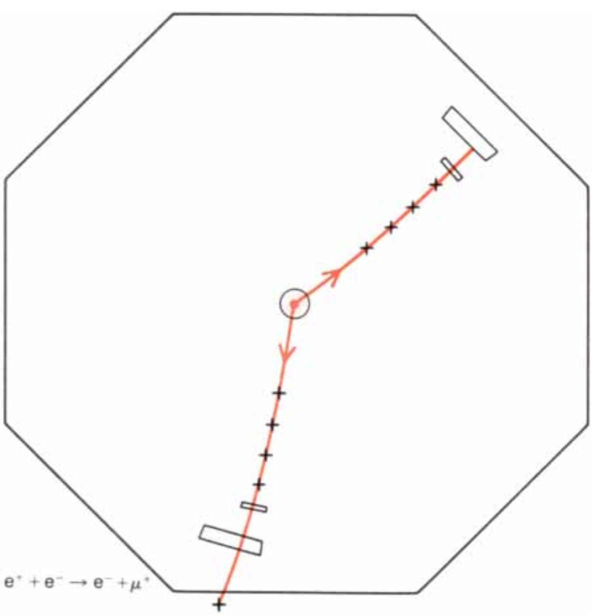


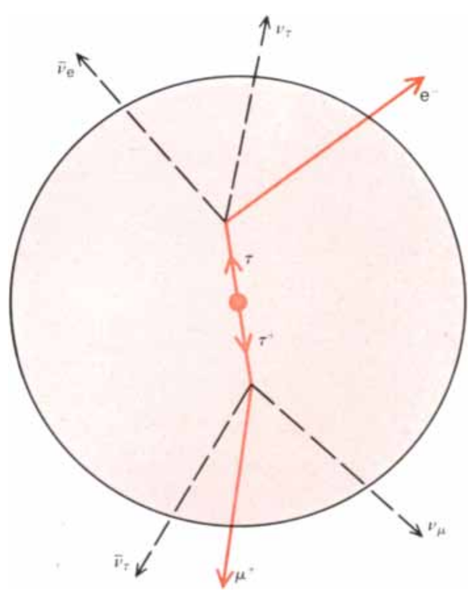
and positively charged particles curve left in the perpendicular magnetic field; hence the particle track in the eight-o'clock position is that of the electron, whereas the one in the two-o'clock position is that of the positron.) The second diagram shows a muon-muon event, in which the newly created particles are a negative muon (in the half past seven position) and a positive antimuon (in the half past one position). Unlike the electron and the positron, which lose most of their energy in generating electromagnetic showers and so cannot penetrate the iron plates, the muons pass freely through the iron and are detected in the outermost spark chambers. In the third diagram the original electron and positron annihilate each other to form a pair of hadrons, in this case a proton and an antiproton. In the fourth diagram the tracks of four hadrons are seen in the detector: two charged pions and two charged kaons. A fifth particle, a neutral pion, is also produced in this annihilation reaction, but since it is uncharged it cannot be detected directly. None of the hadrons are capable of penetrating the iron plates. The formulas with the pictures describe interactions in a shorthand notation commonly used by particle physicists.

though the number of such events was low, we were encouraged, because if the events were real, it meant that the mass of the tau was within the energy range of the SPEAR system.

At that early stage, however, we had to maintain a skeptical attitude, because with just 24 events there were various ways we could be wrong. First of all, our detector was far from perfect in identifying electrons and muons. Indeed, about 20 percent of the time it misidentified a hadron as either an electron or a muon. A careful study of the problem indicated, however, that only five or six of the 24 events could possibly be attributed to a misidentified hadron among the decay products.

A second reason for skepticism was the fact that although we might have found a new particle, it might actually be a new hadron rather than a new lepton. As it happened there was at that time a prime candidate for such a new





TYPICAL ELECTRON-MUON EVENT observed recently in the charged-particle detector at SPEAR is shown in the diagram at the left. The telltale "signature" of such an event is the detection of one particle that traverses the iron plates (the muon) and another that does not (the electron). The event is interpreted in the greatly enlarged view at the right in terms of the hypothesis that the electron and the muon come from the weak decay of a tau-antitau pair. The creation and decay of the heavy leptons happen within a few millimeters of the center of the detector and hence cannot be seen directly. Each tau particle decays into one charged particle and a pair of neutrinos; only the charged particles, in this case a negatively charged electron and a positively charged muon (or antimuon), are detected.

hadron: the charmed hadron called the *D* meson, which had not yet been discovered. It was conceivable that the electron-muon events might be coming from the production and decay of *D* mesons, accompanied by neutrinos and neutral *K* mesons. The neutral *K* meson is peculiar in that half of the time it decays so slowly that it would have escaped from our detector before the decay took place. In that case, if both of the neutral *K* mesons happened to escape, one would see only the electron and the muon in the detector and one might then be misled into thinking that the decay of a tau-antitau pair had actually been observed. As we accumulated more electron-muon events, however, particularly at higher energies, our colleague Gary J. Feldman was able to demonstrate that most of the electron-muon events could not have come from the production of *D*-meson pairs, because there were too few cases in which we observed both the electron-muon signal and the expected decay products of the neutral *K* mesons.

In the two years following our initial discovery we continued to collect more electron-muon events (we now have about 200) and to look for other ways of testing our data. One important test was to see whether the energy distribution of the electron and the muon was what one would expect from the weak decay of a heavy lepton into three particles. We found that the data did fit the three-body hypothesis well but that it did not fit the alternative two-body hypothesis [*see illustration below*].

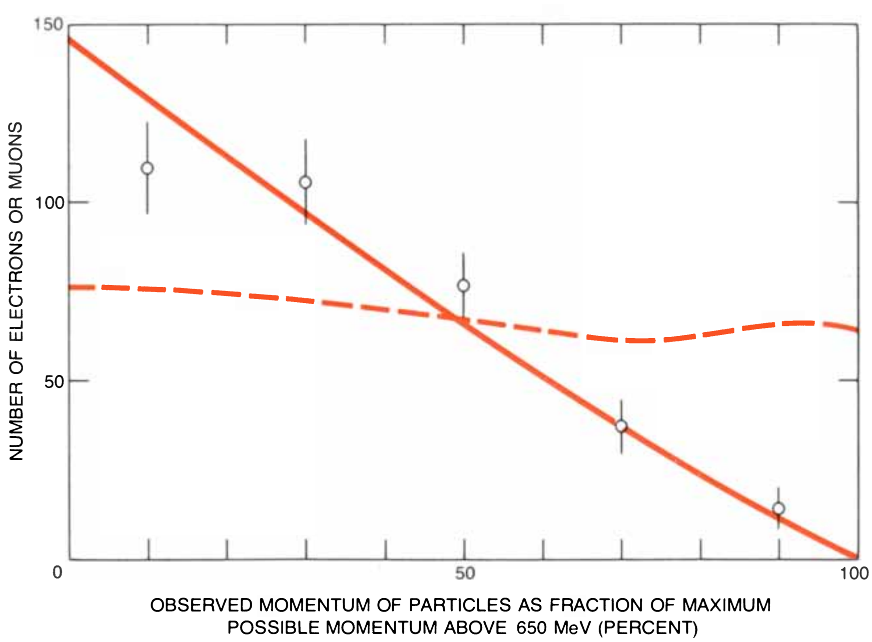
In another test we considered other possible decay modes of the tau-antitau pair. If, for example, one of the tau particles were to decay into a meson (mesons are a subclass of the hadrons) and the other were to decay into a muon, then one would expect to see distinctive muon-meson events. Such events would be distinctive because, like the electron-muon events, they would seem to violate the principle of the conservation of lepton number. A group of physicists from the University of Maryland, the University of Pavia and Princeton University found a few such muon-meson events in an experiment at SPEAR, and later our Stanford-Berkeley group was able to collect about 100 similar events.

Certain other links remained to be closed in the chain of evidence. The German Electron Synchrotron Laboratory in Hamburg has an electron-positron storage ring called DORIS, which has a particle-producing capability similar to SPEAR's. If tau particles were being created in the Stanford device, they should also be created in the Hamburg one. For a year or so there were no reports of heavy leptons being found at DORIS, but then a group at the Hamburg facility began to observe electron-muon events and later muon-meson events; in both cases they had the characteristics expected for the tau-antitau pair. Additional confirmation was provided when another group at DORIS discovered a number of electron-meson events, which are analogous to the muon-meson events.

At an international meeting on the physics of leptons and photons held last August in Hamburg five experimental groups working at SPEAR and two groups working at DORIS reported the results of their independent searches for heavy-lepton production in electron-positron annihilation. All the groups agreed on the following points: (1) that they had evidence for a new particle, (2) that the new particle was not a charmed hadron, (3) that the mass of the new particle was between 1,800 and 1,900 MeV (close to 20 times the muon's mass and 4,000 times the electron's) and (4) that all the measured properties of the new particle were consistent with the properties one would expect for a heavy charged lepton.

One way of showing that the new particle is a lepton and not a hadron is to measure its rate of production as a function of the total energy of the electron-positron collision. From studies of the production of pairs of hadrons it is known that when the total energy of the system is raised above a certain threshold energy, the production rate first increases for a time and then begins to decrease rapidly. The rapid decrease ensues because when there is too much energy available, the hadrons, which are composite particles, cannot hold together. Instead of a pair of hadrons many hadrons are produced. In contrast, since leptons are pointlike particles that do not break up, one would expect the production rate of leptons to rise rapidly from the threshold energy, reach a maximum and then decrease rather slowly as the energy is increased. Using the electron-muon events as a measure of the production rate of tau-antitau pairs, we have determined that the production rate of the new particles changes with increasing energy in the manner predicted for lepton pairs and not in the manner predicted for hadron pairs [*see illustration on opposite page*].

Although all the experiments to date agree that there is a new heavy lepton, there is still much we do not know about it. For example, the existing data are consistent with the idea that the tau is not affected by the strong force, but this insensitivity has not been tested as thoroughly for the tau as it has been for the electron and the muon. In addition many of the elementary particles, including the leptons, are in a state of perpetual rotation about their own axes, like spinning tops, but it has not yet been established that the tau has the same spin characteristics as the electron and

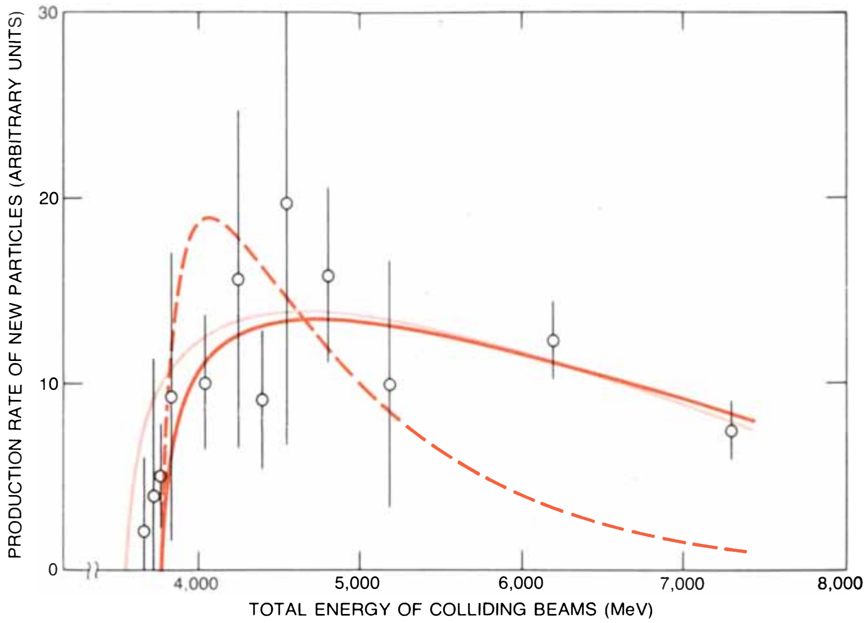


EVIDENCE in support of the conclusion that the newly discovered elementary particle is indeed a lepton and not some novel kind of hadron is presented in this graph of the energy distribution of the particles observed emerging from the distinctive electron-muon events. Only events in which the momentum of both the electron and the muon were above a certain value (650 MeV) were considered. Assuming that the tau is a heavy lepton that decays weakly into three particles (either an electron and two neutrinos or a muon and two neutrinos), the data would be expected to follow the solid colored curve. If the tau were another kind of particle that decays into an electron or a muon, in either case accompanied by a single undetected neutral particle, then the data would follow the broken colored curve. The observed energy distribution is consistent with the three-body hypothesis and not with the alternative two-body hypothesis.

the muon. We also do not know yet whether the tau is the sequential heavy lepton we initially set out to find. In other words, does the tau have its own unique tau lepton number, or taulikeness, or is it an entirely new kind of lepton with even more peculiar properties?

Experiments at other laboratories based on the interactions of muon neutrinos do not seem to have created any tau particles; hence it is unlikely that the tau is muonlike. Nevertheless, there remains the possibility that it is electronlike. (If it is, however, there must be a rather special and complicated mechanism at work to suppress the expected electromagnetic decay of the tau into an electron and a photon.) Furthermore, not all the possible decay modes of the tau have been observed, and it is only through knowing the full set of decay modes that one can know just what kind of lepton the tau is.

Another area that will require further experimental and theoretical study is the question of why the mass of the tau (between 1,800 and 1,900 MeV) is so close to the estimated mass of the charmed *D* meson (1,865 MeV), particularly when there is strong evidence that the tau is itself not a charmed hadron. As it happens, there is another such puzzling correspondence in particle physics: the mass of the muon (106 MeV) is quite close to the mass of the pi meson, or pion (140 MeV). Are these merely chance coincidences or is there some unknown relation between the masses of the leptons and the hadrons?



FURTHER EVIDENCE that the tau is a lepton and not a hadron is presented in this graph, in which the rate of production of the new particle is plotted as a function of the total energy released in the electron-positron collision. If the new particles were hadrons, then as the total energy of the system is raised above a certain threshold energy their production rate would be expected to first increase for a time and thereafter decrease rapidly (*broken colored curve*). If the tau particles were leptons, however, their production rate would rise rapidly from threshold energy, reach a maximum and then decrease slowly as the energy is increased (*solid colored curves*). Again the data are consistent with the heavy-lepton hypothesis and not with the alternative hypothesis. The dark colored curve is based on the assumption that the mass of the new lepton is 1,900 MeV; the light colored curve is based on the assumption that the mass of the new lepton is 1,800 MeV. The data suggest that the actual mass lies between the two values.

There are other questions. For example, what determines the masses of the leptons? It is hard to make sense of the observed mass sequence of the electron, the muon and the tau: .51, 106 and 1,800 to 1,900 MeV. The numbers increase too quickly to be an arithmetic series and too slowly to be a geometric series. Of course, given only three points, there are many empirical formulas that could be found to fit the sequence, but none of them would be based on a fundamental understanding of the leptons, since no one knows what accounts for the individual masses of the leptons.

And what is one to make of the properties electronlike, muonlike and taulike? Perhaps there is nothing further to understand about these properties. Just as physicists have come to accept electric charge as a fundamental property of particles for which they have no deep understanding, so too they may simply have to accept the unique lepton properties electronlike, muonlike and taulike. It is still not known why the total electric charge is conserved in all particle interactions; perhaps it is not possible to know why the total lepton numbers are conserved in all lepton reactions.

Answers may nonetheless be in the offing. Two new high-energy electron-positron colliding-beam accelerators are now under construction: the PEP machine at SLAC and the PETRA machine in Hamburg. Both devices will achieve an energy of 18,000 MeV or more per beam, which means that the search for additional charged leptons can be extended up to masses perhaps five times greater than the present upper limit of about 3,500 MeV. Most of the experiments being planned for PEP and PETRA include in their design ways to search for new heavy leptons, usually by looking for electron-muon events. Whether or not new leptons are found it is certain that the search will be more difficult, if only because the electron-muon events from the decay of the tau particles will be an annoying background. Experimenters will have to distinguish new and interesting electron-muon events from old and uninteresting ones resulting from the tau-antitau decay. On the other hand, the electron-muon events from the taus will be helpful in checking out the new devices.

Finally, there is a deeper question for which no answer exists at present. What is the relation between the two multimember classes of particles that now seem truly elementary, the leptons and the quarks? Before the discovery of the tau particles there were only four known types of lepton and four known types of quark (counting each particle and its antiparticle as one type). There was a nice symmetry here, and certain theories relating the leptons and the quarks made use of it. With the discovery of the new heavy lepton, however, the symmetry is destroyed; there are now more known leptons than there are known quarks, two more if the tau is sequential with its own tau neutrino.

Meanwhile, however, a group of physicists at the Fermi National Accelerator Laboratory (Fermilab) has recently reported evidence of the possible existence of a new, fifth kind of quark. Thus the number of quarks may also be increasing. Some theories preserve the symmetry between the leptons and the quarks, others forgo the need for such a symmetry, but almost all provide for the possibility of additional quarks and leptons. This apparent proliferation in the types of leptons and quarks, although still a matter of speculation, is somewhat alarming. In many ways it would be preferable if the truly elementary particles could remain as few in number as they were in the pre-tau days, or better yet in the pre-tau and pre-charmed-quark days. One cannot, however, dictate to nature what the fundamental constituents of matter should be. One can only hope to be up to the task of finding the constituents and understanding them.

The Electronic Telephone

One of the few components of the telephone system that has been little touched by the revolution in electronics is the instrument itself. A telephone based on integrated circuits is in the offing

by Peter P. Luff

Anyone who has used a push-button telephone and heard the electronic tones it generates might suppose the revolution in electronics had encompassed the telephone long ago. Such is not the case. The telephone of today works according to electromechanical principles that were developed even before the telephone was invented. Although many components of the telephone system have been transformed by the developments of recent years in integrated circuitry, the instrument itself has been left behind. As an engineer who has worked to develop an electronic telephone, I believe I know the reasons for this state of affairs. I can also say with confidence that the electronic telephone will be commonplace within a few years and that it will offer substantial improvements over the service available today.

If technological advances alone were the controlling factor, the electronic telephone would have been in service some years ago. As will become apparent from what follows, the tardiness of its arrival is explained by a complicated mixture of the evolutionary requirements of the telephone system, economic factors and certain human reactions. When the electronic telephone does arrive on the scene, it will provide a quality of sound equivalent to what one hears with today's high-fidelity entertainment systems, it will make possible radical changes in the physical form of telephones and it will offer the user a variety of conveniences that will add a new dimension to the service itself.

Let me set the stage by reviewing the basic principles of the modern telephone system and describing the way it evolved. Then the place of the advanced electronic telephone set will become apparent.

The major objective in telephony is to provide voice communication over large distances. The traditional approach has been to convert the sound waves of the user's speech into electrical energy that can be transmitted efficiently over wires. The electrical waves, which ideally are identical in both form and frequency with the original sound waves, enter the receiver in the earpiece of the listener's set and then are converted back to sound waves corresponding more or less to the original speech pattern.

In addition to enabling the user to talk and listen the telephone set must be able to both select any other set in the system for a connection and indicate to him when it is in use. No telephone can perform these functions without assistance from an interconnection system. Each set in a given geographical area is connected (by way of a pair of copper wires called the loop) to a central office that is a switching center. Telephone calls within the area are routed through that office; calls to numbers outside the area are routed to other offices through interoffice trunk lines.

Picture a user's initiating a telephone call. He lifts the handset of his telephone (set *A*), whereupon the central office returns a dial tone to let him know that a line is available. He then dials a number (the number of set *B*): the central office selects the correct pair of wires and sends a ringing signal to set *B*. If someone answers set *B* by lifting the handset, the central office connects *A* and *B*. Communication is thereby established. The central office also senses the termination of the call and releases the lines for other calls. When a long-distance call is made, the central office automatically records the information needed for billing. The sections of the central office that perform these various functions are termed supervisory circuits.

In North America the basic rotary-dial telephone (known as the 500 set in a series of models that began in the 1920's with the 200 set) represents the starting and finishing point of the voice-communication network. Some 160 million of the sets are in service in North America. A typical instrument consists of a handset, containing a transmitter and a receiver, that is connected to the base with an extendable cord holding four conductors. The base houses a transformer system designed to separate outgoing and incoming speech signals, a switch for connecting and disconnecting the telephone and the loop, a dial pulser to signal the required numbers to the central office and a bell to signal an incoming call to the subscriber.

The transmitter is a microphone controlled by the sound it receives. It operates on the principle that the vibration of a diaphragm can vary the strength of an externally supplied electric current. It does so by changing the resistance of a small enclosure, the "carbon button," that is filled with granules of carbon. As the diaphragm is displaced, the plunger attached to it varies the pressure applied to the granules, so that the total resistance across the button changes by amounts approximately equal to the variations in the incident sound. If a voltage is applied to the carbon button in order to generate a steady electric current, sound waves incident on the diaphragm will result in a modulation of the current.

(The power for the telephone set is provided by a central-office battery system supported by auxiliary generators. That is why telephone service is usually maintained when the supply of electric power for other devices is interrupted.)

The receiver in the earpiece contains a diaphragm and a permanent magnet that has a coil wound around it. The incoming electric signal is applied to the winding of the coil, thereby causing a perturbation of the permanent magnetic field. The perturbation vibrates the diaphragm, which generates sound waves corresponding to the original speech pattern.

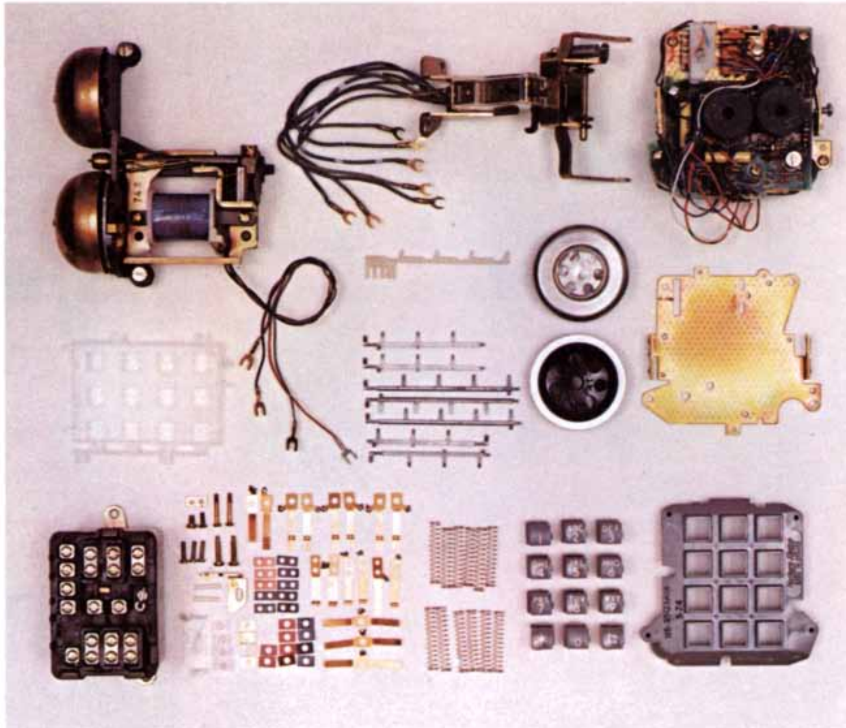
The transmitter and receiver connect with the loop through a speech circuit, which is often referred to as a "two-wire, four-wire" network because the four wires from the handset are reduced to two at the connection with the loop. The primary function of the speech circuit is to separate the circuits for talking and listening in order to limit the amount of the talker's signal (the side tone) in his own receiver. The amount of side tone is important from the viewpoint of human reactions. Too much side tone tires the ear and causes the talker to lower his voice, which reduces the volume of the sound reaching the listener. It also reproduces room noise in the listener's receiver and so degrades the intelligibility of the incoming speech. Too little side tone gives the telephone an unnatural "dead" sound and tends to cause the user to talk too loudly.

Another factor that must be taken into account is the location of the telephone with respect to the central office. The distance between the two can vary from a few feet to a few miles. The effective resistance of the copper-wire loop increases in proportion to its length, thereby attenuating the signals on the line. An additional problem is that the efficiency of the carbon transmitter drops as the current feeding it is reduced by the increased resistance on the line.

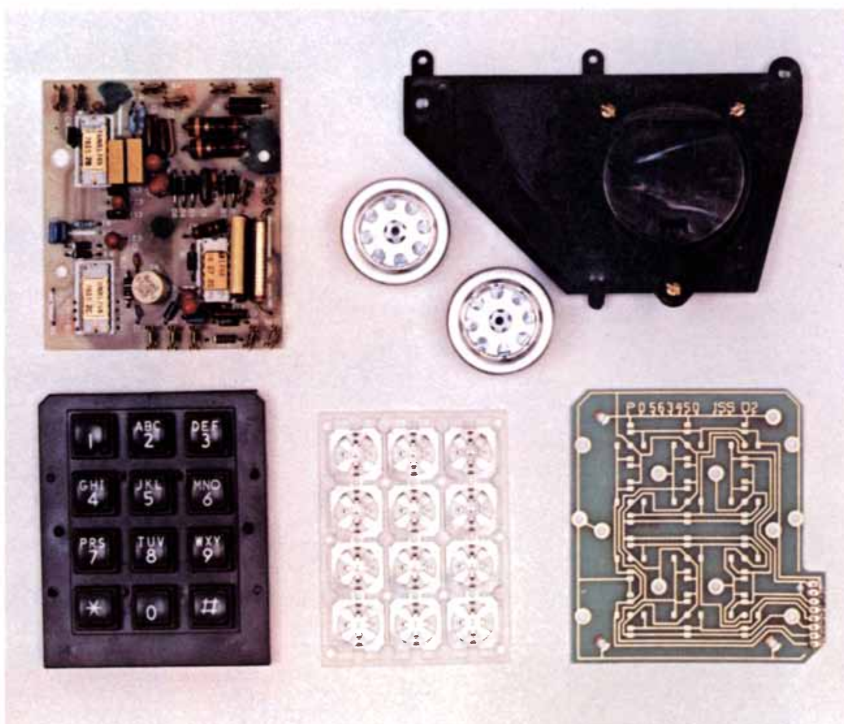
The standard telephone set has been designed to compensate for these losses in some measure. The compensation is termed equalization and is achieved by adding two varistors, or variable resistors, to the speech network. They offer a high resistance to low values of direct current and vice versa. The direct-current voltage from the central office is constant. The role of the varistors is to compensate for the fact that in short loops the current is relatively high, tending to make voices too loud, and that in long loops the current is relatively low, tending to force talkers to raise their voice.

The bell ringer of a telephone is a conventional electromechanical make-and-break circuit that drives a clapper against two bells. The hook switch serves to disconnect the voice network and the dial circuits when the set is not in use. It is held open by the weight of the handset resting in the cradle. When the handset is lifted, the speech network is connected to the loop, causing a direct current to flow. The current actuates a relay in the central office, and the relay connects the loop to the supervisory circuits. The central office returns an audible tone, indicating to the caller that the system is ready to receive the number of the person he wants to reach.

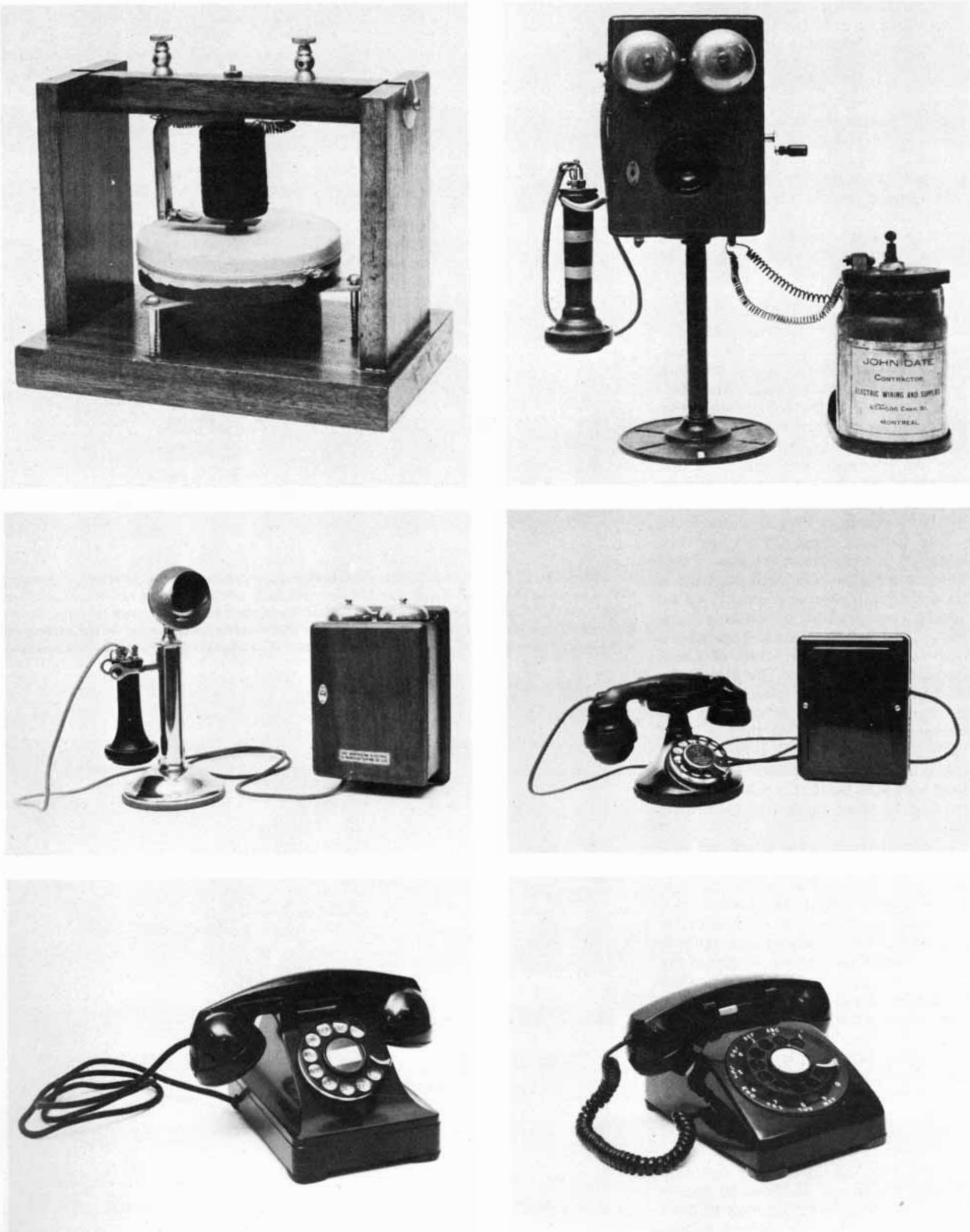
The traditional method of signaling the number is with the familiar rotary dial. If the caller turns the hole aligned with 6 to the end stop, a spring is wound and various electric contacts are actuated. When the dial is released, the speed of return is controlled by a governor so that only six break pulses are conveyed to the central office at the correct rate. The rate is typically 10 pulses per sec-



STANDARD TELEPHONE of the push-button type consists of the internal components shown here. At the top of the photograph one sees from left to right the bell ringer, the hook switch and the oscillator circuit for the electronic tones generated by the push buttons. The circular silver and black objects are the transmitter and receiver transducers. At the bottom left is the speech network. The other objects are some of the 120 parts of the push-button assembly.



ELECTRONIC TELEPHONE now on trial in the Bell Canada system has these major components. The bell ringer, the tone oscillator and the speech network have been replaced by three integrated circuits, which are contained in the unit at the top left. The push-button assembly has far fewer parts, and two identical dynamic transducers serve as transmitter and receiver.



EVOLUTION OF THE TELEPHONE is traced in photographs made at the Bell Canada historical collection in Montreal. Proceeding from left to right and top to bottom one sees a facsimile of the "gallows" telephone of Alexander Graham Bell; an early desk telephone employed before 1890; the two-part "candlestick" telephone; an early handset telephone; the first combined telephone (meaning that the bell and the speech circuit, which in earlier telephones were in a separate box that was usually mounted on a wall near the telephone, were in the housing of the instrument), and the "500 set," a dial instrument that was introduced in 1951 and is still standard in much of North America. Push-button telephone now in use is much the same as the 500 set except that a key pad has replaced the dial.

ond, with an interruption of approximately a sixteenth of a second between one pulse and the next.

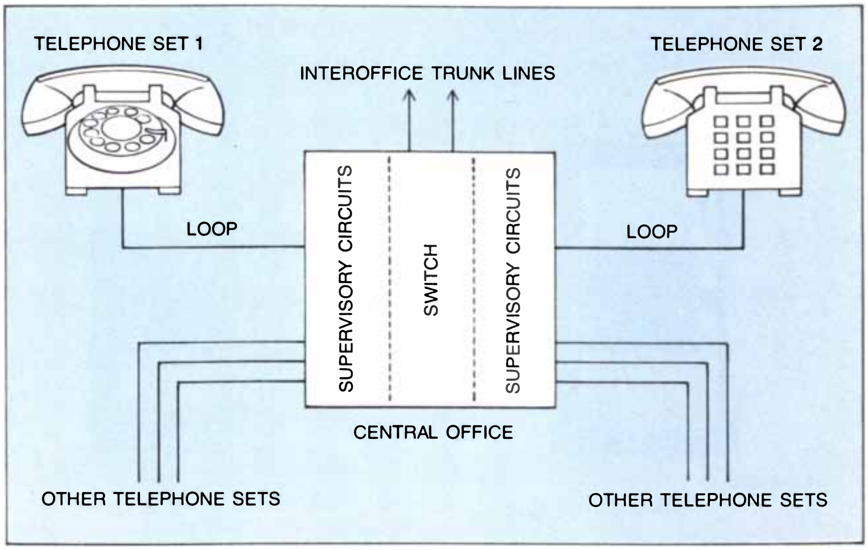
The push-button device works differently. Each time a button is pushed a transistor oscillator is turned on and a pair of tones (a different combination of two tones, selected from eight possible frequencies) is generated. A decoder in the central office detects and translates the tones.

Apart from push-button dialing no change in the basic design or performance of the standard set has been made in North America since 1953. Therefore the appearance of an all-electronic telephone, which is now on trial with Bell Canada, is a significant step forward. In this telephone the conventional electromechanical components such as the bell ringer, the transformer-coupled speech network and the tone-oscillator dial have been replaced by integrated-circuit "chips," one for each function. (Integrated circuits make it possible to interconnect thousands of transistors, diodes and resistors on a single piece of silicon, typically .15 inch square.) The three chips are interconnected with certain external components to make up the basic telephone set. Dynamic transducers replace the carbon transmitter and the magnetic receiver.

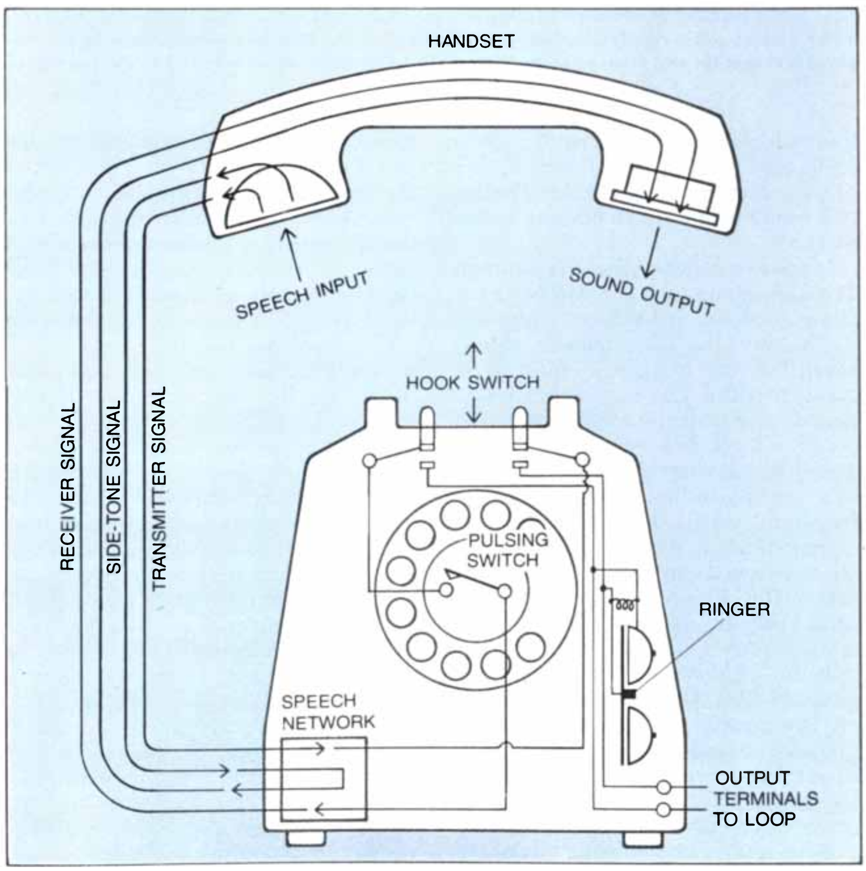
The dynamic transducer works on the principle that an electric conductor vibrating in a magnetic field generates a voltage at its terminals. The transducer consists of a thin diaphragm to which a small coil is rigidly attached. The action of sound waves on the diaphragm causes the coil to move in the radial field of a permanent magnet, thus generating an alternating-current signal. Although the dynamic unit is powered by sound (unlike the carbon microphone, which requires an electric current), the output signal is small; the unit therefore is assisted by an amplifier circuit to provide an adequate volume of sound.

The speech-network chip is designed to work with dynamic transducers, and so it includes amplifiers for both the transducer that serves as the transmitter and the identical transducer that serves as the receiver. The gain of the amplifiers is automatically adjusted in proportion to the variation in loop current between short and long loops. In this way speech levels are equalized to compensate for variations in loop lengths. A balancing circuit in the speech network provides the correct side-tone level. Additional circuits perform certain "handshaking" functions with the tone-generator chip.

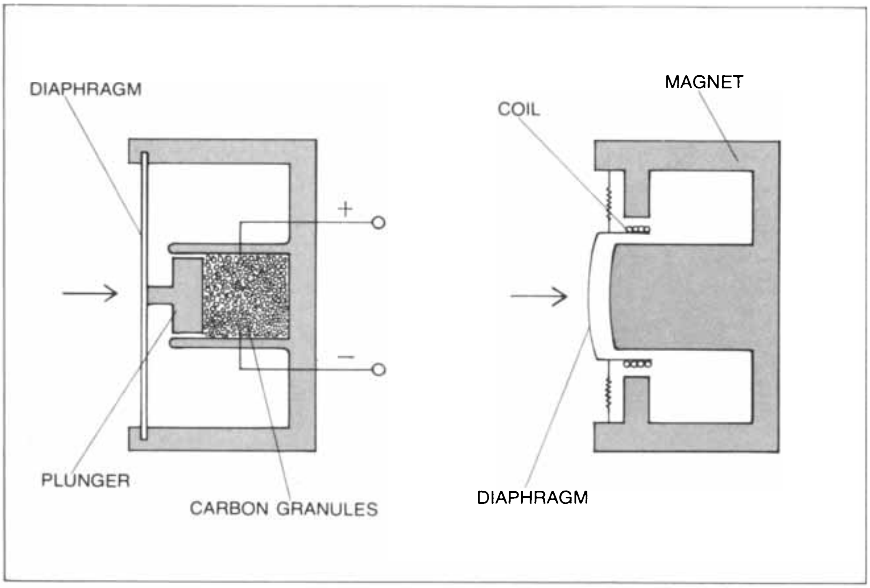
Tone signaling is based on an internationally accepted standard of frequencies that includes four tones of low frequency and three of high frequency. (A



TELEPHONE SYSTEM is depicted schematically and in a simplified way to show the basic functions required to make a telephone call. Each telephone set is connected by a copper wire called a loop to a central switching office. The telephone from which a call is originating signals to the office the number being called. Supervisory and switching circuits in the central office connect the two sets for the conversation and disconnect them when the call is over. The supervisory and switching circuits also provide power for the telephones, detect the demand for service when a handset is lifted, return a dial tone to the caller and send out the ringing signals.



OPERATION OF TELEPHONE is depicted in terms of a 500 set. The instrument consists of a handset, containing a transmitter in the mouthpiece and a receiver in the earpiece, that is connected to the base through a cord that carries four conductors. The base houses a speech network, which isolates the outgoing and incoming signals; a hook switch that connects and disconnects the telephone and the loop; a pulsing switch that is actuated by the rotary dial and signals a number to the central office, and the bell. The function of the side-tone circuitry is to control the amount of extraneous noise, including the talker's voice, reaching the earpiece.



OLD AND NEW TRANSDUCERS are portrayed. The carbon microphone at the left represents a design that was invented in 1876 and is still used in conventional telephone sets. Sound waves from speech cause the diaphragm to vibrate so that the carbon granules are compressed to a greater or a lesser extent depending on the volume of the incident sound. The carbon acts as a variable resistor: if a current is made to flow across it, the sound waves change the resistance of the carbon button and thus the amount of current flowing. The dynamic transducer at the right serves as both transmitter and receiver in the electronic telephone. It is based on the principle that a metallic conductor vibrating in a magnetic field generates a voltage. In the transducer a small coil is rigidly attached to a light diaphragm. Action of sound waves on the diaphragm causes the coil to move in the field of the magnet, thereby generating an electric signal.

fourth high-group frequency has been designated to allow for a fourth column of push buttons in the future. The key pad would then have 16 buttons instead of 12.)

Suppose a caller pushes the 5 button. The high-group tone is 1,336 hertz (cycles per second) and the low-group tone is 770 hertz [*see illustration on opposite page*]. The tone frequencies must be accurate to within 1.5 percent of their stipulated value under all conditions of temperature throughout the life of the set. The chip achieves this level of accuracy by a method of digital division. A high-frequency oscillator that serves as a master clock is divided down by two counters, one counter for each group of tones. The division is controlled by a digital encoder that causes the counters to divide down in a specific sequence in order to generate the correct tone frequencies. The dial pad operates by closing two contacts each time a button is pushed. A decoder on the chip senses the closings and sets up the division programs accordingly. The chip also includes a digital-to-analogue converter that translates the divided-down frequencies from a digital stream of 0's and 1's to an alternating sound signal, which is the tone one hears.

To illustrate the handshaking between the tone-generator chip and the speech-network chip let us follow a typical sequence of events in the generation of a tone signal. When a button on the dial pad is pushed, the keyboard decoder on the tone-generator chip signals a "common switch" on the network chip. This switch, which is so named because it performs a number of identical functions every time a button is pushed, returns power to the tone generator to start up the clock oscillator and the division circuits. Simultaneously it switches out the microphone, since voice signals on the line at the same time as tones could upset the tone receiver in the central office, and directs the tone signals to the amplifier in the transmitter. The common switch also substantially lowers the gain of the amplifier in the receiver, so that the caller does not get a sudden blast of sound through the earpiece. The entire sequence is reversed when the button is released.

The ringer chip does not actually make anything ring. Instead it generates two tones of slightly different frequency. (The selection of the frequencies presents a problem in human reactions: the sound should be subjectively pleasing, but it should also be just irritating enough to encourage the hearer to answer the call instead of lingering to enjoy the tone.)

The electromechanical telephone that serves most of North America is a fairly sophisticated instrument, the product of 100 years of development. A review of the major events in that course of development will indicate why the electronic telephone is only now able to replace the electromechanical device.

One must bear in mind the fact that the ownership of telephone equipment has been largely retained by the telephone companies. As a result the companies have a large investment of capital in their existing plant. Substantial advantages of cost and performance must be shown before any company would be willing to replace its installed equipment. Moreover, any new design must be fully compatible with the existing system.

Another consequence of the heavy capital investment is that equipment must maintain a high degree of reliability for a long time, typically for 40 years. For example, many central offices that were set up in the 1930's are still operating. Similar barriers await the new design at the manufacturing plant, where the obsolescence of established production techniques is not necessarily welcomed. How, then, have these factors been compatible with technological progress in the evolution of the telephone set?

The basic telephone functions of talking, listening, signaling and ringing were all developed before 1890 by Alexander Graham Bell and his contemporaries. The first significant innovation in the telephone beyond these basic designs came some 25 years later in the form of the rotary dial. The change was made not so much for technological reasons (the manual system was actually better than the dial system) but because the rapid expansion of the telephone system was accompanied by a growing shortage of operators.

The one-piece telephone set took even longer. Although it was conceived before 1900, it was not introduced on a significant scale until the late 1920's. Why did it take so long to make such an obvious improvement? The problem was familiar to the children of the time, who with the two-part set could make the telephone "sing" to itself by placing the receiver next to the transmitter. The howl was caused by acoustic coupling between the transducers by way of the side-tone circuit. The effect was intensified when attempts were made to put the receiver and the transmitter in the same handset, because then the sound could be coupled between them directly through the handle. It took considerable research to produce a nonsinging telephone.

Further development, mainly in the output of the transmitter and the receiver, led to the introduction of the 500 set. The push-button set, which was introduced in 1963, has essentially the same components as the 500 set except for the rotary dial. Nevertheless, the push-button set represents the beginning of the electronic-telephone era because it has a single transistor in the tone-generating circuit.

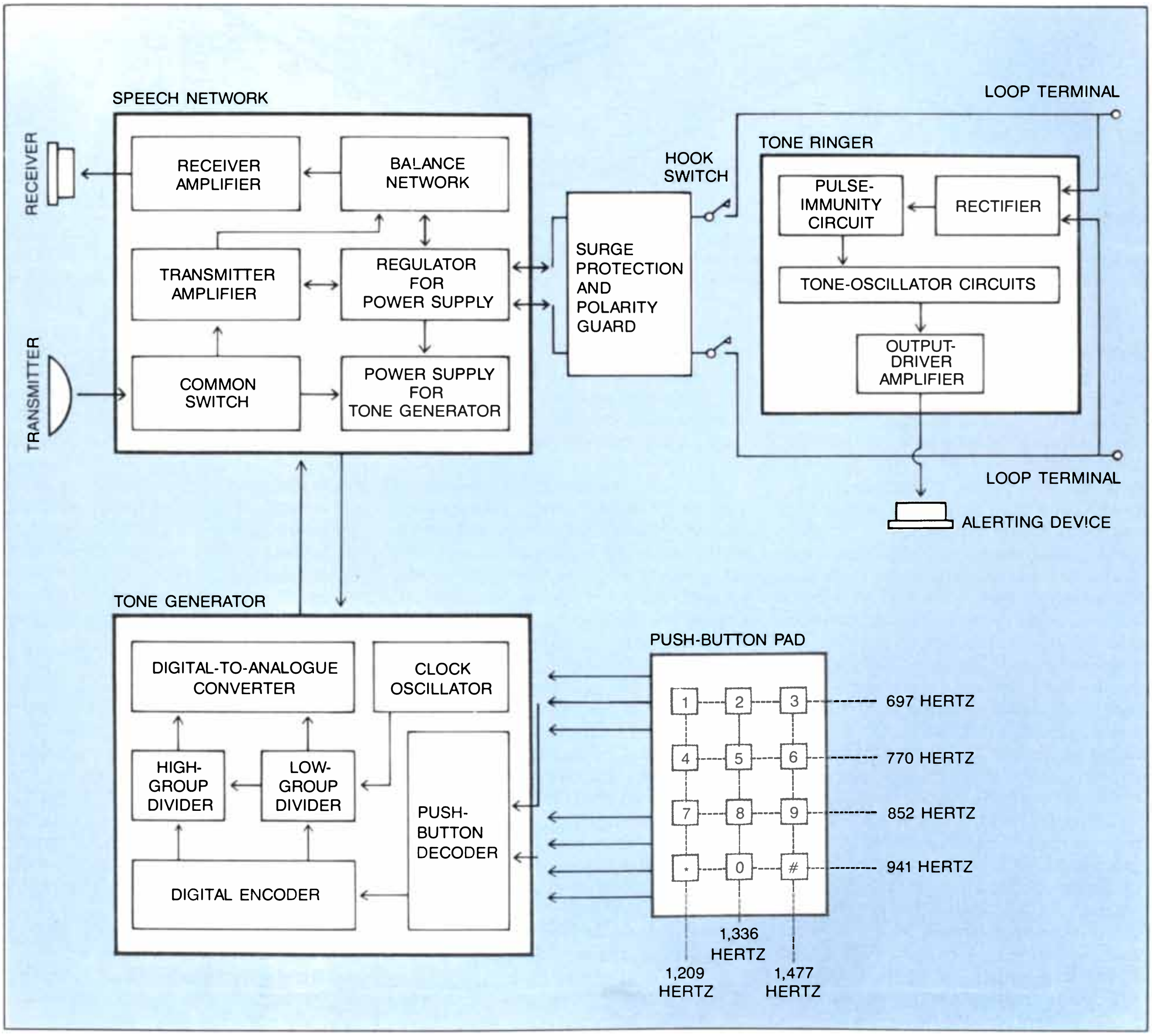
In 1967 a Norwegian design offered a significant departure from the conventional set by replacing the carbon microphone with a dynamic transducer and an electronic amplifier circuit. A further refinement was the electronic speech network incorporated in a Danish telephone introduced in 1973.

Since a new design will not be widely accepted unless it offers substantial advantages in either performance or cost, one is led to consider what advantages are offered by the electronic telephone. There are several. Let me start with performance.

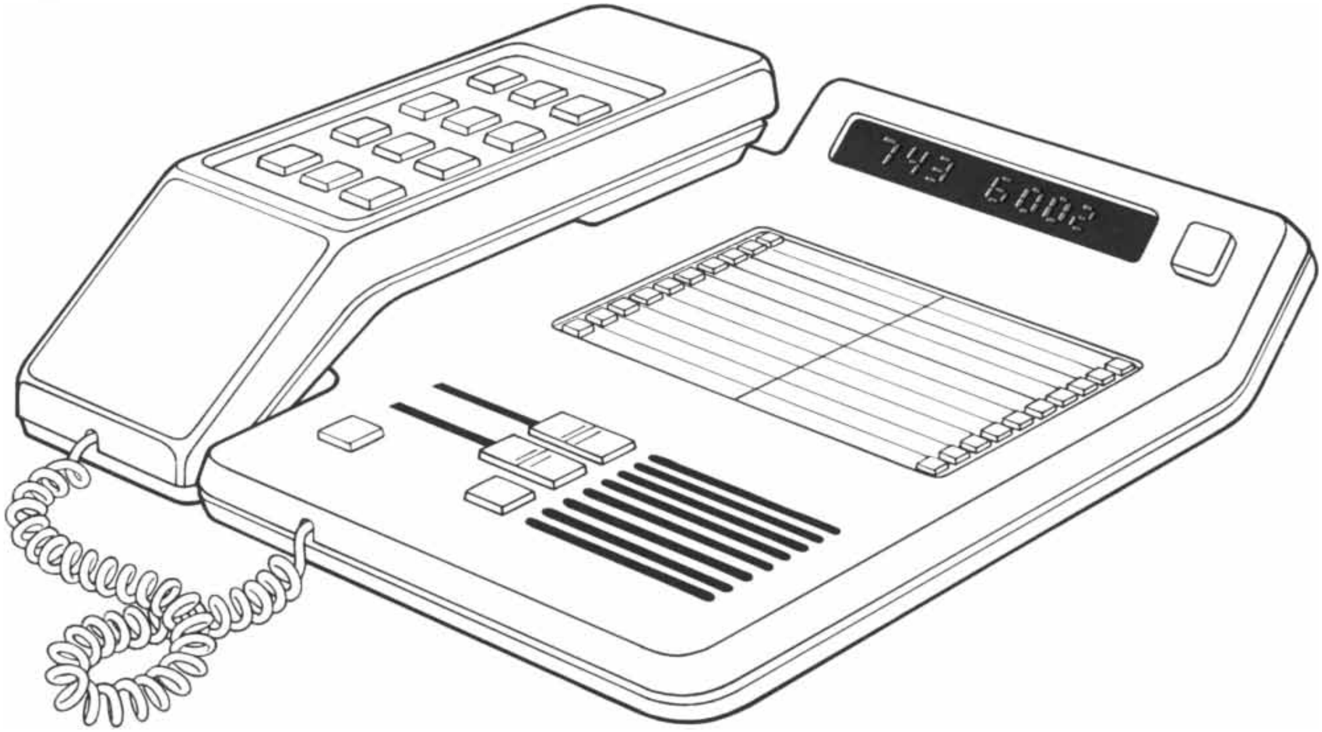
The carbon microphone of the standard set has several drawbacks. The efficiency of its output varies considerably with the loop current, the age of the device and the treatment it has received. (Some people still bang the handset in order to shake up the carbon granules in the hope of removing voice distortion and increasing the sound output of the microphone.) The replacement of the carbon transmitter with a linear dynamic microphone removes all these drawbacks and brings high-fidelity speech quality to the telephone.

The electronic telephone does much better than the 500 set at equalization. The reason in part is that the variation in transmitter efficiency is removed, but the electronic telephone also compensates almost completely for attenuation in long loops. Moreover, the electronic set does better at removing the echo on short loops.

Certain other advantages of the electronic telephone are only potential ones at present. The electronic set receives about five times more power than it needs on a long loop and 100 times more on a short loop. If the supervisory system could be redesigned accordingly, central-office economies would accrue in plant arrangements and in power consumption. As matters stand, however, the installed base of 160 million stan-



FUNCTIONS OF ELECTRONIC TELEPHONE are based on three "chips," or integrated circuits, one for the speech network, one for the tone generator and one for the tone ringer, which replaces the standard bell. The chips are connected to certain external components such as the hook switch and the protection unit. Also evident here is the two-tone system employed in a push-button unit, even one on a nonelectronic telephone. Each number on the dial pad generates a pair of tones, one tone of high frequency and one of low. For example, the tones for the 6 are respectively 1,477 and 770 hertz (cycles per second); the 7 button's two tones are 1,209 and 852 hertz.



NEW DESIGNS are made possible by the reduction in the number and size of the components of an electronic telephone. The design portrayed here is at the prototype stage and is well within the capability of the present technology. The telephone includes a digital display, which could serve among other things to time long-distance calls and to record the steps and results of calculations performed by way of the push buttons. It is also possible to record frequently called numbers, which then can be reached by pushing a single button.

dard telephones in North America will hold back such developments for a good many years.

The immediate economic benefits of the electronic telephone arise from the substantial reduction in the size and number of separate components it requires. The most dramatic example is in the push-button equipment. The standard push-button telephone has more than 120 separate parts in the push-button assembly alone. The reason is that a complicated system involving push rods and springs is needed to provide the common-switch function. The standard push-button unit also needs two ferrite-core inductance transformers for the transistor-oscillator circuit, so that the resulting unit is bulky and expensive. Moreover, the transformers have to be tuned individually during manufacture to bring the tone frequencies into the narrow ranges established for the various numbers. The electronic telephone performs the common-switch function electrically on the network chip. It therefore has a simple dial-pad construction consisting of only a few parts.

The tone ringer, the speech network and the protection circuits of the electronic telephone also have fewer parts than their electromechanical equivalents and are smaller. The total saving achieved by the electronic set is some 180 parts. Since most of the remaining components are capacitors, resistors and integrated circuits, the kind of low-cost, highly automated assembly and test methods developed by the consumer electronics industry can now be applied to the manufacture of telephone sets.

Although the electronic telephone owes much to the recent advances in integrated-circuit technology and in the performance of semiconductors, innovations in other components have also played a part in reducing costs. For example, the frequency of the master-clock oscillator employed in the tone generator must be stabilized to within .2 percent. In the past this strict requirement could be met only with a precisely cut quartz-crystal resonator. The cost of such a crystal ranges from 80 cents to $5, depending on the desired center frequency. Much cheaper ceramic resonators have recently been developed. Such a resonator costs about 30 cents and performs just as well as a crystal in the tone generator. Another component, which protects the telephone against surges of current, has undergone a threefold cost reduction in the past two years.

The electronic telephone offers more benefits than improved performance and reduced manufacturing costs. The shape of the telephone, which must now accommodate the bulky electromechanical components, can be redesigned in a variety of ways, limited almost solely by the imagination of the designer. Many features that can now be obtained only with expensive additional equipment will be integral to the electronic telephone. For example, by means of an inexpensive memory chip it will be possible to reach prestored numbers by pushing a single button. No longer will it be necessary to search for the number of the police station or the fire department in an emergency; each number can be stored. If a number is busy, it will remain in the memory and can be tried again by pushing one button.

For only a few more dollars one could add a display panel resembling the one on an electronic calculator. With such a panel one could time long-distance calls, verify stored telephone numbers and perform calculations on the keyboard. Another option for the user of an electronic telephone is that he can vary both the volume and the sound quality of the tone ringer to suit the situation, achieving, say, a soft, mellow tone in the bedroom and a hard, shrill tone in the workshop. A residential "hold" feature that would enable the user to transfer a call from one extension to another without having to return to the first one later to replace the handset can be incorporated easily in the electronic module. Eventually a "no hands" capability that would enable the user to talk and listen without the handset might be built into every telephone.

THE LEAST REMARKABLE THING ABOUT THE NEW RABBIT DIESEL IS 53 MPG H'WAY, 40 MPG CITY.*

You've read it right, friends.

According to the EPA* estimates the new Rabbit Diesel gets the highest mileage of any car in America. (Of course, mileage may vary depending on how and where you drive, optional equipment, and your car's condition.)

But if by chance you're looking for more, read on. Because the most astonishing news about our economy car isn't the economy. It's the car.

Remarkable thing #1: eye-opening performance.

Are you the kind of person who gets a thrill out of zipping from 0 to 50 in a mere 11.5 seconds? Well, thanks to an efficient use of aerodynamics and weight, you'll be ecstatic in a Rabbit. In fact, it's already set 31 world records for Diesels.

You'll also be thrilled to know the 1978 Rabbit comes with such things as an "independent stabilizer rear axle" which manages to combine the stable tracking of a rigid rear axle with the smoothness of an independent suspension.

"Negative steering roll radius" which helps maintain directional stability even in the event of a front-tire blowout.

"Front-wheel drive" for better tracking, especially in high winds and rotten weather.

Front disc brakes. Radial tires. Rack-and-pinion steering. Breathless cornering.

Do we still sound like an economy car? If so, you're ready for remarkable thing #2:

More room for people than 40 other cars.

A Volkswagen Rabbit looks smaller than other cars, right?

But inside, our engineers cleverly devoted 87% of the interior to functional room.

Open the trunk and (believe it or not!) there's more luggage space than a Cadillac Seville.

Fold down the rear seat and (amazingly!) there's almost as much luggage space as some station wagons.

Then open the door. Your eyes don't deceive you. There's more people space than Chevy Monza, Datsun 510, Pinto Wagon and 37 other cars you could buy.

But wait. While you have the door open, notice remarkable thing #3; a stroke of sheer genius:

The seat belts actually put themselves on.

No fumbling about on the floor trying to find them.

No mumbling about what a pain normal seat belts are.

It's like magic.

Just close the door and they're on.

This type of passive restraint system will be mandatory in 1984. And only a Model "L" Rabbit has it now.

Another stroke of genius: a cooling fan with brains.

When it's freezing out and a fan isn't needed, our cooling fan knows enough to shut itself off. (That saves you noise and energy.)

When it's boiling out, our cooling fan has the good sense to keep running even after the car is shut off. (Because that's an important time to protect your engine against overheating.) Then it automatically stops when the engine is cooled off.

Last but least, a word about money.

Happily, all Rabbits are frugal when it comes to money.

The problem is, which Rabbit should you buy?

A gas-powered Rabbit — which is a wonderful car to begin with.

Or a Rabbit Diesel — which costs about $200 more than our "C" and "L" gas models.

What do you get for $200?

For one thing, diesel fuel costs about 10¢ a gallon less than gasoline.

For another, a diesel engine never needs a major tune-up. Because there are no spark-plugs, points, condensers, or carburetors to tune.

All this, and great mileage, too.

A tough choice, to be sure.

But then, a Rabbit is the only car in its class that gives you a choice at all.

What could be more remarkable than that?

VOLKSWAGEN DOES IT AGAIN

*EPA Gas Mileage Guide

SCIENCE AND THE CITIZEN

Bill of Health

Americans are living longer and paying more for the privilege. The two trends are related: older people are subject to higher rates of illness and disability and require more health services than others, and expenditures for people over 65 have increased in the past 10 years more rapidly than expenditures for younger people, in part because Medicare has made health services increasingly available to the elderly poor. In *Health, United States, 1976–1977,* the second report in an annual series mandated by Congressional legislation in 1974, the Department of Health, Education, and Welfare emphasizes the interrelation of longevity and health-care spending and its significance for the future in view of the projected growth of the elderly population.

The U.S. death rate was 8.9 per 1,000 of population in 1975, the lowest ever in this country. (The age-adjusted rate was 1.8 times as high for men as it was for women, 1.4 times as high for nonwhites as it was for whites; black children less than five years old had a death rate twice as high as that of white children.) On the basis of current death rates the average U.S. life expectancy at birth was 72.5 years in 1975 (up from 70.9 years in 1970); it was 77.2 years for a white female and 69.4 for a white male, 72.3 for a nonwhite female and 63.6 for a nonwhite male. Life expectancy at age 65 increased more between 1950 and 1975 than it had in the first half of the century. It was 11.9 years in 1900, 13.8 in 1950 and 16 years in 1975. Most of the gain was among women. Because the rates at every age are higher for men than they are for women, the sex ratio is skewed: although 105 males are born for every 100 females, among people over 65 there are only 69 men per 100 women. Longevity and a sharp reduction in the birth rate (from 19.4 per 1,000 in 1965 to 14.8 in 1975) have combined to increase the proportion of the population aged over 65: from 9 percent in 1955 to 10.5 percent in 1975 and a projected 11.5 percent in the year 2000. In 1900 there were only 3.1 million people 65 and older in the U.S.; in 1975 there were 22.4 million, or proportionately more than twice as many; by 2030 there will be 55 million, according to current projections, and probably more as mortality rates decline further. Substantial reductions since 1950 in heart disease and cerebrovascular disease (stroke) account for most of the decrease in mortality among the elderly; the death rate from cancer rose 12 percent between 1950 and 1975.

Expenditures for health care more than tripled between the fiscal years 1965 and 1976, from $39 billion (5.9 percent of the gross national product) to $139 billion (8.6 percent of the G.N.P.). About half of the increase can be attributed to inflation in the prices of medical goods and services, 40 percent to greater per capita utilization of health services and 10 percent to population growth. The rate of increase in spending has been more for the elderly than it has for other people. From fiscal 1966 to 1976 the average annual rate of increase in per capita expenditures was 13 percent for those 65 and older and 11 percent for those under 65. The average per capita health-care bill in 1975 was $1,360 for the elderly, $472 for people aged 19 to 64 and $212 for those under 19. The source of these funds changed dramatically with the advent of Medicare. In fiscal 1966 only 30 percent of the money spent for health care for people 65 and over came from public funds; in 1976 the public share was 68 percent.

The Support of Science

"Shortly after taking office," said President Carter in his budget message to Congress, "I determined that investment in basic research on the part of the Federal Government had fallen far too low over the past decade." As a result his new budget calls for an increase of 10.9 percent in the funds for basic science. This represents a real increase of about 5 percent after subtracting an expected 6 percent loss due to inflation. The total figure for research and development in the proposed budget for fiscal 1979 comes to $27.9 billion, in effect a "no growth" increase of 6.1 percent over the $26.3 billion for fiscal 1978. The proposed research-and-development total is divided as follows: $3.65 billion for basic science (up 10.9 percent), $6.7 billion for applied science (up 7.4 percent) and $17.5 billion for development (up 4.6 percent). The small increase in funds for development represents an actual decrease in constant dollars. To achieve the reduction the Administration proposes to eliminate several development projects no longer considered vital, notably the fast breeder reactor and demonstration plants involving costly technologies (such as the gasification of coal).

In reaching $3.65 billion the proposed budget for basic science measurably exceeds for the first time the record Federal expenditure for basic science achieved in 1967 (when the expenditures are compared in constant 1972 dollars). The 1967 expenditure in constant 1972 dollars was $2.177 billion. That sum was barely equaled by last year's budget. The proposed 1979 budget for basic science, in the same constant dollars, is higher by $102 million.

More than 80 percent of the proposed funds for basic science are allocated to five Federal departments or entities: the National Institutes of Health, $856 million (up 12.2 percent); the National Science Foundation, $755 million (up 9.7 percent); the National Aeronautics and Space Administration, $520 million (up 11.1 percent); the Department of Energy, $468 million (up 8.1 percent), and the Department of Defense, $364 million (up 13.4 percent). In the recent past some 47 percent of all Federal funds for basic research have been spent in the nation's colleges and universities. It seems likely, however, that they would receive more than 47 percent of the proposed increment of $360 million in basic science funds and therefore would enjoy an increase of more than 10.9 percent in basic science grants from the Government. (For example, academic institutions would receive virtually all the $160 million increase for basic science that the proposed budget allocates to the National Institutes of Health and the National Science Foundation and a large part of the extra $95 million allocated to the departments of Defense, Energy and Agriculture. According to Frank Press, the President's science and technology adviser, "the growth in basic research support provided in the 1979 budget... is intended to encourage innovative research and to assist in ameliorating some of the problems currently associated with the performance of research in colleges and universities, such as the growing obsolescence of equipment and the lack of opportunities for young investigators."

The new budget identifies several areas of national concern for substantial increases in support. Funding for research on climate would increase nearly 40 percent from $75 million to $104 million. The Department of Agriculture would have $30 million, or 100 percent more than last year, for basic research to raise the productivity of major food crops and to improve knowledge of human nutrition. The budget of the National Institute of Child Health and Human Development would be increased 26 percent to $197 million. The budget of the Department of Energy contains $265 million for construction of the intersecting storage rings (ISABELLE) at the Brookhaven National Laboratory (see "Head-On Physics" below). The National Science Foundation will be enabled to continue construction of the Very Large Array radio telescope in New Mexico. The NASA budget will include funds for only four space shuttles instead of the five originally proposed, thereby making more money available for scientific projects. These include the Solar Polar mission, in which a spacecraft will pass over and explore the sun's poles for the first time, and the Solar Mesopheric Explorer, which will study the effect of solar radiation on the earth's ozone layer. Work will continue on the large space telescope (to be placed in orbit by shuttle), on the *Landsat-D* earth-resources-survey satellite and on a spacecraft to go into orbit around Jupiter and probe its atmosphere.

A New Eminent Domain

The projection of large increases in coal production in the U.S. during the coming decades has brought forward proposals for carrying substantial amounts of coal as a slurry in pipelines. Over a given route the method would be an alternative to, or perhaps a supplement to, transportation by railroad. The pipeline interests, facing difficulty (sometimes from railroads) in obtaining the property rights they need, have urged Congress to enact legislation that would enable pipeline enterprises to obtain rights of way by eminent domain. The Congressional Office of Technology Assessment, asked by several committee chairmen to look into the matter, has concluded that "coal slurry pipelines do represent under some specific circumstances the least costly available means for transporting coal" and that without a right of eminent domain pipelines would have difficulty competing with railroads.

The finding should not be taken as a recommendation, since the office points out in a 190-page report, *A Technology Assessment of Coal Slurry Pipelines,* that a decision depends on the weights that are assigned to other issues. The report identifies three major groups of issues. The first group concerns the desirability from social, economic and environmental standpoints of developing an industry for carrying coal by pipeline. The second concerns the extent to which the present regulatory and institutional arrangements would have to be changed to provide for the allocation of coal traffic between pipelines and railroads in a way that would represent the least cost to society. The third concerns the balance between state and Federal control over such matters as the allocation of water resources, the ownership of land and the protection of the environment.

The study examines the issues under 11 headings. Among other things, it finds that under the assumption that pipelines would carry some 200 million tons of coal per year by the end of the 20th century the railroads might experience slower increases in coal revenues than they could expect otherwise, although the competitive threat would be less than the threats posed by adverse regulatory policy or reduced rates of improvement in railroad productivity. Another finding is that the water-related impacts of coal-slurry pipelines can be diminished if sources of water can be found that are suitable for transporting coal but not for most other purposes. The possibilities include saline ground water, treated sewage effluent and return flows from irrigation. If it is determined that coal-slurry pipelines should be built, the report says, some kind of eminent-domain authority "would appear to be a necessity." If the aim is to encourage the rapid development of pipelines, "Federal legislation has some clear advantages," but if the intent is to let the states keep "the power to protect their interests as they perceive them, then preemption by a Federal statute would be undesirable."

That Old Chondrite Moon

Carbonaceous chondrites are a class of meteorites rich in volatile compounds that are believed to have condensed directly out of the primordial nebula from which the sun and the planets were formed; thus they are samples of the most primitive material in the solar system. Close examination of the larger of Mars's two moons, Phobos, by the spacecraft *Viking Orbiter 1* has now revealed that it may be composed of carbonaceous chondritic material.

According to three reports by a large group of investigators in *Science,* during the last two weeks of February, 1977, *Viking Orbiter 1* was guided so that it had 17 close encounters with the tiny Martian moon, an ellipsoid measuring roughly 27 by 20 kilometers, approaching it to within 300 kilometers each time. During these encounters the spacecraft made some 125 television pictures with a resolution of 30 meters or less. It also observed the surface of Phobos at infrared wavelengths and gathered other valuable data.

During the encounters the physical appearance of Phobos and reflectance spectra of it revealed that the surface of the Martian moon is quite different from that of the earth's moon. The lunar surface consists of basaltic materials. The reflectance spectra of Phobos are similar to reflectance spectra of the asteroids Ceres and Pallas, which are composed of material similar to that of carbonaceous chondrites. Observations of the effect of Phobos' gravity on the trajectory of *Viking Orbiter 1* revealed that the density of the Martian moon is lower than had been supposed: $2.0 \pm .6$ grams per cubic centimeter.

The investigators write: "On the basis of a good match between the reflectance spectra of Phobos and Ceres, and taking advantage of the extensive research done to identify the surface composition of this asteroid, we conclude that the surface composition of Phobos is also similar to that of carbonaceous chondrites." Moreover, the low measured density of Phobos "is consistent with that of type 1 or type 2 carbonaceous chondrites, but it definitely rules out basalts which have densities close to 3 [grams per cubic centimeter]."

The uncertainty in the density value still allows the possibility that Phobos might be made up of type-2 carbonaceous chondritic material, although the weight of the evidence seems to point to type-1 material. Type-2 carbonaceous chondrites contain chondrules: tiny spherical inclusions that are believed to have once been molten droplets. The presence of chondrules requires that at one stage the meteoritic material was subjected to a temperature high enough to melt part of it. Type-1 carbonaceous chondrites contain no chondrules but are classed as chondrites because they are chemically and mineralogically similar to the chondrule-bearing type-2 carbonaceous chondrites.

The evidence that Phobos may be a huge carbonaceous chondrite will bear on hypotheses about the satellite's history and origin. In recent years it has generally been thought that Phobos is a natural satellite of Mars that formed along with the planet. This hypothesis has been popular mainly because of the characteristics of Phobos' orbital motion. "The compositional results . . . may force a reevaluation of the natural-satellite hypothesis of Phobos' origin," the investigators conclude. Calculations by John S. Lewis of the Massachusetts Institute of Technology and others regarding the formation of bodies in the solar system indicate that at Mars's distance from the sun the temperature of the primordial solar nebula may have been too high for volatile-rich carbonaceous chondritic material to have condensed. If the calculations are correct, they imply that Phobos formed in the outer regions of the asteroid belt and was then captured by Mars.

Head-On Physics

Just as a collision between automobiles is most violent when the cars meet head on, so high-energy interactions of elementary particles are observed most efficiently in machines where collisions are arranged between particles moving with equal speed in opposite directions. Three large machines of this kind are operating now, and their success has evidently influenced the design of the next generation of accelerators. Of the high-energy machines now planned or already under construction at least six will provide collisions between opposed beams of particles.

The principal advantage of head-on interactions is that virtually all the energy invested in accelerating the particles is released in the collision. Protons are now routinely accelerated to an energy of 400 billion electron volts (400 giga-volts, or GeV). When such a particle

We're the only U.S. airline that can honestly promise you the world.

It's true for the simplest of reasons: we're the only airline that flies to all parts of the world from America.

84 cities in 58 countries on 6 continents.

Our flights, even to the most far-flung places, are scheduled frequently enough so that every three minutes, somewhere in the world, a Pan Am plane is either taking off or landing.

And more often than not, that Pan Am plane, whether you're a passenger or you're shipping cargo, is a big, roomy 747. The plane, we're told, that's the most preferred by international travelers.

The next time you're traveling abroad, ask your travel agent to put you on Pan Am.

It'll make things a lot easier. Promise.

PAN AM
America's airline to the world.

See your travel agent.

collides with a proton at rest, however, less than 28 GeV is made available for the creation of new particles. If two protons, each with an energy of 400 GeV, could be made to collide head on, essentially all their energy (800 GeV) would be liberated. To reach the same effective energy by bombarding a fixed target would require an accelerator with an energy of some 350,000 GeV (350 teravolts, or TeV).

Head-on collisions are arranged by storing counterrotating beams of particles in concentric or intersecting rings. After the rings are filled by an auxiliary accelerator the beams circulate stably for hours. A price must be paid for the high effective energy of the colliding-beam devices: the rate of interactions is quite low compared with that of a fixed-target accelerator. The beams usually pass through each other without interacting at all, and in existing machines a collision of interest is observed only once every few seconds. For this reason the number of particles stored and the confinement of the beam to a small cross-sectional area are important in determining the performance of the machine.

Two of the large rings that are operating now store counterrotating beams of electrons and their antiparticles, positrons. One is at the Stanford Linear Accelerator Center (SLAC) in California and the other is on the site of the German Electron Synchrotron (DESY) near Hamburg. It is in these machines that the new family of particles with the property called charm have been observed most clearly. The ring at SLAC has also produced events that suggest the existence of a new lepton, or particle related to the electron (see "Heavy Leptons," by Martin L. Perl and William T. Kirk, page 50). Both rings have a maximum energy of about 4 GeV per beam, and both are soon to be replaced by larger machines.

A new German storage ring, to be called PETRA, will employ the existing ring as part of an injector system. The maximum energy available will be 19 GeV per beam, for a total of 38 GeV. Construction is already well under way (an eighth of the ring has been installed and tested), and the project should be completed in about a year. The new electron-positron storage ring at Stanford, which will be called PEP, is being built at the end of the SLAC two-mile electron accelerator. The new ring will have a circumference of 2.2 kilometers, roughly 10 times the size of the present colliding-beam machine. Current plans call for a maximum energy of 18 GeV per beam, but a proposal has been made to raise the limit to 24 GeV. Collisions will take place at six interaction regions, five of which will be available for physics experiments. A few initial experiments have already been approved, and the particle detectors required for them are being built. PEP itself will be completed in 1979 or 1980, depending on the rate at which money is made available.

Smaller storage rings for electrons and positrons have been operating for some time at the Frascati laboratories near Rome and at the Serpukhov Institute for High-Energy Physics near Moscow. Three more small rings are now under construction. At the Orsay laboratories near Paris a machine with two tangent rings will be installed. At Cornell University an electron synchrotron is being converted to store beams of electrons and positrons with energies of from 1.5 to 10 GeV. A ring called VEPP-4, capable of 7 GeV per beam, is nearing completion at Novosibirsk in the U.S.S.R.

All the foregoing machines store electrons and positrons; the one remaining large ring that is operating now stores beams of protons. This is the facility called the Intersecting Storage Rings (ISR) at the European Organization for Nuclear Research (CERN) near Geneva. It is capable of a maximum energy of 28 GeV per beam.

A much larger machine of the same kind, called ISABELLE, will be built at the Brookhaven National Laboratory on Long Island. ISABELLE will consist of two interlaced rings, each ring capable of supporting a proton beam that can be adjusted continuously over a range of energies from 30 to 400 GeV. Not only these higher energies but also the number of protons stored will represent an improvement over the performance of earlier accelerators. The circulation of the protons in each ring will be equivalent to an electric current of some eight amperes. The beams will be confined to the rings and will be focused by 1,084 superconducting magnets installed in a tunnel 3.7 kilometers in circumference.

At the Fermi National Accelerator Laboratory (Fermilab) near Chicago, which now has only a fixed-target accelerator, two colliding-beam projects are under consideration that might reach energies even higher than those of ISABELLE. A new ring of superconducting magnets (capable of accelerating protons to 1 TeV) is being installed in the same tunnel as the existing proton synchrotron. By storing protons in both the new ring and the old one collision energies of 1.3 TeV might be achieved. Another possibility is to simultaneously store protons and antiprotons in the superconducting ring. In principle an energy of 2 TeV could then be attained, but techniques for creating stable, intense beams of antiprotons are still being investigated.

A proton-antiproton storage ring is also under study at CERN, along with several other proposals. One of these is notable for being the largest of the contemplated storage devices. It would be an electron-positron ring an order of magnitude larger than PEP and PETRA, with a circumference of more than 40 kilometers and a maximum energy of 200 GeV per beam.

It is impossible to say what will be found with the new instruments, but there are two matters left unsettled by studies at lower energy that will surely be given high priority. One is the study of quarks, the elementary particles thought to make up the structure of the proton and many related particles; the new electron-positron rings might settle the question of how many kinds of quark there are. Another is the search for the particles labeled *W* and *Z*, which are thought to transmit the "weak" force of nuclear beta decay. The *W* and *Z* are perhaps the most highly prized trophies of contemporary physics. With estimated masses of from 50 to 100 GeV they are probably beyond the range of PEP and PETRA, but they should be seen with ISABELLE or with the storage rings being considered at Fermilab and CERN.

Microbial Magnets

A new lead in the case of the north-seeking bacteria was disclosed recently by two workers engaged in the investigation. The story goes back to the fall of 1975, when Richard P. Blakemore, a microbiologist now at the University of New Hampshire, reported a curious discovery. As he was examining samples of mud collected from freshwater and marine ponds in the vicinity of Woods Hole, Mass., Blakemore noted in his microscope that certain kinds of bacteria consistently swim north when they are separated from the sediments. He found that their direction of swimming could be readily changed by moving a small bar magnet up to the microscope slide.

Blakemore joined forces with Adrianus J. Kalmijn, a specialist in the sensory biophysics of marine vertebrates at the Woods Hole Oceanographic Institution, to establish that the bacteria also orient toward the north in an artificially generated uniform magnetic field comparable in strength to the earth's magnetic field. Furthermore, Blakemore and Kalmijn were able to show that when the bacteria are outside the laboratory environment (in a nearby woods free of man-made distortions of the geomagnetic field), they actually follow the steeply inclined lines of force of the geomagnetic field. The north-seeking tendency of the bacteria was particularly evident when the ambient magnetic field was reversed with the aid of large Helmholtz coils: the single-cell organisms immediately changed course, making 180-degree *U*-turns several cell diameters wide. They realigned themselves with the field in a matter of seconds, this time heading in the opposite direction. On the basis of the mag-

IF YOU'RE GOING TO BUY AN EASY-TO-USE CAMERA, MAKE SURE IT'S REALLY EASY TO USE.

Like many of the new, compact 35mm reflex cameras, the Minolta XG-7 is automatic. You simply point, focus and shoot. The XG-7 sets the shutter speed up to 1/1000th of a second. And you get perfectly exposed pictures, automatically.

But easy operation is more than just automatic exposure. Here's what to look for when you compare cameras at your photo dealer.

Easy focusing. The XG-7's viewfinder is big and bright, even in the corners. Your subject snaps into critical sharpness.

It's easy to be creative. You can make the automatic exposure setting brighter or darker for creative effects.

An easy-to-understand electronic viewfinder. Light emitting diodes tell how the XG-7 is setting itself and warn against under- or over-exposure.

An easy-to-see electronic self-timer. The self-timer lets you get into your own pictures. It's a large flashing light mounted on the front of the camera. The flashing speeds up when the picture is about to be taken.

An easier-to-use auto winder. It automatically advances film, as fast as two pictures a second. You attach the optional Auto Winder G without having to remove (or lose) any caps from the XG-7.

The easier-to-be-creative flash. The optional Minolta Auto Electroflash 200X synchronizes continuously with the winder. This exclusive feature allows you to take a sequence of up to 36 flash pictures.

The important "little" extras. The XG-7 has a window that shows when film is advancing properly. A memo holder holds the end of a film box as a reminder. There's even an optional remote control cord.

Fast, easy handling. The way a camera feels has a lot to do with how easy it is to use. Is it comfortable or awkward? Are the controls placed where your fingers naturally fall, or are they cramped together? The Minolta XG-7 is human engineered for comfort and smooth handling. It's quiet, with a solid feeling you find only in much more expensive equipment.

Easy-to-change lenses. Remove or attach lenses with less than a quarter turn. And a system of almost 40 different lenses, from fisheye to super-telephoto, makes the XG-7 a key to virtually unlimited creativity.

Try the Minolta XG-7. At your dealer. Or write for literature to Minolta Corp., 101 Williams Dr., Ramsey, N.J. 07446. In Canada: Minolta Camera (Canada) Inc., Ont.

Minolta products are featured in the new 20th Century-Fox motion picture "HIGH ANXIETY" starring Mel Brooks.

netic behavior of the bacteria and Blakemore's finding that they contain iron-rich particles, the two investigators tentatively assumed that their specimens had a permanent magnetic-dipole moment; in other words, they resembled a miniature bar magnet.

The latest word from Kalmijn and Blakemore is that they have managed to substantiate this novel hypothesis experimentally by subjecting the bacteria to a comparatively strong magnetic pulse. The pulses were superimposed on the earth's steady magnetic field, and the vertical component of the earth's field was nulled in order to have the bacteria for convenience move only in the horizontal plane. When the pulse was antiparallel to the horizontal component of the earth's field (and was sufficiently strong), the experimenters observed that the applied field instantly caused the bacteria to turn around and swim in the opposite direction (that is, straight to the south), which is what one would expect from the reversal of their intrinsic magnetic-dipole moments.

It was found that to affect 50 percent of the freshwater specimens a pulse strength of approximately 375 to 400 gauss was needed. For the marine species the corresponding value was significantly higher: between 525 and 550 gauss. Depending on the strength of the applied magnetic pulse, the bacteria either remained northbound or became southbound; none were completely depolarized. In that respect, Kalmijn and Blakemore point out, the test organisms exhibit the properties characteristic of single magnetic domains in a ferromagnetic material. They conclude that although the biological role of this response remains to be determined, the magnetic behavior of these bacteria provides direct evidence of ferromagnetic orientation in nature.

The Pharaoh's Wine Cellar

The Egyptians were growing grapes and pressing wine long before Bacchus brought his gift to the Greeks; sealed wine jars are among the royal treasures of the First Dynasty (about 3000 B.C.). Some 1,500 years later, by the time of Amenhotep III (1413–1379 B.C.), the wine-jar inscriptions had become even more informative than the labels on the best bottles of today. Most of the inscriptions identified the year of pressing (modern equivalent: vintage year), the geographic locale (modern equivalent: *appellation contrôlée*), the vineyard ownership (modern equivalent: chateau) and the chief vintner responsible for the product (modern equivalent: none).

So concludes Leonard H. Lesko of the University of California at Berkeley, by profession an Egyptologist and by avocation a student of wine. Responding to the upsurge of interest in ancient Egypt accompanying the touring exhibition of the treasures of Tutankhamen, Lesko recently published a 46-page study of Egyptian wine making, "King Tut's Wine Cellar," which features his analysis of the inscriptions on 24 wine jars recovered from the pharaoh's tomb.

When Howard Carter entered the burial chambers of Tutankhamen (reigned 1361–1352 B.C.) in 1922, Lesko notes, he found in the storage annex 36 wine jars, 26 of which were inscribed. Two of the inscriptions were incomplete, but most of the other labels, written in hieratic script, specified the year, the district, the vineyard owner and the chief vintner. Additional information on four of the jars indicated that the wine was sweet, suggesting that Egyptian wine was usually dry.

Of the 24 complete labels 22 either identify the grape district as being the "Western River," the part of the Nile delta that today surrounds Alexandria, or specify locations elsewhere in the fertile delta area. One of the two other jars is almost certainly from the delta; the label indicates that the wine in it was pressed in the same year under the supervision of the same chief vintner as the wine in one of the jars with a Western River label.

The labels name a total of 15 chief vintners. One of them, Kha'y (the name is evidently Syrian rather than Egyptian), produced in two consecutive years the wine in six of the jars stored in the annex; the wine in five of the six had been pressed from grapes grown on delta land identified as Tutankhamen's Western River estate. Only three of the other 14 vintners are represented by more than one jar. One had worked at his profession for at least five years, the interval between the first and the second jar he supplied. The other produced both of his jars in a single year. At least one chief vintner in addition to Kha'y was an immigrant; this was 'Aperershop, evidently another Syrian, who produced one of the jars of sweet wine.

The pharaoh's vintners fermented more than grapes. Three of the wine jars in the annex, one of them bearing the name of Kha'y, contained a beverage identified as "Sdh," which was probably made from fermented pomegranates. The writer of the labels was at pains to note in each instance that the Sdh was of "very good quality." Comment of this kind, except for the distinction between sweet and regular wine, is absent from all but one of the remaining 21 wine labels. On a jar that contained the only nondelta pressing, brought to the tomb by an overland journey of 150 miles from the Western Desert oasis of Kharga, the label writer added the notation: "Wine of good quality." His judgment will never be put to the test. When Carter uncovered King Tut's wine cellar 56 years ago, millenniums of evaporation had emptied the jars.

The Surface of Mars

The Viking spacecraft have provided an unparalleled view of it from orbit and from the ground, adding much evidence on how it has been shaped by volcano, meteorite impact, water and wind

by Raymond E. Arvidson, Alan B. Binder and Kenneth L. Jones

The two Viking landers have now been on the surface of Mars for nearly a full Martian year; the two Viking orbiters have been circling and photographing the planet for slightly longer. Ever since the landers touched down in the summer of 1976 they have been gathering data on the characteristics of the Martian atmosphere, rocks and soil. Meanwhile the orbiters have been monitoring the water-vapor content of the atmosphere, mapping the temperature of the surface and photographing the surface with unprecedented resolution and clarity.

Together the photographs made and analytical experiments done by the four spacecraft reveal that Mars is a planet with an even more complex history than had been suspected. They have provided evidence that even the most ancient cratered terrain has been modified by volcanic activity, that early in the planet's history flowing water was a significant agent in shaping its features and that since then surface material has been extensively redistributed by high-velocity winds. Surprisingly, the surface has been little eroded by such aeolian activity. In appearance the surface of Mars is more like rocky volcanic deserts on the earth than it is like the highly cratered surface of the moon, yet Mars, once visualized as being largely a world of gently rolling dunes, seems to possess little sand. The Viking mission has also provided evidence that has both strengthened and altered hypotheses concerning the early Martian atmosphere and climate proposed after the *Mariner 9* mission of 1971 and 1972.

Mars before Viking

Mariner 9 photographed practically the entire surface of Mars at a resolution of a few kilometers and a small fraction of the surface at a resolution of a few hundred meters. The global coverage of Mars by *Mariner 9* indicated that the planet was divided into two distinctly different hemispheres: a rugged, heavily cratered southern hemisphere traversed by huge channel-like depressions and a smooth, lightly cratered northern hemisphere dotted with extinct volcanoes. The two types of terrain are divided roughly along a great circle inclined to the equator by about 30 degrees.

The craters in the southern hemisphere range in size up to the basin Hellas, which is 1,600 kilometers across. The abundance of craters in some of the more heavily cratered regions is comparable to the abundance of craters in the bright highlands of the moon. Samples of rocks and soil collected on the lunar highlands during the Apollo missions have been dated, and their ages indicate that most of the craters on the lunar highlands were excavated between four and four and a half billion years ago. At that time the moon was torrentially bombarded by the interplanetary debris left over from the formation of the solar system. Since the abundance of craters on Mars is similar to that on the moon, it is believed the Martian cratered terrain is approximately the same age as the lunar highlands. In other words, almost half of the surface of Mars is ancient terrain where many of the larger land forms have remained unchanged for some four billion years.

Mariner 9 revealed that the sparsely cratered areas in the northern hemisphere of Mars are plains of lava that extensively flooded the surface at different times after the heavy bombardment of the planet ceased. The few craters on the plains record the occasional impact of a stray asteroid or comet. Although there is no way as yet to estimate with confidence the absolute ages of the plains, the abundances of craters in different regions vary widely, which implies that the ages of the plains range from hundreds of millions to some billions of years.

The channels on Mars photographed in such detail by *Mariner 9* seemed to imply that the past climate of the planet must have been quite different from the present one. If all the water in the Martian atmosphere today were condensed in one place, it would form a body of water no larger than Walden Pond. In fact, the abundance of water in the atmosphere today is so low that rain is an impossibility.

By the time the Viking landers touched down most investigators were convinced that the largest channels, which are tens of kilometers wide, were carved by water derived from the melting of ice below the surface. Some investigators believe the ice is a remnant from a denser atmosphere the planet may have had in its first billion years. According to calculations made by Fraser P. Fanale of the Jet Propulsion Laboratory of the California Institute of Technology, the early Martian atmosphere may have contained ammonia and methane in addition to carbon dioxide and water vapor. According to James B. Pollack of the Ames Research Center of the National Aeronautics and Space Administration, such an atmosphere would have acted to trap infrared radiation (the "greenhouse effect") and would have been warm enough to hold substantial amounts of water vapor. At some time in the past various reactions would have removed the ammonia and methane from the atmosphere. The atmosphere would then have become more transparent to infrared radiation. As a consequence the atmosphere would have cooled and the water vapor in it would have condensed and precipitated onto the surface, where it would have made its way to the polar regions and into the planet's impact-fractured crust and regolith (the loose material lying on the crust).

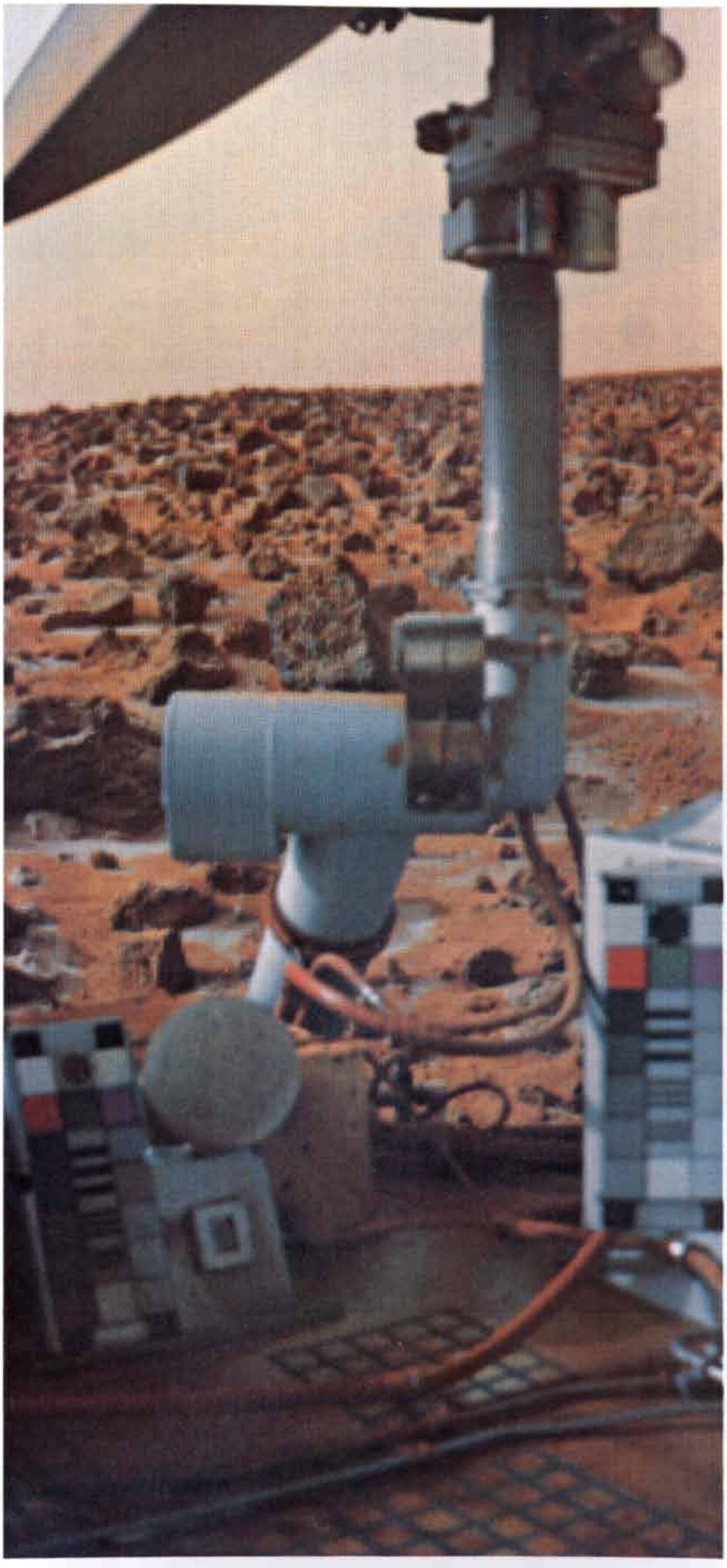
Mariner 9 also revealed that both the north and south poles of Mars are covered with ice and windblown dust as much as several kilometers thick. In the oldest deposits the ice and dust are nonlayered; in the younger ones they alternate in layers, each layer some tens of meters thick.

The large quantities of dust in these deposits provide more evidence that at one time the atmosphere of Mars was denser than it is today; the present atmosphere does not seem capable of transporting such quantities of material to the poles. In fact, the most recent process has been one of partial erosion of the deposits by the wind, moving debris away from the polar deposits to form a mantle blanketing much of the surface at high latitudes. The layered deposits imply that the past climate of Mars not only was different but also may have been subject to periodic changes.

On the basis of the photographs from *Mariner 9* a number of hypotheses were proposed for the origin of the polar deposits. Some investigators suggested that the deposits of ice and dust accumulated mostly during an early period when the Martian atmosphere was denser. Others suggested that the polar deposits have been accumulating throughout Mars's history and that their accumulation has been modulated by the amount of solar radiation received by the planet, particularly at the poles. Calculations by Carl Sagan and his colleagues at Cornell University and by William K. Hartmann of the Planetary Science Institute in Flagstaff, Ariz., indicated that the amount of heat received by Mars could have been changed by the requisite amount if the luminosity of the sun has varied. Such variations are thought to occur over a period of between a million and 100 million years.

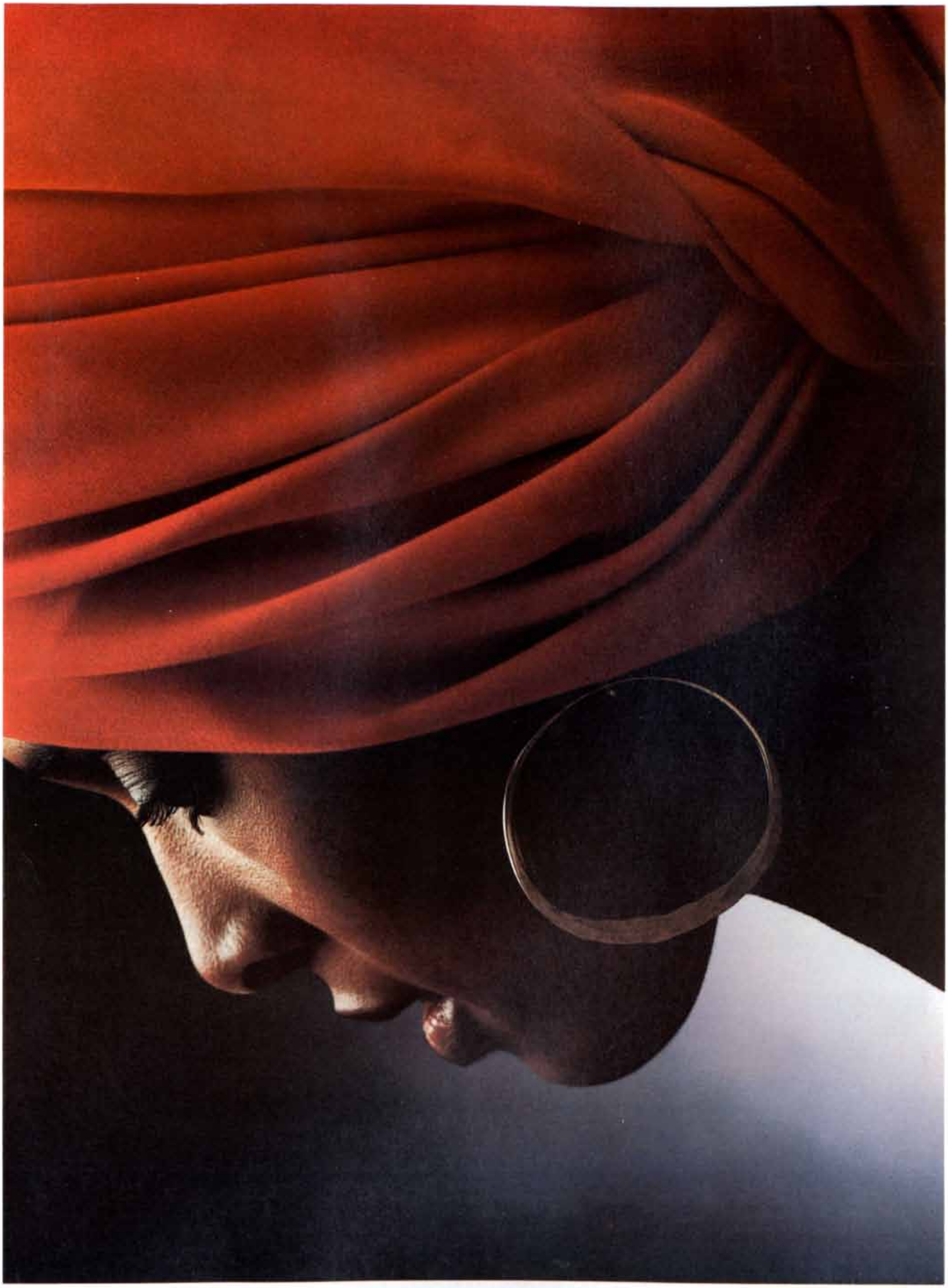
Moreover, William R. Ward of the Jet Propulsion Laboratory showed that the tilt of Mars's axis of rotation changes considerably with time because of the pull of the sun on the planet's equatorial bulge. Currently the axis of Mars is tilted from the vertical with respect to the plane of its revolution around the sun by about 25 degrees. Over a period of between 100,000 and a million years, however, the tilt of the axis changes from a minimum of 15 degrees to a maximum of 35. To add an even greater degree of complexity, more recent calculations by Ward, Joseph A. Burns of Cornell and O. Brian Toon of the Ames Research Center show that the tilt of the axis could have occasionally reached 45 degrees during an earlier period, before the formation of the great volcanic bulge Tharsis.

A radiometer aboard the early *Mariner 7* spacecraft revealed that the seasonal ice caps on Mars were composed

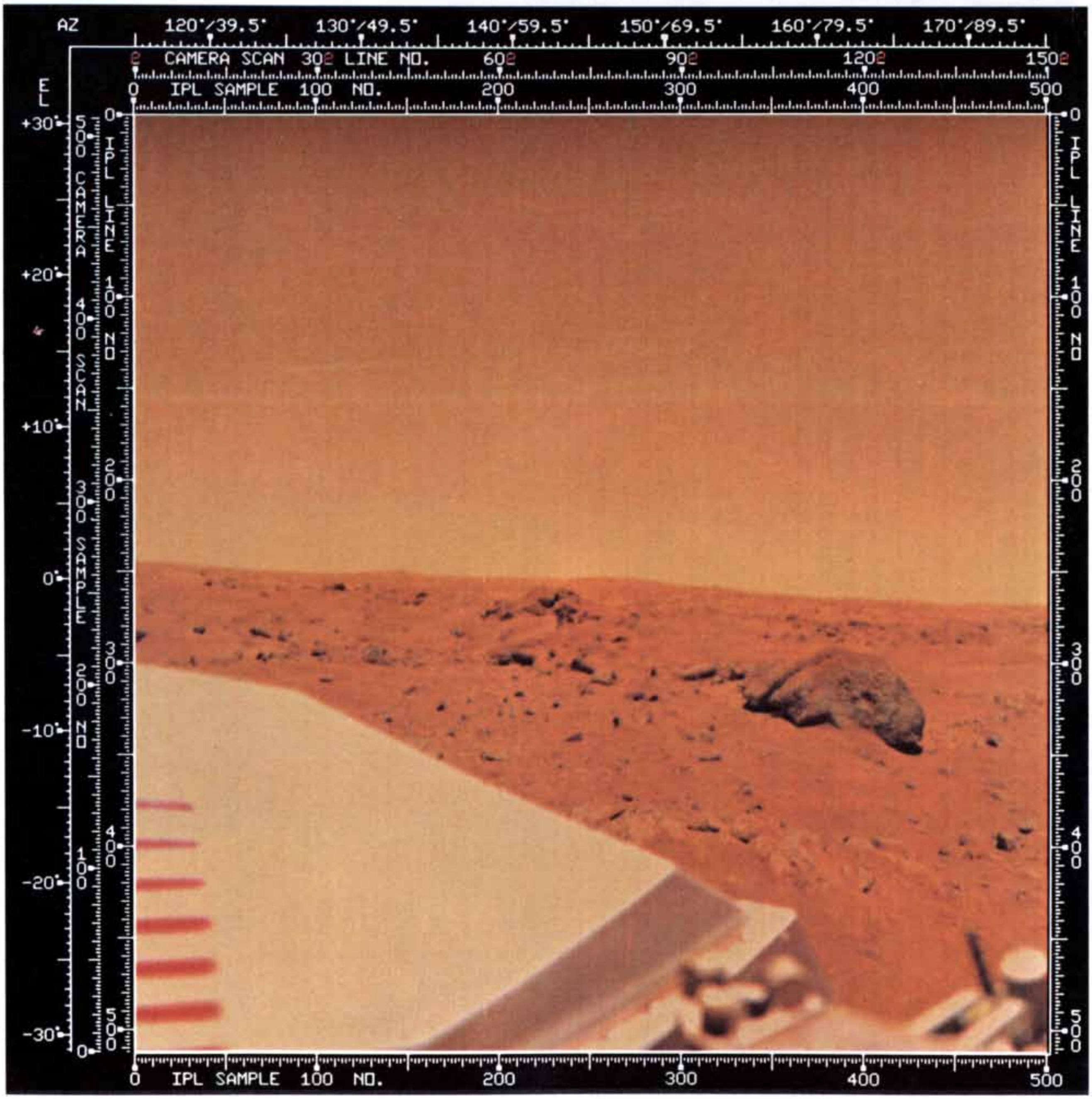


FROST ON MARS was photographed by *Viking Lander 2* near local noon on September 13 of last year, which was late winter in the Martian northern hemisphere. The lander is located in Utopia Planitia at a latitude of 48 degrees. The view is to the southwest over the vehicle's deck. The frost is the white patches at the base of the rocks. It is probably a remnant of a thin deposit of water ice, perhaps mixed with a clathrate in which carbon dioxide ice is caged inside water ice. The frost had slowly accumulated earlier in the winter as water was transferred to the northern hemisphere from the southern hemisphere. Analysis of the color of the surface indicates that the frost had been on the ground for at least 100 days before this photograph was made.


Polaroid introduces a whole new dimension to instant film: 8 x 10.
Now photographers using 8 x 10 view cameras can make large-format pictures on the spot– by using Polaroid's new 8x10 Polacolor® 2 instant film. In just 60 seconds this new professional film delivers all the exceptional detail and sharpness inherent in 8 x 10 photography. And its brilliant metallized dyes are among the most stable and fade-resistant in photography. The portrait on the right was reproduced directly from an unretouched original print. The new 8 x 10 system is available now from Polaroid® professional product dealers.
For more information or the names of dealers, call us toll free: 800-225-1618. (In Massachusetts call collect: 617-547-5177.)



of frozen carbon dioxide. It seemed likely that the residual ice cap that remained during the summer was also composed of carbon dioxide ice. If the permanent ice cap consisted of carbon dioxide ice, a relatively small increase in the atmospheric temperature should increase the pressure of the Martian atmosphere from its present value of some two to 10 millibars to somewhere between 30 millibars and one bar: the atmospheric pressure at sea level on the earth. The estimate of 30 millibars, calculated by Bruce C. Murray and his colleagues at the California Institute of Technology, was based on the amount of solar radiation received by the ice caps when Mars's spin axis was at its maximum tilt of 35 degrees, and on the assumption that the atmosphere transported a minimum amount of heat to the poles. The estimate of one bar, calculated by Sagan and his co-workers, was an upper limit resulting from a runaway greenhouse effect, whereby an initial increase in temperature leads to an increase in atmospheric pressure, which allows more heat to be transported to the poles, which in turn allows even more carbon dioxide ice to evaporate.

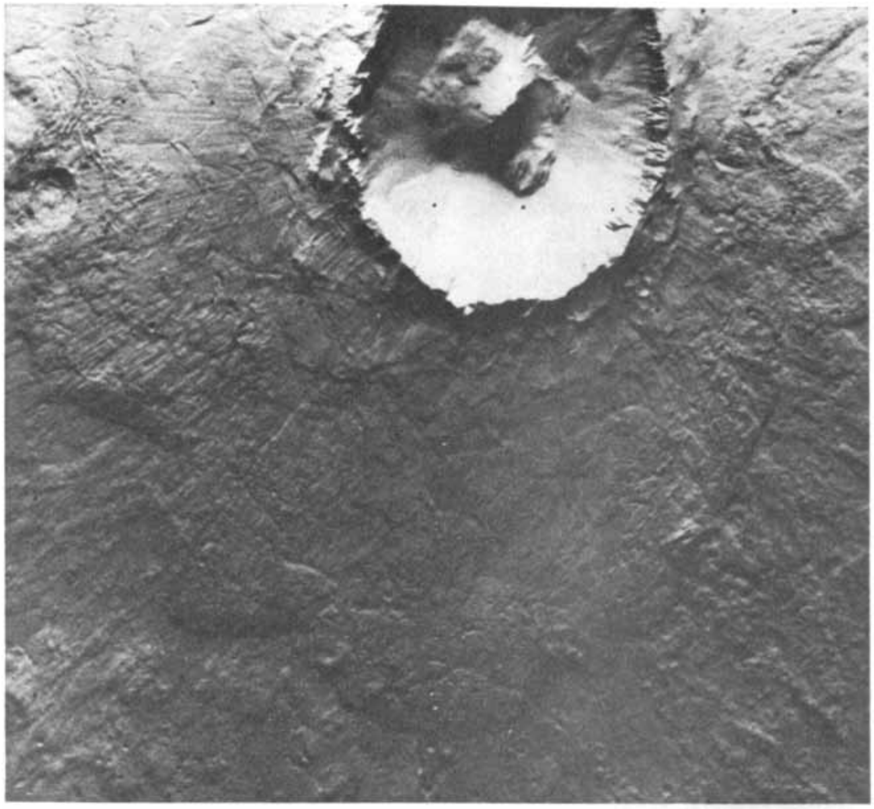


"BIG JOE" is the name given the large dark boulder to the right of the center of this photograph made by *Viking Lander 1* at local noon on August 23, 1976. The boulder is some two meters across and is about 10 meters away from the lander; it is one of a number of boulders that can be seen lying among drifts of fine-grained material. It is likely that Big Joe was once covered by drifted material, some of which still seems to be clinging to the top of it as a yellowish brown deposit. The picture was made in the summer of the northern hemisphere, a season noted for its lack of dust storms, yet the yellowish brown color of the sky is due to suspended particles of dust. Apparently some dust is present in the Martian atmosphere most of the time. The profile of an impact crater some 300 meters from the lander can be seen on the horizon in the middle of the picture. The numbers bordering the picture aid in the interpretation of the photograph. The numbers in the vertical column at the far left show the elevation in degrees above and below the camera's horizontal plane; the numbers in the top two horizontal rows show azimuth in degrees respectively in the coordinate system of the camera and in that of the lander. Due north is at an azimuth of 130 degrees. The other numbers on photograph refer to the scan lines by which the camera built up the image.

In warmer periods, or during periods when the poles of Mars point more toward the sun than they do at present, the density of the Martian atmosphere would have been higher and more dusty material would have been eroded at low latitudes and transported to the poles. As the atmosphere cooled over the poles and condensed to form the ice cap, dust suspended in it would have been deposited along with the condensates. In cooler periods, however, so much of the atmosphere would have condensed over the poles that at lower latitudes its gaseous remnant would have become quite thin and would have transported far less dust to the poles.

The Viking Orbiter Observations

The Viking orbiters, which began their coverage of Mars early in the summer of 1976, photographed the planet with much greater clarity than *Mariner 9* did. The difference was due largely to the fact that *Mariner 9* approached Mars when the planet was still enveloped in a dust storm. Moreover, after the storm enough dust remained suspended in the atmosphere to significantly obscure the surface of Mars for several terrestrial months. The Viking orbiters began making images of Mars when the atmosphere was relatively free of dust. Moreover, although the Viking orbiter cameras were Vidicon systems similar to those carried on *Mariner 9,* they had considerably better resolution.



FLOW OF DEBRIS surrounds the 25-kilometer crater Arandas, photographed by *Viking Orbiter 1* on July 22, 1976. Although the rim of the crater is clearly defined, the material around it appears to have flowed along the ground instead of being ejected as the meteorite struck the surface. At the top left the flow may have been deflected around a small crater. The radial grooves on the surface of the flow may have been eroded into it during the last stages of the impact process. Flow was probably created when heat from the meteorite's impact melted ice below the surface, and water and steam transported material away from crater in a coherent flow.

The Viking orbiter photographs show that much of the surface of Mars retains crisp topographic detail: lava flows, wrinkle ridges and crater ejecta stand out in sharp relief. In addition the photographs show numerous lava flows in even the most primitive cratered terrain. From these last features Michael H. Carr of the U.S. Geological Survey and his colleagues suggest that early in Mars's history even much of the ancient cratered terrain in the southern hemisphere was flooded with lava. Such volcanic leveling would explain why the highly cratered terrain on Mars is relatively smooth compared with the mountainous highlands of the moon.

The fact that the most ancient features on Mars are still sharply defined also indicates that during the planet's history there has been relatively little breakdown of rock and redistribution of debris. The only clear evidence for wind erosion on a large scale is found in regions composed of older sedimentary deposits, such as those near the poles. It is likely that the polar deposits are only partly consolidated and are therefore easily eroded by the wind.

The Viking orbiters have shown that the northern plains at high latitudes are more than simple mantles of debris. They are a complex of lava flows and deposits of windblown dust that have themselves been partially stripped by the wind. Closer to the north pole the orbiter cameras found large fields of dunes girdling the residual ice cap; the dunes seem to be composed of particles the size of grains of sand. Sand-sized particles, which on Mars probably range from .1 millimeter to several millimeters, are too large and heavy to remain suspended in the rarefied atmosphere and be carried away, but they are sufficiently small and light to be rolled and lifted for short distances by the wind. Dust particles, however, which on Mars are smaller than about .1 millimeter, are sufficiently small and light to be carried away in suspension.

The bulk of both the debris on the northern plains and the dunes closer to the poles apparently consists of material eroded from the polar deposits themselves. When the deposits were stripped by the wind, the embedded dust was carried away and any sand lagged behind and accumulated in dunes close to the poles. A puzzle here is that although the polar deposits contain dust, they seem to be yielding both dust and sand. The answer may be that the sand-sized particles are actually dust-sized particles that have been cemented together by oxides, salts or perhaps even ice.

Interpretation of the data from the radiometers on the Viking orbiters by Hugh H. Kieffer of the University of California at Los Angeles and his associates showed that during the summer in the northern hemisphere of Mars the temperature over the residual north-polar ice cap was some 205 degrees Kelvin (68 degrees below zero Celsius). This result was striking. Since the atmospheric pressure at the surface of Mars is only about six millibars, the temperature would have to be less than 148 degrees K. to maintain a permanent cap of carbon dioxide ice. Even if the ice in the ice cap were a clathrate compound in which carbon dioxide ice was caged inside water ice, it could not exist at a temperature higher than about 155 degrees K. The only condensate that can remain stable at 205 degrees K. is water ice.

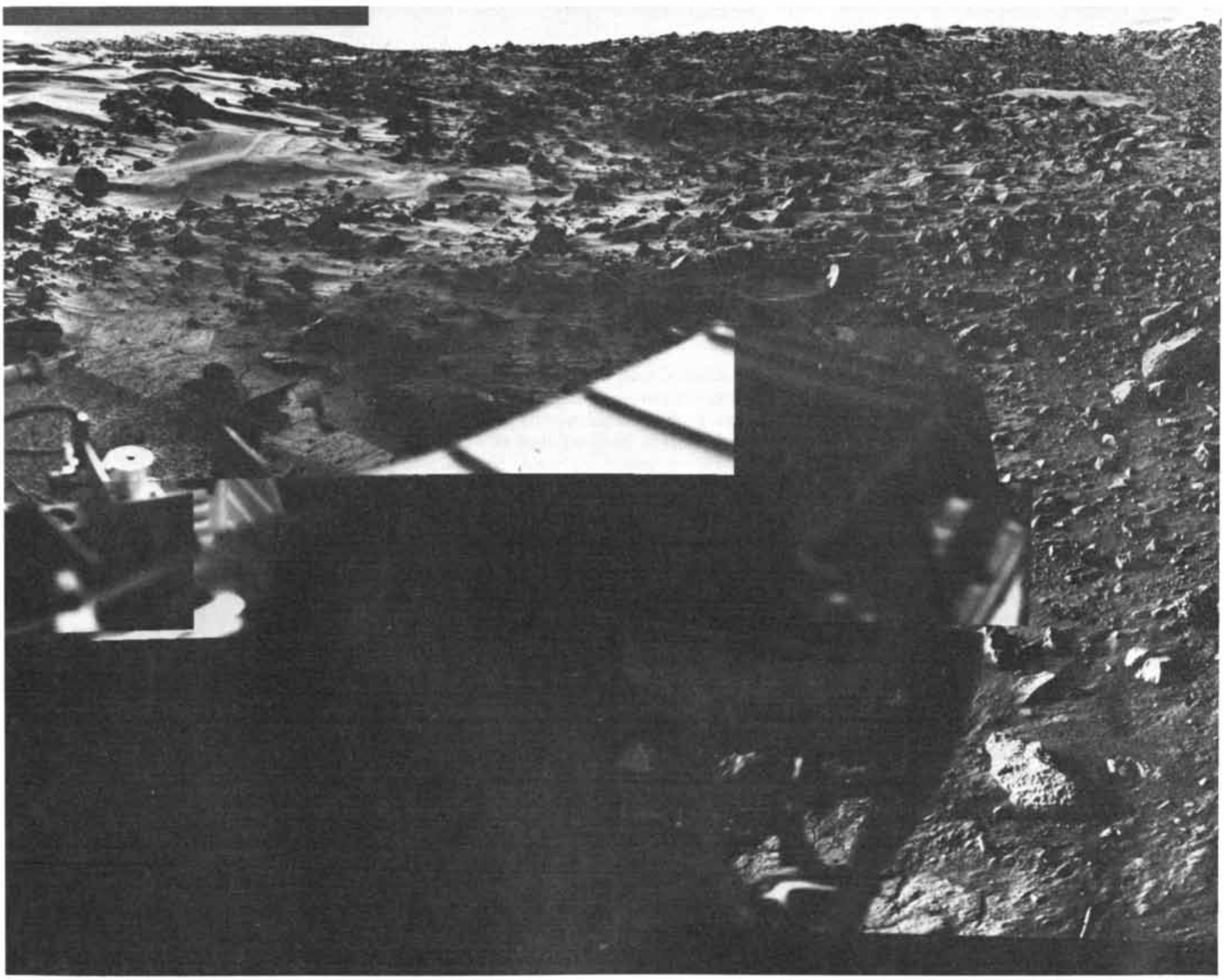
Further support for the view that the residual polar cap consists only of water ice comes from an analysis of data from the Viking orbiter spectrometers made by Crofton B. Farmer and his associates at the Jet Propulsion Laboratory. They find that during the summer at northern latitudes the amount of water vapor in the atmosphere is such that the temperatures must be higher than 200 degrees K.

Observations of the residual south-polar cap during the summer in the southern hemisphere are difficult to interpret because of the effect of a global dust cloud that may have modulated atmospheric temperatures. The residual southern ice cap is probably water ice too, although measurements of its temperature are ambiguous, and there is a slim chance that its major constituent is carbon dioxide ice.

Since it seems that there is water ice at both poles, hypotheses involving water as an active agent in Mars's past have gained measurably more acceptance. Moreover, the Viking orbiter photographs have revealed that the channels on Mars are even more abundant than was indicated by the *Mariner 9* photographs and that the channels extend to smaller sizes and form a much more integrated drainage system than had been previously perceived. The runoff of rain from an ancient dense atmosphere may have carved some of the treelike networks of channels. Other channels may have been formed when underground ice was melted by heat released in volcanic activity. Then the channels would have been created when the ice melted, the ground above it slumped and the water flowed away. In this case the formation of the channels would seem to be directly linked to the thermal history of the planet.

Intriguing evidence for the presence of water ice in the crust and regolith of Mars has also been obtained from examining pictures of the peculiar terraces, ramparts and lobes characteristic of the ejecta of many of the large craters. On the moon the ejecta from impact craters appear to have been blocks of material that were hurled outward by the original impact and fell back to the surface, excavating myriads of secondary impact craters. On Mars the ejecta deposits surround the impact craters almost like a solidified flow. The probable explanation is that the heat of the impact melted and vaporized water ice trapped in the crust, and the liquid water and steam transported the ejecta away from the crater in a coherent, ground-hugging flow of debris. In some photographs one can even see where the flow was diverted around obstacles in its path.

Although water is now being recognized as an important agent in Mars's past, the discovery that the polar regions are probably dominated by water ice instead of carbon dioxide ice places severe constraints on the intensity of past climatic fluctuations. If a water-ice cap on Mars received more sunlight, it would begin to evaporate, subliming directly from the solid ice to a vapor. In order for the vapor pressure of water on Mars to reach the pressure required for liquid



CHRYSE PLANITIA, the site of *Viking Lander 1*, is a rolling plain littered with blocks. Just visible to the right of the center are several areas of exposed bedrock. Below them small linear deposits or streaks of sediment can be seen extending away from the rocks. The streaks extend roughly in a north-to-south direction. Toward the upper left is a large field of drifted material where deposits accumulated during a period when winds were also blowing from north to south. Because of the panoramic geometry of the image it seems that the streaks change direction with azimuth. The center of the mosaic is pointing toward the southeast. The entire mosaic covers 160 degrees in azi-

water to exist the temperature would have to be raised by at least 70 degrees K. Such a dramatic rise in temperature would be most unlikely to occur even if the luminosity of the sun varied by the maximum amount allowed by theory or if the inclination of the planet's axis periodically changed to the maximum extent. Indeed, the fact that the most ancient surfaces of Mars are so well preserved is consistent with the hypothesis that the atmospheric conditions on Mars have not fluctuated greatly over most of the planet's history. It seems likely that the bulk of the polar deposits formed very early and have since been eroded by the wind. The exact time they formed and the reason for their formation, along with the history of any early, dense atmosphere that may have been supported by greenhouse effects, however, remain a mystery.

Viking Lander 1 touched down on the western slopes of the region named Chryse Planitia (22.5 degrees north latitude, 47.8 degrees west longitude). From orbit the landing site looks much like the surface of a lunar mare, or "sea": it is a smooth volcanic plain sparsely cratered and lined with a series of wrinkle ridges. The walls of the craters seem nearly pristine and the wrinkle ridges are also well preserved. The amount of erosion must be very slight or confined to a scale on the order of meters for the morphology of the ejecta deposits and the ridges to be so little changed.

The landing site is some 130 kilometers east of Lunae Planum, one of the most heavily cratered plains on Mars. Lunae Planum is about a kilometer higher in elevation than Chryse Planitia, and the boundary between the two regions is marked by an irregular escarpment. A number of large channels course through Lunae Planum, emerge from the escarpment and extend eastward across Chryse Planitia toward the landing site. The channels were most likely created by ground water from ice trapped within Lunae Planum. At some time in the past volcanism and geothermal heating melted some of the underground ice and the water escaped to the surface to create one or more torrential floods that cut into the Lunae Planum deposits and poured onto Chryse Planitia. The flow of water from Lunae Planum breached several craters and wrinkle ridges to the west of the landing site.

The Surface of Chryse Planitia

To judge from correlations between pictures from *Viking Lander 1* and pic-

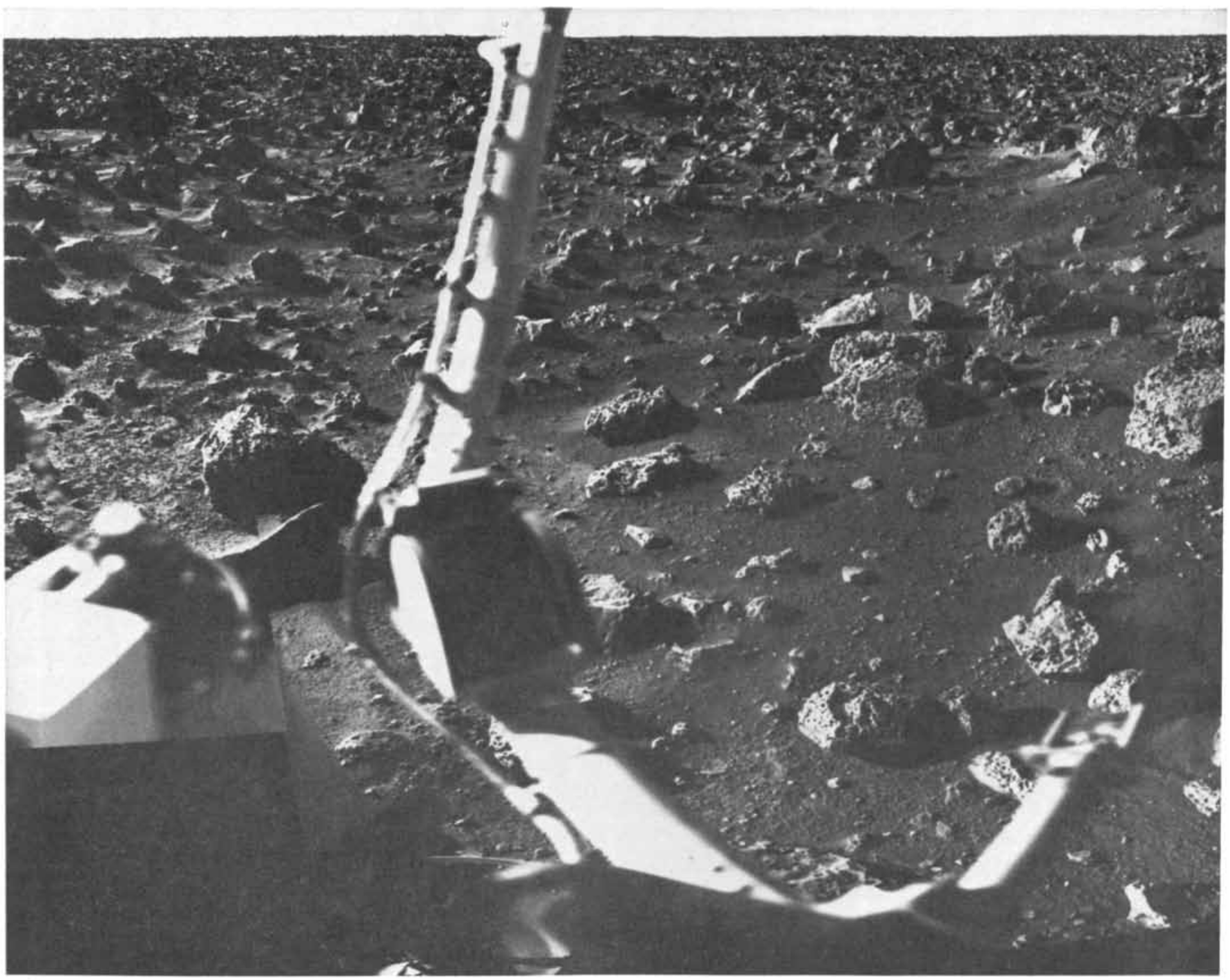
muth. The local topography is such that the distance of the horizon ranges from several tens of meters at the left to several kilometers at the right. The terrain at the bottom is only a couple of meters from the camera. Scale is provided by the surface sampler assembly pin, the bright cylindrical object lying in the soil in the bottom middle portion of the mosaic; it is 10 centimeters long. The lander's footpad is at the bottom right, and other parts of the vehicle can be seen to the left. Under the spacecraft the retrorocket exhaust blew loose material away, exposing a fractured, crusted type of soil called duricrust; a few chunks of such soil are visible to the left of the assembly pin.

tures from the Viking orbiters, the lander is sitting on the flank of a wrinkle ridge. From the lander the site strikingly resembles many rocky deserts on the earth, particularly those with exposures of volcanic rock. The gently rolling landscape is yellowish brown in color, strewn with rocks and dotted with drifts of fine-grained material. Within 30 meters of the spacecraft several outcrops of bedrock can be discerned. No positive evidence of the flood from Lunae Planum is visible on the photographs from the lander: scoured features, channels or fluvial deposits cannot be detected. Apparently the flood either did not reach the landing site or had largely dissipated by the time it passed over it. Alternatively the surface may have been so modified since the flood that any fluvial features are no longer discernible.

The resemblance of the landing site to rocky deserts on the earth was somewhat surprising on the basis of what most investigators expected from the moonlike orbital pictures of Chryse Planitia. From orbit the dominant features of the region are craters. From the ground there are only a few obvious craters to be seen in the immediate vicinity of the lander. Based on the population of large craters visible from orbit, Edward A. Guinness, Jr., of Washington University calculated that if Mars were like the moon, some 35 craters with diameters ranging between 25 and 50 meters should be in the lander's field of view.

A relative deficiency of craters on Mars smaller than 50 meters across was actually predicted in 1970 by Donald E. Gault and Barrett S. Baldwin, Jr., of the Ames Research Center. They calculated that even though the atmosphere of Mars is thin, it is dense enough to ablate and break up small incoming meteoroids before they reach the surface. As a result the Martian surface is not subjected to the repetitive high-velocity impact of small objects and the consequent "gardening" of the top few meters of the surface. The only substantial population of craters with diameters of less than 50 meters should be secondary craters produced by the impact of debris thrown out during the formation of a large crater (one that is tens of kilometers in diameter). In contrast, the surface of the moon shows a continuous spectrum of crater sizes, and the impact of small objects over millions of years has created a dusty soil. On Mars large impacts would fracture the surface and strew over it a discontinuous layer of relatively large blocks, just as at the landing site. The soil on Mars must have been created by other processes.

In addition to the rocks and outcrops visible on the ground at Chryse Planitia, fine-grained material is abundant in the form of streaks on the lee side of most of the rocks; the streaks are several centi-



UTOPIA PLANITIA, the site of *Viking Lander 2,* is superficially like Chryse Planitia in that the terrain is littered with blocks. The terrain is remarkably flat, however, and the horizon is several kilometers away. Scale is provided by the largest block in the middle of the image, which is 2.75 meters from the spacecraft and is 35 centimeters wide. No bedrock seems to be exposed. A prominent troughlike depression

meters in depth and range from 10 centimeters to a meter in length. To the northeast of the lander there is a complex of drifts in a field of boulders; these drifts probably were formed when windblown material was trapped among boulders that were sufficiently large to reduce the velocity of the wind locally. Several of the large drifts appear to be layered or laminated. On the earth layers are usually not seen in actively growing or moving drifts; they are visible only when the drifts are stabilized by vegetation or cementation and are being excavated by wind erosion. It appears that on Mars the drifts observed were deposited some time ago, that they were partially lithified (formed into sedimentary rock) and that they have recently been eroded.

On the average the long axes of the windblown streaks behind the rocks point almost due south. Photographs made by *Mariner 9* of regions just to the north of the landing site show that large streaks extending from craters also point roughly south. Both sets of streaks seem to indicate that the prevailing winds near the surface had blown from north to south during the period when the streaks were formed. Moreover, the pattern of exposed layers on the drifts indicates that when the drifts were deposited, the wind direction was also north to south.

Another feature at the landing site was revealed by the landing itself. Close to the spacecraft the retrorocket exhaust blew away loose material, exposing a crust of soil fractured into a polygonal pattern. Such soil, known as duricrust, is similar in appearance to the deposits called caliche in the U.S. Southwest and Mexico. On the earth duricrust is formed when dilute solutions of salt migrate up through the soil; the water evaporates from the solutions, and the salts and other substances collect just below the surface. The same process probably has operated on Mars. It is not known whether on Mars the water comes from relatively large pores below the surface or from thin films of water between grains of material. It probably comes from the thin films, because variations in the water-vapor content of the lower atmosphere suggest that water is regularly cycled between the surface and the atmosphere.

Direct evidence that salts are present in the Martian duricrust was obtained by sampling the soil in front of the lander and analyzing its chemistry with the X-ray-fluorescence spectrometer on the lander. Priestley Toulmin III of the U.S. Geological Survey and the members of his X-ray team found that the amount of sulfur, a likely candidate for being bound in salt minerals, was somewhat greater in clods of soil than it was in loose soil. The clods are abundant around the lander, and they probably

about a meter wide and 10 centimeters deep cuts horizontally through the middle of the picture. Several drifts of material occupy its bottom. Linear deposits or streaks extend downwind of the rocks at the far left. The streaks have approximately the same azimuth as those behind the rocks at Chryse Planitia. The bright plateau on the horizon to the right is approximately in direction of the large crater Mie.

are chunks of duricrust broken up by the landing and also by the natural wind.

The Surface of Utopia Planitia

The search for the landing site for *Viking Lander 2* was based partly on the desire of investigators to have the second lander in a region distinctly different from that of *Viking Lander 1* but still smooth enough for a successful landing. The overriding concern, however, was to find a location with a high water-vapor content in order to maximize the probability of finding evidence of life. The region chosen was the surface of Utopia Planitia, which is part of the mantle of debris on the vast plains of the northern hemisphere. Observations from orbit showed that the surface is cut by fractures that break the terrain into polygonal forms kilometers on a side. The landing site (48 degrees north latitude, 225.6 degrees west longitude) is 200 kilometers southwest of the large crater Mie, 100 kilometers across.

Soon after *Viking Lander 2* had settled on the surface it was apparent that from the ground the site of *Viking Lander 2* was superficially similar to the site of *Viking Lander 1:* the surface is a rock-strewn plain with duricrust. That, however, is where the resemblance ends. *Viking Lander 2* came down on a flat plain, where most of the topographic relief is created by troughlike depressions that divide the terrain into polygons. No outcrops of bedrock can be discerned. Instead boulders and cobbles either are partly embedded in a matrix of fine-grained material or are resting on top of it. Although other explanations for the appearance of the landing site are possible, it is likely that *Viking Lander 2* came to rest on a lobe of the debris that flowed out from the crater Mie. In flows of debris on the earth large rocks are commonly carried on top of the flow, leaving a field of boulders partly embedded in finer-grained material.

The troughlike depressions visible from *Viking Lander 2* are approximately a meter wide and 10 centimeters deep, and the edges of the troughs are slightly turned up. The troughs visible from the lander are much smaller than the troughs seen from orbit, but it is likely that the features have a similar origin. The scale and the form of the troughs seen from the lander make them a good physical analogue of the "patterned ground" that is found in cold regions on the earth. On the earth patterned ground forms in ice-saturated soil where low temperatures cause the ground to contract and fracture; the fractures usually cut the ground in a polygonal pattern. In the spring the frozen soil thaws and the fractures fill with water. In the fall the water freezes, and in the following winter the ice, which is weaker than the frozen soil, fractures in the same pattern. Repeated cycles of the process create a terrain that is cut by troughs in a polygonal pattern, with the troughs occupied by wedges of ice.

One problem with the proposal that this process is currently operating on Mars is that the temperatures at Utopia Planitia are always below the freezing point of water. Benton C. Clark of the Martin Marietta Corporation has estimated that even if the ice contained salts, the freezing point of the solution would be depressed only 10 or 20 degrees K. below the melting point of pure water. At the temperatures on Utopia Planitia the ice would still never melt.

Patterned ground can be created by one other process: the desiccation of clays. Clays, which are the leading candidate for the major constituent of the Martian soils sampled at the landing sites, expand and contract by as much as 20 percent as they absorb water and then lose it. If the ground on which the *Viking Lander 2* is resting was once saturated with water and then dried out, cracks may have formed, creating the troughs in a polygonal pattern.

Small streaks can be seen extending downwind from some of the rocks at the site of *Viking Lander 2,* just as they do at the site of *Viking Lander 1.* In addition several small drifts occupy the floor of a large trough in front of the lander. Again the drifts indicate that the prevailing wind direction is roughly from north to south; thus Chryse Planitia and Utopia Planitia may have been subject to the same wind system. Since the sites are halfway around the planet from each other, such a wind system would have to have been global. Most likely both sets of wind streaks and drifts formed during dust storms at Mars's closest approach to the sun, when the flow of the atmosphere could have been in a north-to-south direction in the northern hemisphere. The Viking lander meteorology stations, which monitored winds during the 1977 dust storms, did not, however, show any predominance of a north-to-south flow.

The Martian Soil

The X-ray-fluorescence spectrometer on each lander analyzed soil samples and determined the abundance of a number of elements with atomic numbers higher than that of sodium (atomic number 11). Toulmin and his colleagues have examined the data and have shown that the overall composition of the soil at one site is much the same as that at the other. The composition is different from that of any single known mineral or type of rock, indicating that the soil is probably a complex mixture of materials.

The surface of Mars seems to be composed of soil derived from mafic igneous rocks, that is, rocks that have crystallized from a melt that was rich in mag-

TWO DRIFTS OF FINE-GRAINED MATERIAL are 15 meters from *Viking Lander 1.* The drift at the right is composed of darker material than the drift at the left. Spectra of the dark drift indicate that it is the only example of such material at either landing site. All the other soil areas seen from the landers are quite similar. They are probably the same as the soils that dominate the bright areas of Mars.

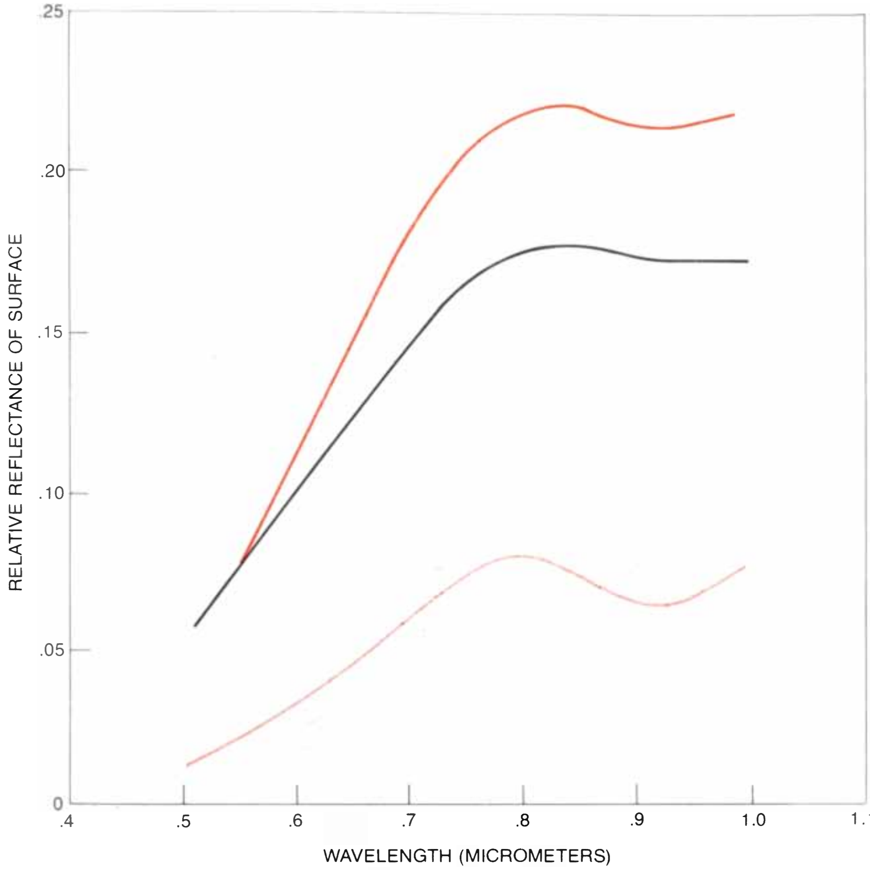
nesium and iron. Compared with rocks on the earth in general, they are rich in magnesium, iron and calcium and poor in potassium, silicon and aluminum. Such abundances are compatible with the kind of materials one would expect from a partial melting of the Martian mantle: the deep layer below the crust.

It is likely that the soil analyzed by the Viking X-ray spectrometers is a mixture of iron-rich clay minerals, iron hydroxides, sulfate minerals and carbonate minerals. Such a deduction is consistent with the results from the combined gas-chromatograph and mass-spectrometer experiments on each lander, which found that when samples of soil were heated, water vapor and carbon dioxide were released. The soil contains about 1 percent water by weight, some of which is probably in hydrated minerals.

On the earth mafic materials are chemically altered by water and give rise to iron-rich clayey soils. The same kind of process could have operated in the Martian past, when liquid water was prevalent. Some of the soil could also have formed when hot magmas penetrated the ice-laden crust and regolith, erupting explosively to form clay tuffs. In addition, if a sufficient quantity of water was available, the heat from impacts may have been sufficient to alter volcanic rock materials into clays.

An intriguing alternative hypothesis has been proposed by Robert L. Huguenin of the University of Massachusetts. He suggests that the soil has largely been produced by oxidation stimulated by ultraviolet irradiation of the rocks. Ultraviolet radiation from the sun is not absorbed by the Martian atmosphere because the atmosphere does not have an ozone layer. In the presence of small amounts of water vapor such radiation is capable of breaking down aluminosilicate minerals by causing ions such as those of iron to migrate toward the surface, thus disrupting the minerals' crystal structure. The extent to which the soil might have been created by such a process is not known.

Two magnets were mounted on the hoe that was part of the soil-acquisition scoop on each lander. One of the magnets was mounted flush with the hoe's surface; the other was embedded in the metal so that its effective strength was a twelfth that of the first. At both sites equal amounts of material clung to both magnets after the hoe had been inserted into the soil. According to Robert B. Hargraves of Princeton University and David W. Collinson of the University of Newcastle upon Tyne, for equal aggregates of material to cling to both magnets between 3 and 7 percent of the material by weight must be magnetic. If the soil had a lower concentration of magnetic material, the aggregates would not cling to the weaker magnet. Hence magnetic material is present in significant amounts in the Martian soil.



REFLECTANCE SPECTRA are shown for the bright drift and the dark drift at the site of *Viking Lander 1* and also for a trench dug there by the surface sampler. The spectra were produced at the Langley Research Center of the National Aeronautics and Space Administration with a specialized technique devised by Friedrich O. Huck and Stephen K. Park. The spectra show the fraction of sunlight reflected at wavelengths between .5 micrometer (in the blue) and 1.0 micrometer (in the near infrared). The shapes of the spectra for the bright drift (*dark color*) and for the trench (*light color*) are similar. Both show an absorption band at .93 micrometer, which implies that they have a similar composition. The only major difference between them is in the intensity of the light reflected. (The difference is actually to be expected because the soil in the trench was disturbed when the trench was dug, thereby increasing the soil's microtopography, increasing its degree of scattering and shadowing and decreasing its reflectance.) The reflectance spectrum of the dark drift (*black*), however, has a different shape. The absorption band at .93 micrometer is gone, making the spectrum look flat, and there may be a shallow band near 1.0 micrometer. The difference between the bright drifts and the dark ones may mimic the difference between large bright and dark areas on Mars as it is seen from the earth.

The color of the material clinging to the magnets is the same as the color of the surface, which indicates that the magnetic materials are covered or stained with the same materials coloring the Martian soil. Reasonable candidates for the magnetic minerals are magnetite and maghemite, both of which are iron oxides, and metallic nickel-iron. Maghemite is yellowish brown or reddish in color, and if it is present, it may contribute to the color of Mars. The abundance of the magnetic material and its probable nature is also consistent with a source that is mafic, such as mafic basalts.

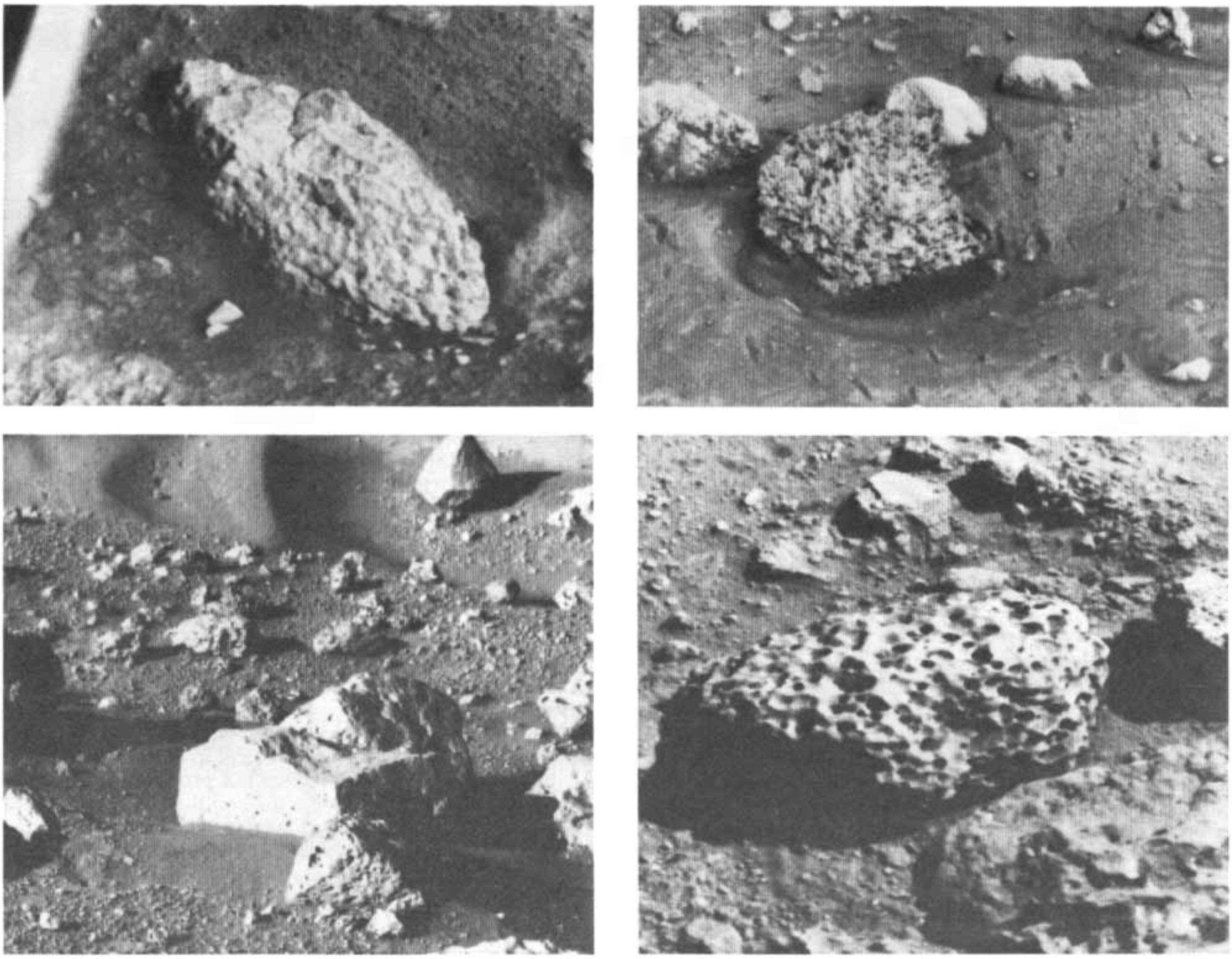
The cameras on the Viking landers made pictures of the surface and the sky in six regions of the spectrum, ranging from about .5 micrometer (in the blue) to 1.0 micrometer (in the near infrared). The color pictures were generated by first determining the spectral irradiance of Mars in each of the regions and then computing the hue, brightness and saturation of color for the range of wavelengths to which the human eye is sensitive. The spectral irradiance of the surface of Mars is a product of the irradiance from the sun, the reflectance of the surface of Mars and the character of the light scattered by the atmosphere. The color pictures show that the surface is yellowish brown. If the effects of the color introduced by the atmosphere are removed, the color of the surface tends even more toward brown.

The reflectance spectra derived from the camera data for various soil areas at both landing sites are remarkably similar and seem to vary only as the lighting changes. The spectra are similar in appearance to spectra made of bright areas on Mars with instruments on the earth. Those areas are thought to be covered with a fine-grained, chemically altered soil. An explanation fitting the available data is that most of the soil exposures seen at the landing sites and sampled for analysis consist of a bright chemical weathering product that has been mixed on a global scale by the wind.

The one area seen at the landing sites that has a reflectance spectrum appreciably different from the spectrum of all the other areas is a drift of material some 15 meters away from *Viking Lander 1*. This drift, which is significantly darker than the surrounding soil, may have been left behind after the wind carried away brighter material that was more easily moved. The spectrum for the dark drift is similar to the spectra of the dark areas on Mars that have been made with instruments on the earth. The drift may be composed of partially altered iron-rich igneous rocks that have been broken down into soil particles. Thus although in general the similarities between the soils at the two sites are strong, differences do exist that may mimic on a small scale the differences between the classic bright and dark areas of Mars seen from the earth.

Most of the rocks and bedrock outcrops in the color pictures made by the landers appear to be darker than the soil on which they rest, but when they are viewed under comparable illumination, most of them are actually brighter. In the majority of the photographs the angle between the sun, the surface and the camera is larger for the sides of the rocks than it is for the soil on the ground, and hence the brightness of the rocks appears lower. The rocks and soil have similar spectral shapes, indicating that the rocks are covered with a smooth stain or dusting of material similar in composition to the soil. The general appearance of most of the rocks suggests that they are volcanic in origin, although some have been eroded significantly by the wind. Most of the rocks at the site of *Viking Lander 2* and some of those at the site of *Viking Lander 1* are pitted like the ones that form from gas-rich lavas on the earth. Gases dissolved in the lava because of the confining pressures at great depths are liberated when the lava reaches the surface and form pockets of gas. When the molten rock cools and hardens, the pockets are preserved as bubbles or vesicles in the rock. Unfortunately distinguishing between pits produced by this process and those produced by the wind erosion of softer minerals is extremely difficult.

One of the more intriguing aspects



CLOSEUP VIEWS OF MARTIAN ROCKS illustrate the variety of rocks on Mars. The top two photographs show rocks from Chryse Planitia; the bottom two photographs show rocks from Utopia Planitia. The rock at the top left is about 25 centimeters long; its pitted or mottled nature is reminiscent of igneous rocks on the earth that have been abraded by the wind. Duricrust exposed by the retrorocket exhaust can be seen to the lower left. The rock at the top right appears to be a breccialike volcanic rock about 20 centimeters across. A moat scoured by the wind can be seen on one side of the rock; a raised deposit of windblown material extends from the other side. The elongated pits in the soil were made when clods hurled out by the landing hit the ground. The rock at the bottom left is a rectangular block about 40 centimeters wide. The angular form of the rock indicates that its appearance is dominated by fracture planes at right angles to one another. Drifted material is visible in the background, along with a conical rock that may have been shaped by wind erosion. The peanut-shaped rock at the bottom right may be a piece of lava some 30 centimeters wide. The pits in it may have been created when gas dissolved in the lava formed bubbles or vesicles and then escaped, or they may have been pockets of softer minerals preferentially eroded by the wind. The landing sites display a variety of rock forms, some eroded, some appearing pristine. Most seem to be of volcanic origin.

of both landing sites, and probably of Mars in general, is the fact that there seems to be a marked deficiency of sand compared with a typical desert on the earth. The lack of sand at the landing sites has been inferred by Henry J. Moore II of the U.S. Geological Survey and his associates from photographs of the walls of the trenches that were dug by the soil samplers of both Viking landers. Their results indicate that the bulk of the small particles on Mars are smaller than .1 millimeter. Such particles would be difficult to erode from the surface, and once eroded they would probably be carried away by the wind.

The Winds of Mars

Particles the size of those of sand exist on Mars, probably in the dunes that girdle the Martian ice caps; perhaps they even make up the dark drift exposed at the site of *Viking Lander 1* and parts of the classic dark areas that cover the middle latitudes of the southern hemisphere. The Martian soil does not, however, contain the kind of sand found on the earth. Most terrestrial sand consists of quartz and feldspar minerals that have been weathered from acidic igneous and metamorphic rocks. Quartz and feldspar, which dominate acidic rocks, are resistant to chemical weathering and mechanical attack, and the two minerals are major constituents of sedimentary rocks on the earth. Mars has not differentiated to the extent that acidic rocks have been created in large quantities; the planet is probably dominated by mafic basalts. In such basalts the major minerals are olivine, pyroxene and plagioclase feldspar. Huguenin has shown that even in the cold, arid environment of Mars these minerals are rapidly weathered by ultraviolet-stimulated oxidation.

If, as is suspected, the bright soil at the landing sites is made up of clay minerals, the soil would be very fine-grained. Particles the size of grains of sand could exist, however, as aggregates of smaller particles. Such sand-sized aggregates would have a short lifetime in the Martian aeolian system. Grains bouncing along the surface travel roughly at the speed of the wind. Since Mars's atmosphere is only about a hundredth as dense as the earth's, winds of about 50 meters per second (180 kilometers per hour) would be needed to erode sand-sized particles from the Martian surface. Particles of terrestrial sand carried at that velocity would be powerfully erosive, but aggregates of clayey material would rupture on impact into harmless motes of dust. Even mineral grains such as olivine, pyroxene or feldspar would be likely to rupture if they were carried at such velocities. One consequence of this situation on Mars is that the rate of aeolian erosion is probably low compared with the rate on the earth. The low rate of erosion probably explains why much of the Martian surface looks so crisp and pristine. The loose material can be moved around readily, but the rock erodes very slowly.

The skies at the landing sites are yellowish brown, and have remained that color over the Martian year the landers have been on Mars. This observation is somewhat unexpected. According to calculations by Pollack, the color is due mainly to particles of dust suspended in the atmosphere up to a height of 40 kilometers. It was expected that large amounts of dust would be suspended in the atmosphere after the major dust storms, which occur when the planet passes perihelion, the point on its orbit where it is closest to the sun. The lander photographs showed, however, that the skies were yellowish brown even when Mars was at aphelion, the point on its orbit where it is farthest from the sun. At that time major dust storms are rare. Dust either is often raised on Mars or is dynamically supported by atmospheric turbulence for long periods of time. Even on the dustiest days, however, if all the dust were precipitated onto the surface, the layer would be only a fraction of a millimeter thick.

The landers have gathered data during two large dust storms, both of which began in the southern hemisphere as Mars neared perihelion in 1977. The first storm began in February and the second in May. The dust cloud was quickly distributed around the planet by high-altitude winds. Each storm took several months to subside. At the landing sites in the northern hemisphere the wind only rarely reached velocities sufficiently high to disturb the soil on the ground. It seems probable that after both storms a thin layer of dust accumulated on the landers and on the surfaces around them. The layer of dust indicates that not all the material raised from the southern hemisphere is returned. Over a period of time the dust storms should denude the middle-southern latitudes of dust, exposing the bedrock and leaving behind deposits of darker, less weathered material. Perhaps such deposits are the classic Martian dark areas seen from the earth.

The latitude on Mars that is directly below the sun at perihelion is the point at which most dust storms begin. This perihelion subsolar point slowly varies with time because the axis of Mars precesses. Martian precession should cause the latitude of the perihelion subsolar point to migrate between +25 degrees and −25 degrees in latitude with a period of 50,000 years. Lawrence A. Soderblom of the U.S. Geological Survey has pointed out that if the dark areas on Mars are regions stripped of a relatively large fraction of the bright layer of mobile dust, they may migrate around the equator with the same period. In other words, some 20,000 years ago most of the perihelion dust storms may have begun at the latitude of *Viking Lander 1*, and Chryse Planitia may have been partially stripped of its bright deposits. If that is the case, the drifts and streaks seen at the site of *Viking Lander 1* may be younger than 20,000 years.

The Future

It is clear that the Viking mission has greatly increased our knowledge of the geology of Mars. We now have a good understanding of what the surface looks like and of the kind of surface materials present. The discovery that at least one of the residual ice caps consists of water ice has significantly added to our knowledge of the intensity of the planet's climatic fluctuations. There remain important questions about the evolution of the Martian surface. The ages of the planet's various terrains are not known with much certainty. The structure and composition of the interior of Mars are still largely a mystery. Without such knowledge it is not possible to construct a unique theoretical model of how the planet formed and evolved.

Some of the questions may be answered by further analysis of the Viking data. Others will have to await future missions. One possible mission is an orbiter, much more sophisticated than the terrestrial satellite Landsat, capable of mapping detailed chemical and mineralogical characteristics of the Martian surface. Another possibility is an unmanned surface rover capable of traversing hundreds of kilometers over a period of several years, which would be able to analyze the regolith in more detail than any satellite in orbit could. A third possibility is a series of penetrator rockets launched from an orbiting platform that could embed themselves in the planet's surface at various points and provide a network of meteorological and seismological sensors. A fourth possibility is a mission that would return samples of Martian material, affording the kind of data that can be obtained only in a terrestrial laboratory. One need only consider the vast amounts of data that were obtained by analyzing the samples returned from the moon in the Apollo missions to appreciate the extent of the information that could be gained from samples returned from Mars.

The future exploration of Mars can be justified from a number of points of view. Perhaps the most important is that the earth and Mars seem to have followed evolutionary tracks that are sufficiently different for the two planets to have had unique histories. Their evolutionary tracks have been sufficiently similar, however, to provide for meaningful comparisons of data on their atmosphere, surface and interior. There can be little doubt that understanding the history of Mars will increase our understanding of the earth.

The Honda Civic. The car we

designed around a shopping bag.

It's true. We had supermarket shopping bags flown to our factory when we designed the Honda Civic® Hatchbacks. Measuring the bags helped us determine the size of the rear cargo area.

This may seem like a lot of fuss over a small detail. But at Honda we do everything that way. For all its simplicity, the Honda Civic has been planned with meticulous care.

So you see? It's not by accident that four full-sized shopping bags fit inside our hatch.

Of course, four also happens to be the number of full-sized people that fit inside our Civic's passenger compartment. And if we went to all that trouble to make a shopping bag comfortable, just imagine what we must have done to make you comfortable.

HONDA
We make it simple.

The Flow of Energy in a Forest Ecosystem

Life on the earth is sustained by the finite amount of solar energy that is fixed by green plants. Quantitative study of the energetics of a forest has shown how this energy is partitioned and controlled

by James R. Gosz, Richard T. Holmes, Gene E. Likens and F. Herbert Bormann

Ecology, like economics, concerns itself with the movement of valuable commodities through a complex network of producers and consumers. Just as an economy runs on money, so does an ecosystem run on energy, all of which comes initially from the sun. A basic functional unit of nature, an ecosystem comprises a group of living organisms and the physical and chemical environment in which they live. For our purposes it can be thought of as being made up of plants, animals, organic debris, available nutrients, soil minerals, water and gases, all linked by food webs and flows of energy and nutrients. Producing and consuming organisms interact in a self-regulating manner, usually in relation to the total amount of energy available in the ecosystem.

Although the natural world would seem to receive a virtually limitless influx of energy capital in the form of solar radiation, its energy budget is actually quite small. This paradox can be explained by the fact that living organisms do not utilize solar energy directly; they have access to only the small portion of it that is converted by green plants into a stored chemical form through the process of photosynthesis. The organic matter fixed by plants is utilized by animal consumers; the plants are also consumers to the extent that they utilize some of the energy stored in their tissues for their own maintenance. Animals are associated with a grazing food web (in which living plant tissue is the source of energy) or a detritus food web (in which dead tissues are the source).

Because most consumers are inefficient in their utilization of chemical energy the finite amount of fixed solar energy in an ecosystem places stringent limits on the number of plants and animals it can support. These organisms have accordingly had to evolve diverse strategies to obtain their share of the available energy. The analysis by our group of the energy budget of a forest ecosystem in the northeastern U.S. has begun to reveal in some detail how living organisms partition and control the movement of energy through the system.

There are three general ways to investigate the dynamics of energy in nature, each of which has its own inherent strengths and weaknesses. The most specific approach is population analysis, which estimates the energy budget of a particular population in the wild. The analysis is done by determining the amount of energy in the form of organic matter that is consumed, assimilated and excreted by individual organisms, together with the amount of energy required for growth and for maintenance activities such as respiration (the oxidation of foodstuff). These data, obtained in the laboratory, are then combined with demographic surveys of the population in its natural habitat to yield an estimate of how much energy the population utilizes. The population approach is time-consuming, however, and as a result only a limited number of species have been studied.

The second approach is food-chain analysis, in which the population energetics of only one or two species at different levels of a food chain is investigated. For example, a plant species may serve as food for a species of insect, which in turn serves as food for a species of bird. Each species is viewed as forming a single link in the food chain, and alternative pathways for energy flow are ignored. This approach can yield detailed information about how species interact to partition available energy resources, but it provides little useful information about the partitioning of energy within the complex food webs of an ecosystem.

The third and the broadest approach is ecosystem analysis, in which the amount of energy transferred between the consumer compartments in an ecosystem is quantified. Since the inputs and outputs of each compartment can be calculated, this approach makes it possible to draw up a balance sheet for energy flow. To provide a bookkeeping structure for the balance sheet topographical boundaries are defined for the ecosystem under investigation, so that energy flow can be expressed per unit of land or water per unit of time. The energy is usually expressed in units of kilocalories per square meter per year. A kilocalorie is 1,000 calories; one calorie is defined as the amount of energy required to raise the temperature of one gram of water one degree Celsius. One kilocalorie is also equivalent to 3.97 British thermal units (B.t.u.'s).

The ecosystem we have studied is in the Hubbard Brook Experimental Forest, which is operated by the U.S. Forest Service in the White Mountain National Forest of New Hampshire. Within the experimental forest are a number of watersheds underlain by a bed of impermeable bedrock and divided by ridges (topographic highs). One of them, designated Watershed 6, is maintained in a natural condition so that it can serve as a base-line reference for experimental studies of hydrology and biogeochemi-

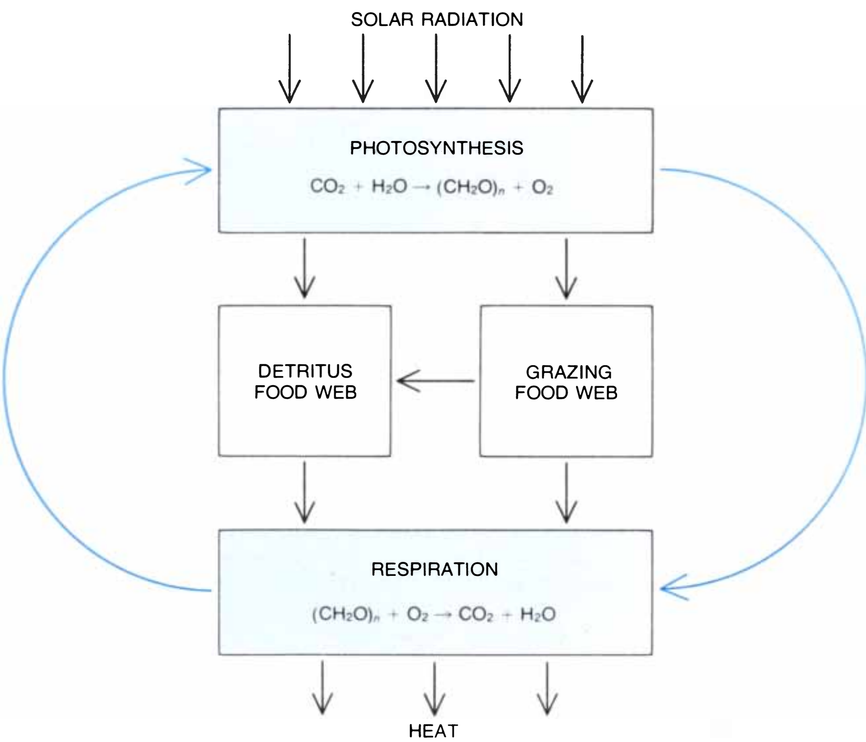
ABSORPTION OF SOLAR RADIATION by a hardwood forest at Temperate Zone latitudes is indicated in the aerial photograph on the opposite page, which was made with a special emulsion sensitive to near-infrared wavelengths. Plant foliage absorbs light primarily at blue and red wavelengths and reflects the rest; thus the healthy trees in this photograph appear red because they reflect the near infrared. The trees that appear green have been partially stripped of leaves by gypsy-moth caterpillars, reducing their ability to absorb and fix solar radiation. In years when defoliating insects are particularly abundant they may have a major impact on the functioning of the forest ecosystem by affecting the rate and amount of energy fixation. The photograph was provided by Robert L. Talerico of the Northeastern Station of the U.S. Forest Service.

cal cycling. The analysis of energy flow was done on this watershed, which has a total area of 132,300 square meters. The vegetation of the natural watershed is a well-developed, second-growth forest of sugar maple, beech and yellow birch, and it has been undisturbed by fire or cutting since 1919, when much of the first growth was removed by lumbering.

Although Watershed 6 has definite topographical boundaries defined by the underlying bedrock and the ridge divides, it is "open" in the sense that survival depends on the continuous input and output of energy in the form of both radiation and organic matter. Energy-containing organic matter can be transported across ecosystem boundaries by meteorological forces such as precipitation or wind, by geological forces such as running water and by biological vectors such as the movement of animals. All these avenues of transport must be taken into account in quantitatively describing the flow of energy through the ecosystem.

Measuring the movement of energy across ecosystem boundaries is usually a difficult task, but it was simplified in our watershed ecosystem by several factors. For one thing, since the ecosystem is underlain by impenetrable bedrock the geological output of organic matter could be easily quantified because it inevitably turned up in the streams draining the watershed. Moreover, biological transport across the ecosystem boundaries was considered negligible because the animals randomly consumed and discharged organic material as they moved in and out of the ecosystem from similar forests in surrounding regions. Measurement of the solar-energy input, the meteorological input and the geological output was therefore sufficient for us to estimate the total flux of energy through the ecosystem.

The amount of solar radiation received by an ecosystem during any given day is determined by the latitude, the climate and the time of year. There may also be variations from year to year as a result of changing weather patterns. We based our analysis of energy flow on solar-radiation data obtained from June 1, 1969, through May 31, 1970, a typical year for the Hubbard Brook region. During that period a total of 1,254,000 kilocalories per square meter of solar radiation reached Watershed 6, of which 10,400 kilocalories per square meter (.8 percent) was fixed in chemical form by plants through photosynthesis. The plants of this forest need 55 percent of the energy they fix for their own maintenance, so that the actual amount of new plant tissue formed that year (the net primary production) was about 4,680 kilocalories per square meter, or some .4 percent of the total annual input of solar energy. This percentage is the net primary production efficiency.



BASIC PATHWAYS of energy flow through an ecosystem all originate with solar radiation, a small percentage of which is converted by green plants into organic material through the process of photosynthesis. This organic material may be consumed either as living plant tissue by herbivores in the grazing food web or as dead tissue by decomposers in the detritus food web; these animals in turn provide the food base for a variety of carnivores. Assimilated organic matter that is not utilized for the growth of individual organisms or populations is "burned" by the process of respiration to power the metabolism of plants and animals. Because respired energy is ultimately lost as heat, an ecosystem requires a continual influx of solar energy.

The estimate for the net primary production efficiency is inaccurate for two reasons. First, it is misleading to express plant production efficiency in terms of total annual solar radiation, since the plants in this region are leafless or dormant for about eight months of the year. Although there is a low level of photosynthetic activity in some species during the winter months, virtually all the growth comes in the late-spring and summer months. Thus it is preferable to calculate the efficiency of photosynthesis on the assumption that plant growth is limited to the period from June through September. The total solar radiation during this four-month period was 480,000 kilocalories per square meter, meaning that the net primary production of 4,680 kilocalories per square meter would amount to about 1 percent of the solar-energy input, which is somewhat larger than our initial estimate of .4 percent.

Moreover, although the energy value for the total solar radiation includes the energies of all the wavelengths of electromagnetic radiation that reach the earth from the sun, the photosynthetic process does not utilize all wavelengths with equal efficiency. Chlorophyll absorbs light strongly in the blue and red portions of the visible spectrum and reflects most of the green light, whereas the other pigments in plant foliage, such as the carotenes and the xanthophylls, absorb the green wavelengths and transfer some of the absorbed energy to the chlorophyll molecules, where it can be utilized in photosynthesis. The combined effect of these various pigments is that the energy in about half of the wavelengths in the solar radiation is available for conversion into chemical form. Taking this into consideration in our calculations, it would appear that photosynthesis is slightly more efficient than it is described as being above: during the growing season about 2 percent of the solar energy of the appropriate wavelengths is converted into organic matter.

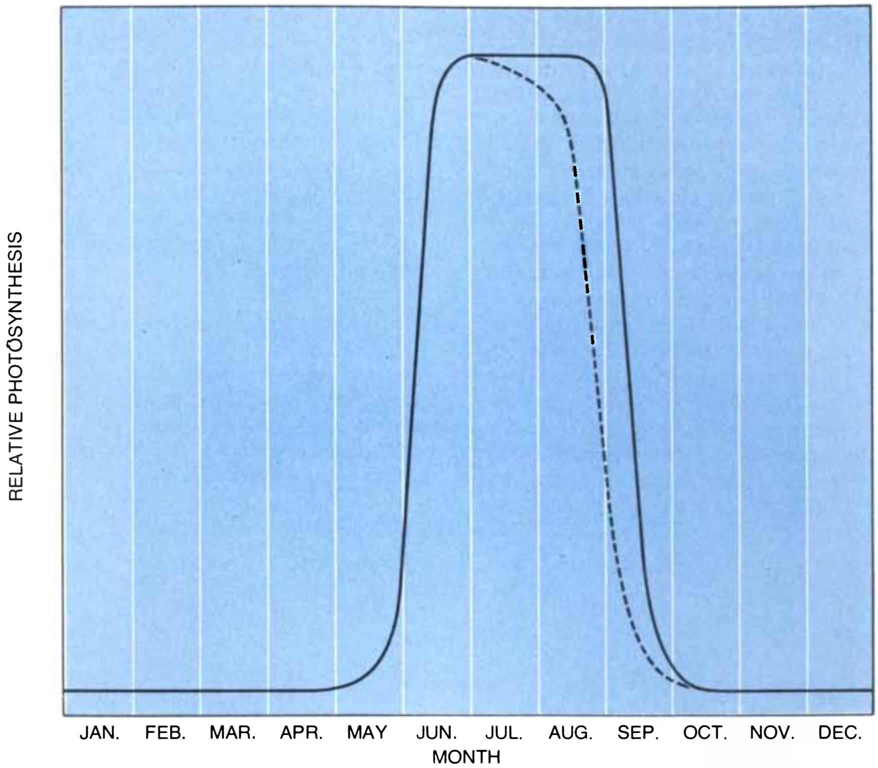
The fact that only a tiny fraction of the solar radiation reaching the ecosystem is converted into new plant tissue does not mean that the plant community is wasteful or that the remainder of the energy is not utilized. Much of the radiation from the sun generates heat when it is absorbed by the ground or other components of the ecosystem, and without heat to maintain normal levels of metabolism, photosynthesis could not operate. Another energy-dependent process crucial to the functioning of plants is transpiration: the movement of water from the soil through the plant and its evaporation at the surface of the leaf. Transpiration has been shown to be an important factor for normal plant growth, and it removes water without eroding the landscape or carrying away essential nutrients. The total amount of water removed from the ecosystem by nonbiological evaporation and transpiration together (evapotranspiration) can be calculated in our watershed by subtracting the output of water in the stream from the input of water in precipitation.

We have estimated that each year evapotranspiration removes a volume of water equivalent to a layer 49.4 centimeters (about 20 inches) deep over the entire area of the watershed. At the normal temperatures in our forest approximately 585 calories is required to evaporate a gram of water, so that evapotranspiration requires 288,990 kilocalories per square meter per year: about 23 percent of the total annual input of solar energy. In the growing season evapotranspiration removes 34.2 centimeters of water and consumes 200,000 kilocalories per square meter, or about 42 percent of the solar radiation for the four-month period. Thus with the exception of the green and infrared wavelengths that are reflected by plants most of the incoming solar radiation is responsible for creating the environmental conditions that enable the plant community to fix 4,680 kilocalories per square meter of energy into organic matter per year.

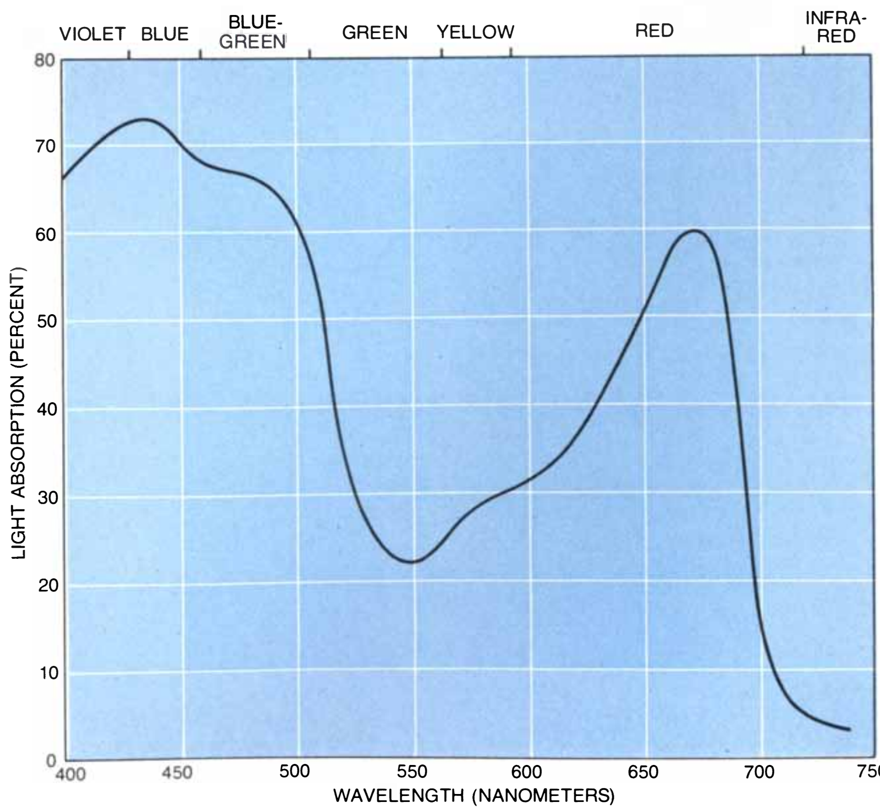
The other major input of energy into the ecosystem is the meteorological transport of organic matter, either in particulate form by the wind or in chemical form dissolved in rain and snow. Precipitation falling on the Hubbard Brook watersheds contains an average of about three milligrams of organic carbon per liter, which corresponds to an energy input of about 35 kilocalories per square meter per year. Although the meteorological input is small compared with the solar-radiation input, it is more than five times the energy contained in the organic matter leaving the ecosystem in stream water. Moreover, since the energy is in chemical form, it is directly available to consumers.

Once energy has been converted into chemical form by photosynthesis or transported into the ecosystem by wind or precipitation it may enter the grazing food web through the direct consumption of living plant tissue or enter the detritus food web of the soil. Alternatively it may leave the ecosystem through animal movement such as bird migration, as dissolved organic matter in the stream, as volatile organic matter in the air or as heat. If the plant production is larger than the amount of organic matter entering the various food webs and export pathways, the excess will be accumulated in the living-plant biomass: the total amount of living plant tissue in the ecosystem.

We have calculated that of the total yearly primary production of 4,680 kilocalories per square meter 3,481 kilo-



PHOTOSYNTHETIC PRODUCTION by plants is not uniform throughout the year but is largely concentrated in the four-month period from June through September; during the other eight months most of the trees in northern forests are in a leafless or dormant state. The solid curve shows the relative photosynthesis in the ecosystem under favorable summer conditions; the broken curve shows it under stressed conditions, such as low moisture. The efficiency with which plants convert solar energy into organic matter should therefore be calculated on the basis of the amount of solar energy received by plants in growing season and not the entire year.



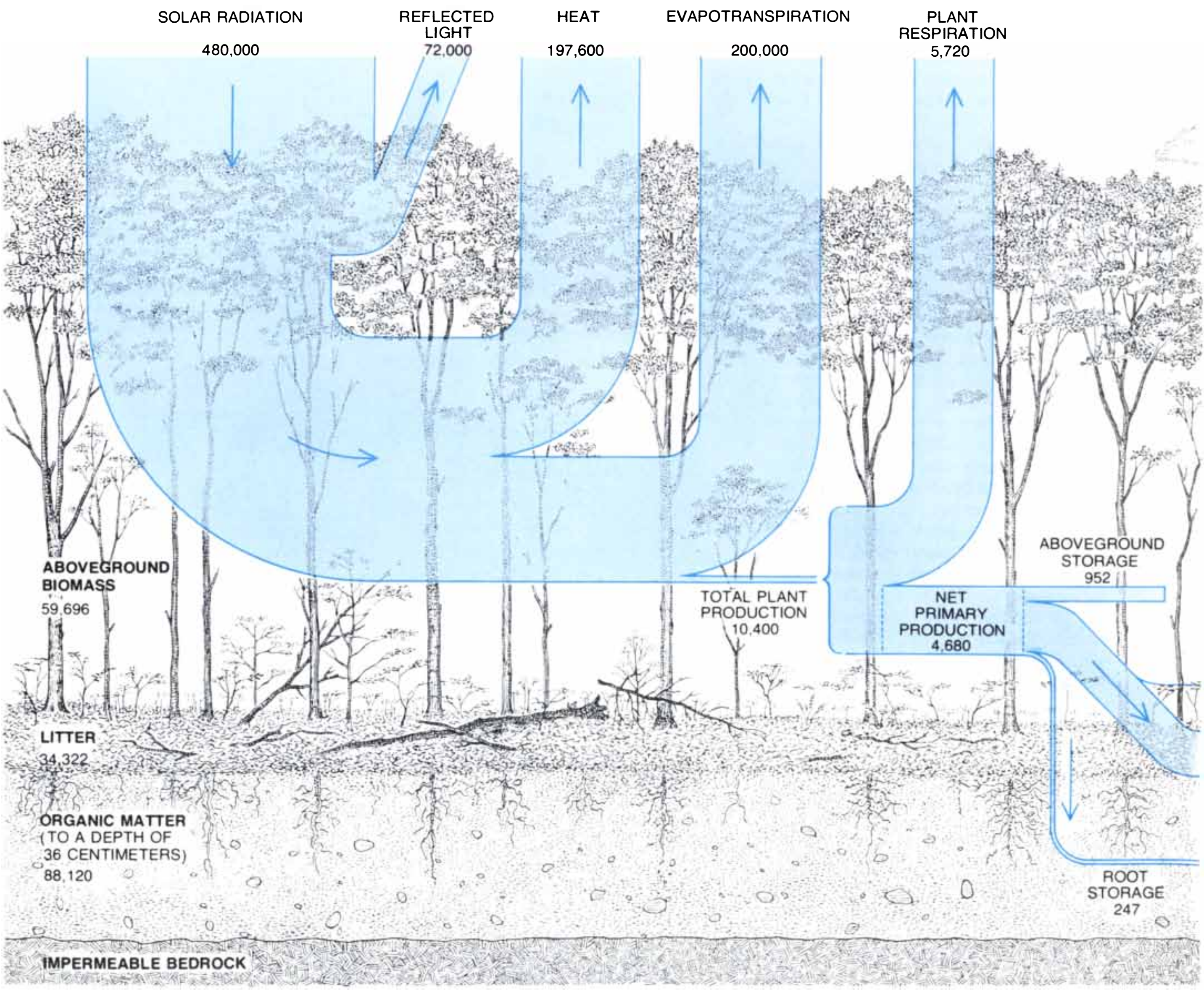
TOTAL ABSORPTION SPECTRUM for all the various pigments present in plant foliage indicates that about half of the wavelengths making up solar radiation are utilized in photosynthesis. Foliage reflects light most strongly in the green and infrared portions of the spectrum.

calories per square meter enters the detritus and grazing food webs and 1,199 kilocalories per square meter is stored in the vegetation biomass. Of the amount of energy stored 952 kilocalories per square meter is aboveground and 247 kilocalories per square meter is belowground. Thus it is clear that this second-growth forest is still growing and increasing its biomass following the lumbering operations early in the century.

The total living-plant biomass of the forest contained 71,420 kilocalories per square meter in 1970, with 59,696 kilocalories per square meter aboveground and 11,724 kilocalories per square meter belowground. Although the living biomass of the forest may appear to be the largest storage form of organic matter and energy, it is not; the organic matter lying on the forest floor and mixed into the soil contains 1.7 times as much energy. The floor of the forest contained 34,322 kilocalories per square meter in 1970; the organic matter in the soil to a depth of 36 centimeters contained 88,120 kilocalories per square meter. These figures indicate the magnitude of the detritus system in this forest and probably in most forests at Temperate Zone latitudes.

The living plant tissue that is available to consumers in the forest ecosystem is of three types: seeds and fruits, foliage and woody tissue. Seed production (80 kilocalories per square meter per year) and foliage production (1,485 kilocalories per square meter per year) are fractions of the year's net primary production. These tissues are consumed by organisms in both the grazing and the detritus food webs. The woody tissue produced in the current year's growth is normally stored in the vegetation biomass; the portion of woody tissue that is utilized by consumers is primarily dead stem, branch and root tissue that represents plant production from previous years.

The major consumers in the grazing food web are leaf-eating insects, primarily caterpillars. Although mice, chipmunks, deer and snowshoe hares consume some leaf tissue at or near the ground, the wide variation in the amounts consumed in different years is related mainly to fluctuations in the caterpillar population. Of the 1,485 kilocalories per square meter per year net pro-

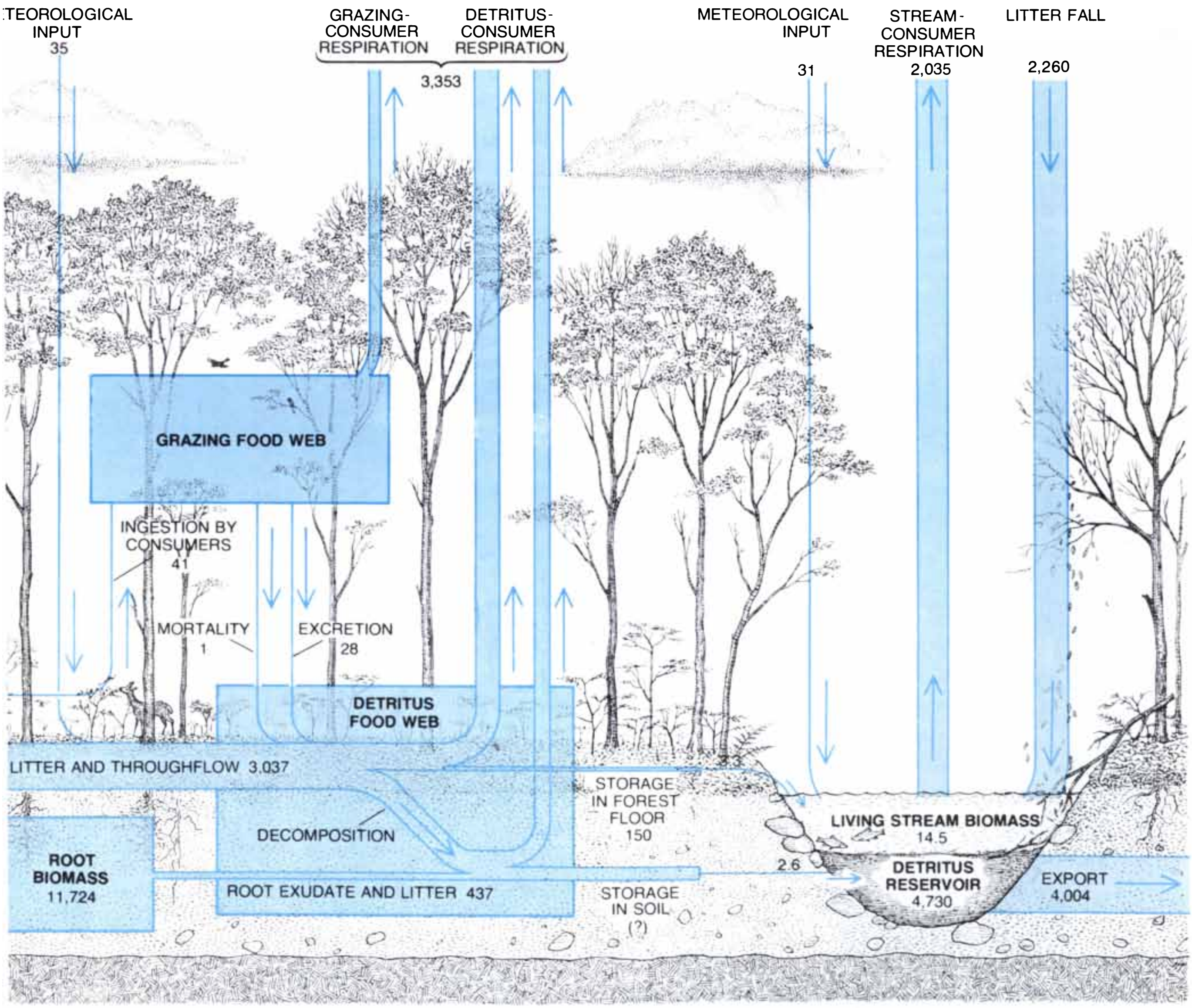


ENERGY IN A HARDWOOD FOREST is partitioned as is indicated in this diagram of an ecosystem within the Hubbard Brook Experimental Forest in New Hampshire. The ecosystem is a watershed that is bordered by ridge divides and underlain by impermeable bedrock. The numerals represent inputs and outputs of energy in units of kilocalories per square meter per year, with relative energies indicated by the width of the tubes (the width of the total plant production tube is arbitrarily expanded at the bracket). Energy inputs include the amount of solar radiation that is converted into organic matter by photosynthesis, together with direct inputs of organic matter carried by precipitation. Outputs include the portion of solar energy reflected or lost as heat; the energy consumed by the evaporation of wa-

duction of foliage in the Hubbard Brook forest these animals consume amounts ranging from less than 1 percent in most years to more than 40 percent in others. For example, in 1969 through 1971, when a species of defoliating caterpillar, the saddled prominent (*Heterocampa guttivita* Walker), was extremely abundant, about 44 percent of the total leaf tissue at Hubbard Brook was consumed. In some areas within the watersheds and in other parts of the northeastern U.S. trees and patches of forest were totally stripped of leaves. Such severe outbreaks occur only occasionally in northern forests, but they represent times when leaf-eating insects can have an important impact on the functioning of the ecosystem by affecting the rate and amount of energy fixation.

In most years about 75 percent of the net annual production in our ecosystem is not consumed by animals in the grazing food web or accumulated in the living-plant biomass. Instead it falls to the forest floor and enters the detritus energy pathway. The average annual amount of energy entering this pathway is estimated at 3,505 kilocalories per square meter. The fall of leaves, branches and trunks from the forest canopy is the major route of energy transfer, accounting for 83 percent of the total. The remaining 17 percent is derived from a variety of sources: root death (12 percent), litter fall from lower vegetation (2 percent), organic matter carried in precipitation or washed from foliage (2 percent), fecal material from animals in the grazing food web (.9 percent) and organic material exuded by roots (.1 percent). Only about one kilocalorie per square meter per year (.03 percent) is added from the death of animals in the grazing food chain.

The organic matter falling to the forest floor is utilized by consumers in the detritus food web, including bacteria, fungi and a variety of invertebrates such as protozoa, millipedes and certain insect larvae. These consumers in turn serve as prey for carnivorous invertebrates (such as centipedes, beetles and spiders) and certain vertebrates (such as shrews, salamanders, rodents and some birds). Since 75 percent of the energy in the net primary production enters the detritus pathway each year, an important question is whether the organisms of the detritus food web consume the



ter (evapotranspiration); the energy contained in the organic matter transported out of the system in the stream, and the energy lost as heat following the respiration of organisms in the ecosystem. An excess of inputs over outputs results in an increase in the total amount of energy stored in the ecosystem (the net ecosystem productivity). The majority of the energy contained in the annual production of plant tissue enters the detritus food chain either directly or indirectly, resulting in a large buildup of energy in and on the forest floor. Energy is also transferred from forest compartments of the ecosystem into the stream by seepage and litter fall. Because the ratio of land area to stream area is 570:1, small transfers of energy from each square meter of land amount to large inputs of energy into stream.

entire annual input of energy or whether there is an accumulation of organic matter on the forest floor. An analysis of litter fall and forest-floor characteristics for several northern hardwood forests in New Hampshire suggests that after logging activity there is a net decline of organic matter for 15 to 20 years, after which a net accumulation occurs for about 150 years. Although our forest ecosystem has been undisturbed for some 60 years, the forest floor is accumulating about 150 kilocalories per square meter per year. The sum of the increase in the living-plant biomass (1,199 kilocalories per square meter per year) and the energy content of the forest floor (150 kilocalories per square meter per year) represents the total net production of organic matter by the ecosystem.

One of the major outputs of energy from the ecosystem is the loss of organic matter to the streams draining the watershed. Since this factor can be quantified, it enables us to evaluate the energy dynamics of the aquatic portion of the ecosystem. There are three pathways for energy to enter the stream. One is by the primary production in the stream itself of aquatic plants, which utilize solar radiation directly. Our stream, however, is characteristic of most small headwater streams in densely forested regions in that it is heavily shaded and contains a very small amount of plant life, mostly mosses. As a result the contribution of energy by aquatic plants is only 10 kilocalories per square meter of streambed per year, about .2 percent of the total input of energy into the stream.

The two remaining pathways, the transfer of organic matter into the stream from the surrounding forest and forest floor, and the movement of particulate and dissolved organic matter into the streambed in drainage water, account for the majority of the energy entering the stream: a total of 6,039 kilocalories per square meter of streambed per year. Expressing the energy input per square meter of stream area is useful in considering the energy dynamics of the stream, but it is somewhat misleading in discussing the total transfer of energy from the terrestrial portion of the ecosystem, since the streambed occupies only about .2 percent of the area of the watershed. Although the energy contributions by litter fall from the forest canopy and precipitation are direct in that one square meter of stream surface is functionally equal to one square meter of land, litter that is blown into the stream and water draining into the stream are subject to a concentrating effect. Because of the high ratio of land area to stream area relatively small losses from each square meter of terrestrial surface make large contributions to the energetics of the stream. In the Hubbard Brook ecosystem the total input of energy to the stream from the forest is about 1 percent of the net productivity of the vegetation in the forest.

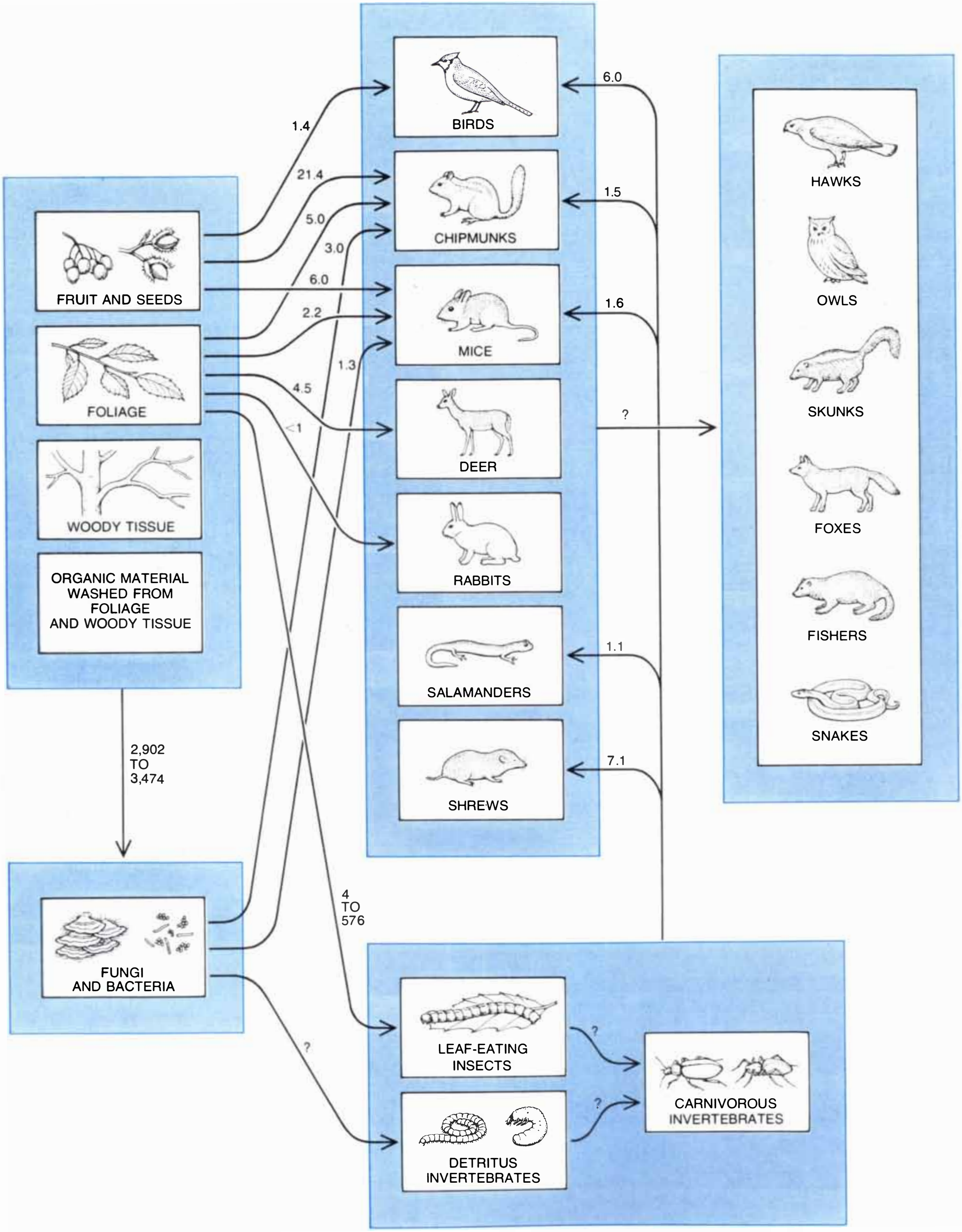
The output of energy from the stream flows primarily through two pathways: the conversion of chemical energy into heat by the respiration of stream organisms and the downstream movement of organic matter in the water. About a third of the energy passing through the stream ecosystem as organic matter is utilized by stream organisms, such as bacteria, insects and trout. Because the loss of energy by the respiration of these organisms is much larger than the energy fixed by the photosynthesis of aquatic plants, the consumers in the stream are dependent on the energy input from the surrounding land areas. The energy-containing organic matter that is carried away by the flowing water (4,004 kilocalories per square meter per year) apparently serves as an energy source for aquatic organisms farther downstream, where inputs from the adjacent land areas are proportionately smaller.

Most of the energy stored in the stream at any one time is held in the organic matter of the detritus reservoir (4,730 kilocalories per square meter of streambed), with only 14.5 kilocalories per square meter in the living biomass. The energy stored in the stream is small compared with the amount of energy moving through it during the course of a year. The ratio between the amount of energy present in a segment of the stream as detritus and biomass and the annual flux of energy passing through that segment is called the residence time. In our stream the residence time is .8 year, that is, on the average the energy contained in organic matter remains in a given portion of the stream for less than 10 months. As the organic matter decomposes or is flushed out it is replaced by transfer from the adjacent land areas or upstream areas, usually within the same year. (Debris dams, whose structural framework is provided by large logs, are an exception and may remain for several years.) The residence time for energy in the forest is generally much longer: about 19 years for the energy fixed by photosynthesis. This figure is of course an average, since some energy is lost quickly (such as that used up in plant respiration) and some is stored in tree biomass and released many years later when the tree is decomposed.

SADDLED PROMINENT (*Heterocampa guttivita*) is a defoliating caterpillar indigenous to the Hubbard Brook region. In years when it is unusually populous it may consume more than 40 percent of the annual production of foliage. The caterpillars assimilate only about 14 percent of the energy in ingested leaf tissue; the rest is excreted and enters the detritus food web.

The amount of energy in food that is converted into the body tissue of a consumer population depends on several factors. After consumption a certain amount of energy is not assimilated by the organism and is excreted as fecal material. A portion of the assimilated energy is then utilized by the organism for maintenance activities such as respiration; that energy is ultimately converted into heat and lost from the system. The portion of energy in excess of the maintenance requirement is converted into the growth of the population, either as an increase of body weight or as additional offspring. The utilization of energy in a consumer population for growth or increased reproduction is referred to as secondary production.

Each species of consumer in the eco-

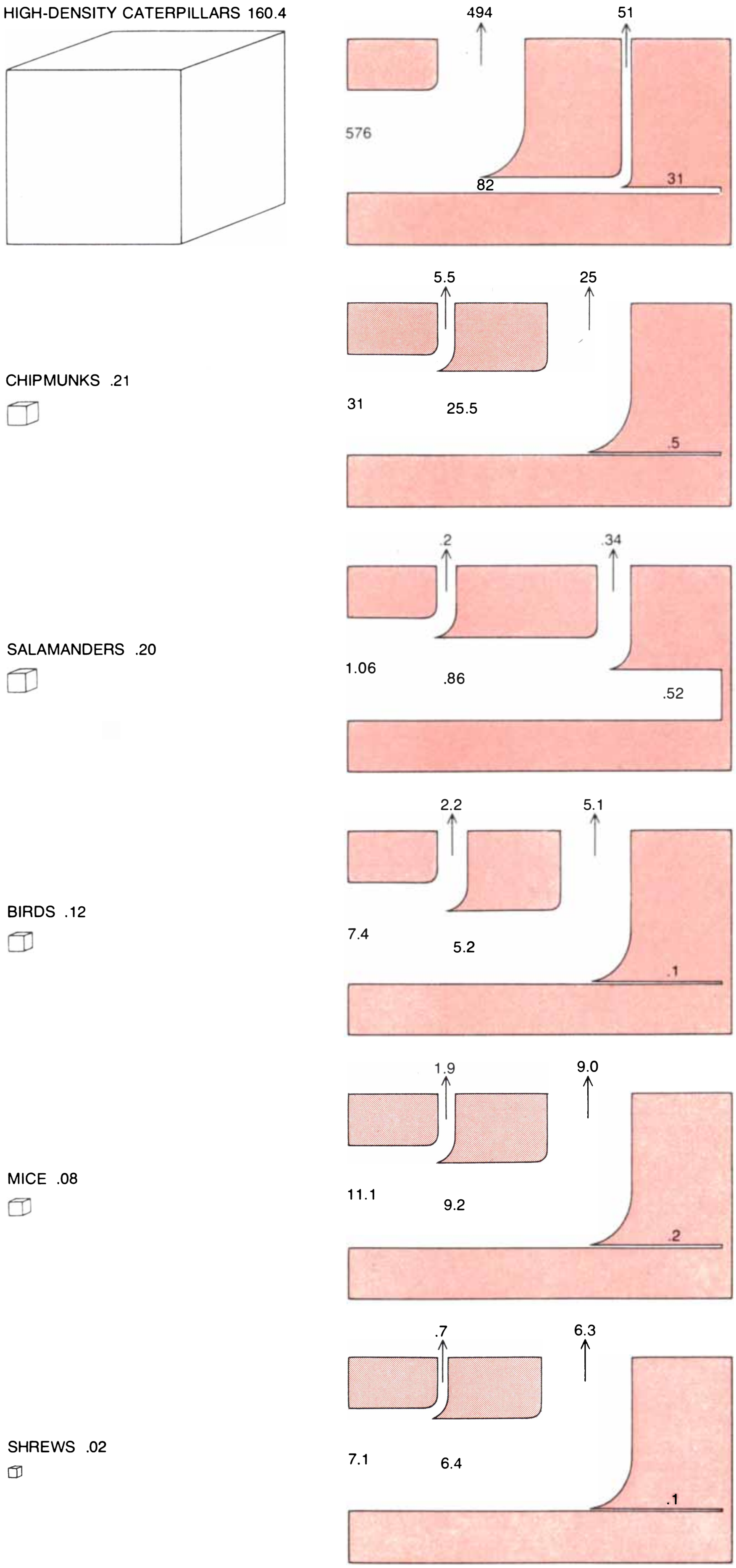


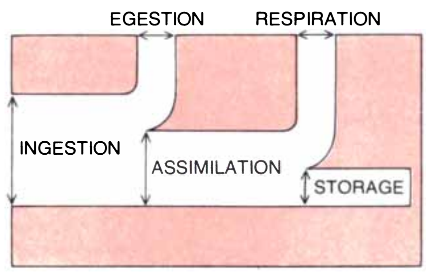
FOOD WEBS of consumers in the Hubbard Brook ecosystem are diagrammed, together with the consumption rate of organisms in each population. All numerical values are in units of kilocalories per square meter per year. The consumption of leaf tissue by herbivorous insects varies greatly from year to year and may have widespread effects on the utilization of energy by other consumers. In most years a large amount of organic material falls to the forest floor and enters the detritus food web, where it is utilized by fungi, bacteria and certain invertebrates. These organisms serve as food for carnivorous invertebrates, salamanders, shrews and some animals primarily associated with the grazing food web. Birds participate in the grazing food web by eating berries and caterpillars, but they also are able to tap the large detritus energy pool characteristic of northern hardwood-forest ecosystems by feeding on insects whose larvae feed on detritus.

system differs in its abundance and in the efficiency with which it utilizes energy. Although detailed information on the energy requirements of the decomposers, such as bacteria and fungi, and their invertebrate predators is not yet available, we do have estimates for several animal species in the Hubbard Brook ecosystem.

Caterpillars may consume large quantities of foliage, but the rate at which they extract energy from their food is low. Our measurements indicate that *Heterocampa guttivita* assimilates only about 14 percent of the energy in the leaf tissue it ingests. The remainder is excreted as fecal material and falls to the forest floor, where it enters the detritus food web. Of the energy assimilated by the caterpillar about 60 percent is utilized for respiration and about 40 percent is incorporated into new tissue. This tissue production forms a portion of the energy base for a variety of consumers, including invertebrate predators, birds, and mammals of both the detritus and the grazing food webs.

Shrews and salamanders are the two most important groups of vertebrate animals that capture and eat the invertebrates of the forest floor. There are two species of shrews and five species of salamanders within our ecosystem, all of which rely primarily on invertebrates for food. Since shrews maintain a high and relatively constant body temperature, they are active throughout the year, even under the winter mantle of snow. They utilize roughly 98 percent of their assimilated energy for respiration and only about 2 percent for growth or





ENERGY BUDGETS for the major aboveground consumer organisms in the Hubbard Brook forest are represented schematically. The size of the cubes is proportional to the biomass of each population at midsummer, when it is maximal. The tubes passing through the boxes represent the efficiency of energy utilization by each organism. Energy ingested in the form of organic matter is either passed through the digestive tract and excreted or assimilated by the organism. Assimilated energy is then utilized in respiration or channeled into the production of new tissues through growth or reproduction. The "cold-blooded" animals, such as insects and salamanders, are more efficient at channeling energy into increased biomass, whereas "warm-blooded" vertebrates, such as mammals and birds, utilize about 98 percent of their assimilated energy for respiration and only about 2 percent for growth.

reproduction. With such a large metabolic requirement shrews must eat often to survive. An adult short-tailed shrew (*Blarina brevicauda*) at Hubbard Brook weighs about 15 grams and requires some 15.2 kilocalories of food energy per day. The prey of the shrews (spiders, beetle larvae and other ground-dwelling invertebrates) contain on an average about .11 kilocalorie per individual prey item, so that a typical shrew would need 138 such "meals" per day. Assuming that the shrew is active for 12 hours of the day, it would have to capture and eat one prey item about every five minutes. These values would vary somewhat depending on the density and the type of the prey, on the temperature and on other environmental variables, but it is clear that there is little time in a shrew's life for anything but eating.

In contrast to shrews, salamanders are cold-blooded amphibians that use proportionately less energy in respiration, particularly during the cold months of the year, when their metabolic processes are significantly slowed. The salamander population consumes only about a sixth as much energy as the shrew population (1.1 kilocalories per square meter per year compared with 7.1 kilocalories), but approximately 60 percent of its assimilated energy is channeled into the production of new individuals. Thus salamanders are very efficient at transforming assimilated energy into increased biomass. It is interesting that shrews and salamanders, both of which are vertebrate animals occupying similar habitats within the ecosystem, have evolved such divergent strategies.

Birds are also an important group of insect predators in the forested watersheds at Hubbard Brook. They vary in abundance from a maximum of 322 individuals per 10 hectares (24.7 acres) in midsummer, when the young have just left the nest, to a minimum of 14 per 10 hectares in midwinter. By combining information on population densities and biomass fluctuations throughout the year with laboratory estimates of energy utilization per unit mass we have estimated that the bird population at Hubbard Brook assimilates 5.2 kilocalories per square meter per year. As with the shrews, 98 percent of this energy is utilized for respiration, with the remaining 2 percent going into the production of new tissue.

The bird population of the forest is very mobile, and many species migrate south before the winter season, thereby transporting energy out of the system. Our data indicate that the biomass of the bird population in the forested ecosystem is fairly constant from year to year, suggesting that the production of new tissue (.1 kilocalorie per square meter per year) balances the loss of bird biomass to mortality. It is difficult to assess how much of the mortality is inside the ecosystem and how much is outside, but in any case the amount of energy involved is quite small. Even if we assume that all the mortality in migrating birds occurs outside the ecosystem, the maximum possible amount of energy lost to migration would be .1 kilocalorie per square meter annually, a minute fraction of the total energy flux through the system.

Of the food ingested by birds in the forest 81 percent is derived from animal sources, almost exclusively insects; the remainder is plant matter, mostly fruits and seeds obtained during the fall and winter. In years when leaf-eating caterpillars are abundant these insects make up the bulk of the bird population's diet, and birds are hence clearly linked to the grazing food web. In years when caterpillars are not abundant, however, we have found that many species of forest birds eat predominantly adult insects, mostly flies, wasps and beetles, all of which are linked in some way to the detritus food web. For example, the adult flies are largely represented by groups whose larvae feed on dead organic matter on the forest floor (the houseflies, the Muscidae, and the fungus gnats, the Mycetophilidae) or in the streams (the black flies, the Simuliidae).

The same is true of the adult beetles, whose larvae prey mainly on soil-dwelling organisms, and the adult wasps (mostly the Ichneumonidae), whose larvae are parasitic on insects in the detritus food web. It therefore seems that although birds exploit leaf-eating insects when they are available, a significant portion of their summer diet—and consequently of their total energy intake—consists either of adult insects whose larvae feed on detritus or of predators on these insects. In this way birds are able to tap the large detritus energy pool characteristic of forest ecosystems at Temperate Zone latitudes. This finding also underlies the variety of routes that energy can take in such an ecosystem.

The rodent fauna of the Hubbard Brook forest consists mainly of the eastern chipmunk and three species of mouse: the deer mouse, the woodland jumping mouse and the red-backed vole. These rodents derive energy directly from the primary producers (leaves, shoots, seeds and buds), from other consumers (invertebrates in the forest litter and occasionally birds' eggs and young) and from decomposers (fungi); the prey therefore include members of both the grazing and the detritus food webs. Collectively the mice consume an estimated 11.1 kilocalories per square meter per year, whereas the chipmunk, because of its larger body size and higher population density, consumes nearly three times as much. In fact, the single species of chipmunk in our ecosystem utilizes more energy than all the mice, shrews, birds and salamanders combined, making it potentially more important in the pattern of energy flow than the other animal species.

The overall diagram of energy flow through the consumer compartments of the ecosystem is now fairly well worked out. It is clear that the largest proportion of the energy fixed by plants in this northern hardwood forest is channeled to the decomposers in the forest floor. As we pointed out above, however, some of this energy is still available to the larger consumers, since it supports many populations of insect-eating invertebrates, amphibians and mammals. Even the birds in the forest canopy manage to exploit the detritus energy pool by feeding on adult insects whose larvae feed on detritus.

Although we have made considerable progress toward understanding the energetic relations and hence the functional relations within the forest ecosystem, the relations among the consumer compartments and the factors that regulate the flow of energy within the system are just beginning to be elucidated in their full complexity. For example, how important are the leaf-eating insects in affecting the rate of energy fixation by plants? As we have mentioned, in some years insect defoliators consume as much as 44 percent of the green leaf tissue in the ecosystem, which suggests that they have a major impact on primary production, but the long-range effects of insect defoliation on plant productivity still remain to be determined. Similarly, what influence do seed predators (bird, mammal and insect) have on the reproduction of forest trees? And how do organisms in the detritus food web regulate the rate of nutrient cycling and consequently the productivity of the forest? Since many soil organisms feed on living root tissue, they may act as consumers in the grazing food web and hence affect the ability of plants to absorb nutrients and fix energy.

These and other questions must be answered before we shall have a complete understanding of the regulatory effects of consumers on the functioning of forested ecosystems. Additional questions arise when we consider the influence of man's activities on natural systems. For example, atmospheric pollutants, such as ozone and acidic chemicals dissolved in precipitation, are known to have adverse effects on vegetation, but little is known about their effects on consumer or decomposer organisms. Do the consumers have a greater impact on energy flow when the plant community is stressed by an external toxicant? Such questions are not merely academic. An in-depth understanding of the natural regulatory mechanisms influencing ecosystem functioning is essential to the development of sound forest-management plans.

Get your musical act together with a Sansui GX-7 deluxe rack.

Combine the superior performance of Sansui's high quality audio components with the beauty and convenience of Sansui's deluxe new GX-7 rack. Then sit back and enjoy a unique musical experience.

With the versatile GX-7, which comes with three adjustable shelves, you can create the high fidelity system that's ideal for you. Choose from Sansui's full line of high performance receivers, cassette decks and turntables. We've chosen, for example, an economical mid-power semi-automatic system including the G-3000 low-distortion FM/AM receiver, the SR-232 belt-driven turntable with very low noise and rumble, and the SC-1100 cassette deck which features Sansui's exclusive Direct-O-Matic loading and Tape Lead-In.

Whatever Sansui system you choose, you'll always enjoy the many advantages of the GX-7. It comes with separators and plenty of room for your favorite records. The tempered glass door, which opens at the touch of a finger, protects your equipment from dust. And so that you can move your system from place to place, the GX-7 stands on casters.

Visit your franchised Sansui dealer today for a look at the elegant GX-7. We're sure you'll appreciate the superior craftsmanship and the handsome simulated walnut grain finish.

The Sansui deluxe GX-7 rack. It's the ideal way for you to enjoy the very finest in high fidelity.

Sansui. A whole new world of innovation.

SANSUI ELECTRONICS CORP.
Woodside, New York 11377 • Gardena, California 90247
SANSUI ELECTRIC CO., LTD., Tokyo, Japan
SANSUI AUDIO EUROPE S.A., Antwerp, Belgium • In Canada: Electronic Distributors

Sansui

How Cells Make ATP

The prevailing theory is the "chemiosmotic" one. Light or oxidation drives protons across a membrane; then the energy-rich compound ATP is formed as the protons flow back through a complex of enzymes

by Peter C. Hinkle and Richard E. McCarty

Energy acquired by living cells is conserved in useful form mainly as molecules of adenosine triphosphate, abbreviated ATP. Whether the energy comes ultimately from light or from the oxidation of organic compounds, most of it is invested in the manufacture of ATP, which then serves as an "energy currency" that can be spent in powering the other functions of the cell.

The overall chemical reactions that lead to the synthesis of ATP are now well understood. In the chloroplasts of green plant cells hydrogen atoms—or electrons and protons—are extracted from water. The protons are released into solution and the electrons are driven by the energy of light through a sequence of carrier molecules. Eventually the electrons and the protons combine with carbon dioxide to form organic molecules. In the mitochondria of all cells with nuclei electrons donated by organic molecules are passed along a similar chain of carrier molecules and are ultimately accepted by oxygen, forming water. Many of the intermediate stages in these energy transformations have been worked out, but one crucial step has remained puzzling. It has not been clear how the transfer of electrons through the series of carrier molecules is coupled to the synthesis of ATP. Enough energy for ATP formation is made available by the electron transfer, but the mechanism of ATP synthesis has proved difficult to characterize.

A hypothetical mechanism for coupling electron transfer to ATP synthesis was proposed in 1961 by Peter Mitchell of the Glynn Research Laboratories in England. Mitchell suggested that the flow of electrons through the system of carrier molecules drives positively charged hydrogen ions, or protons, across the membranes of chloroplasts, mitochondria and bacterial cells. As a result an electrochemical proton gradient is created across the membrane. The gradient consists of two components: a difference in hydrogen ion concentration, or *p*H, and a difference in electric potential. The synthesis of ATP is driven by a reverse flow of protons down the gradient. Mitchell's proposal is called the chemiosmotic theory. In the past 15 years work by Mitchell and his colleague Jennifer M. Moyle and by many other workers has shown that the basic postulates of the chemiosmotic theory are almost certainly right, although some of the details of the theory are still controversial.

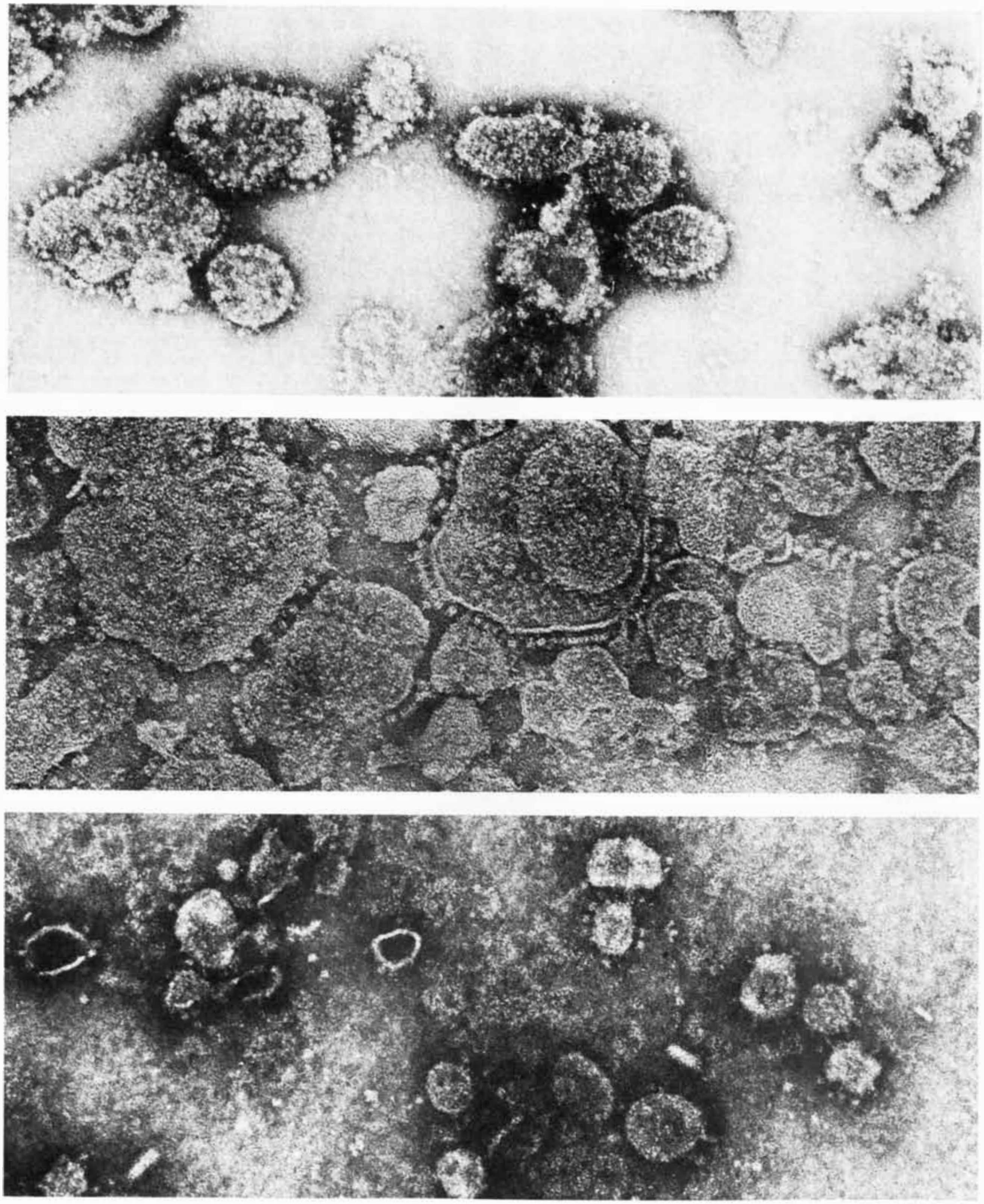
In the chemiosmotic theory there is no isolated molecular engine where the flow of energy from light or from oxidation is coupled to ATP synthesis. Instead the crucial role is played by a membrane that divides one region from another. The membrane provides much more than shelter, confinement and a controlled internal milieu. It is an asymmetrical arrangement of the carrier molecules across the membrane that allows the proton gradient to be established. Even the topological properties of the membrane are important: it must form a closed envelope if the proton gradient is to be maintained, so that in energy metabolism the most important distinction is the one between inside and outside.

An essential concept for an understanding of ATP synthesis is that of free energy, which measures the amount of energy in a chemical system available for doing work. Any chemical reaction, whether or not it takes place in the living cell, can proceed only in the direction of lower free energy.

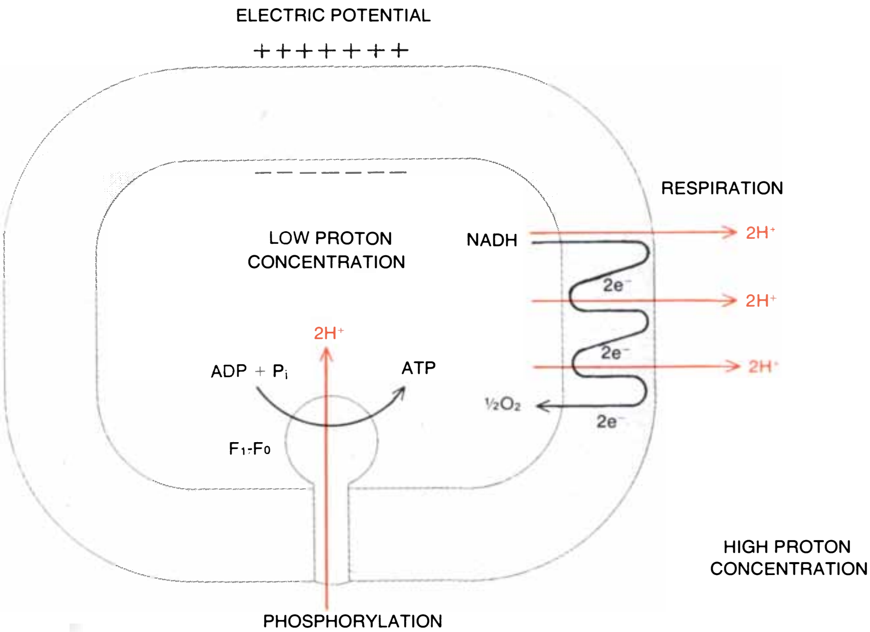
The reaction of carbohydrate with oxygen (to yield carbon dioxide and water) releases a large quantity of free energy. Hence the reaction, which takes place in the physiological process called respiration, is thermodynamically favored. In photosynthesis the same overall reaction proceeds in reverse: carbon dioxide and water are combined to yield carbohydrates and molecular oxygen (O_2). By itself this reaction would call for an increase in free energy, and so it can take place only when energy is supplied from an external source. The energy, of course, comes from sunlight; several photons, or quanta of light, must be absorbed for every molecule of carbon dioxide converted into carbohydrate. The photons are absorbed by pigment molecules, which are thereby promoted to an excited state of high free energy. The overall reaction, in which the excited pigments are considered among the reactants, proceeds as required by the laws of thermodynamics in the direction of lower free energy.

Energy constraints of the same kind apply to the synthesis of ATP. The ATP molecule consists of a nitrogen-containing organic base (adenine), a sugar (ribose) and a string of three phosphate groups. It is one of the class of molecules called nucleotides. In most of the reactions where ATP serves as an energy currency only the terminal phosphate group is involved. ATP is formed by attaching a third phosphate to adenosine diphosphate (ADP) with the elimination of water. The reaction does not proceed spontaneously; on the contrary, a substantial quantity of energy must be supplied. Much of that energy can then be recovered in the reverse reaction, in which ATP is split into ADP and inorganic phosphate. By donating the terminal phosphate group to other molecules, ATP creates phosphorylated species of high free energy, which can participate in reactions that otherwise could not take place.

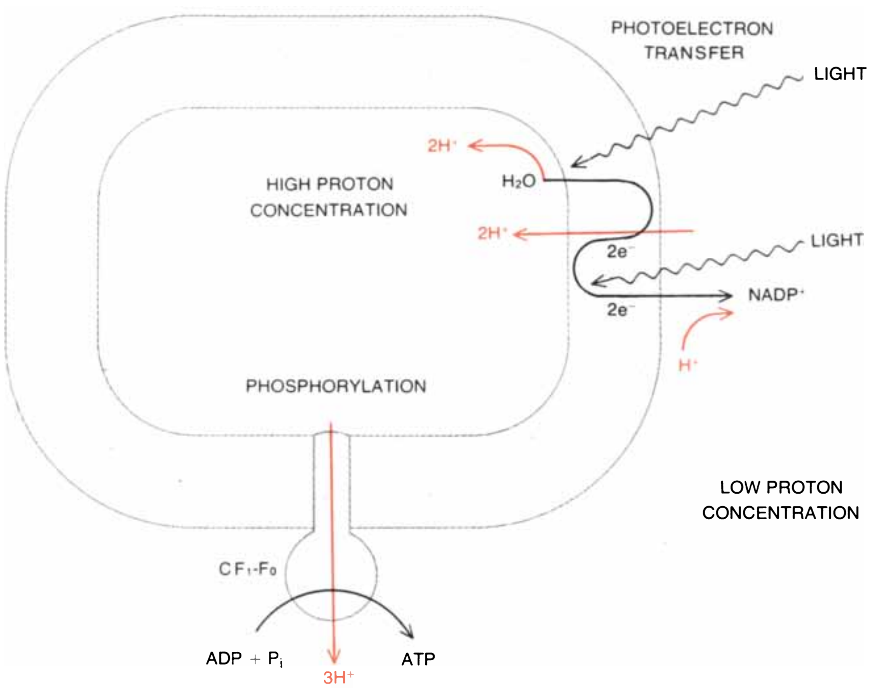
The energy transactions of biochemistry are conveniently measured in units of kilocalories per mole. One kilocalorie is the quantity of energy needed to raise the temperature of a kilogram of water one degree Celsius. A mole is the amount of a substance in grams that is numerically equal to its molecular weight; a mole of the sugar glucose, for example, weighs 180 grams. The complete oxidation of glucose yields about 700 kilocalories per mole. The amount of energy needed to form ATP depends on the chemical environment, but it is never more than about 15 kilocalories per mole. Hence each mole of glucose oxidized provides enough energy for the



MEMBRANE VESICLES display "knobs" on their surface that have been identified as part of an enzyme complex capable of synthesizing adenosine triphosphate, or ATP. The vesicles at the top were prepared from the inner membranes of the subcellular organelles called mitochondria; those in the middle are from the inner membranes of chloroplasts, the photosynthetic organelles found in green plant cells; the vesicles at the bottom are from membranes of the bacterium *Escherichia coli*. All the membrane systems are capable of making ATP, and their similarity in appearance suggests that they employ similar mechanisms. The knobs are made up of several protein subunits, known collectively as F_1; they are attached to a membrane protein called F_0. In the vesicles shown here the F_1 protein protrudes from the outer surface of the membrane, but that is its natural orientation only in chloroplasts. In mitochondria and bacteria the protein is ordinarily inside; it is outside here because the disrupted membranes form inside-out vesicles. The mitochondrial and chloroplast vesicles were prepared by the authors and the bacterial vesicles by E. Hertzberg. The micrographs were made by John N. Telford of Cornell University.



OXIDATIVE PHOSPHORYLATION couples the release of energy by the oxidation of molecules derived from carbohydrates and fats to the synthesis of ATP. In cells with nuclei oxidative phosphorylation takes place only in mitochondria. The basic events of the mitochondrial process are depicted here as they are interpreted in the chemiosmotic theory. Electrons and protons (or hydrogen atoms) from carbohydrates and fats are conveyed by molecules of the hydrogen carrier NADH to a system of enzymes in the mitochondrial membrane. In respiration pairs of electrons cross the membrane three times, each time transporting two protons from inside the mitochondrion to outside. The result is a gradient in proton concentration and electric potential that tends to force the protons back through the membrane. The energy of the gradient drives the process of ATP synthesis. For each two protons traversing the F_1-F_0 complex one molecule of ATP is formed from adenosine diphosphate (ADP) and inorganic phosphate (P_i).



PHOTOSYNTHETIC PHOSPHORYLATION in chloroplasts derives the energy needed for making ATP from light. As in oxidative phosphorylation, hydrogen ions are transported across the membrane to create a proton gradient, and ATP is synthesized as the protons flow back across the membrane down the gradient. In chloroplasts, however, the direction of proton flow is reversed: the light-driven movement of electrons pumps protons inward, making the interior acidic, and phosphorylation is driven by an outward flow. Moreover, the stoichiometry, or ratio of reactant molecules and ions, is different from that in mitochondria. Each two electrons cross the membrane only twice, translocating only four protons, and for each molecule of ATP formed three protons must pass through the enzyme complex, which is designated CF_1-F_0.

synthesis of about 46 moles of ATP. Actually no more than 36 moles of ATP, and perhaps as few as 25, are formed for each mole of glucose oxidized. The coupling of these reactions is clearly allowed thermodynamically.

If a chemical process leads to a state of lower free energy, the reaction can proceed, but it should not be assumed that it necessarily will. Glucose and oxygen, for example, are quite stable at room temperature; they react (the glucose burns) only when they are heated. The applied heat represents an activation energy, which is recovered along with the 700 kilocalories per mole evolved by the oxidation. In biological systems the need for activation energy is reduced by the catalytically active proteins called enzymes. Because virtually all biochemical reactions are mediated by enzymes they can take place at physiological temperatures and pressures and they are confined to particular chemical pathways. It is important to note that enzymes do not affect the direction in which a reaction proceeds, only the rate of the reaction. The action of an enzyme is somewhat like that of a lubricant: it cannot make the reaction go uphill, but it can make it roll downhill faster.

In oxidative metabolism almost 95 percent of the energy captured in ATP is derived from the transfer of electrons from carbohydrates or other substrates to oxygen. In photosynthesis all the energy is derived from the transport of electrons in the other direction, from water to carbon dioxide. The chemiosmotic theory is concerned with those stages of metabolism that begin with electron transfer and end with ATP synthesis. In mitochondria and bacteria these processes are called oxidative phosphorylation; in chloroplasts and photosynthetic bacteria they are called photosynthetic phosphorylation. Many of the details of both systems remain uncertain. What we shall present here is the theory of phosphorylation that now seems to be most nearly in accord with experimental observations.

The overall chemistry of oxidation and of photosynthesis consists in a transfer of hydrogen, but it is not necessary at each stage of the process to transport complete hydrogen atoms. Indeed, in the chemiosmotic theory hydrogen carriers alternate with molecules that carry only electrons. The use of electron carriers is possible because protons are soluble in water and in the aqueous medium of the cell, whereas electrons are not. When a molecule that bears a whole hydrogen atom interacts with another molecule that accepts only electrons, the proton is released into solution. When an electron carrier then donates its electron to a hydrogen carrier, the hydrogen atom is reconstituted by withdrawing a proton from the medium.

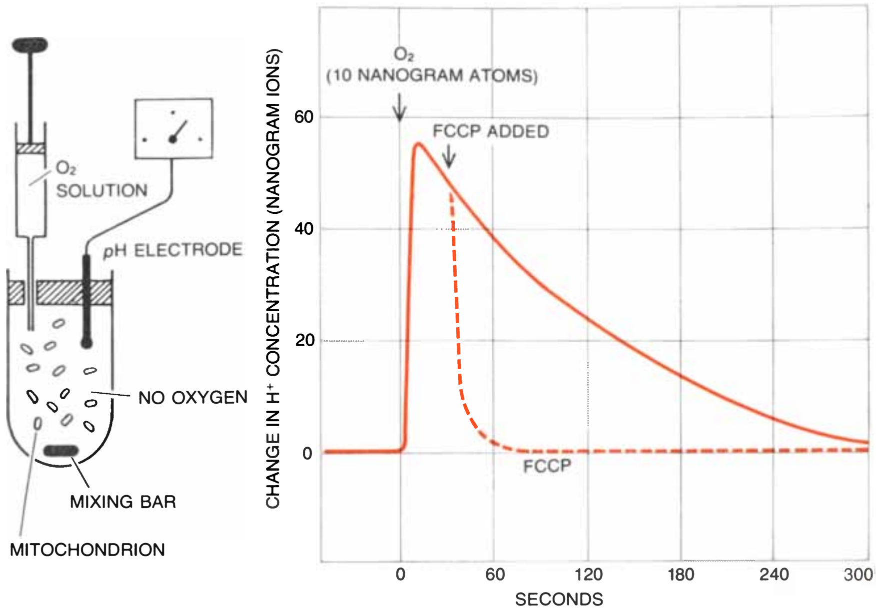
The transfer of an electron or a hydrogen atom from one molecule to another is called an oxidation-reduction reaction. The molecule that donates the electron or the hydrogen is said to be oxidized by the molecule that accepts it; the electron acceptor or hydrogen acceptor, conversely, is said to be reduced. Whenever one substance is oxidized, another must be reduced. In this context the term oxidation sometimes causes confusion. Oxygen is indeed a strong oxidizing agent, but the chemical process of oxidation is a general one that can proceed in the absence of oxygen. Any acceptor of electrons can be regarded as an oxidant.

In the mitochondrion hydrogen is extracted from carbohydrates through the complicated series of chemical transformations called the citric acid cycle. The details of the cycle will not concern us here; it suffices to report that its overall effect is to break down the carbon chain of glucose to carbon dioxide and to deliver the liberated hydrogen atoms to the molecule nicotinamide adenine dinucleotide, or NAD^+. The plus sign has been included in the abbreviation to show that the molecule ordinarily carries a positive electric charge. Each molecule of NAD^+ accepts two electrons and one proton. The proton and one electron bind to a carbon atom in the NAD^+ molecule; the other electron neutralizes the positive charge. This reduced form of NAD^+ is designated NADH. NADH is the principal intermediary between the citric acid cycle and the enzymes in the inner mitochondrial membrane that eventually deliver electrons to oxygen, forming water. The overall process is called respiration, and the hydrogen carriers and electron carriers in the membrane are called the respiratory chain.

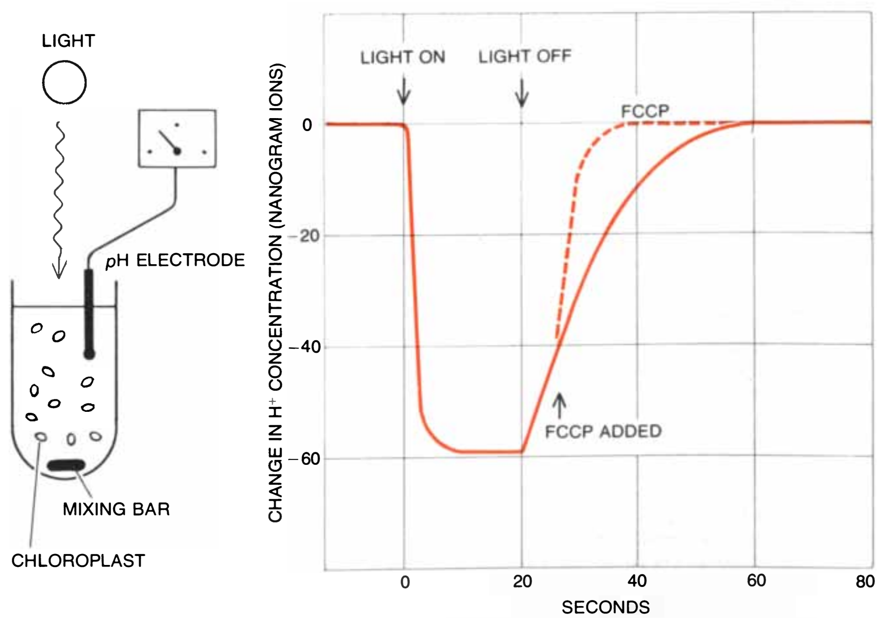
In the version of the chemiosmotic theory of mitochondrial respiration that now seems most plausible, each pair of electrons transferred from NADH to oxygen results in the outward translocation of six protons across the membrane. The ratio is expressed in terms of pairs of electrons because the electrons appear in pairs both at the beginning of the respiratory chain (at NADH) and at the end (where they reduce an oxygen atom to water). Within the respiratory chain the electrons are transported one at a time by some carriers and in pairs by others.

NADH donates its two electrons and one proton to a carrier group called flavin mononucleotide, or FMN. In this process NADH is oxidized, or restored to the form NAD^+, and FMN, having accepted two electrons and a proton, takes up an additional proton from the medium inside the membrane and is thereby reduced to $FMNH_2$. The FMN molecule is attached to a large protein, which is embedded in the mitochondrial membrane and probably extends all the way across it.

By a mechanism that has not yet been elucidated $FMNH_2$ transfers the two hydrogen atoms from the interior surface of the membrane to the exterior surface. There the atoms are ionized and the protons are released in the extramitochondrial medium, so that the first two protons have been transported across the membrane. In one hypothetical transport mechanism the protons are released by the flavin group inside the membrane, and they reach the external



PROTON TRANSPORT by active mitochondria can be observed as a change in the concentration of hydrogen ions in the medium surrounding the organelles. The mitochondria were suspended in a medium that contained a source of electrons for respiration but had been depleted of oxygen. When a pulse of oxygen was introduced into the suspension, the *p*H of the medium declined sharply, a change that corresponds to an increase in the concentration of protons. When the oxygen was used up, protons slowly leaked back into the mitochondria. FCCP, a molecule that makes membranes permeable to protons, dissipated the proton gradient quickly.

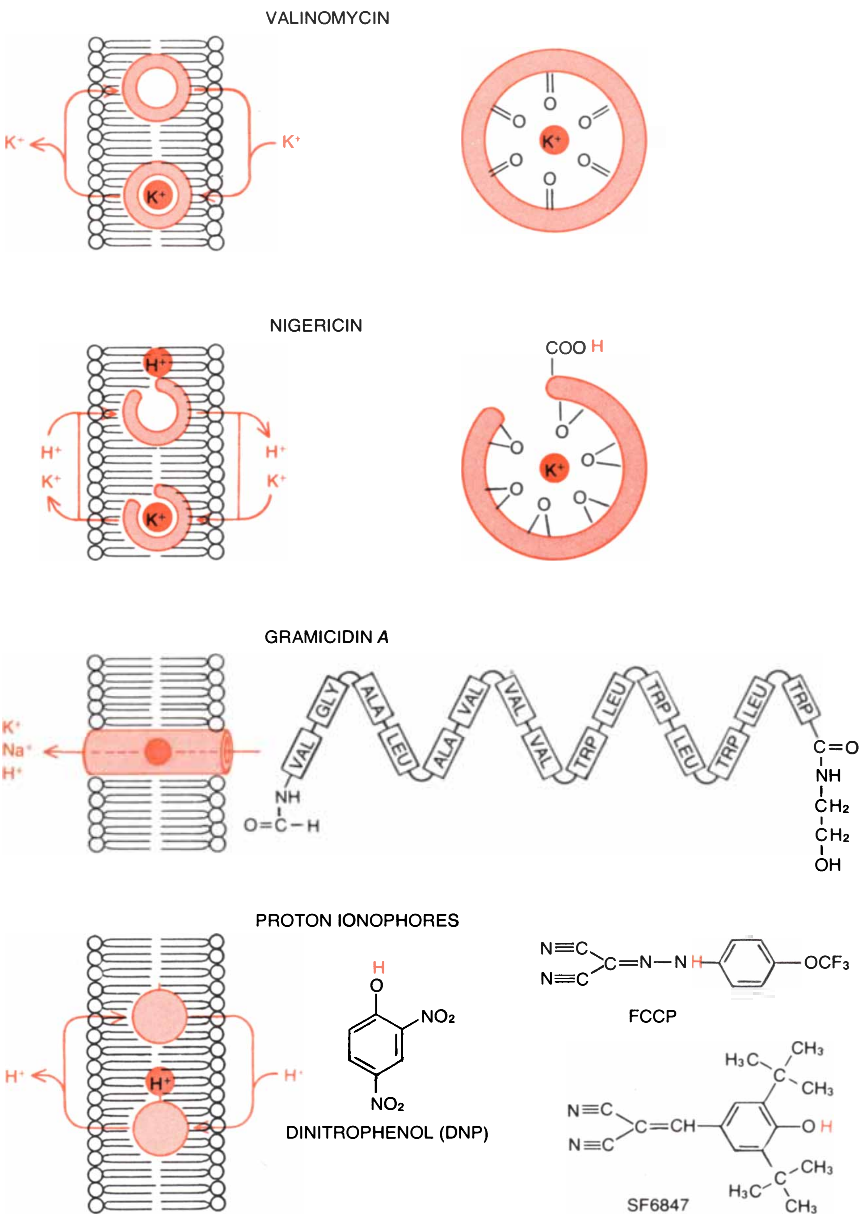


CHANGE IN *p*H in a suspension of chloroplasts is in the direction opposite to that observed with mitochondria. The chloroplasts were suspended in a medium containing an acceptor of hydrogen atoms and were kept in the dark. When the suspension was illuminated, the concentration of hydrogen ions declined (indicating that protons were being taken up by the organelles), and when the light was turned off, the gradient decayed slowly. Again FCCP accelerated the leakage of protons, which in this case was out of the organelle and into the medium.

medium through a channel in the protein molecule.

As the two protons are released by FMN the two electrons are passed on to other proteins, which because of the elements associated with them are called iron-sulfur proteins, or FeS. Having given up two protons and two electrons, $FMNH_2$ is restored to its original form of FMN and can be reduced again by NADH.

Unlike FMN, the iron-sulfur proteins carry only electrons, not complete hydrogen atoms; moreover, they carry the electrons one at a time rather than in pairs. There are several types of iron-sulfur proteins in this part of the respiratory chain, and they may be arranged across the membrane so that they can transfer the electrons back to the interior side. The electrons then enter the most complicated part of the respiratory chain, and the part where the chemiosmotic theory is the most speculative. Mitchell has recently proposed a mechanism for electron transfer through this section of the chain, which we shall present, although definitive experimental support for the idea is still lacking.



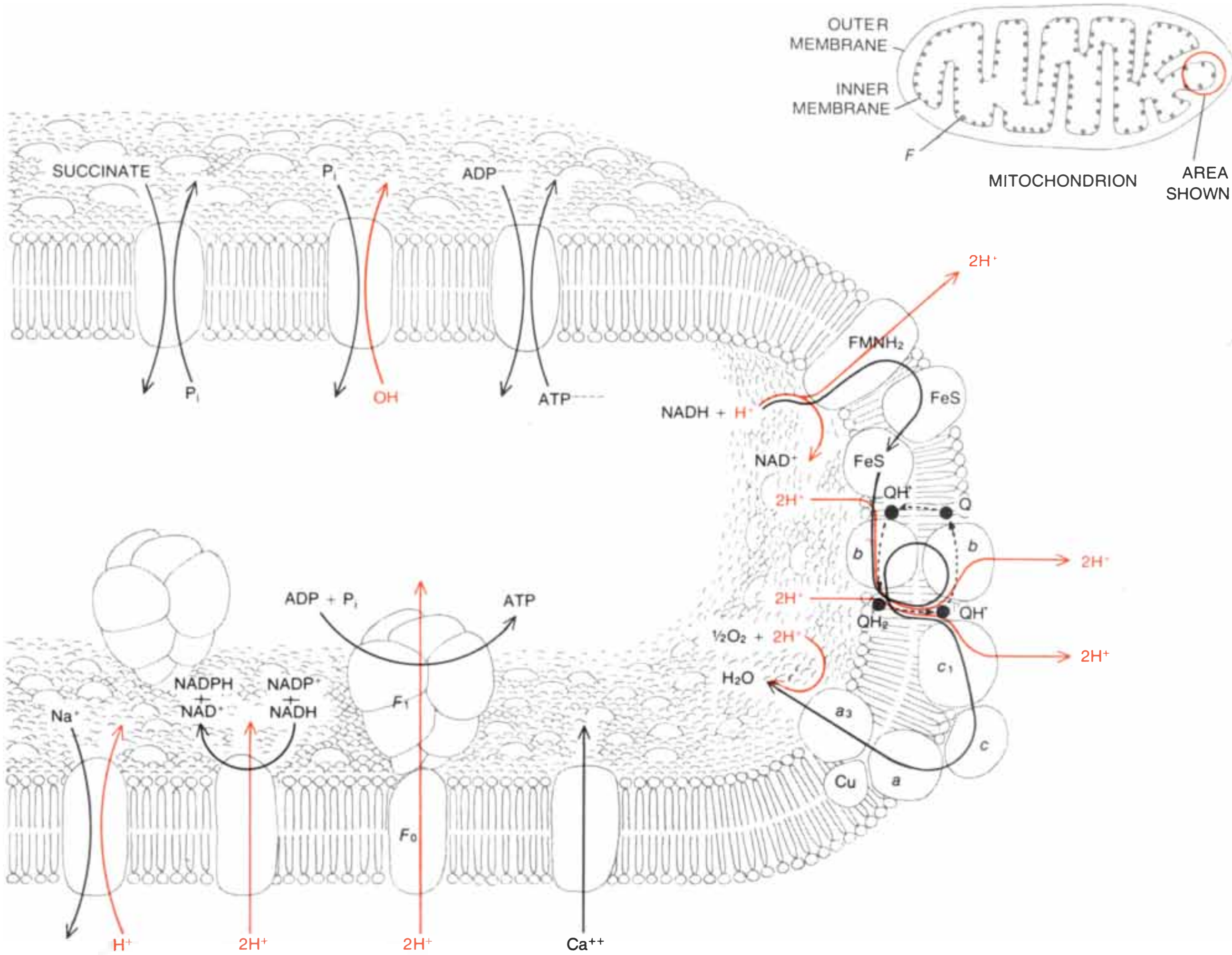
IONOPHORES, molecules that carry ions across membranes, are essential tools in the study of chemiosmotic mechanisms. Most simple ions are soluble in water but not in the lipid matrix of membranes, and so the membranes are impermeable to them. The ionophore valinomycin is a cyclic molecule that is soluble in membranes and selectively carries potassium ions (*K*+). Nigericin also transports potassium, but only in exchange for protons; it is insoluble in the membrane unless one ion or the other is bound. Gramicidin *A* forms a pore in the membrane, made up of two helical molecules, that is permeable to ions with one positive charge. The proton ionophores were first studied as molecules classified functionally as "uncouplers": they inhibit phosphorylation but stimulate respiration. That mode of action is explained by their ability to dissipate a proton gradient. The proton ionophores are all weak acids that readily gain or lose a proton (*at the positions shown in color*) and can diffuse across the membrane in either state.

The carrier to which the iron-sulfur proteins donate the pair of electrons is a small molecule called ubiquinone, or coenzyme Q; it is usually abbreviated Q. Mitchell calls his proposed mechanism the Q cycle. Ubiquinone is a ring-shaped molecule with two oxygen atoms attached to it and with three possible oxidation states. In the fully oxidized, or quinone, form both oxygen atoms are connected to the ring by double bonds. Attaching a hydrogen atom to one of the oxygens creates the semiquinone form, or QH•. In the fully reduced form, called hydroquinone or QH_2, both oxygens have hydrogen atoms attached.

In the proposed Q cycle the iron-sulfur proteins donate an electron to each of two ubiquinone molecules, which take up protons from the internal medium to form two molecules of the semiquinone, QH•. Two more electrons are then supplied by molecules of another protein in the respiratory chain, cytochrome *b*. (The source of these electrons will be discussed below.) By taking up two more protons from inside the mitochondrion two molecules of fully reduced QH_2 are formed.

Ubiquinone is soluble in the lipid matrix of the membrane, and it therefore may be mobile. Unlike the other carriers, which presumably remain stationary and serve merely as conduits for hydrogen atoms or electrons, ubiquinone may migrate as a molecule from one side of the membrane to the other. In support of this hypothesis is the observation that there are about 10 times as many ubiquinone molecules in the membrane as there are molecules of other components of the respiratory chain. In the proposed Q cycle the two molecules of hydroquinone cross from the inner surface of the membrane, where they acquired a total of four hydrogen atoms, to the outer surface. There each QH_2 molecule gives up an electron to the next protein in the respiratory chain, cytochrome c_1, and releases a proton outside the mitochondrion. Hence the total number of protons transported so far is four.

Each of the ubiquinone molecules is now in the semiquinone, or QH•, state. They complete the cycle by returning to the fully oxidized form. Each gives up its remaining proton to the external medium and transfers the associated electron to cytochrome *b*. All six protons have now been moved across the membrane and expelled from the mitochondrion, but the fate of four electrons remains to be determined: two were deposited with cytochrome c_1 and two

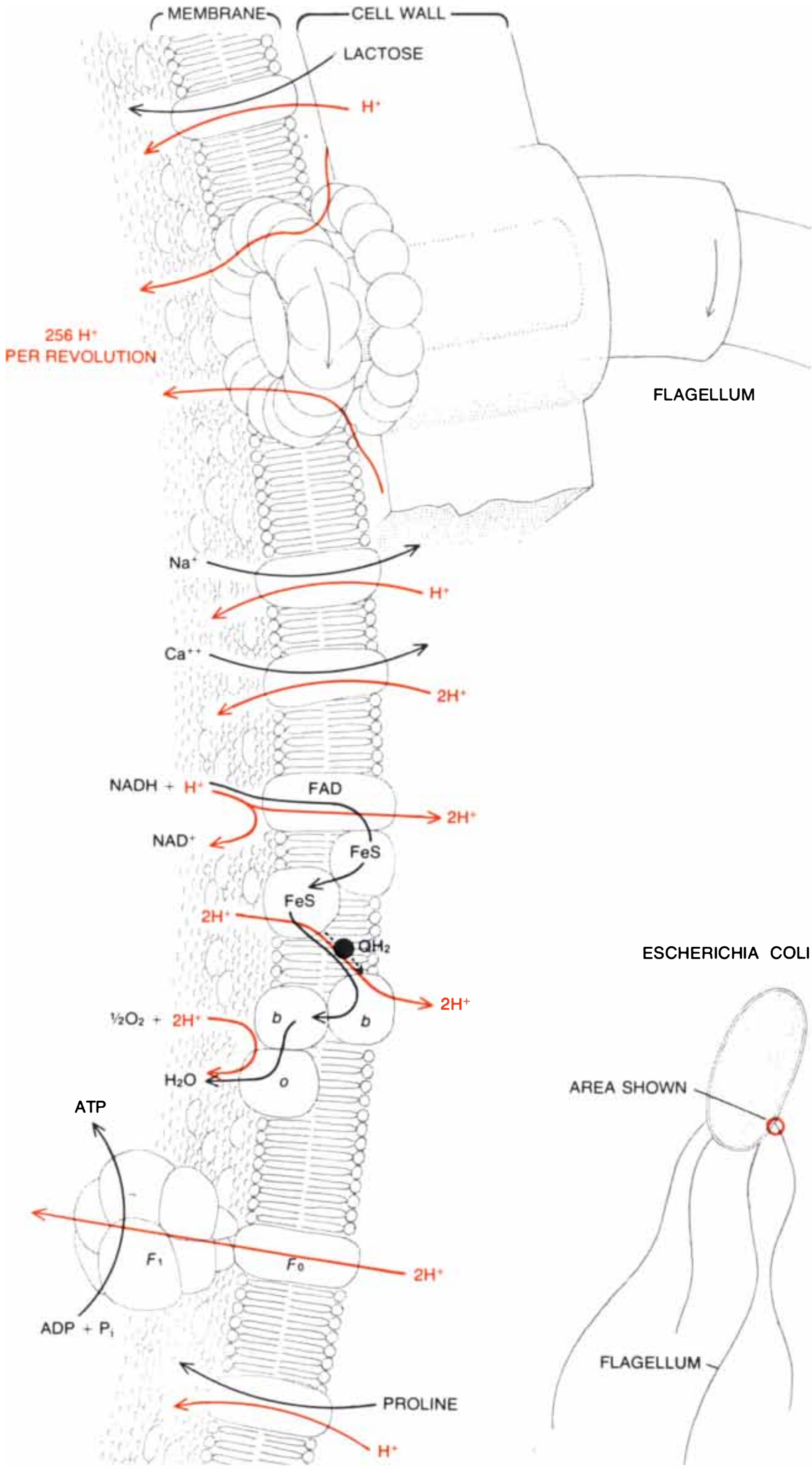


MEMBRANE OF THE MITOCHONDRION has embedded in it the enzymes and other components of the respiratory chain. The arrangement of the molecules, however, is not yet certain, and the model presented here is somewhat conjectural. Respiration begins with NADH, which gives up two electrons and a proton to flavin mononucleotide (FMN); another proton is picked up from the internal medium, so that the reduced form of the molecule ($FMNH_2$) carries two complete hydrogen atoms. The protons are expelled and the electrons return through an iron-sulfur protein (FeS) to the inner surface of the membrane. There the two electrons are donated to two molecules of ubiquinone (Q), each of which acquires a proton to form the semiquinone (QH•). Unlike the other components of the respiratory chain the quinones probably migrate as molecules through the membrane (*broken lines*). The semiquinone takes on two more electrons from cytochrome *b* and with two more protons from inside the mitochondrion is converted into the fully reduced hydroquinone (QH_2). Each hydroquinone gives up one electron to cytochrome c_1 and releases the corresponding proton outside. The remaining two electrons are then returned to the cycle through cytochrome *b* and the last two of six protons are released. Finally the two electrons deposited with cytochrome c_1 pass through cytochromes *c*, *a* and a_3 to oxygen, which is thereby reduced to water. The proton circuit is completed by the F_1-F_0 complex, where each two protons driven inward bring about the synthesis of one ATP molecule. Other processes are also powered by the proton gradient. They include the reduction of $NADP^+$ by NADH, the transport of calcium (Ca^{++}) and sodium (Na^+) ions and the exchange of ADP for ATP. The illustrations on this page and the next two pages are based on drawings made by Maija V. Hinkle.

with cytochrome *b*. The cytochromes are all proteins with an embedded heme group: a large ring structure with a central iron atom that can readily accept or donate an electron. The two electrons on cytochrome *b* molecules are returned across the membrane to the Q cycle: they ultimately reduce two more molecules of QH• to QH_2. It was from this closed loop that the electrons were obtained to reduce the semiquinones discussed above.

The two electrons deposited with molecules of cytochrome c_1 continue through the respiratory chain to its end. They are passed from cytochrome c_1, which is embedded in the exterior surface of the membrane, to cytochrome *c*, which lies on the exterior surface. The two electrons then go to cytochrome *a* and, crossing the membrane for the last time, are transferred to cytochrome a_3. Finally, cytochrome a_3 is oxidized by molecular oxygen. The two electrons are donated to an oxygen atom and two protons are picked up from the interior medium of the mitochondrion to form a water molecule.

In this long sequence of oxidation-reduction reactions the pair of electrons makes three round-trip crossings of the membrane and exports two protons each time. The first passage is made through reduced flavin mononucleotide, $FMNH_2$. After returning through the iron-sulfur proteins the electrons traverse to the exterior side again with hydroquinone. The second return trip is made through cytochrome *b*, where the electrons reduce another pair of ubiquinone molecules and migrate to the outside surface for the third and last time. Finally they return to the inside of the mitochondrion through the system of cytochromes (c_1, *c*, *a* and a_3) and are delivered to oxygen. The flow of electrons in the mitochondrion is an electric current. The potential difference between NADH and oxygen is about 1.2 volts, and the total current in the mitochondria of a human being at rest is



MEMBRANE OF THE BACTERIUM *E. coli* has a system of respiratory proteins similar to the system in mitochondria. Electrons and protons are donated to flavin adenine dinucleotide ($FADH_2$), which exports the two protons; the electrons are returned to the inner surface through iron-sulfur proteins (FeS). The electrons, together with two protons from the internal medium, then reduce a single molecule of ubiquinone to hydroquinone (QH_2), which diffuses across the membrane and releases the protons outside. Finally the electrons proceed through cytochromes *b* and *o* to water. Just four protons are translocated for each electron pair instead of six. As in mitochondria, phosphorylation requires two protons for each molecule of ATP. Sodium and calcium transport are coupled to the movement of protons down the gradient, and so is the uptake of certain sugars, such as lactose, and amino acids, such as proline. The rotation of the flagellum is also powered by the influx of protons. At the root of the flagellum is a ring of 16 proteins, apposed to a similar ring in the cell wall. If a proton must pass through each protein to rotate the flagellum a sixteenth of a turn, 256 protons would be consumed in each revolution.

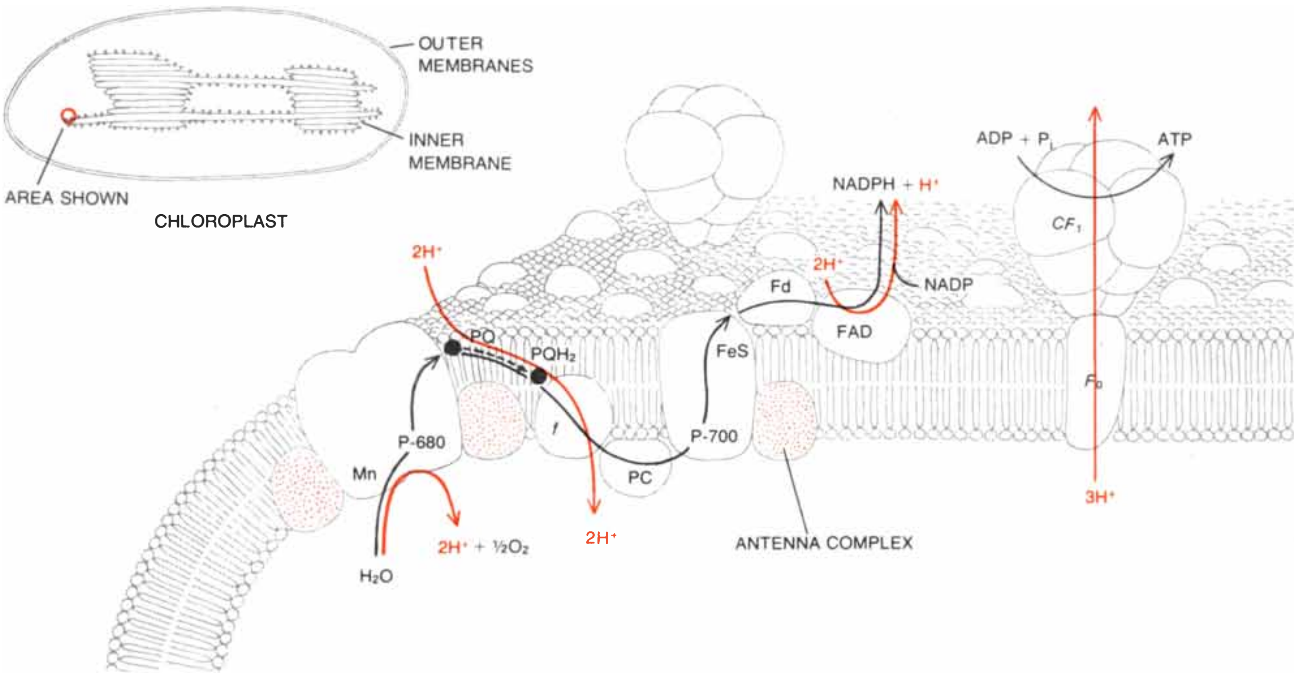
about 100 amperes, so that 120 watts of power are generated.

It must be emphasized that this reconstruction of the mitochondrial membrane proteins and their interactions is tentative. Only for a few of the molecules is there any evidence bearing on their position in the membrane. FMN must be exposed on the interior surface since NADH reacts with it only there. Conversely, cytochrome *c* must be external since it can be removed by washing mitochondria with salt solutions. Two complexes of cytochrome molecules, one made up of *b* and *c* and the other of *a* and a_3, have both been shown to span the membrane, although the exact position of each cytochrome has not yet been established. Other arrangements of the same molecules, however, could lead to the same result: that the transfer of two electrons from NADH to oxygen is coupled to the export of six protons from inside the mitochondrion to outside it.

The respiratory chain of the bacterium *Escherichia coli* incorporates several components analogous to those found in mitochondrial membranes and operates according to the same principles. There is a major difference, however: for each pair of electrons passing through the respiratory chain only four protons (instead of six) are exported from the bacterial cell. The number of protons transported also varies with the conditions under which the bacteria are grown.

Similarities between bacteria and mitochondria are not unexpected. Both mitochondria and chloroplasts are believed to have evolved from bacteria that may have entered cells as parasites or symbionts early in the history of the nucleated cells and only later became captive organelles. Mitochondria and chloroplasts have their own genetic material, which in form resembles that of bacteria. Many proteins of mitochondria and chloroplasts are more like bacterial proteins than they are like cellular proteins of equivalent function.

In *E. coli* the respiratory chain is apparently simpler than it is in mitochondria. Electrons are carried to the respiratory proteins in the cell membrane by NADH and are donated to a flavin group analogous to FMN; the group is flavin adenine dinucleotide, or FAD. Like FMN, FAD is associated with a protein, and it carries two hydrogen atoms to the outer surface of the membrane, where the protons are released. The electrons are passed on to iron-sulfur proteins, which transport them back across the membrane. As in mitochondrial respiration, the second hydrogen carrier is ubiquinone, but at present there is no evidence for a Q cycle in *E. coli*. The two electrons are donated to a single quinone molecule, which takes up two protons from inside the cell to form



MEMBRANE OF THE CHLOROPLAST incorporates a light-driven system of pigments and other molecules that translocates protons inward. Two photons, or quanta of light, must be absorbed for each electron carried from water to the ultimate acceptor of electrons, $NADP^+$. For each pair of electrons passing through the chain of carriers three protons are taken up outside the chloroplast and four appear inside. The first photon is absorbed by an array of chlorophyll molecules, the antenna complex associated with a specialized chlorophyll designated P-680. Two electrons freed from P-680 molecules cross the membrane and are replaced by electrons taken from a molecule of water. The electrons from P-680, with two protons from outside the membrane, reduce plastoquinone (PQ) to PQH_2. The protons are released inside when the electrons are transferred to cytochrome *f*. The electrons then proceed through plastocyanin (PC) to a second photocenter, P-700. With the absorption of additional photons the electrons complete their journey through an iron-sulfur protein (FeS), ferredoxin (Fd) and flavin adenine dinucleotide ($FADH_2$) to $NADP^+$. The synthesis of ATP in the CF_1-F_0 complex is the only process in chloroplasts known to require a proton gradient; three protons apparently cross the membrane for each ATP molecule formed.

the hydroquinone. After it has migrated across the membrane the hydroquinone releases the two protons outside and passes the two electrons on to two cytochrome *b* molecules. Finally the electrons flow back across the membrane through cytochrome *b* to cytochrome *o*, which is oxidized by molecular oxygen. In summary, the pair of electrons crosses the membrane twice in each direction and translocates four protons outward. In other bacteria the respiratory chain takes other forms, and in some it is similar to the mitochondrial chain. It should also be pointed out that several compounds in addition to NADH can be oxidized by mitochondria and bacteria.

In chloroplasts the electron-transfer chain has some carrier molecules similar to those of the respiratory chains of mitochondria and bacteria, but the flow of electrons is in the opposite direction. The chain begins with water and ends with a phosphorylated form of NAD^+, nicotinamide adenine dinucleotide phosphate, or $NADP^+$. What is even more remarkable, the direction of proton translocation is also opposite to that of mitochondria. Chloroplasts have a triple system of membranes: the outer two define the organelle, and the inner one, which is very convoluted, contains the apparatus of photosynthetic phosphorylation. In chloroplasts protons are translocated inward across the inner membrane and are accumulated inside.

For electrons to flow from water to $NADP^+$ the free energy of the initial state must be raised. The required energy is supplied by the absorption of light by pigments stationed at two points in the electron-transfer chain. Each photon absorbed drives an electron across the inner membrane.

The first photon is absorbed by an array of molecules that includes several hundred chlorophyll molecules and other pigments, all of which are bound to proteins. This dense collection of light-absorbing substances, called the antenna complex, is embedded in the inner surface of the membrane. The energy of excitation is communicated rapidly from the antenna complex to a specialized chlorophyll molecule called P-680. ("P" stands for "pigment" and the number refers to the wavelength of light, expressed in nanometers, that the molecule absorbs.) Chlorophylls have a ringlike structure similar to that of the heme group, except that the central cavity is occupied by a magnesium atom instead of an iron atom. In P-680 light energy alters the distribution of electrons in the ring and makes one electron available for transfer.

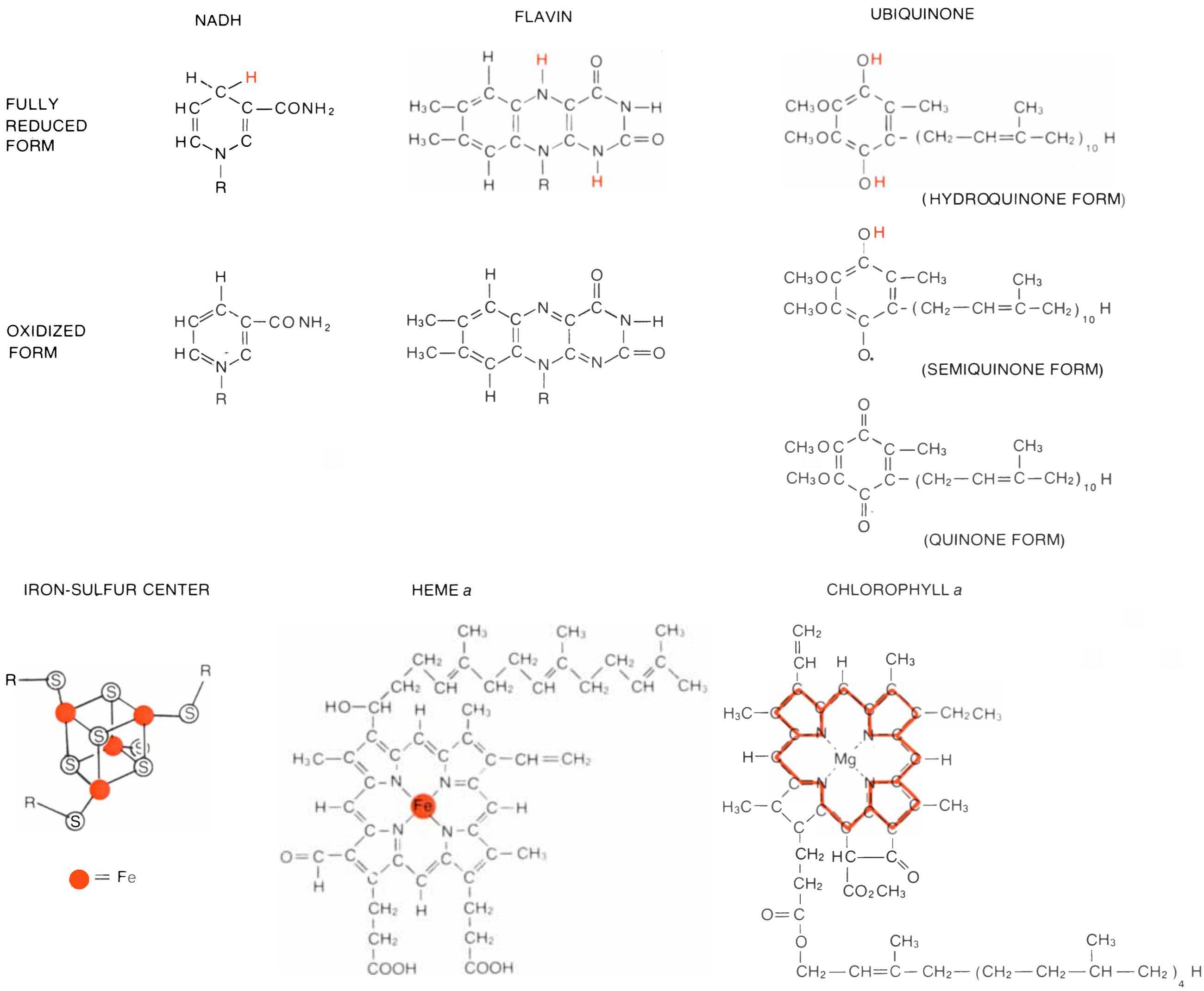
The electron from P-680 is apparently transferred to an electron acceptor on the outer surface of the membrane immediately after a photon is absorbed. On the absorption of another photon the process is repeated. The two oxidized P-680 molecules are reduced by two electrons derived from water with the help of an enzyme that contains manganese. The oxygen atom from the water molecule is released and diffuses out of the chloroplast and the two protons go into solution inside the inner membrane.

The two electrons that cross the membrane from P-680 are picked up at the outer surface by a hydrogen carrier similar in structure to ubiquinone but called plastoquinone, or PQ. The electrons, with two protons extracted from the external solution, reduce PQ to PQH_2, which migrates back across the membrane to the inner surface. There the PQH_2 releases the protons internally and transfers the electrons to cytochrome *f*, a protein of the same type as cytochrome *c*.

The next electron carrier in the chain is the copper-containing protein plastocyanin, and it turns the electrons over to the second photochemically active site. The most important component of this system is another specialized chlorophyll molecule, designated P-700. When P-700 is excited by light absorption in the antenna complex, the electrons are once again moved to the outer surface of the membrane, where they are accepted by an iron-sulfur protein. The route then continues along the outer surface from that iron-sulfur protein to another one, ferredoxin. From there the electrons go to a proton containing FAD, which withdraws two protons from the external medium to form $FADH_2$. Finally $NADP^+$ acquires the electrons and a proton from $FADH_2$ to form NADPH. When the passage of two electrons through the chain is complete, three protons have disappeared outside the membrane and four protons have appeared inside.

This model of the chloroplast inner membrane, like the model of the mitochondrial and bacterial membranes discussed above, is a tentative reconstruction based on fragmentary evidence. Structural studies of the membrane have shown that ferredoxin is on the outside and that functional plastocyanin is on the inner surface. Other features of the model are more speculative. H. T. Witt and his colleagues at the Max Volmer Institute in Berlin have shown that an electric potential is created across the chloroplast inner membrane in less than 10^{-8} second after illumination. The speed of the reaction is strong evidence that the electrons released by the specialized chlorophylls are transferred across the membrane; less direct methods for creating a potential would be too slow.

The proton gradient established by photochemical or oxidative electron transport represents a store of free energy, much like water pumped up a gravitational gradient to an elevated reservoir. The energy can be recovered by allowing the protons to flow back across the membrane through an appropriate "mill." The main work done by the flow of protons down the gradient is the phosphorylation of ADP to form ATP,



CARRIER MOLECULES in both respiratory and photosynthetic chains are specialized for the transport of electrons or hydrogen atoms. The active centers of several carriers are shown; *R* stands for the rest of the molecules. The hydrogen atom (or atoms) donated to the oxidized forms of the carriers to generate the reduced forms are in color. NAD^+ accepts two electrons and a proton to form NADH. One electron neutralizes the charge on the ring nitrogen atom, and the other combines with a proton to hydrogenate a carbon atom. A very similar molecule, $NADP^+$, is a carrier in chloroplasts. The flavins FAD and FMN are the active parts of flavoproteins. Two hydrogen atoms are donated to FAD and FMN to form $FADH_2$ and $FMNH_2$. Iron-sulfur proteins are electron carriers rather than hydrogen carriers; the active center of one type is a cagelike structure in which the electron's charge is distributed among the iron and sulfur atoms. Some of the sulfur atoms are sulfide (S^{--}), and the remainder are from the amino acid cysteine in the protein (*R*). Ubiquinone is a hydrogen carrier in mitochondrial and bacterial membranes. In its fully reduced and protonated form the oxygen atoms on the benzene ring have hydrogen atoms attached to them. In the fully oxidized state the oxygen atoms are linked to the carbon ring by double bonds. Ubiquinone can also exist in a half-reduced or semiquinone form, which is not shown in the illustration. The semiquinone is a free radical, and the unpaired electron is delocalized by resonance over the carbon ring. Plastoquinone, a hydrogen carrier in chloroplasts, is very similar to ubiquinone. Chlorophyll and the heme group of cytochromes are also similar in structure. In cytochromes electron is carried primarily by iron atom (Fe) at the center of the large heme ring; the positive charge of oxidized chlorophyll is probably distributed throughout the ring.

but other processes such as the active transport of some ions are also powered by the gradient.

Energy is stored in the proton gradient in two forms, or in other words the gradient has two components. One component is the difference in concentration or chemical activity of protons on opposite sides of the membrane. The protons tend to diffuse from a region of high concentration to a region of low concentration when a pathway through the membrane is provided. Proton concentration is measured in terms of *p*H, which is defined as the negative logarithm of the hydrogen ion concentration. The energy of the concentration gradient is determined by the difference in *p*H across the membrane and is independent of the absolute magnitude of the *p*H.

The electric charge carried by the proton contributes the second component to the energy of the gradient. The net movement of charges across the membrane creates a difference in electric potential, and all charged particles are affected by the resulting electrostatic field. The total energy of the proton gradient is the sum of the concentration (or osmotic) component and the electric component.

Because of the difference in concentration and in electric potential a proton that has been expelled from a mitochondrion experiences a force tending to draw it back across the membrane. The movement of the proton in response to that force can be made to do work, such as the work of phosphorylation.

The enzymes that couple the diffusion of protons back through the membrane to the synthesis of ATP appear to be remarkably similar in mitochondria, bacteria and chloroplasts. They are conspicuous in electron micrographs as globular bodies protruding from the surface of the membrane. The protruding "knob," designated F_1, was isolated in 1960 by Efraim Racker, Maynard E. Pullman and Harvey Penefsky of the Public Health Research Institute of the City of New York. It is a soluble protein made up of five kinds of subunit, some of which are present in multiple copies. F_1 is attached to the membrane through another set of proteins designated F_0, which is embedded in the membrane and is thought to pass all the way through it. F_1 is readily removed from the membrane, but F_0 can be removed only when the membrane itself is destroyed with detergents. The entire system of enzymes is called the F_1-F_0 complex.

In mitochondria and in bacteria the F_1-F_0 complex is oriented with the knobs protruding into the interior matrix from the inner surface of the membrane. The evidence indicates that for every two protons passing inward through the complex one molecule of ADP is combined with inorganic phosphate to form ATP. An important observation is that the reaction is reversible: under appropriate circumstances the F_1-F_0 complex can split ATP molecules and employ the energy made available to pump protons out of the mitochondrion or bacterial cell. Like all enzymes, the F_1-F_0 complex controls the rate of the reaction, but the direction of the reaction is determined by the balance of free energy.

In chloroplasts the equivalent enzyme system is called the CF_1-F_0 complex, and it is oriented in the opposite direction, that is, the knob protrudes from the outer surface of the membrane. The orientation is exactly what would be expected, since the direction of the proton gradient in chloroplasts is also opposite to that in mitochondria and bacteria. Protons flow outward through the CF_1-F_0 complex, and the ATP is formed in the space between the inner and the outer chloroplast membranes. Another difference is more difficult to account for: the chloroplast CF_1-F_0 complex apparently requires three protons per ATP synthesized instead of two.

The mechanism of ATP synthesis at the active site of the F_1-F_0 complex is not understood. Several hypotheses have been formulated, but the evidence from studies of the enzyme is not yet sufficient to decide among them.

The active site is apparently associated with the F_1 part of the complex, the knobs seen in electron micrographs. Even when the F_1 components are isolated and in solution, they catalyze the splitting of ATP to yield ADP and phosphate, the reverse of the phosphorylation reaction driven by a proton gradient. The forward reaction cannot be observed with isolated F_1 molecules, of course, because a proton gradient cannot be established in solution. Of the five kinds of protein subunit in F_1, only the two largest are required for ATP-splitting activity. The smallest of the subunits is an inhibitor of catalytic activity that presumably serves as a check valve: it prevents the reaction from proceeding at all when the proton gradient is low, since under those circumstances the reaction would proceed in reverse, consuming ATP.

The F_0 part of the complex, which is left behind in the membrane when F_1 is removed, exhibits no catalytic activity with ATP. On the other hand, the removal of F_1 makes the membrane highly permeable to protons, and these leaks can be plugged by rebinding F_1 or by certain inhibitors. These observations have led to the view that F_0 serves as a channel through which protons cross the membrane to F_1.

One of the hypothetical catalytic mechanisms was proposed by Mitchell. In this scheme a phosphate group binds to the enzyme at an active site within the F_1 part of the complex but near the end of the proton channel through F_0. Two protons driven down the channel by the membrane potential and the *p*H gradient attack one of the phosphate oxygens, pulling it loose to form a molecule of water. The unattached phosphorus bond created in this way converts the phosphate group into a highly reactive species that can bind directly to ADP.

In the main alternative hypothesis protons have a less direct role. This indirect hypothesis actually embraces several possible mechanisms. The idea common to all of them is that the passage of protons through the F_1 part of the complex could change the conformation of protein. For example, the binding of a proton to a group in the protein could cause it to move as a result of electrical interactions with other charged regions. Paul D. Boyer of the University of California at Los Angeles has proposed that proton-induced changes in protein conformation near the active site could result in ATP synthesis. In one such mechanism the dissociation of the completed ATP molecule from the enzyme is the crucial, energy-requiring step. Although free ADP cannot combine with phosphate without an input of energy, the reaction might take place spontaneously if both molecules were bound to a protein. The resulting ATP would remain bound to the enzyme and could be dissociated from it only by the addition of energy. That energy would be supplied by a change of conformation, driven in turn by the translocation of protons.

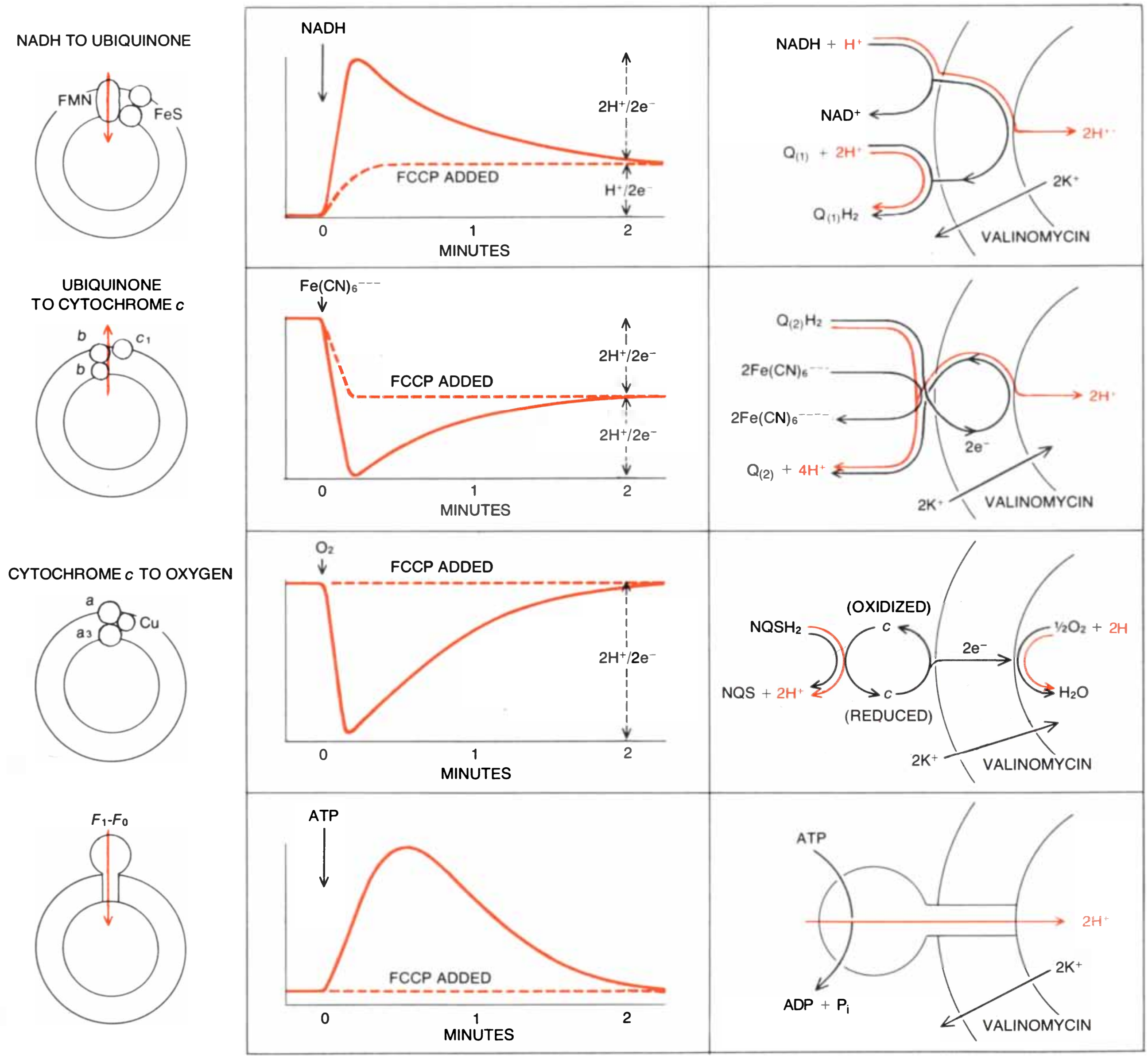
In 1971 I. J. Ryrie and André T. Jagendorf of Cornell University observed changes in the conformation of CF_1 bound to chloroplasts when the chloroplasts were illuminated or when the *p*H of the medium in which they were suspended was rapidly changed from acidic to basic, creating a momentary *p*H gradient. R. P. Magnusson, J. Fagan and one of us (McCarty) showed that the reactivity of a chemical group in one subunit of CF_1 is enhanced when chloroplasts are illuminated. Penefsky and others have shown that the F_1 protein of mitochondria also undergoes structural changes when it is exposed to an electrochemical proton gradient. In 1973 E. C. Slater and his colleagues at the University of Amsterdam found that CF_1 and F_1 have ADP and ATP firmly bound to them. In chloroplasts the ADP bound to CF_1 can be exchanged for nucleotide molecules in the medium when a proton gradient is formed. The bound ADP can be phosphorylated, but only at a low rate. It is probably not bound at a catalytic site in the enzyme but may regulate the structure or the functioning of F_1 molecules.

These findings can be interpreted as evidence for an indirect mechanism of phosphorylation, in which conformational changes have an essential role. Mitchell's direct hypothesis cannot be ruled out, however. Conformational changes in enzymes are hardly unusual and may be universal in enzymes. The changes in F_1 could be associated with a direct mechanism or they could be incidental to the catalytic activity.

The movement of molecules or ions across a membrane can proceed spontaneously only in a direction that tends to reduce a concentration gradient or a gradient in electric potential. In the cell and in subcellular organelles many substances must be assimilated or expelled against a gradient, and so most transport requires energy. The movement of some ions and molecules is driven directly by the proton gradient.

An example of proton-linked transport is the flow of sodium ions (Na^+) out of the mitochondrion in exchange for hydrogen ions. In this system one sodium ion is expelled from the mitochondrion for each proton that crosses the membrane in the opposite direction. The exchange is assumed to be mediated by a membrane protein.

Sodium transport is driven by only one component of the electrochemical proton gradient, namely the difference in proton concentration, or *p*H. The membrane potential has no effect, since while it accelerates the proton's inward passage it retards the outward movement of the sodium ion and the two effects exactly cancel. The reaction comes to equilibrium when the sodium gradient is equal to the *p*H gradient in proton concentration. An equivalent transport system powers the uptake of inorganic



SEGMENTS OF THE ENZYME SYSTEM from mitochondrial membranes can be isolated and assembled in artificial membrane vesicles; when each segment is supplied with appropriate substrates, it is seen to transport protons across the membrane. The first segment consists of those components of the respiratory chain extending from NADH to ubiquinone, including FMN and the iron-sulfur proteins. The second segment includes all the molecules from ubiquinone to cytochrome *c*. The third segment completes the respiratory chain, extending from cytochrome *c* to oxygen. The fourth isolated system is the F_1-F_0 complex, which is operated in reverse, so that it splits ATP to form a proton gradient. The respiratory systems are driven by adding a donor and an acceptor of electrons; they can be natural substrates, such as NADH or oxygen, or artificial ones, such as ferricyanide ($Fe(CN^{---})_6$), ubiquinone analogues (Q_1 or Q_2) or the reductant naphthoquinol sulfonate ($NQSH_2$). Proton transport is measured by monitoring the *p*H of the external medium. In some cases the reaction causes a permanent change in the *p*H, which is independent of the gradient, but since that offset is not affected by the proton ionophore FCCP it can be measured separately. When corrections are applied for the permanent offset, so that only the effects of the proton gradient are measured, it is seen that each segment of the respiratory chain transports two protons for each electron pair.

phosphate, which enters the mitochondrion as a negatively charged ion. The movement of the phosphate ions is balanced by a counterflow of hydroxyl ions (OH^-).

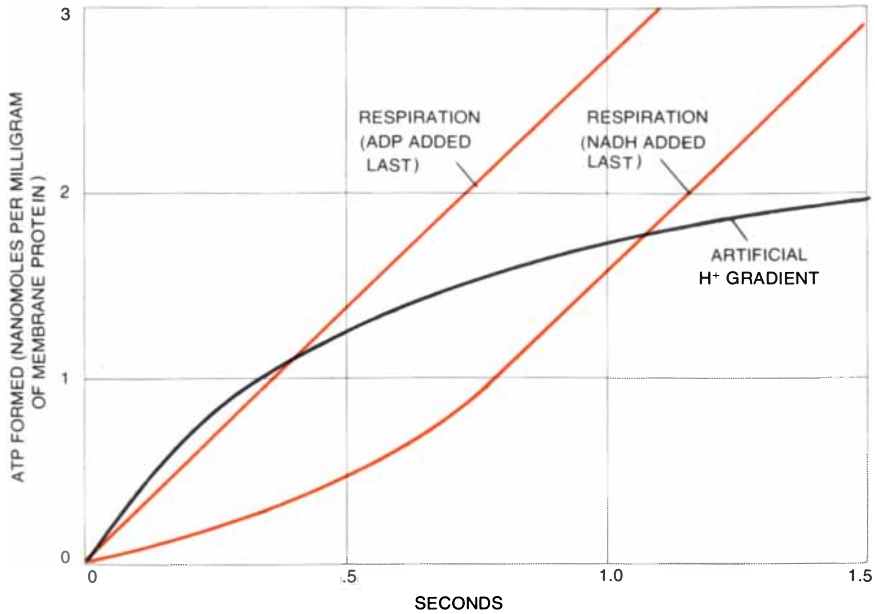
The other component of the proton gradient, the membrane potential, can also drive active transport. The uptake of calcium ions (Ca^{++}) by the mitochondrion, for example, is powered by the membrane potential and is independent of the *p*H gradient. Calcium probably enters without a countercurrent of any other ion, its transport being motivated entirely by electrostatic forces that pull it toward the negatively charged inner surface of the membrane. The membrane potential also provides the motive force for the exchange of ADP and ATP molecules, ensuring a continuing supply of substrate for phosphorylation and exporting the product. ADP carries a charge of minus 3 and ATP a charge of minus 4, so that the exchange is equivalent to the outward movement of one negative charge or the inward movement of one positive charge.

The proton gradient in chloroplasts is apparently not coupled to any function other than phosphorylation, with the possible exception of magnesium ion transport. In bacteria, on the other hand, a great variety of metabolic processes seem to derive their energy from the proton gradient. Indeed, Mitchell was studying transport across bacterial membranes when he proposed the chemiosmotic theory in 1961.

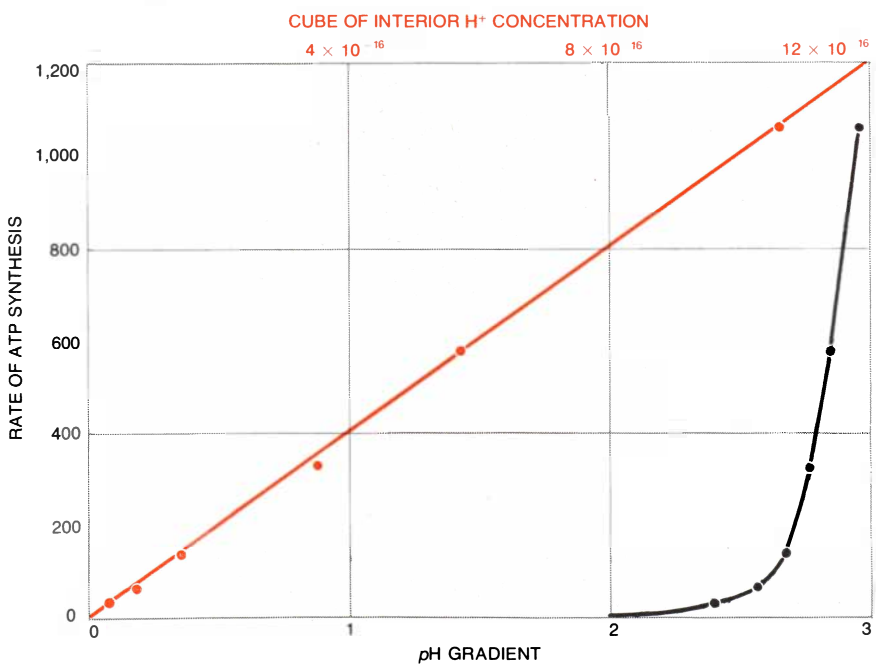
The mechanism of sodium transport in *E. coli* is apparently similar to that in mitochondria. Calcium transport, however, is in the direction opposite to that in mitochondria, that is, *E. coli* excretes calcium. The bacterial calcium transport is probably an electrically neutral exchange of two protons for each Ca^{++} ion, a movement powered by the *p*H gradient.

In 1963 Mitchell proposed that uptake of the sugar lactose by *E. coli* could be coupled to an influx of protons, even though lactose is an electrically neutral molecule rather than an ion. Subsequent studies have shown that lactose transport can be driven by an artificially formed proton gradient and, conversely, that the influx of lactose into starved cells is accompanied by an influx of protons. The transport system appears to be symmetrical: the sugar can be concentrated on either side of the membrane, depending on the polarity of the proton gradient. Similar systems for the uptake of certain amino acids have been discovered in *E. coli*.

It should be pointed out that these are not the only systems of active transport in bacteria; there are at least two others that are not directly coupled to the proton gradient. Glucose and some other sugars are brought into the bacterial cell



RATE OF PHOSPHORYLATION appears to be limited by the magnitude of the proton gradient. The rate of ATP synthesis was measured in vesicles made from mitochondrial membranes. When an artificial electrochemical proton gradient was imposed, the initial rate of synthesis was even greater than that observed under natural conditions, but the rate declined as the gradient decayed. When the gradient was created by respiration, a constant rate of synthesis was eventually attained, but the initial rate depended on the experimental conditions. If all the reactants needed for respiration were supplied but ADP was withheld, phosphorylation began immediately at a high, steady rate. If the missing reactant was NADH, however, there was a lag of about a second before the maximum rate of synthesis was achieved. Presumably the time was required for the respiratory chain to build up a proton gradient. These findings indicate that a proton gradient can supply energy fast enough to meet the demands of phosphorylation.



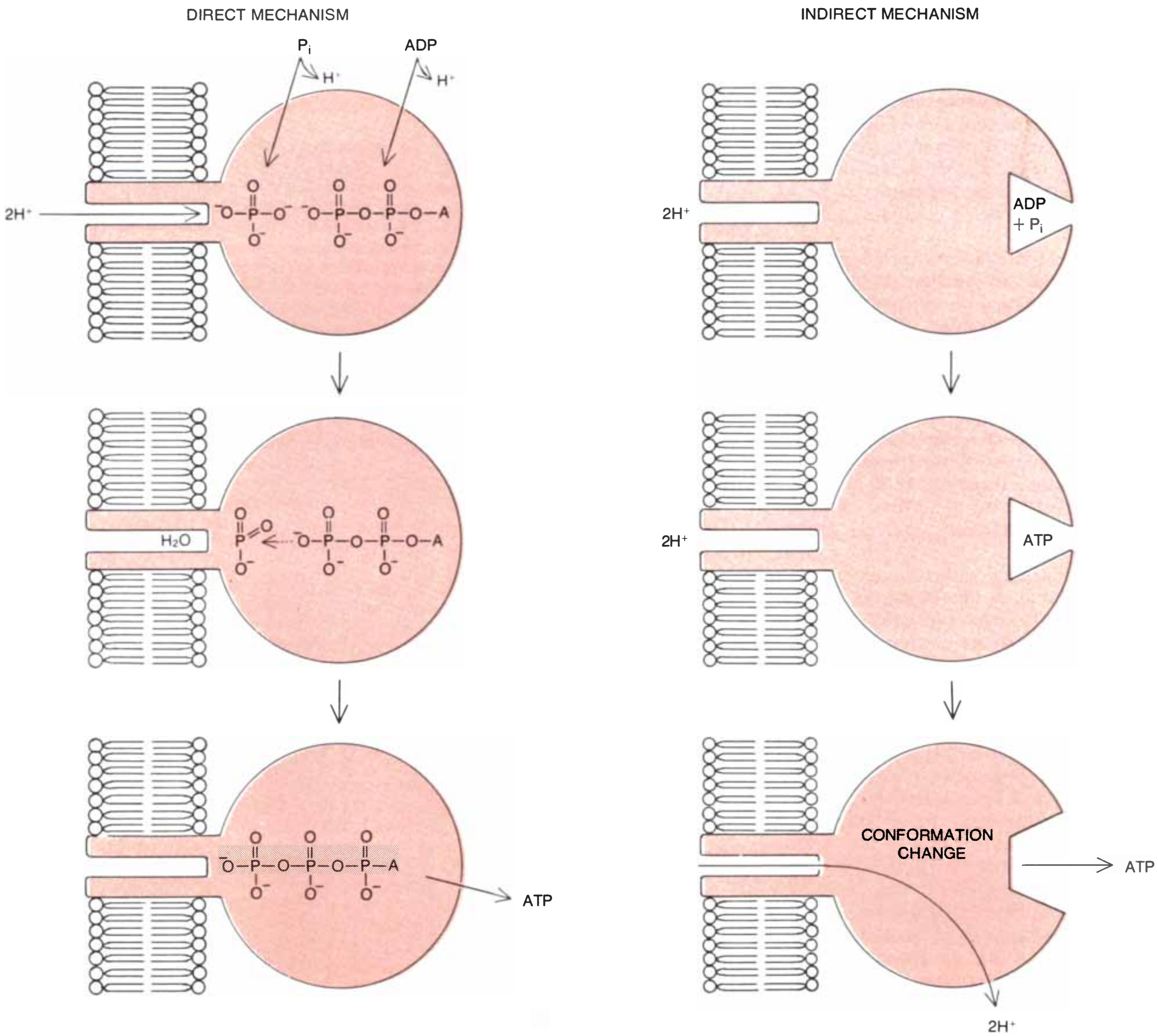
KINETICS OF PHOTOPHOSPHORYLATION were studied in functioning chloroplasts. The rate of ATP synthesis and the magnitude of the proton gradient were determined separately but under similar conditions in chloroplasts illuminated with various intensities of light. The relation derived from these measurements (*black curve and scale*) shows that the rate of phosphorylation is strongly dependent on the *p*H gradient. In particular no ATP is made until the difference in *p*H reaches about two *p*H units, presumably because protons moving down a smaller gradient have too little energy to drive the reaction. The rate measurements can be made to fall along a straight line by replotting them as a function of the cube of the internal proton concentration (*colored curve and scale*). This third-power relation supports other evidence suggesting three protons are required for each molecule of ATP formed in chloroplasts.

by a system in which the essential preliminary event is the phosphorylation of the sugar. Other molecules cross the membrane by first binding to soluble proteins in a system that derives the energy needed for active transport from the splitting of ATP.

An extraordinary mechanical linkage driven by the proton gradient is the flagellum responsible for motility in *E. coli* and other bacteria. It now appears that the flagellum moves by rotating like a propeller. The shaft at the root of the flagellum passes through a bearing in the rigid cell wall and terminates in a rotor in the cell membrane. The rotor appears to be a ring of 16 proteins, and it is apposed to another protein ring that may act as a "stator" in the cell wall.

Howard C. Berg and his co-workers at the University of Colorado have recently determined that the flagellar rotor is powered by the electrochemical proton gradient. Berg's calculation of the energy required for rotation suggests that about 250 protons would have to flow inward to power one revolution if the gradient is a typical one for *E. coli*. The protons presumably flow through the protein ring that forms the rotor, generating force against the adjacent ring in the cell wall. An obvious hypothesis is that the passage of one proton through each of the 16 subunits in the ring rotates the flagellum through a sixteenth of a turn. A complete revolution would then require 16 × 16, or 256, protons.

When Mitchell proposed the chemiosmotic theory, the translocation of protons in mitochondria and chloroplasts was unknown. (Indeed, he first suggested that respiration would pump protons into mitochondria rather than out of them.) The existence of a proton-transport system coupled to respiration was established by Mitchell and Moyle, who employed a comparatively



POSSIBLE MECHANISMS of phosphorylation in the F_1-F_0 complex fall into two categories, called direct and indirect. In the direct mechanism (*left*) first event is binding of a phosphate ion and ADP to the F_1 part of the enzyme complex. Protons move through a channel in the F_0 part and then attack one of the phosphate oxygens, which is removed as a molecule of water. Finally an oxygen of ADP attacks the phosphorus atom, forming ATP. The ATP molecule then dissociates from the enzyme. A variety of indirect methods are possible; one example is shown here (*right*). At the active site of the enzyme it might be possible for ADP and inorganic phosphate to combine spontaneously, without an input of free energy. The resulting ATP molecule would be securely bound to the enzyme, however, and energy would be needed to release it. The energy might be supplied by protons that bind at a position other than the active site and that thereby change the conformation of the enzyme. The protons would then be released into the solution on the F_1 side of the membrane.

simple technique to measure the number of protons transported for each pair of electrons. Mitochondria are suspended in a medium that contains no oxygen; then the concentration of hydrogen ions is monitored as a known quantity of oxygen is injected into the solution. The proton concentration is observed to rise immediately and rapidly; when the oxygen is exhausted, the concentration slowly returns to normal as protons leak back across the membrane. Adding a substance called a proton ionophore immediately dissipates the gradient; such ionophores are known to make membranes permeable to protons. The inverse behavior can be demonstrated in chloroplasts. After a brief flash of light the proton concentration in the external medium falls rapidly and then slowly regains its original level as protons leak out of the chloroplasts. Again proton ionophores collapse the gradient.

Measuring the magnitude of the gradient is complicated by the need to determine the difference in *p*H and the membrane potential separately. The gradient in proton concentration can be determined by measuring the uptake of weak acids into mitochondria or bacteria; in chloroplasts a weak base is employed instead of an acid. The membrane potential is revealed by the uptake of an ion (other than H^+ or OH^-) that is free to pass through the membrane and reach an equilibrium concentration gradient.

In mitochondria the proton gradient has been found to consist of a *p*H difference of about 1.4 *p*H units (acid outside) and a membrane potential of some 140 millivolts (positive outside). Vesicles prepared from *E. coli* membranes have a *p*H gradient of two units and a membrane potential of about 70 millivolts, with the same polarity as in mitochondria. The way in which the electrochemical gradient is partitioned into its two components depends on what ions are present in the solution and on the external *p*H. In chloroplasts, of course, the gradient is oriented in the opposite direction. What is more, in the steady state almost the entire gradient in chloroplasts is expressed as a difference in proton concentration. Under these conditions the membrane potential is negligible because the membrane is permeable to chloride ions, but the *p*H gradient reaches a magnitude of about 3.5 units. As a result the interior matrix of chloroplasts can become strongly acidic, reaching a *p*H of about 4. In effect the chloroplast accumulates hydrochloric acid inside. It should be noted in passing that the enzymes of carbon dioxide fixation in chloroplasts are all between the inner and outer membranes and thus are not exposed to this harsh environment. The interior of mitochondria and bacteria, on the other hand, is the site of many enzymatic reactions. Mitochondria and bacteria have mechanisms, such as the exchange of protons for sodium ions, that maintain the internal *p*H close to neutrality.

The entire electrochemical gradient, regardless of how it is partitioned into components of *p*H difference and membrane potential, can be measured in terms of energy. It is the energy acquired by a proton when it is transported across the membrane against the gradient, or the energy given up by a proton when it is carried back across the membrane with the gradient. In chloroplasts the electrochemical gradient is equal to about 4.8 kilocalories per mole of protons; in vesicles made from *E. coli* membranes it is 4.4 kilocalories per mole. Mitochondria have the largest total gradient, about 5.3 kilocalories per mole of protons.

The detection of a proton gradient generated by respiration is by no means a demonstration that phosphorylation is driven by the gradient. The proton ionophores, however, provide evidence bearing on this point. The first of the ionophores to be discovered, dinitrophenol, was found to have an effect on oxidative phosphorylation long before its mode of action was understood. In 1948 William F. Loomis and Fritz A. Lipmann of the Harvard Medical School showed that dinitrophenol inhibits ATP synthesis but stimulates electron transport; in other words, it uncouples respiration from phosphorylation. In 1963 Mitchell proposed that dinitrophenol acts by ferrying protons across membranes, and he showed that the addition of dinitrophenol to bacteria equilibrates a *p*H gradient across the membrane. It has since been shown that the effectiveness of various uncouplers is correlated with their capacity to act as ionophores. One alternative hypothesis is that the uncouplers act by interfering with the F_1-F_0 complex, but the most potent uncoupler, a molecule designated SF6847, is completely effective at a concentration of about one molecule for every five F_1-F_0 complexes.

For complete uncoupling both the *p*H gradient and the membrane potential must be dissipated. The proton ionophores can accomplish this because they carry the protons themselves across the membrane. Each of the two components of the gradient can also be abolished individually by other ionophores. Valinomycin, an ionophore that carries potassium ions (K^+), dissipates only the membrane potential. In the presence of valinomycin a countercurrent of potassium ions neutralizes the difference in electric potential, but with continued respiration there is a compensating increase in the *p*H gradient. Nigericin, another ionophore, brings about an electri-

Who do you call to help the Space Shuttle program get off the ground?

You should see what's taking off at the Kennedy Space Center these days.

The whole place is flying.

The first reusable space vehicle gets a recycled launch site.

To get ready for the first Space Shuttle launch by 1979, over 21 facilities and operations at the Center must be adapted and refitted.

The result?

An efficient *recycled* launch site for an efficient *reusable* space vehicle.

It's a complex assignment entailing hundreds of design analyses and systems modifications—and it's a project monitored for NASA by a team from PRC.

To this very special task, Planning Research Corporation brings uncommon resources—the unique strength of 19 technical service companies in one. And the worldwide professional skills of more than 5,000 experts in over 100 technical disciplines.

The largest diversified professional services company in the world.

In fact, for every problem you can think of, chances are there's a PRC

To help launch the Space Shuttle era at the Kennedy Space Center, PRC has provided cost analyses and design engineering, and scheduled and monitored construction of hundreds of system and site modifications at more than 21 Center facilities.

team of specialists already working on answers.

Planning a major dam system in the Philippines.

Monitoring solar energy projects for ERDA.

Designing a criminal justice data system for New Zealand.

Developing the first national real estate computer service in the U.S.

The list of PRC projects goes on. And on.

Planning, engineering, architecture. Management consulting. Information services. We're here to help.

The point is, there's no problem too technical, too complex or too big for Planning Research Corporation to handle.

As a matter of fact, what could PRC do for you? Planning Research Corporation, 1850 K Street, N.W., Washington, D.C. 20006. Call (202) 293-4700.

Call

prc
Planning Research Corporation

Saving the Main Frame.

The Harris 1600 Remote Communications Processor

Some central computers get bogged down by nickel and dime jobs from distant plants and offices. With Harris "intelligent" terminals and controllers, remote locations can handle routine processing on the spot, saving main frame time for main frame work.

Write: Harris Corporation,
55 Public Square, Cleveland, Ohio 44113.

HARRIS

COMMUNICATIONS AND INFORMATION HANDLING

cally neutral exchange of protons for potassium ions. It thereby acts to eliminate the pH gradient, so that continued respiration enhances the membrane potential. Phosphorylation can continue in the presence of either ionophore. In the presence of both ionophores, however, both components of the proton gradient are collapsed and phosphorylation is strongly inhibited.

When a proton ionophore is present, the energy of respiration is dissipated as heat. In the mitochondria of brown fat tissue this effect is employed to maintain body temperature. In these mitochondria a protein in the membrane short-circuits the proton gradient; fatty acids may also act as proton ionophores.

A persuasive demonstration of the role of the proton gradient in phosphorylation was achieved in 1966 by Jagendorf and Ernest G. Uribe of Johns Hopkins University. They showed that imposing an artificial pH gradient can result in ATP synthesis. Chloroplast inner membranes were suspended in a solution at pH 4 (moderately acidic), then the solution was rapidly brought to pH 8 (slightly basic), so that protons were driven outward. ATP was formed in amounts corresponding to 100 molecules for each CF_1 complex. Mitchell later reported the synthesis of ATP in mitochondria exposed to a transition in the opposite direction, from basic to acidic. William S. Thayer of Cornell and one of us (Hinkle) have also elicited ATP synthesis by imposing artificial gradients on inverted vesicles formed from mitochondrial membranes. The maximum yield of about 10 ATP for each F_1 was obtained when a pH gradient was combined with a membrane potential.

If an electrochemical proton gradient is to serve as the intermediate state coupling respiration to phosphorylation, then it must be energetically and kinetically competent to do so. Enough protons must be translocated to account for the amount of ATP synthesized, the protons must have sufficient energy to provide the free energy of synthesis, and the rates of proton translocation and phosphorylation must correspond.

A basic assumption of the membrane models presented above was that two protons are translocated for each "coupling site," or each passage of a pair of electrons across the membrane. Thus in the mitochondrion, which is thought to have three coupling sites, six protons are thought to be translocated for each pair of electrons passing from NADH to oxygen. The mitochondrion is said to have an H^+/O ratio, or an $H^+/2e^-$ ratio, of 6. *E. coli* and chloroplasts, which are thought to have two coupling sites, have an $H^+/2e^-$ ratio of 4.

The study of such ratios in chemical reactions is called stoichiometry. The stoichiometry of proton translocation can be determined experimentally in mitochondria and bacteria by the oxygen-pulse method and in chloroplasts by measuring proton translocation following a light flash. Recently Martin D. Brand, Baltazar Reynafarje and Albert L. Lehninger of the Johns Hopkins School of Medicine have reported experimental findings that more than six protons are translocated by mitochondria for each electron pair passing from NADH to oxygen. If their findings turn out to be correct, significant modifications of Mitchell's theory will be necessary. One of us (Hinkle) has studied proton translocation in vesicles formed from mitochondrial membranes and in artificial membrane vesicles incorporating mitochondrial respiratory enzymes. The $H^+/2e^-$ ratio measured in these systems agrees with the findings of Mitchell and Moyle, but the controversy is not yet resolved.

The stoichiometry of proton translocation by the F_1-F_0 complex is most conveniently measured by driving the process in reverse, that is, by splitting ATP to create a proton gradient. What is measured is the number of protons translocated for each ATP molecule split, which is presumably equal to the number of protons required to synthesize a molecule of ATP. Through an ATP-pulse method, analogous to the oxygen-pulse method, Mitchell determined that mitochondria translocate approximately two protons for each ATP, a finding confirmed by Thayer and one of us (Hinkle) in inverted mitochondrial vesicles. As we noted above, the corresponding enzyme of chloroplasts, CF_1-F_0, seems to require three protons per ATP, but direct measurements of this ratio are difficult to make. Values ranging from 2 to 4 have been reported, and the ratio of 3 was deduced indirectly.

From these ratios the overall stoichiometry of oxidative and photosynthetic phosphorylation can be estimated. If six protons are translocated for each pair of electrons in mitochondria, and two protons are required to flow back across the membrane for each ATP formed, then no more than three ATP molecules can be synthesized for each electron pair. The ratio is only an upper limit, since fewer ATP molecules would be produced if some protons were diverted into other processes, such as active transport. The overall ratio has been determined experimentally many times by measuring the number of ATP molecules synthesized for each oxygen atom reduced to water. In mitochondria the experimental evidence is generally considered to indicate a ratio of three ATP molecules for each pair of electrons passing from NADH to oxygen. Several recent studies, however, have suggested that the ratio may be as low as two. A possible mechanism that might account for the lower ratio, based on the need for energy to transport ADP and ATP, will be discussed below.

In chloroplasts, where only four protons are translocated for each electron pair and where three protons are required for each ATP synthesized by CF_1-F_0, the maximum stoichiometry of synthesis should be four ATP molecules for every three electron pairs. Most observed ratios are consistent with that value, although ratios greater than 4/3 have occasionally been reported.

From these stoichiometries an energy balance sheet for ATP synthesis can be drawn up. As we calculated above, the electrochemical proton gradient in chloroplasts represents an energy store of about 4.8 kilocalories per mole of protons. Since three protons flow back across the membrane for each ATP formed, an energy of 3×4.8, or 14.4, kilocalories is available for phosphorylation. Other estimates suggest that the maximum free energy required for ATP synthesis in chloroplasts is 14.5 kilocalories per mole, and so the energy budget is very close to balancing.

The energetic demands of mitochondrial phosphorylation are somewhat more complicated. The protons crossing the mitochondrial membrane, it will be remembered, have an energy of 5.3 kilocalories per mole, but only two protons are consumed for each ATP formed. The free energy made available, then, is 10.6 kilocalories per mole, which is insufficient since a maximum of about 15 kilocalories is required for phosphorylation. There is an important difference, however, between phosphorylation in chloroplasts and in mitochondria, since in the latter the reaction takes place inside the inner membrane. In that environment the high concentration of reactants (ADP and phosphate) and the low concentration of product (ATP) shifts the equilibrium point of the reaction and effectively reduces the free energy required. Inside the mitochondrion ATP synthesis demands only about 11 kilocalories per mole, which is within range of the calculated energy available.

The reduction in the energy requirement of phosphorylation in mitochondria does not come without a cost. The cost is the energy needed to concentrate ADP and phosphate in the interior cavity and to remove ATP. From a study of the transport reactions Martin Klingenberg and his colleagues at the University of Munich have concluded that the counterflow of ADP and ATP ions involves the net movement of charge across the membrane. The exchange of reactants and products could therefore be powered by the membrane potential. Similarly, phosphate is absorbed in exchange for OH^- ions and its transport is therefore driven by the pH gradient. The combined effect of these transactions is to transport an additional proton for each ATP synthesized and exported. Hence a total of three protons is expended for each ATP appearing outside the mitochondrion, and the total energy available is 3 × 5.3, or 15.9, kilocalories per mole. The stoichiometry of mitochondrial phosphorylation is also changed: the $H^+/2e^-$ ratio is reduced from 3 to 2, in agreement with recent findings.

If a proton gradient is to be a satisfactory intermediate between electron flow and phosphorylation, then it not only must provide sufficient free energy but also must be kinetically competent. The rate at which protons are produced must at least equal the rate at which they are consumed.

Thayer and one of us (Hinkle) made a stringent test of the chemiosmotic theory by comparing the rate of ATP synthesis in inverted mitochondrial vesicles driven by respiration with that driven by an artificial electrochemical proton gradient. In the first tenth of a second the artificial gradient drove ATP synthesis faster than the oxidation of NADH did, but thereafter synthesis driven by the artificial gradient declined as the gradient decayed. The implication is that the rate of phosphorylation is controlled by the magnitude of the gradient. A difference in the initial rate of synthesis was also observed in two systems of vesicles with naturally generated gradients. One suspension of vesicles had all the reactants needed for oxidative phosphorylation except NADH, which was added at the last moment to start the reaction; in the other suspension the one missing component was ADP. The system triggered by adding NADH started much slower, presumably because time was required for the proton gradient to build up. In the system lacking ADP respiration could establish the gradient in advance and phosphorylation could begin immediately when ADP was added.

More recently Boyer and his colleagues have shown that the rate of ATP synthesis in chloroplasts exposed to an acid-to-base transition is equal to the maximum rate driven by light. A. R. Portis, Jr., and one of us (McCarty) have found that the rate of phosphorylation in chloroplasts is critically dependent on the magnitude of the pH gradient. Indeed, the rate is proportional to the cube of the internal proton concentration, a finding consistent with other indications that three protons pass through the CF_1-F_0 complex for each molecule of ATP formed.

One of the most powerful techniques of biochemistry consists in taking apart a complex system of enzymes and reassembling selected parts of it so that the operation of each part can be examined in isolation. This method can be applied to the study of oxidative and photosynthetic phosphorylation by embedding selected proteins in an artificial membrane.

In 1971 Yasuo Kagawa and Racker, who had by then moved to Cornell, developed a method of incorporating isolated transport enzymes into artificial vesicles made of phospholipids, the fat-like molecules that form the basic matrix of all biological membranes. The vesicles were made by mixing the phospholipids and the selected proteins with a detergent and then removing the detergent by dialysis. The molecules assemble themselves spontaneously into closed vesicles.

The first vesicles made by Kagawa and Racker incorporated the mitochondrial F_1-F_0 complex. They were inverted with respect to normal mitochondria, that is, the F_1 knobs were on the outer surface rather than the inner one. As a result the splitting of ATP by the F_1-F_0 complex transported protons inward.

Since then one of us (Hinkle) and his co-workers have employed the same method to study each of the three segments of the mitochondrial respiratory chain separately. Each of the systems could be driven by adding a different combination of oxidants and reductants. In each case the resulting electron flow was found to be coupled to the transport of protons with a stoichiometry similar to that observed in mitochondria. Both a proton concentration gradient and a membrane potential could be formed.

Racker and his co-workers have constructed vesicles incorporating both the F_1-F_0 complex and segments of the respiratory chain. Care must be taken in preparing such particles to see that all the components in a single vesicle have the same sidedness. When all the functional proteins are properly oriented, the vesicles are capable of oxidative phosphorylation. With Walther Stoeckenius of the University of California at San Francisco, Racker has also combined the mitochondrial F_1-F_0 complex with a light-driven proton pump, called bacteriorhodopsin, from a salt-loving bacterium. The hybrid vesicles exhibit light-driven phosphorylation.

These reconstituted systems offer persuasive evidence that each segment of the respiratory chain transports protons and that a proton gradient alone, generated by any means, will drive phosphorylation by the F_1-F_0 complex. They strongly suggest that a proton gradient is not only necessary for ATP synthesis but also sufficient. The investigation of the chemiosmotic theory can thus move on from testing basic postulates to examining detailed mechanisms.

The Combinatorial Mathematics of Scheduling

What is the best way to organize work so that it is finished in the shortest possible time? Discoveries in mathematics and computer science reveal the values and limits of various scheduling methods

by Ronald L. Graham

Scheduling problems arise in almost all areas of human activity. The Viking mission to Mars called for coordinating the activities of more than 20,000 people. Meeting the daily manufacturing quota in an automobile plant can depend on the precise allocation of manpower and tools. Even the preparation of a multicourse dinner can present a nontrivial scheduling problem. It might appear that there are natural algorithms, or step-by-step procedures, for constructing highly efficient schedules. That, however, is not the case. Apparently logical ways of constructing schedules cannot be counted on to perform equally well in different situations. For example, in some instances increasing the number of workers on a job can actually increase the time required to meet a schedule. Some of the commonest and most intuitive scheduling procedures can give rise to unexpected and even seemingly paradoxical results.

Over the past few years several mathematical models of scheduling processes have been devised and closely analyzed in an effort to find the causes of such anomalous results. It was necessary to determine how inefficient different scheduling procedures could be and how poor performance could be eliminated or at least minimized. The research has shown that in most cases perfect results cannot be expected of efficient scheduling algorithms. Nevertheless, general guidelines have emerged for avoiding the pitfalls in scheduling problems and for finding acceptable solutions to particular problems. Here I shall describe some recent results of this kind that apply to one of the most basic scheduling models. The model deals with the behavior of the finite sets of quantities that are involved in scheduling problems. The analysis of the model demonstrates the productive interaction that often occurs between computer science and mathematics. In this instance the theory of algorithms is paired with combinatorial theory (in particular the study of finite sets) to yield important results in scheduling theory.

In order to provide insight into scheduling problems a model must isolate the essential parts of real scheduling situations. The basic scheduling model consists of a system of m identical processors $P_1, \ldots, P_m$ and a set of tasks $A, B, \ldots, T$ to be performed by the processors. (The processors could be workers on an assembly line or physical devices such as electronic microprocessors.) With each task T there is associated a positive number $\tau(T)$; it is the amount of time required by the processor to execute T, and so it is called the execution time of T. Although each processor is capable of executing each of the tasks, no processor can execute more than one task at a time. Moreover, when a processor begins to execute a task T, it must continue executing the task until T is completed.

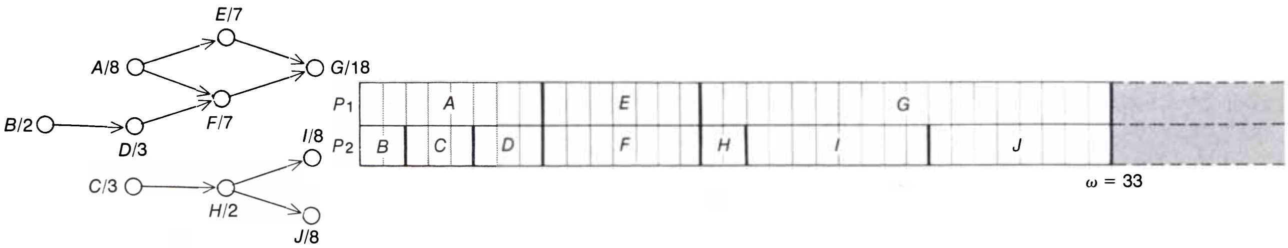
A scheduling algorithm is the set of rules by which tasks are assigned to the individual processors; changing the rules changes the algorithm. In the model the order in which tasks are selected for execution depends on two factors: a collection of precedence constraints and a priority list. If task R must be completed before task S can be started, R is called a predecessor of S. This relation is called a precedence constraint and is written $R \rightarrow S$. The priority list L is an ordering of the tasks according to the preferences of the scheduler: $L = (B, C, R, \ldots)$. The list is not required to be consistent with any precedence constraints that may exist. This model is often called the priority-list scheduling model because changing the priority list is the only way to alter the schedules the model constructs.

The operation of the model begins at time $t = 0$ with all the processors scanning the priority list of the set of tasks from the beginning in search of "ready tasks." A task is said to be ready for execution if all its predecessors have been completed and no processor has begun to execute it. Of course, at time $t = 0$ the ready tasks are those with no predecessors. If two or more processors are competing for a task, the task is assigned to the processor P_i that has the lowest index number i. When a processor cannot find a ready task, it stops scanning and becomes idle. The processor remains idle until a task is completed somewhere in the system. The task might be a predecessor of other tasks that would become ready on its completion. Therefore whenever a task is completed, all idle processors instantaneously begin to scan the priority list from the beginning. The schedule ends when the final task has been completed; the finishing time is denoted ω. (For the purpose of simplification the time spent in scanning the list is not included in the operating time of the model.)

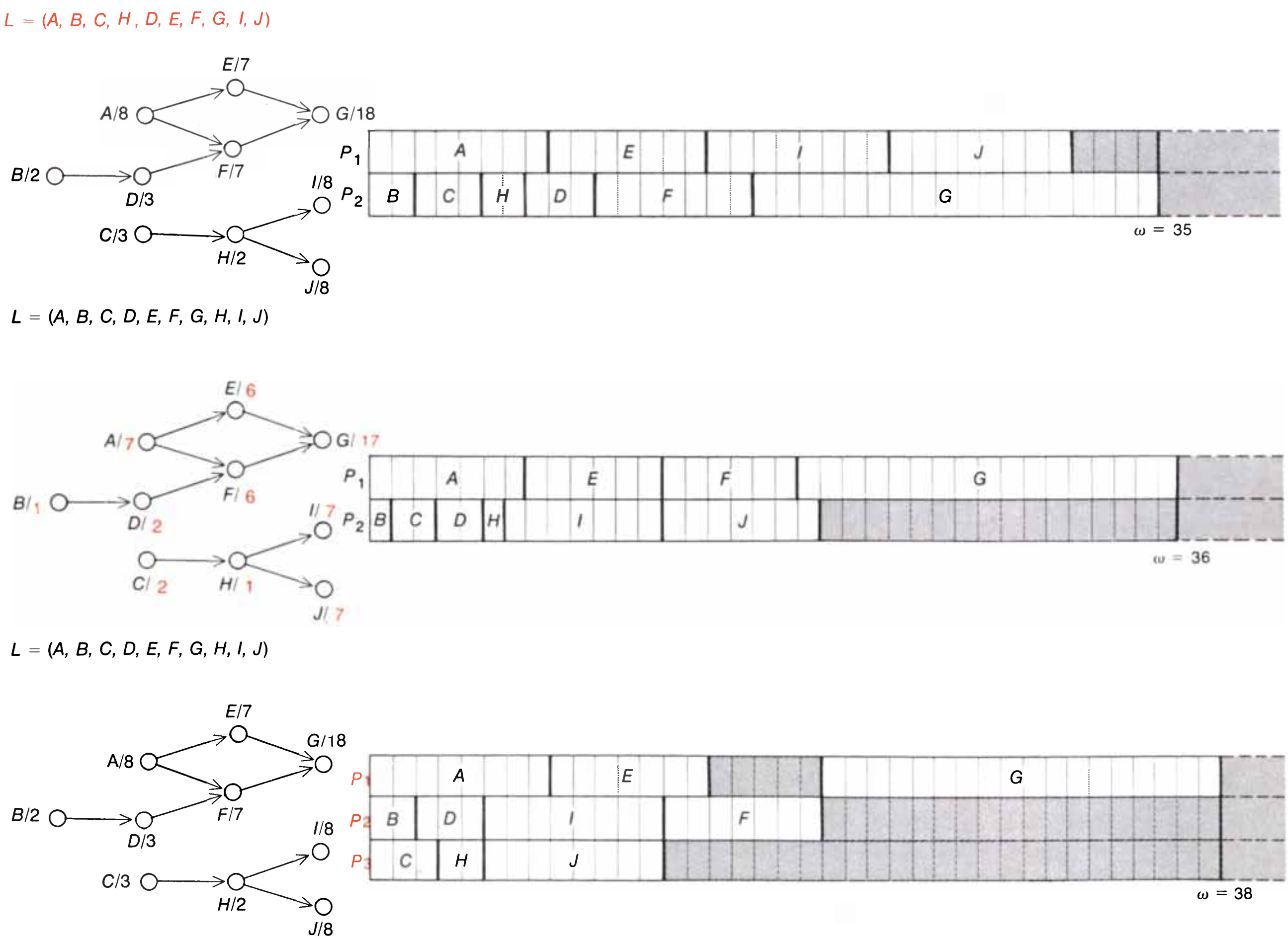
The model, which describes a typical scheduling situation, was studied to find answers to the following questions: How does the finishing time depend on the choice of the priority list? How does it depend on the precedence constraints? How does it depend on the execution times? How does it depend on the number of processors? In particular, how can the list be chosen in order to minimize the finishing time? In other words, the model serves to evaluate the scheduling algorithm by assessing its performance under different conditions, that is, with faster processors, fewer tasks and so on.

The model does not apply to all scheduling situations. For example, there is no consideration of the probabilistic aspects of scheduling in which, say, the execution times of tasks are not fixed but random, according to some distribution of probabilities. The model also regards finishing time as the only measure of algorithm performance, when actually there can be several others. On the other hand, the model does incorporate a number of features that are common to many real situations.

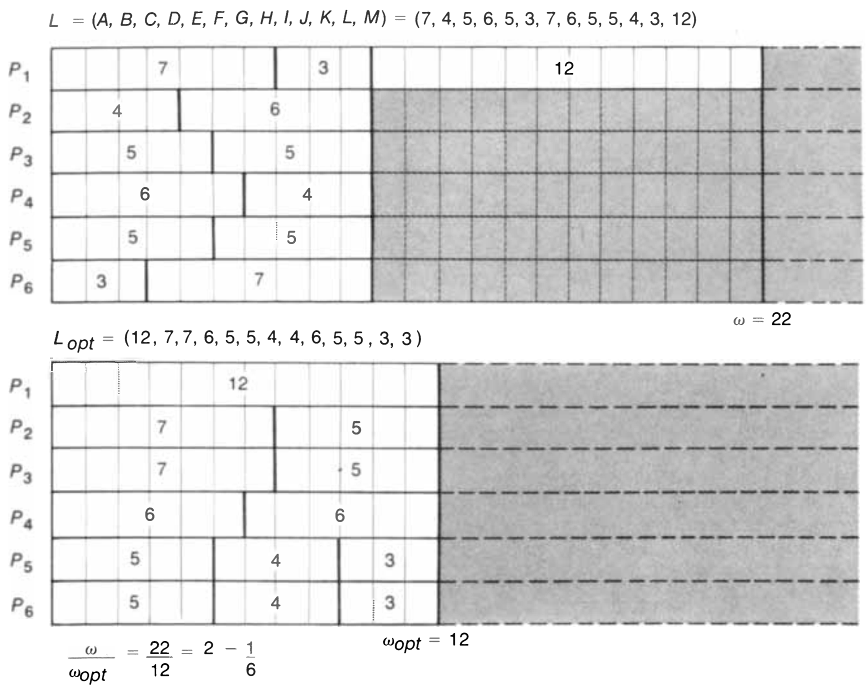
L = (A, B, C, D, E, F, G, H, I, J)



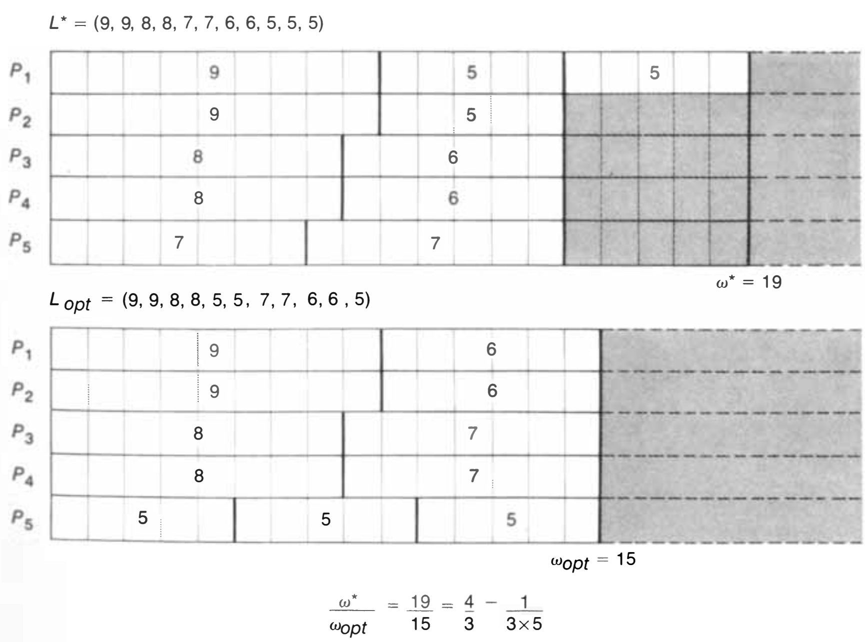
A SCHEDULING MODEL isolates the essential elements of real scheduling situations. It can be used to study the behavior of algorithms, or step-by-step procedures, for constructing efficient schedules. The basic scheduling model consists of a set of identical processors and a set of tasks to be performed by the processors according to certain rules. With each task T there is associated an execution time $\tau(T)$, equal to the amount of time required to execute T. In the example shown in this illustration there are 10 tasks, $A, B, \ldots, J$, and two processors, P_1 and P_2. The tasks to be executed are related by various precedence constraints. Task Q is said to be a predecessor of task S if Q must be completed before S can be started; this relation is called a precedence constraint and is written $Q \rightarrow S$. A scheduling algorithm is the set of rules by which tasks are assigned to processors. The assignment of tasks in the model is determined by the use of a priority list, an ordering of the tasks according to the scheduler's preferences. It is not required to be consistent with any precedence constraints that may exist. The priority list L for the example is shown at the top left; the graph at the bottom left displays the tasks (*open circles*) and the precedence constraints among the tasks (*arrows*). Each task T in the graph is labeled with its name and execution time: $T/\tau(T)$. The distribution of the tasks begins with all processors scanning the priority list from the beginning in search of ready tasks, that is, tasks whose precedence constraints have been satisfied. When a ready task is located, it is always assigned to the processor P_i with the lowest index i. In this case A is the first ready task in the list L and so it is assigned to the processor P_1. After a task has been assigned to a processor the unoccupied processors resume scanning the list. The timing diagram at the right shows the schedule, or allocation of processing time, determined by the rules of the model. The shaded areas in the diagram represent periods when a processor was idle. In the model a processor may become idle only if there is no ready task in the system, and whenever a task is completed, all idle processors begin scanning the list again. The schedule is completed when all the tasks in the set have been executed. In this example the finishing time ω is 33.



CHANGING THE PARAMETERS of the basic scheduling model can have unpredictable and often undesirable results. For example, when the priority list L of the scheduling problem in the illustration at the top of the page is rearranged (*top*), the finishing time increases to 35. A more surprising schedule results when all the execution times in the problem are decreased by 1 (*middle*): the finishing time increases to 36. Moreover, with the decreased execution times the finishing time is always at least 36, no matter what priority list is chosen. Even when third processor is added to original system (*bottom*), finishing time does not decrease but increases to 38 for any choice of list.



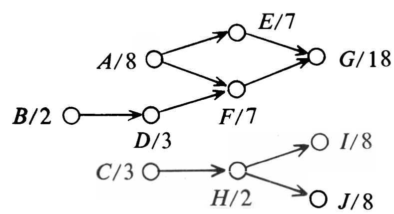
A SCHEDULE IS OPTIMUM for a particular set of tasks if its finishing time is the shortest that can be achieved by any rearrangement of the priority list for the tasks. In the scheduling problem shown here there are no precedence constraints among the tasks, and so the tasks are distributed to the processors in the order in which they appear in the priority list *L*. The finishing time of the resulting schedule (*top*) is $\omega = 22$; during the schedule many of the processors are idle for long periods of time. A different ordering of the tasks results in a schedule with a shorter finishing time (*bottom*). This schedule is clearly optimum because no processor is idle at any point before the finishing time $\omega_{opt} = 12$. The ratio of the two finishing times is the largest possible because it has been proved that ω/ω_{opt} is never greater than $2 - 1/m$, where *m* is the number of processors utilized. Tasks in illustration are identified by their execution times; since tasks are independent (there are no precedence constraints), no confusion results.



IN THE DECREASING-LIST ALGORITHM demonstrated here the more time-consuming tasks are executed as early as possible in the schedule so that there is less chance of some of the processors being idle while others are still operating near the end of the schedule. In the algorithm a priority list L^* is formed by arranging the tasks in decreasing order of execution times. When there are precedence constraints among the tasks, the algorithm can perform quite badly, but when the tasks are independent, the decreasing-list algorithm is guaranteed to give results that are well within the $2 - 1/m$ bound. In fact, for any list of tasks *L*, ω^*/ω_{opt} (ratio of finishing time of decreasing-list schedule for a set of tasks to optimum finishing time for the set) is never more than $4/3 - 1/3m$. Five-processor example shown here attains bound exactly.

The processors of the model could be typists working for a company and the tasks could be a set of reports to be typed. Although the company president's report might be at the head of the priority list, his report would probably depend on the results of subordinates' reports, that is, those reports would be predecessors of the president's. The processors of the model could also represent minicomputers in a multiprocessing computer system that is executing the various subroutines of a complex program. Although the model is rather simple, it does have sufficient structure to exhibit almost the full range of difficulties encountered in general combinatorial scheduling problems. An example will help to demonstrate this fact.

Assume that there are two processors, P_1 and P_2, and 10 tasks to be executed, arranged in a list: $L = (A, B, \ldots, J)$. The precedence constraints among the tasks and the execution times can be conveniently displayed in a diagram in which each open circle represents a task *T*, each arrow represents a precedence constraint and the tasks are labeled $T/\tau(T)$:



At time $t = 0$ the two processors begin to scan *L* and immediately arrive at task *A*, which, having no predecessors, is ready to be executed. According to the rules of the model, *A* is assigned to the processor P_1 because it has the lower index number. The next ready task is *B*, and so it is assigned to P_2. At time $t = 2$, *B* is completed and P_2 scans *L* again, finds that task *C* is ready and begins to execute it. The process continues until time $t = 33$, when all 10 tasks are completed [*see top illustration on preceding page*]. Since neither processor was idle at any time before $t = 33$, the list *L* is optimum in the sense that no matter how the tasks are arranged in a list it is not possible to finish all of them in a shorter amount of time.

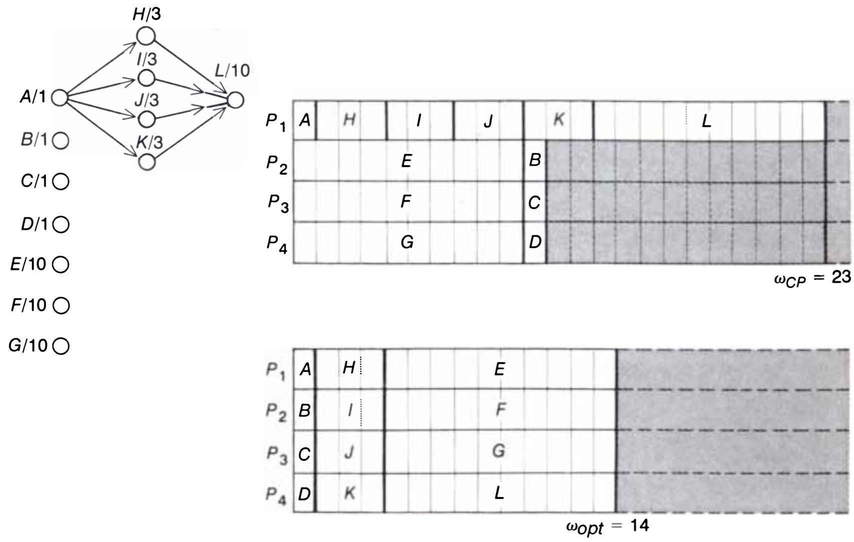
The finishing time $\omega = 33$ cannot be improved by changing the list. What happens when the other parameters of the model—the execution times and the number of processors—are varied? The results are surprising [*see bottom illustration on preceding page*]. If all the execution times in the example are decreased by 1, the finishing time increases to 36. It is tempting to try to explain this odd result by postulating that the priority list *L* is simply an extremely poor choice for the revised set of execution times and that with a better priority list the finishing time would be reduced. A little experimentation shows, however, that with the new execution times no matter how the tasks are arranged the finishing time is never less than 36.

Suppose another processor is added to the original system. With a 50 percent increase in processing capability it is not unreasonable to expect a comparable decrease in the finishing time. In this instance, however, the finishing time is 38, no matter what priority list is chosen.

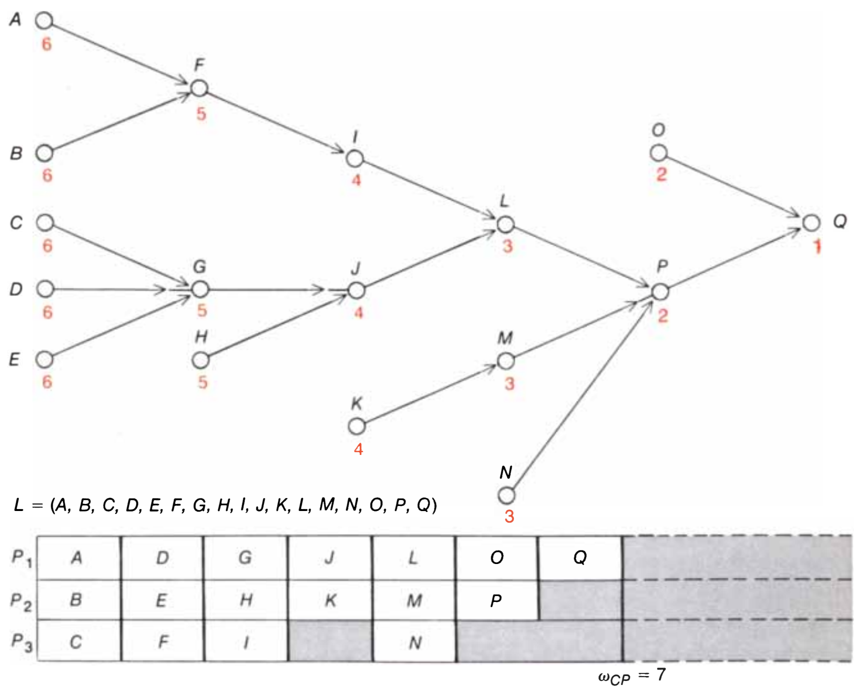
Two elements of the scheduling model are responsible for this unpredictable and undesirable behavior. First, no processor is allowed to be idle if there is a ready task in the system; second, once a processor begins to execute a task it cannot stop until the task is completed. As a result the processors are forced to begin the execution of relatively unimportant tasks (tasks that are short or that are involved in few precedence constraints), and once they have begun they cannot interrupt execution to begin more urgent tasks that subsequently become ready. Of course, the structure of the model could be altered, but the fact is that it reflects many real scheduling situations. Hence it is of interest to know just how much of an effect these aspects of the model can have on the schedules that are created. There is an unexpectedly simple answer to the question.

Imagine that a set of tasks is executed twice: on one occasion with m processors, a particular priority list, a set of precedence constraints and a set of execution times, and on another occasion with m' processors, a different list, a weaker set of precedence constraints and a set of reduced execution times. The finishing time is denoted by ω for the first schedule and by ω' for the second. The example with two processors and 10 tasks showed that ω'/ω, the ratio of the finishing times, can be greater than 1, that is, improving the values of the model parameters can cause an unavoidable increase in the finishing time. There is, however, a limit to the ratio and thus a limit on the adverse effects of the model structure. Some years ago I was able to show that ω'/ω is always less than or equal to $1 + (m-1)/m'$, that is, $\omega'/\omega \leq 1 + (m-1)/m'$. Moreover, there are examples where ω'/ω is equal to the upper bound $1 + (m-1)/m'$, and so the bound cannot be improved by substituting a smaller quantity on the right-hand side of the inequality. (If a smaller bound were used, the examples would contradict the inequality.)

In instances where the schedules that are produced by different priority lists are compared, the number of processors remains the same. Therefore m is equal to m' and the inequality takes on its most elegant form: $\omega'/\omega \leq 2 - 1/m$. For example, if there are three processors, m equals 3, and so ω'/ω is never more than 5/3. This result means that for any set of tasks scheduled on three processors even utilizing the worst possible list



CRITICAL-PATH SCHEDULING is one of the commonest scheduling algorithms, although it does not always perform well. In a critical-path schedule such as the one shown at the top of the illustration tasks are distributed to processors according to the relative urgency of the tasks, that is, according to the length (the sum of the execution times) of the various processing chains each task heads in the unexecuted part of the precedence-constraint diagram. The longest chains in the diagram are called the critical paths because those chains are most likely to cause bottlenecks in the execution of the set of tasks. A task that heads a current critical path is always chosen as the next task to be executed in critical-path scheduling. At time $t = 0$ in this example there are four critical paths, each of length 14: $A \to H \to L$, $A \to I \to L$, $A \to J \to L$ and $A \to K \to L$. Therefore task A is the first task to be executed, and it is assigned to processor P_1 because P_1 has a lower index number than P_2, P_3 or P_4. In some instances critical-path schedules are extremely efficient, but for this particular problem ω_{CP}, finishing time of critical-path schedule, is almost twice ω_{opt}, finishing time of optimum schedule shown at bottom of illustration.



CRITICAL-PATH SCHEDULES ARE OPTIMUM for scheduling problems in which all the execution times are equal and the set of precedence constraints is treelike, that is, no task has more than one successor. Because all the execution times are equal (say 1) in this special case, critical-path scheduling can be accomplished by assigning to each task T a number $L(T)$ equal to the number of tasks in the longest chain headed by T. $L(T)$ is called the level of T, and the schedule is constructed by always choosing a ready task with the highest level as the next task to be executed. In the three-processor example shown here the levels are indicated in color. At time $t = 0$ the ready tasks are A, B, C, D, E, H, K, N and O. Tasks A, B, C, D and E have the highest level, 6, and A is chosen to be the first task executed. T. C. Hu of University of California at San Diego has shown that the level algorithm always creates optimum schedules in this special case, and so $\omega_{CP} = 7$ is the best possible finishing time for this particular problem.

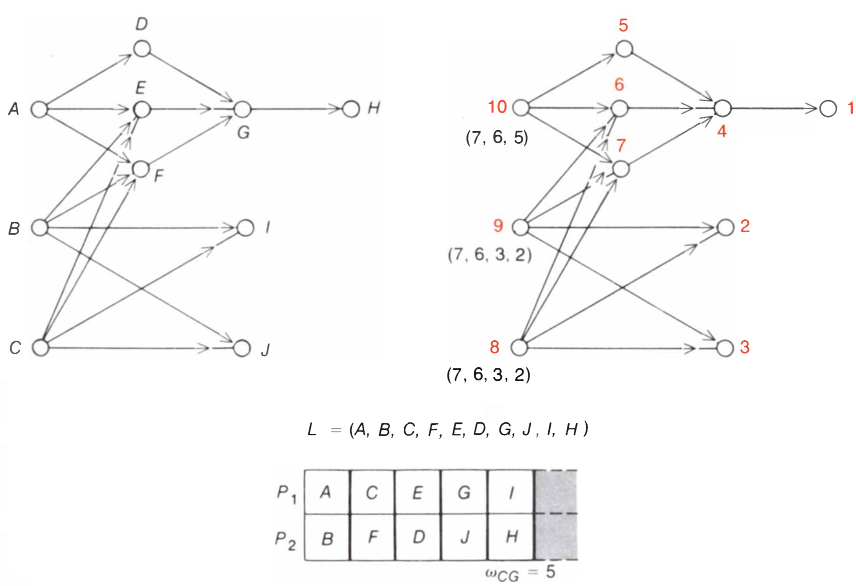
can increase the finishing time obtained with the best possible list by no more than 66⅔ percent. The bound $2 - 1/m$ is a performance guarantee. It ensures that no matter how complicated or exotic a system of tasks, precedence constraints and execution times may be and no matter how cleverly or carelessly a list is chosen, the ratio of finishing times ω'/ω is still never greater than $2 - 1/m$.

It is obvious that within this bound a good or bad choice of a priority list can still make a great difference in the finishing time of a schedule. Common sense suggests that the best lists might be those in which the tasks with the longest execution times appear near the beginning of the list. In that case only relatively small additions to processor usage would be made at the end of the schedule and there would be less chance that some of the processors would be idle while others were still operating.

One of the commonest scheduling algorithms consists in forming a priority list L^* by placing the tasks in decreasing order of execution times and then executing the tasks according to the rules of the model. The finishing time for this decreasing-list algorithm is denoted ω^* [*see bottom illustration on page 126*]. How good is the algorithm? In other words, how close is ω^* to ω_{opt}, the optimum finishing time? When there are precedence constraints among the tasks, the algorithm can create the worst possible schedule, that is, the ratio of ω^* to ω_{opt} can attain the bound $2 - 1/m$.

When there are no precedence constraints, the tasks are distributed to the processors in the order given in the priority list. In that case the tasks are said to be independent, and it has been shown that the decreasing-list algorithm will always give results that are well within the $2 - 1/m$ bound. In fact, for a set of independent tasks the ratio ω^*/ω_{opt} is never more than $4/3 - 1/3m$, which is substantially less than $2 - 1/m$, when m becomes large. Since there are instances where the ratio attains the bound, the bound cannot be improved. The inequality $\omega^*/\omega_{opt} \leq 4/3 - 1/3m$ guarantees that applying the decreasing-list algorithm to a set of independent tasks will never create a schedule whose finishing time is more than 33⅓ percent over the optimum.

At present there is no algorithm for scheduling sets of tasks with precedence constraints that compares with the decreasing-list algorithm for independent tasks. In fact, there is no known efficient procedure for constructing lists of nonindependent tasks whose schedules for three or more processors are guaranteed to finish in $2 - \epsilon$ times the optimum finishing time for some positive number ϵ. The inequality $\omega'/\omega \leq 2 - 1/m$ shows, however, that the schedule for even the worst possible list has a finishing time that is less than twice the optimum. Hence there is clearly much progress to be made in this area.



THE *CG* ALGORITHM creates optimum schedules for the special scheduling problem in which all the execution times are equal and only two processors execute the set of tasks. In this algorithm priority numbers (*color*) are assigned to each task so that tasks heading long processing chains or having many successors receive higher priority. Before the *CG* algorithm can be implemented all extraneous precedence constraints must be removed from the diagram of tasks to be executed. (For example, if $A \rightarrow B$ and $B \rightarrow C$, then $A \rightarrow C$ should be eliminated.) A reduced diagram is shown at the top left; the execution times have been omitted from the diagram because they are all equal, say, to 1. The *CG* algorithm begins by numbering the tasks as is shown at the top right. First, the number 1 is assigned to some task that has no successors, and if there are other tasks without successors, they are numbered 2, 3 and so on. In the illustration task *H*, which has no successors, was numbered 1 and tasks *I* and *J* were respectively numbered 2 and 3. Thereafter for each task for which all the successors have been numbered the decreasing sequence of all the successors' numbers is formed. The next task to be assigned a number is always the one whose sequence is first in the dictionary order (5, 3, 2 comes before 6, 1; 5, 4, 3; 5, 3, 2, 1, and so on) of the established sequences. For example, 7, 6, 3, 2, the sequence for tasks *B* and *C* in the illustration, comes before 7, 6, 5, the sequence for task *A*. Therefore the numbers 8 and 9 are assigned to *B* and *C*, and 10 is assigned to *A*. When all tasks have been numbered, priority list *L* is formed by arranging tasks in decreasing order of assigned numbers. Timing diagram at bottom of illustration shows that *CG* schedule for tasks *A*, *B*, . . . , *H* is indeed optimum.

The preceding discussion raises an obvious question: Why should so much effort be devoted to creating schedules that are good but not the best? Why not try to find the optimum schedule for a set of tasks? One way to do so would be to examine all the possible schedules for the tasks and then choose the one with the shortest finishing time. The trouble with this brute-force approach is that as the number of tasks in a set becomes large the number of possible priority lists (and thus the number of schedules) grows so explosively that there is no hope of examining even a small fraction of them. If there are n tasks in the set, the number of different lists is $n!$, or $n(n-1)(n-2)\cdots 1$, a very large number even for relatively small values of n. For example, when there are 20 tasks, even if a computer could check as many as a million schedules per second, it would take more than 70,000 years to check all 20! lists.

The number of possible lists $n!$ (and hence the number of operations and the computer time needed for checking the schedules) is an exponential function of the number of tasks n. Exponential functions increase rapidly as the value of the variable n increases. A polynomial function, however, say n^2, does not explode as rapidly as n increases. It would be practical to insist on optimum schedules if an algorithm for finding optimum schedules could be found in which the number of computational steps grows as a polynomial function of the number of tasks.

It seems highly unlikely that such an algorithm will be found. This gloomy prospect is the result of the fundamental work of Stephen A. Cook of the University of Toronto, who in 1971 introduced the concept of *NP*-complete, or nondeterministic-polynomial-time-complete, problems [see "The Efficiency of Algorithms," by Harry R. Lewis and Christos H. Papadimitriou; SCIENTIFIC AMERICAN, January]. Hundreds of problems notorious for their computational intractability are now known to belong to this class of problems. *NP*-complete problems have two important properties. First, all methods, or algorithms, currently known for finding general solutions to these problems require exponentially increasing amounts of time and thus are extremely inefficient. Sec-


THE NEW 510.
BEST ALL-AROUND
DATSUN YET.
BIGGER THAN THE SMALL CARS. SMALLER THAN THE BIG CARS. The 510 Hatchback. Datsun ingenuity at its best... Big where you need it. In leg room. Hip room. Family room. Cargo room. And an abundance of big car comforts (even side-window defoggers) at no extra cost. Yet the new 510 Hatchback is small where it counts. The styling is sporty and compact. MacPherson strut front suspension for nimble handling. A 5-speed stick for agile performance. And yes, it's all yours at a Datsun price. See the whole series of perfectly new 510s: Wagon, 2- and 4-Door Sedans and a Hatchback that got 35 mpg highway, 25 mpg city.* It's Datsun, driven to its all-around best.
R 2 4
1 3 5
* 1978 EPA estimates based on standard manual transmission. Your mileage may differ, depending on how and where you drive, your car's condition and optional equipment. California high-altitude mileage lower
NOBODY DEMANDS MORE FROM A DATSUN THAN DATSUN.
DATSUN.
WE ARE DRIVEN.
"You're perfect, 510!"

Physics...analytical math... astronomy and cosmology... chemistry...computer sciences... structural geology...undoubtedly the best books in every scientific area that interests you.

Take any 3 books, values to $68.95, all for only $3.95 as a member of

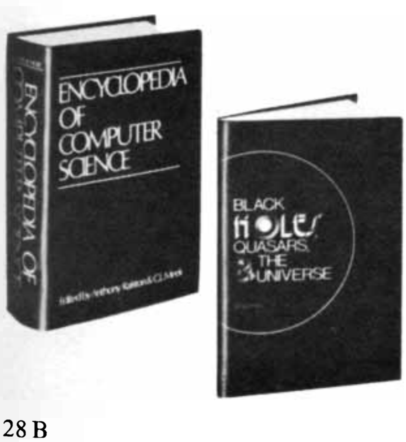
The Library of Science

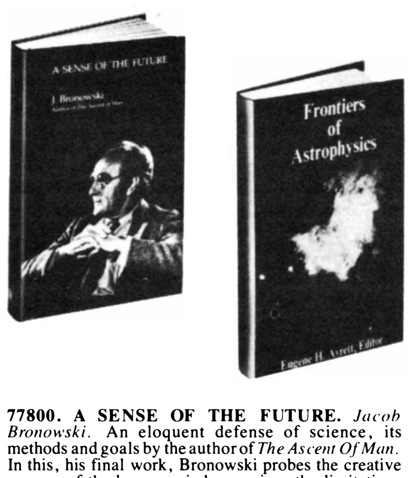
if you will join now for a trial period and agree to accept only 3 more books—at members' discount prices—over the next 12 months

(Publishers' Prices shown)

44900-3. THE ENCYCLOPEDIA OF COMPUTER SCIENCE. *Edited by Anthony Ralston and Chester L. Meek.* A monumental 1,550-page volume, offering condensed, accurate information on essential topics ranging from theory to applications, from program structures to the logical design of central processing units, from Boolean algebra to microprogramming. *Counts as 3 of your 3 books.* **$60.00**

36395. BLACK HOLES, QUASARS AND THE UNIVERSE. *Henry J. Shipman.* Black holes, quasars, white dwarfs, neutron stars, supernovae, Seyfert Galaxies—explores the frontiers of astronomy. **$12.95**



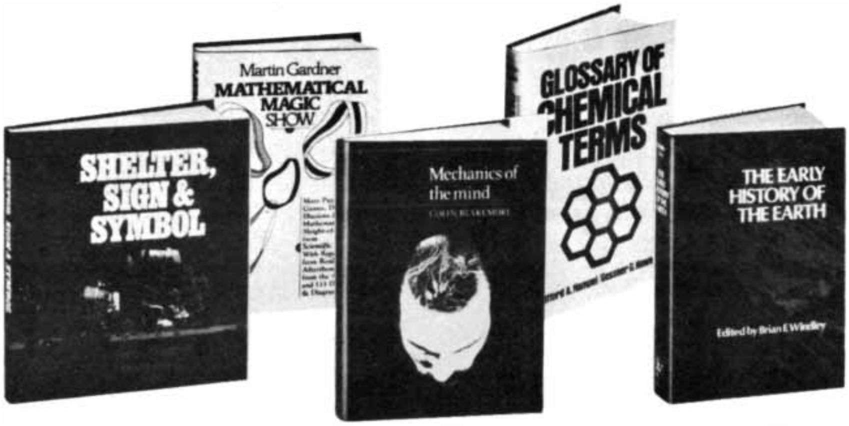


77800. A SENSE OF THE FUTURE. *Jacob Bronowski.* An eloquent defense of science, its methods and goals by the author of *The Ascent Of Man*. In this, his final work, Bronowski probes the creative process of the human mind, examines the limitations and values of science, and speculates on future developments in the sciences, especially biology. **$12.50**

49267. FRONTIERS OF ASTROPHYSICS. *Edited by Eugene Avrett.* A contemporary source of astrophysical topics from recent solar research to neutron stars, supernovae and intergalactic matter. **$20.00**

82775. TEN FACES OF THE UNIVERSE. *Sir Fred Hoyle.* The noted British astronomer and author's personal view of the universe as seen through the eyes of the mathematician, the physicist, biologist, and everyman. A powerful display of Hoyle's keen scientific insights and literary skills that is at once informative and entertaining. Handsomely illustrated. **$10.95**

34140. APPLIED ANALYTICAL MATHEMATICS FOR PHYSICAL SCIENTISTS. *James T. Cushing.* Deals with topics ranging from spectral analysis of linear operations to complex variables, to group theory. **$19.95**



34670. ASTRONOMY AND COSMOLOGY. *Sir Fred Hoyle.* Hoyle combines nearly 600 illustrations with a lucid text as he tours the large-scale structures of the universe, the history of matter, the origins of the planets and the origin of life. **$16.50**

52292-2. A HANDBOOK OF NUMERICAL AND STATISTICAL TECHNIQUES. *J. H. Pollard.* Here are all the statistical and numerical methods you need to solve the mathematical problems on a programmable calculator, minicomputer or interactive terminal. Examples are drawn mainly from the life sciences. *Counts as 2 of your 3 books.* **$24.95**

50551. GRANTS: How to Find Out About Them and What to do Next. *Virginia P. White.* A "must" for anyone who wants to get, keep, or increase grant funding. **$19.50**

78235. SHELTER, SIGN, AND SYMBOL. *Edited by Paul Oliver.* A guided, thorough illustrated tour of the wonders of such exotic locales as Kathmandu; the cave city of Cappadocia, Turkey; and the Greek island of Chios. Also provides a revealing look at the rituals, customs, ancient traditions, and religions of their inhabitants. **$22.50**

43065-2. THE EARLY HISTORY OF THE EARTH. *Edited by Brian F. Windley.* A highly readable, comprehensive look into our planet's past that covers a wide range of topics such as the early earth-moon system, geochronology, archaean tectonics, the atmosphere, the oceans, and life forms. Contains over 390 illustrations, photographs, and tables. *Counts as 2 of your 3 books.* **$45.00**

65640. ON STONEHENGE. *Sir Fred Hoyle.* Was Stonehenge man's earliest astronomical observatory? The mystery of the British Isles' most celebrated megalith is tackled by the world renowned astronomer, who proposes that the 5,000-year-old structure was built to predict lunar eclipses. Absorbing reading for both scientist and layman. **$9.75**

69305. PLANETARY GEOLOGY. *Nicholas M. Short.* A comprehensive, non-technical account of solar system geology. Data from the Apollo landings and recent unmanned probes provide clues to questions concerning lunar and planetary origin, evolution, and composition. Lavishly illustrated with more than 500 photographs, line drawings and graphs. **$21.95**

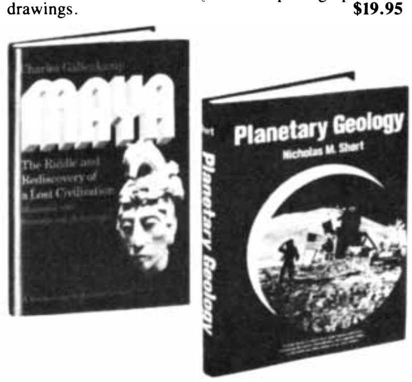
54995. THE ILLUSTRATED ENCYCLOPEDIA OF ARCHAEOLOGY. *Glyn Daniel, General Editor.* Worldwide coverage of the people, places, and things of the past with over 1,000 entries and more than 100 illustrations. The significant sites, cultures, civilizations, terms, tools, and techniques packed into one all-inclusive volume. **$17.95**

87071. THE WHITE LIONS OF TIMBAVATI/THE LAST PLACE ON EARTH. *First,* a young zoologist and his family protect to maturity two rare white lion cubs. Then a close look at Serengeti, where great herds of wildlife roam free. *The 2 count as one book.* **$20.95**

73280. THE RACE BOMB: Skin Color, Prejudice and Intelligence. *Paul R. Ehrlich and Shirley Feldman.* A penetrating look at the volatile issue of race-related inheritance by the author of *The Population Bomb.* Myths of racial superiority and inferiority are put to the test as the authors question the meaning of race, intelligence tests, and group differences. **$9.95**

61610. MAYA. *Charles Gallenkamp.* Mayan civilization from its glory to extinction. A compelling view of the life, arts, religion, and science of this unique civilization. **$12.95**

61660. MECHANICS OF THE MIND. *Colin Blakemore.* Reports discoveries now being made about the human brain: why we have an awareness of self, how we feel and see the world we live in, why we sleep, how we remember. 200 beautiful photographs and drawings. **$19.95**



40167-2. THE CONDENSED CHEMICAL DICTIONARY. *Revised by Gessner G. Hawley.* Not just a dictionary, it covers technical data and descriptive information covering thousands of chemicals and chemical phenomena to meet today's needs. Over 18,000 clear, accurate entries. *Counts as 2 of your 3 books.* **$32.50**

61300. MATHEMATICAL MAGIC SHOW. *Martin Gardner.* Delightful collection of brain teasers and inspired dissertations by the author of *Scientific American's* popular Mathematical Games column. Includes word squares, double acrostics, card tricks, sophisticated jigsaw puzzles and much much more. **$8.95**

50380. GLOSSARY OF CHEMICAL TERMS. *Clifford A. Hampel and Gessner G. Hawley.* Over 2,000 concise definitions: all major chemical groups, important functional terms, basic phenomena and processes, all the chemical elements, and more. **$14.95**

59800. MAN DISCOVERS THE GALAXIES. *Richard E. Berendzen, Richard C. Hart & Daniel Seeley.* Topical, current look at modern astronomy dealing mostly with the discovery and arrangement of our Home Galaxy, the Milky Way System. **$15.95**

61530. MATHEMATICS DICTIONARY. Fourth Edition. *Edited by Robert C. James and Edwin F. Beckenbach.* Revised and expanded edition of this leading mathematics dictionary offers clear, accurate definitions of more than 8,000 terms. **$17.95**

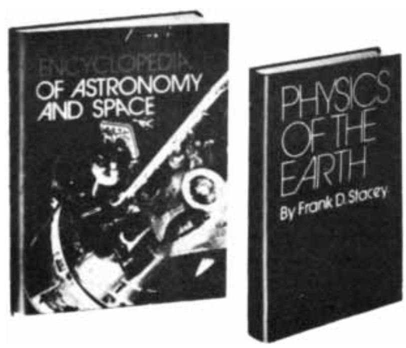
55000. THE ILLUSTRATED ENCYCLOPEDIA OF ASTRONOMY AND SPACE. *Ian Roxburgh.* Covers the full range of astronomical knowledge and space exploration, from ancient times to now. **$16.95**

73240. QUASARS, PULSARS, AND BLACK HOLES. *Frederic Golden.* A thorough review of the spectacular discoveries in astronomy. Illustrated with drawings and color photographs. **$7.95**

75710. SCIENCE BRAIN-TWISTERS, PARADOXES AND FALLACIES. *Christopher P. Jargocki.* Over 160 puzzles, questions, problems and situations, each based on scientific principles, with detailed explanatory solution. Great fun. **$9.95**

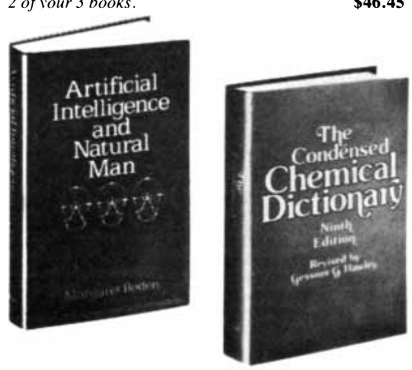
40690. COSMOLOGY. *Michael Rowan-Robinson.* Superbly illustrated with space photographs and numerous star charts and graphs, this fascinating blueprint brings together all the important ideas and developments in the new astronomy. Includes twenty selected controversies in cosmology today. **$13.95**

68780. PHYSICS OF THE EARTH. *Frank D. Stacey.* Covering rotation, gravity, internal structures and plate tectonics, the completely rewritten second edition of this highly acclaimed work also includes the first comprehensive look at rotation and excitation of the wobble, tidal friction, earthquake prediction, thermodynamics of convection and melting and anelasticity. **$16.95**



34430. ARTIFICIAL INTELLIGENCE AND NATURAL MAN. *Margaret Boden.* Brings you up-to-date on the latest work in artificial intelligence as it confronts the complex philosophical, psychological, and social issues raised by man's continued use of [illegible] computers. Informative and thoughtful reading for both scientist and layman. **$17.50**

39746-2. COMPANION TO CONCRETE MATHEMATICS. Volumes I and II. *Z. A. Melzak.* Mathematical techniques and various applications, with historical notes, solutions, on geometry, iteration, series and products, integrals, more. *The set counts as 2 of your 3 books.* **$46.45**



If reply card is missing, please write to
The Library of Science, Dept. 2-A8A
Riverside, N.J. 08075 to obtain membership information and an application.

The TI Programmable 59 has it all!

Solid State Software™ modules. Magnetic card storage capability. Printer/plotter. Professional Program Exchange. Quality. Value. And much more.

Evaluate programmable calculator systems on the basis of power, flexibility, total software support and price/performance. You'll find there's one clear-cut answer: The TI Programmable 59.

Leading the field in performance, quality and value, the Programmable 59 can get you into the power of programming now at a price you can afford. Then, you can add capability as you need it.

With up to 960 program steps, or up to 100 memories, the computer-like TI Programmable 59 can access an additional 5,000-step library of pre-written programs just by dropping in a tiny interchangeable *Solid State Software*™ module. And, with seven different library modules available now, two additional modules this spring and more on the way, whole new worlds are open to you through the power of programming—even if you've never programmed before. *Magnetic card* read/write capability means you can also record your own programs and make them part of your library. A *Professional Program Exchange* makes hundreds of additional programs available to you and TI's new *Specialty Pakettes* offer programs of interest to groups of specialists in a variety of fields. The TI Programmable 59 is only $299.95*. The non-card programmable TI-58 is also available at $124.95. Add the optional *PC-100A printer/plotter*, for an additional $199.95*, and you can print, list, or trace your programs; plot curves and histograms, and print out alpha headings.

This enormous capability at your fingertips can help you maximize your problem-solving abilities. TI believed this strongly enough to invest millions in a program to provide 59's to our own professional employees worldwide. 8,000 calculators were distributed by the end of 1977. Results over the first 6 months of the program show time devoted to creative work increased 25%, with a corresponding reduction in fact gathering and repetitive tasks!

Free literature.
We want to tell you more about the TI Programmable 59 and our other programmable calculators. And, about our software and the PC-100A printer/plotter. You'll discover how easily and economically you can increase your own personal productivity... with a little help from Texas Instruments. Circle the reader service number shown below. Or write: Texas Instruments, Programmable Calculators, P. O. Box 53, Lubbock, Texas 79408. *U.S. suggested retail price.

TEXAS INSTRUMENTS ...INNOVATORS IN PERSONAL ELECTRONICS

TEXAS INSTRUMENTS
INCORPORATED

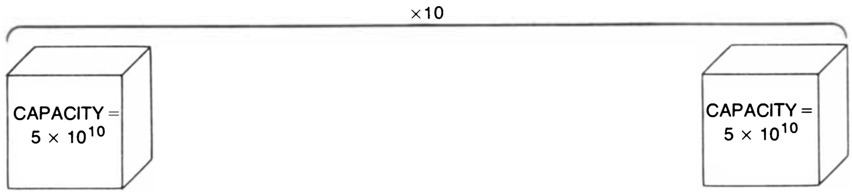
ond, if any one of the *NP*-complete problems had an efficient, or polynomial-time, solution, then all of them would. It seems highly probable, but it has not yet been proved, that the difficulty in finding efficient procedures for solving these problems is inherent in *NP*-complete problems; it appears that no such procedures can exist.

Most scheduling problems are *NP*-complete. In fact, even the comparatively simple situation in which there are no precedence constraints and only two processors presents an *NP*-complete problem. The discoveries about *NP*-completeness changed the direction of research on scheduling. Earlier efforts were directed at finding optimum, or exact, solutions to scheduling problems, but now most attention has been turned in the more fruitful direction of determining approximate solutions easily, of finding efficient methods that are guaranteed to give close to optimum results. The decreasing-list algorithm for scheduling independent tasks, guaranteed to finish within 33⅓ percent of the optimum finishing time, exemplifies this new approach.

The decreasing-list algorithm achieves its close to optimum finishing time by doing a certain amount of work. The n tasks to be executed must be sorted into a decreasing list before they are distributed to the processors. That sorting can be accomplished in an amount of time, or a number of computational steps, that is proportional to $n \log_2 n$. As n increases, $n \log_2 n$ increases only slightly faster, and so the algorithm is acceptably efficient.

SUM = 5×10^{11}

1415926535	5820974944	8979323846	5923078164	2643383279
8214808651	4811174502	3282306647	8410270193	0938446095
4428810975	4564856692	6659334461	3460348610	2847564823
7245870066	7892590360	0631558817	0113305305	4881520920
3305727036	0744623799	5759591953	6274956735	0921861173
9833673362	6094370277	4406566430	0539217176	8602139494
0005681271	1468440901	4526356082	2249534301	7785771342
4201995611	5187072113	2129021960	4999999837	8640344181
5024459455	7101000313	3469083026	7838752886	4252230825
5982534904	8903894223	2875546873	2858849455	1159562863
0628620899	5028841971	8628034825	6939937510	3421170679
8521105559	5058223172	6446229489	5359408128	5493038196
4543266482	3786783165	1339360726	2712019091	0249141273
4882046652	9628292540	1384146951	9171536436	9415116094
1885752724	8193261179	8912279381	3105118548	8301194912
2931767523	6395224737	8467481846	1907021798	7669405132
4654958537	7577896091	1050792279	7363717872	6892589235
2978049951	5981362977	0597317328	4771309960	1609631859
5875332083	3344685035	8142061717	2619311881	7669147303
9550031194	8823537875	6252505467	9375195778	4157424218



BIN-PACKING, another type of scheduling problem, involves a set of items, or weights, and a collection of identical bins with a fixed weight capacity; the problem consists in packing all the weights of the set into a minimum number of bins. The difficulty in obtaining precise solutions to bin-packing problems is demonstrated in this example. The total of the 100 weights shown is 5×10^{11}. Can the weights be packed into 10 bins of capacity 5×10^{10}? The number of possible packings for this relatively small group of weights is so large that even if all the computing power in the world were applied, it is extremely unlikely that an answer to the question would be found. Most scheduling problems are similarly complex, and so many algorithms are designed to create packings or schedules guaranteed only to be reasonably close to optimum.

By doing more work it is possible to obtain schedules that are even closer to the optimum than those constructed with the decreasing-list algorithm. Consider a two-processor system. One way to obtain such superior schedules is to choose, for some integer k, the $2k$ longest tasks, construct the optimum schedule for these tasks and then schedule the remaining tasks arbitrarily. If the finishing time of this schedule is denoted ω_k, then it can be shown that $\omega_k/\omega_{opt} \leq 1 + 1/(2k + 2)$. For a set of n tasks the entire procedure can be accomplished in at most $2kn + 2^{2k}$ operations. (The term $2kn$ comes from choosing the $2k$ longest tasks and the term 2^{2k} from examining all possible schedules for the $2k$ tasks on the two processors.) Since k is a fixed number, the function $2kn + 2^{2k}$ is a polynomial function of n, not an exponential one, and it grows moderately as n increases. For example, if k is 3, then $\omega_3/\omega_{opt} \leq 9/8$ and the amount of work required is proportional to at most $6n + 64$. In general any degree of accuracy can be obtained with sufficient work. Of course, the amount of work required can rise rapidly; for example, obtaining a value of ω_k that is guaranteed to be within 2 percent of the optimum can take an amount of time proportional to $48n + 2^{48}$, which would be enough to exhaust more than a few computer budgets.

This type of behavior should not be too surprising. Since an exponentially increasing amount of time seems to be necessary for finding an optimum solution, it makes sense for the cost of approximate solutions to behave in the same way as the guaranteed accuracy of the solutions increases. What is surprising is that the exponential increase in time can be avoided; there is a method for constructing schedules for independent tasks that are guaranteed to be quite close to the optimum that requires polynomially increasing amounts of time. Oscar H. Ibarra of the University of Minnesota and Chul Kim of the University of Maryland have recently developed an efficient algorithm for constructing schedules for two processors that have a finishing time ω_k for which $\omega_k/\omega_{opt} \leq 1 + 1/k$. Implementing the algorithm requires an amount of time proportional to $n + k^4 \log n$. (When n and k are large, the value of $n + k^4 \log n$ is usually much smaller than 2^{2k}.) Sartaj K. Sahni of the University of Minnesota has extended the procedure to create efficient algorithms that can be applied to more than two processors. These procedures involve a clever combination of techniques that are beyond the scope of this discussion, but this kind of approximation may well be able to guarantee close to optimum results at a cost of reasonable amounts of time.

Although research into *NP*-completeness indicates that in general no efficient techniques will be found for constructing optimum schedules, there are many interesting special cases of scheduling problems that are not *NP*-complete and that it is possible to construct optimum schedules for in polynomial time. Much of the complexity in scheduling problems is derived from the intricate structure of precedence constraints and from the complicated relations among the execution times. Limiting one or both of these factors can result in the kinds of special cases for which optimum schedules can be found efficiently.

For example, assume that there is a scheduling situation with an arbitrary number of processors where all the processing times are equal and the set of precedence constraints is treelike, that is, every task has at most one successor. In this instance "critical path" scheduling, one of the commonest scheduling methods, will always create optimum schedules [*see top illustration on page 127*]. In critical-path scheduling the tasks are assigned to processors according to the length of the various precedence chains they head in the diagram of precedence constraints. The longest chains in the unexecuted part of the diagram are the ones that have the greatest sum of execution times; they are called the critical paths because their tasks are the ones most likely to be the bottlenecks in the execution of the set of tasks. In critical-path scheduling a task that heads a current critical path is always chosen as the next task to be executed.

T. C. Hu of the University of California at San Diego proved in 1961 that critical-path schedules are optimum for the special case of treelike precedence constraints and equal execution times. Hu's result was one of the first in scheduling theory. Because in this special case all the execution times are equal, the critical-path scheduling consists in assigning to each task T a "level" $L(T)$ equal to the number of tasks in the longest chain headed by T [*see bottom illustration on page 127*]. Whenever a processor is available, a ready task with the highest level is assigned to it.

In another special case of scheduling no limit is placed on the structure of the precedence constraints but all the processing times must be equal and only two processors execute the set of tasks. There are now several methods for determining optimum schedules in this situation. One of them, sometimes called the *CG* algorithm, was developed in 1972 by Edward G. Coffman of the University of California at Santa Barbara and me. (*CG* stands for Coffman and Graham.) It is much in the spirit of the level algorithm for the case of treelike precedence constraints. In the *CG* algorithm, however, the order in which the tasks are executed depends on all the chains headed by each task rather than on a single longest chain, as is the case in the level algorithm.

Before the *CG* algorithm can be applied it is necessary to remove all extraneous precedence constraints from the diagram of the tasks to be executed. For example, if $A \rightarrow B$ and $B \rightarrow C$, then the precedence constraint $A \rightarrow C$ can be eliminated. This process can be accomplished in at most $n^{2.81}$ operations for a set of n tasks. The *CG* algorithm begins by numbering the tasks in the set [*see illustration on page 128*]. First the number 1 is assigned to some task that has no successors. Thereafter for each task for which all the successors have been numbered the decreasing sequence of all the successors' numbers is formed. The next task to be given a number is always the one whose sequence is first in the "dictionary order" of the established sequences. (In dictionary order 5, 3, 2 comes before 6, 1; 5, 4, 2; 5, 3, 2, 1, and so on.) After all the tasks have been numbered from 1 to n the priority list is constructed by placing the tasks in decreasing numerical order. It has been shown that the schedule constructed with this list is always optimum in the special case of two processors and equal execution times. Basically the *CG* algorithm works because it gives larger numbers and thus higher priority to tasks that either head long chains or have many successors. The numbering can be done in approximately n^2 operations, and so the algorithm is quite efficient.

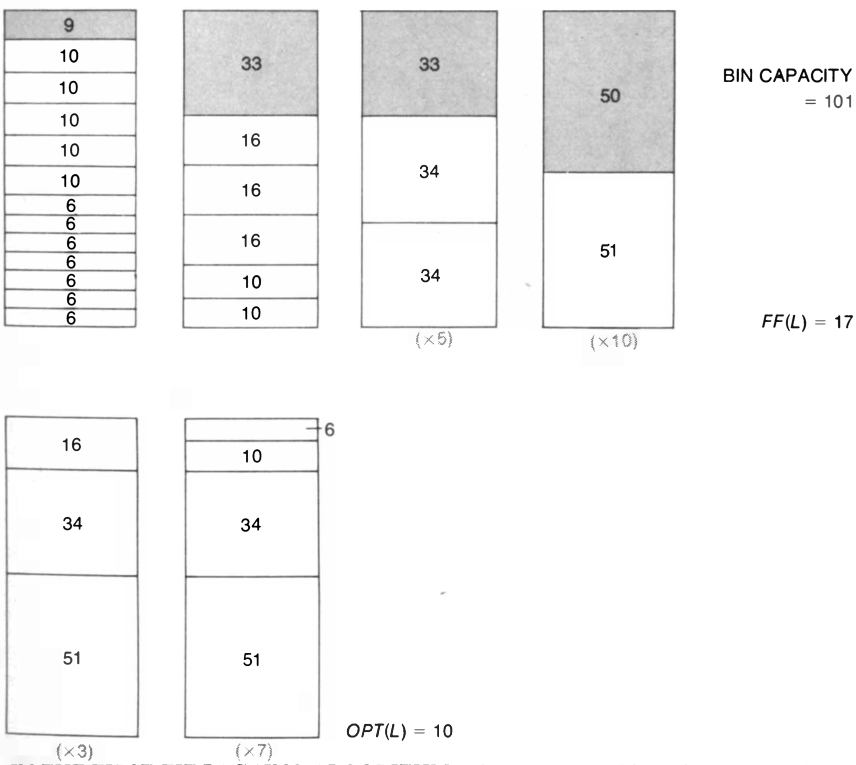
The *CG* algorithm is only one of a variety of techniques that can be applied to special scheduling problems to get optimum results. Perhaps extensions of these techniques will lead to equally successful algorithms for similar problems, such as the problem of three processors with a set of tasks that have equal execution times. It should be noted, however, that even the slightly less special case of two processors with a set of tasks that have execution times of either one unit or two units has recently been shown to be *NP*-complete.

So far I have discussed only one type of scheduling problem, but scheduling problems arise in many places and in many different forms. One of the most interesting problems turns the basic scheduling model around; instead of fixing the number of processors and trying to minimize the finishing time, the problem is to try to complete the execution of a set of tasks by a fixed time with a minimum number of processors. In other words, the problem asks how few processors will suffice to execute a given set of tasks by a fixed deadline.

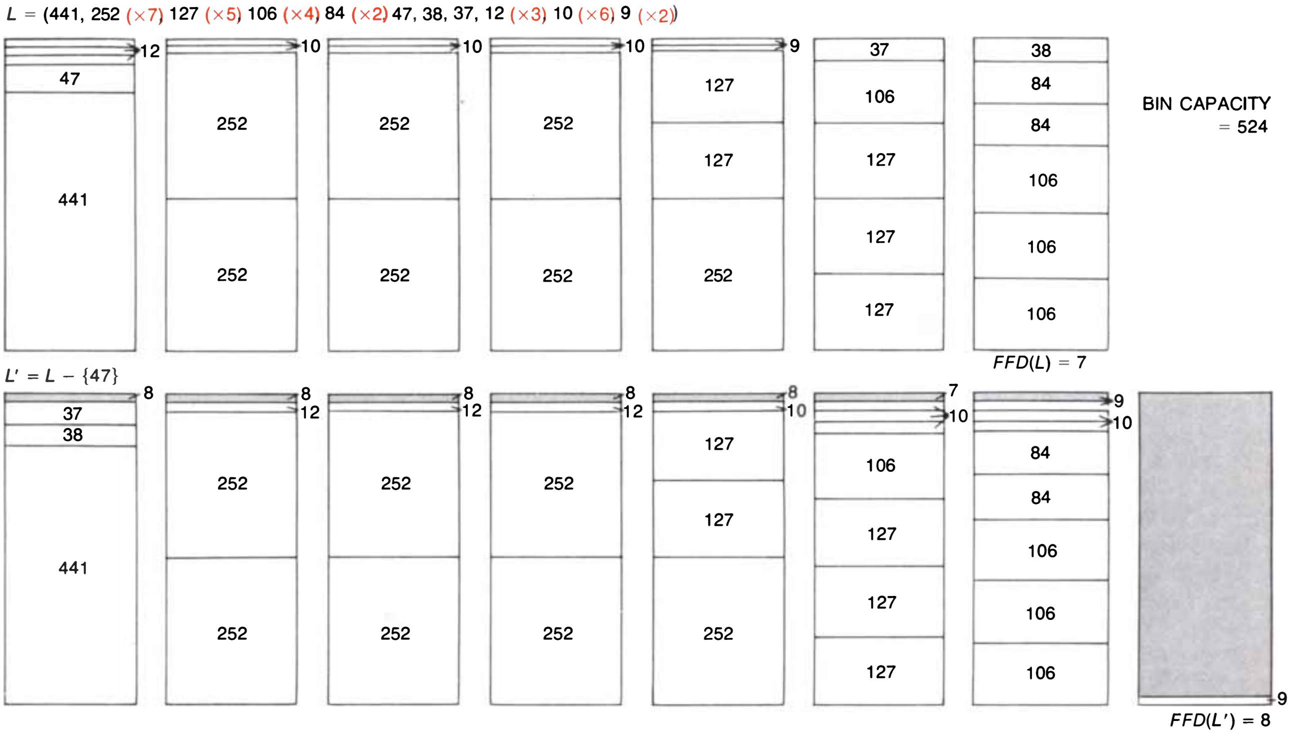
When the tasks are independent, this scheduling problem is stated in a different way and is called the bin-packing problem. In a model of the standard bin-packing problem there is a set of items $I_1, \ldots, I_r$; each item I_k has a weight w_k. The problem is to pack all the items into a minimum number of bins $B_1, B_2, \ldots$ so that the total weight of the items in each bin does not exceed some fixed weight W. (In the terminology of the basic scheduling model the items are tasks, the weights are execution times, the bins are processors and the fixed weight is a fixed finishing time.)

The bin-packing problem arises in a variety of guises in many practical situations. A plumber must cut a set of pipes of different lengths from a minimum number of standard-length pipes; a television network would like to schedule its commercials of varying lengths in a

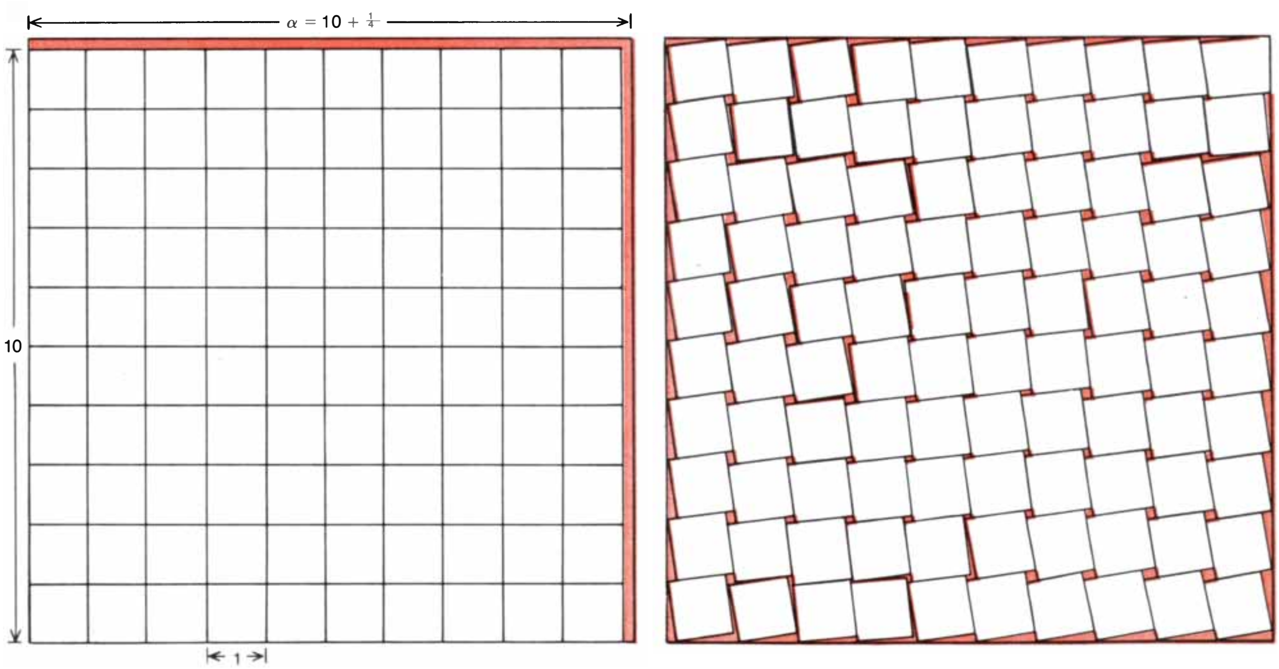
L = (6, 6, 6, 6, 6, 6, 6, 10, 10, 10, 10, 10, 10, 10, 16, 16, 16, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51) = (6(×7), 10(×7), 16(×3), 34(×10), 51(×10))



IN THE FIRST-FIT PACKING ALGORITHM weights are packed into bins $B_1, B_2, \ldots$ in the order in which the weights appear in their priority list. (If there is no priority list, the weights are arbitrarily arranged into one.) Each weight is placed in the first bin into which it fits. In other words, the weight is packed in the bin B_i with the smallest index i where the addition of the weight does not make the total of the weights in the bin exceed the fixed weight capacity. The first-fit packing of the list L shown at the top of the illustration is fairly efficient: $FF(L)$, the number of bins required, equals 17. A more efficient packing of L is shown at the bottom. This 10-bin packing is clearly optimum, since there is no wasted space in any of the bins. The example demonstrates worst possible performance of first-fit packing algorithm because it attains bound that has been established for algorithm: $FF(L) \leq (17/10)OPT(L)$, for any list L for which $OPT(L)$ is a multiple of 10. Numbers of multiple weights and bins are indicated in color.



FIRST-FIT DECREASING PACKING ALGORITHM improves on the first-fit packing algorithm by packing larger weights earlier. In the first-fit decreasing packing algorithm the list of weights to be packed is rearranged so that the weights are in decreasing order, and then the first-fit packing algorithm is applied to the altered list. In the first-fit decreasing packing of L (*top*) the number of bins required, $FFD(L) = 7$, is clearly optimum since each bin is filled to capacity. Like most scheduling algorithms, packing algorithms are subject to unpredictable behavior. When a weight is removed from L (*bottom*), additional bin is required for first-fit decreasing packing of smaller list.



IN TWO-DIMENSIONAL BIN-PACKING PROBLEM a list of planar regions, possibly of different sizes and shapes, must be placed without overlapping on a minimum number of identical regions. Placing sewing patterns on pieces of material is a common instance of this type of problem. Solutions to two-dimensional bin-packing problems are elusive, even when the shapes involved are highly regular. This fact is demonstrated by the following problem: How many squares with sides of unit length can be placed inside a larger square with sides of length α? If α is an integer, the problem is simple, but if α is not an integer (say α equals $N + 1/4$ for some integer N), the problem is more interesting. One obvious solution to the problem is to pack the $\alpha \times \alpha$ region by filling an $N \times N$ square with N^2 unit squares and sacrificing the uncovered area (*color*) of nearly $\alpha/2$ square units as unavoidable waste (*left*). After experimenting with other packings (*right*) it is tempting to conclude that no improvement can be made on the obvious packing, but surprisingly this is not the case. Paul Erdös, Hugh Montgomery and the author have recently proved that when α becomes large, there are packings for any $\alpha \times \alpha$ square that leave no more than $\alpha^{.634...}$ square units of uncovered area, significantly less than the $\alpha/2$ square units wasted in obvious packing. It has not yet been determined how small an area can be left uncovered when α becomes very large, although $\alpha^{.5}$ seems to be a likely possibility.

minimum number of program breaks; a paper manufacturer must furnish his customers with rolls of paper of different widths that he slices from a minimum number of standard rolls. In general bin-packing problems are extremely difficult to solve. At present the only known methods for producing optimum packings (those that require the minimum number of bins) involve examining essentially all the possible packings and then choosing the best. Like most scheduling problems, bin-packing is *NP*-complete, and so it is likely that any general algorithm for producing optimum packings will be similarly flawed. Therefore many bin-packing algorithms are designed to create packings that are reasonably close to the optimum.

In considering the bin-packing problem it is convenient to arbitrarily arrange the weights of the items into a list: $L = (w_1, w_2, \ldots, w_r)$. Since there are no precedence constraints, no confusion will arise from identifying an item with its weight, and L can be regarded as a list of the items to be packed. One obvious way to pack the weights of L is called the first-fit packing algorithm [*see illustration on page 130*]. Under the rules of this algorithm the weights are placed in bins in the order of their appearance in L: w_1 first, w_2 second and so on. When it is w_k's turn to be packed, it is put into the first bin in which it fits, that is, into the bin B_i, with the smallest index i that can accommodate the weight. (A weight w_k fits into a bin if the addition of w_k to the weights already in the bin does not make the total of the weights exceed W.)

How good a scheduling algorithm is first-fit packing? In other words, if $FF(L)$ denotes the number of bins required when the first-fit packing algorithm is applied to L and $OPT(L)$ denotes the number of bins required in an optimum packing of the weights of L, how much larger than $OPT(L)$ can $FF(L)$ be? In 1973 Jeffrey D. Ullman of Princeton University discovered that for any list L, $FF(L) \leq (17/10)OPT(L) + 2$. Ullman also showed that the coefficient 17/10 cannot be improved. If $OPT(L)$ is a multiple of 10, however, the constant 2 can be dropped from the inequality: $FF(L) \leq (17/10)OPT(L)$. It is conjectured that this simpler bound applies in all cases.

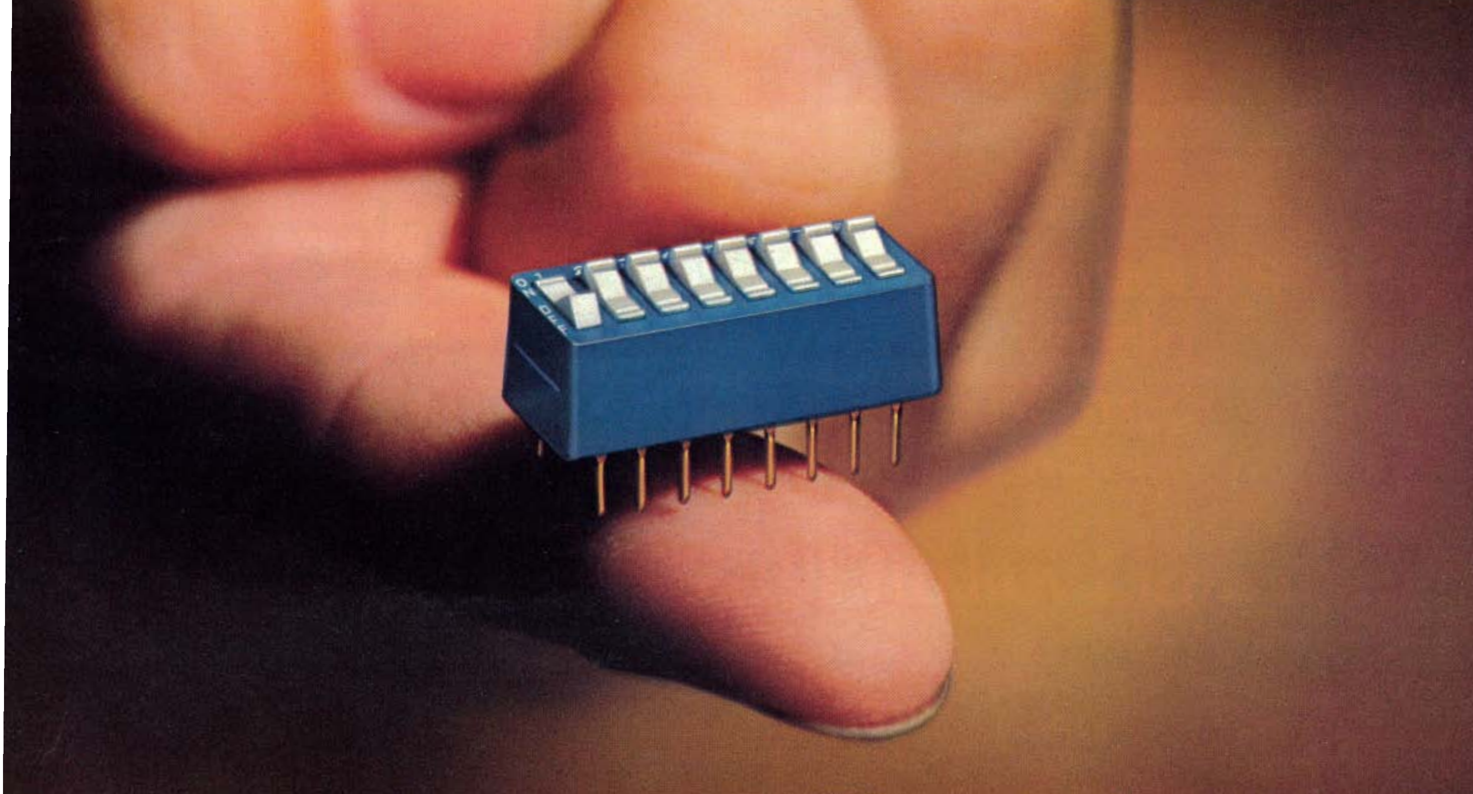
This bound shows that first-fit packing can perform rather poorly: it can require as much as 70 percent more than the optimum number of bins. Experimenting with the first-fit packing algorithm shows that the results are worse when large weights appear at the end of the list, requiring that new bins be started even though a great deal of space remains in partly filled bins. It makes sense to rearrange the list, putting all the large weights near the beginning so that the small weights at the end will be placed in odd gaps in nearly filled bins. This notion suggests a new packing procedure called the first-fit decreasing packing algorithm. The weights of L are arranged in a decreasing list and then the first-fit packing algorithm is applied. The new algorithm turns out to be quite effective [*see top illustration on preceding page*]. If $FFD(L)$ denotes the number of bins required for the first-fit decreasing packing of L, then it can be shown that $FFD(L) \leq (11/9)OPT(L) + 4$ for any list L. It has been shown that the coefficient 11/9 cannot be improved.

The expression $(11/9)OPT(L) + 4$ looks deceptively simple. Substantial difficulties are encountered in trying to prove that it is indeed the upper bound for $FFD(L)$. The only proof known at present is one devised by David S. Johnson of Bell Laboratories and it is more than 75 pages long.

When a large number of bins is required in a packing problem, the constant 4 in the inequality becomes relatively insignificant. In that case the first-fit decreasing packing algorithm is guaranteed to pack the weights of any list into no more than about 22 percent over the optimum number of bins. This result is certainly much better than the 70 percent increase over the optimum number of bins that can occur in the first-fit packing of a particularly unwieldy list.

Within their established bounds the first-fit and first-fit decreasing packing algorithms, like other scheduling procedures, are subject to unexpected behavior as the parameters of the model are varied. For example, removing a weight from a list for the first-fit decreasing packing algorithm can increase the number of bins needed. If the reduced list is denoted by L', it is still not known how large the ratio $FFD(L')/FFD(L)$ can be and whether $FFD(L)$ can increase when the largest element of L is removed. Once again it is the requirement that a weight be placed in the first available bin that is responsible for the unpredictable behavior of packing algorithms, but such behavior is not uncommon in the more complex scheduling situations of the real world.

The simple scheduling model I have described has provided a great deal of information about problems of realistic complexity. Many extensions of the basic scheduling model are possible. The model can be modified to allow interruption of tasks before completion or to allow unforced idleness. It can include the resources other than processors that are required for the execution of tasks, the probabilistic execution times of tasks, various measures of model performance and so on. By subjecting these extended models to the type of analysis I have discussed here investigators today are rapidly gaining insight into the difficult problems of scheduling.



How an 80¢ AMP switch cut manufacturing time and costs. And created a new technology.

Many companies don't call AMP until they need electrical connectors or switches. But why wait till then? If you involve us early enough, we can help you save money. A lot of money.

Here's an example:

Programming printed circuit boards used to involve installing an intricate network of jumper wires and plugs.

Until a customer asked us to find a better way.

We invented the DIP switch, which made programming p-c boards as easy as pushing number-coded actuators.

As a result, our customer was able to produce standardized boards that they could easily and quickly customize according to orders.

They saved hundreds of thousands of dollars in manufacturing and inventory costs. And a whole new technology was born.

Can we produce the same kind of savings for you?

If your company manufactures electrical/electronic products, the answer is probably yes. But it's important for you to call us in early.

That's when our experience in developing connectors, switches and application equipment for hundreds of industries worldwide can help you the most.

Early Involvement.

At AMP, we call this approach to solving our customers' problems "Early Involvement." It's our better way. And it's what makes an 80¢ AMP switch worth so much more than 80¢.

Ask for a copy of our brochure, "AMP Has a Better Way." Call (717) 564-0100, Ext. 8420.

Or write to AMP Incorporated, Harrisburg, PA 17105.

AMP
INCORPORATED

AMP has a better way.

Barcelona • Brussels • Buenos Aires • Frankfurt • Harrisburg • Helsinki • 's-Hertogenbosch • London • Luzern • Mexico City • Montreal
Paris • Puerto Rico • Sao Paulo • Stockholm • Sydney • Turin • Toronto • Tokyo

Pieter Bruegel the Elder as a Guide to 16th-Century Technology

The great Flemish artist had a deep interest in the scientific concepts and the machines of his time. As a result many of his works offer rare glimpses of practical knowledge 400 years ago

by H. Arthur Klein

Pieter Bruegel the Elder, the great Flemish artist of the 16th century, is noted among other things for the copiousness and clarity of detail in his paintings and in his designs for engravings on allegorical subjects. This characteristic of his work provides a rare glimpse of 16th-century science and technology. Clues of this kind can be found in certain paintings among some 50 surviving from Bruegel's hand, in his drawings, of which nearly 80 remain, and in several of 170 graphic works (copperplate engravings) following designs drawn by Bruegel. To find the clues, however, one must examine the pictures closely.

Two paintings done by Bruegel in 1563 show 16th-century techniques of large-scale construction. Both are titled *Tower of Babel*. The "small *Babel*" is today a principal treasure of the Museum Boymans–van Beuningen in Rotterdam. The "large *Babel*" is one of some 14 Bruegel masterpieces in the Kunsthistorisches Museum in Vienna.

Bruegel began work on this biblical theme at least a decade before he finished the paintings. The Book of Genesis relates that after the Flood, Noah's descendants in the land of Shinar set out to build a tower whose top would reach "unto heaven." God, however, ended this impious effort by making the builders speak so many different languages that they could no longer communicate with one another. They were scattered over the earth, and the tower remained unfinished.

To picture this account of the origin of humankind's many tongues Bruegel drew on at least three sources outside the Bible. The first is the Colosseum in Rome, which he saw during a visit to Italy as a young man. The second is the writings of Herodotus. Finally, Bruegel incorporated a number of the machines and methods employed in the main harbors of the Netherlands in his lifetime.

The Babel paintings boldly confront the problem of how such a lofty tower could have been planned. They depict it as being huge but incomplete. Its finished portions already loom above the clouds. Its internal structure is revealed as being a series of arches, closely akin to the Colosseum in design. That ancient amphitheater had been formed from an intricate interlocking complex of more than 1,000 stone arches. Of these 240 could be seen from the outside, arranged in three great elliptical tiers. Another 240 major arches were disposed radially around interior ellipses. The interior and exterior ellipses were linked together solidly by a system of radial arches, 480 in number.

The interior anatomy of the Colosseum was adapted by Bruegel to account for the sheer height and rigidity of his imaginary Babel. On the other hand, both Babel paintings show a broad spiral ramp winding around the rising tower, whereas the Colosseum had no such ramp. According to Herodotus, however, just such a ramp wound around the exterior of the great Babylonian ziggurat, or sacred tower, called Etemenanki. Herodotus could hardly have seen the ziggurat himself, but from inhabitants who said they had seen it he compiled a description of a "tower of solid masonry upon which was raised a second tower, and on that a third and so on up to eight." The way from base to summit, he related, took the form of a spiral ramp that "winds around all the towers."

The Babel paintings by Bruegel show (the one in Rotterdam quite clearly, the one in Vienna less so) such a great ramp spiraling around the massive tower. The fact that it will have made eight turns when the indicated construction is completed suggests the link between Herodotus and Bruegel.

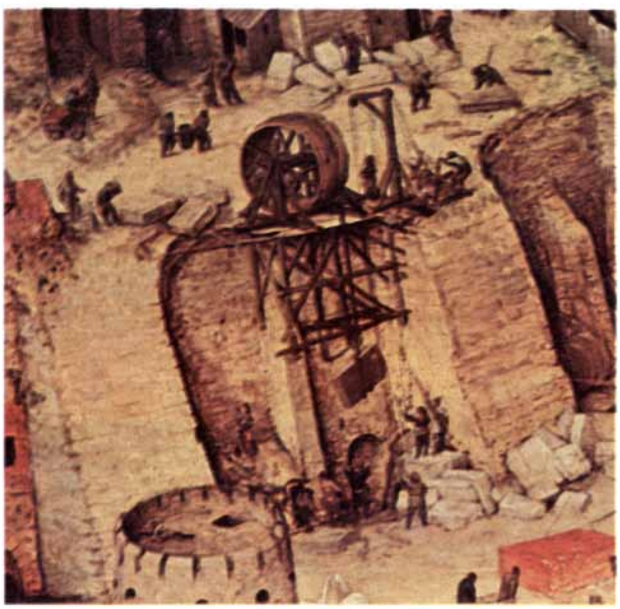
The most significant aspect of the paintings in the context of this discussion is the extent to which Bruegel has interwoven techniques and devices from his Flemish-Dutch environment. Both giant towers stand beside an active harbor, thereby answering the question of how all the building materials could have been assembled. The harbors lie beside some great body of water or estuary, as did the active and expanding harbor of Antwerp, the city in which Bruegel lived for a number of years.

On the waterfront in the small *Babel* one sees, amid the welter of goods and materials, the hulking mass of a great machine that slants upward and outward beyond its side wheel, looking at first like some monstrous ancient siege gun. Closer examination reveals it to be a huge wood crane or hoist. The great wheel at its side is a rotary treadmill, a kind of squirrel-cage arrangement in which human laborers furnished the power for operating the crane. In this way Bruegel indicated how the stones of the lofty tower were lifted into position.

The authenticity of the apparatus as a 16th-century lifting device is supported in a painting by another artist of the period, Pieter Pourbus the Elder. The painting is a portrait of Jan Fernaguut, a merchant in Bruges. Over the shoulder of the merchant, as if through an open window, one sees a section of the loading dock whose commerce helped to enrich him. Rising up in front of the buildings on the waterfront is the same kind of slanting wood structure, shaped like some strange pecking bird. The large wood wheel at its side, apparently almost 20 feet in diameter, is partly covered by a slanting roof, which would enable the human treadmill workers to labor on in bad weather.

In the large *Babel* Bruegel placed a massive man-powered crane of this type on the second level of the spiral ramp. The crane has a treadmill wheel on each side. Close examination of the tiny details reveals that at least six and possibly

"TOWER OF BABEL," painted by Pieter Bruegel the Elder in oil on an oak panel, depicts the Babel story told in the Book of Genesis and also shows a number of 16th-century construction techniques. They include the man-powered lifting devices on the right side of the tower at the first and second levels of the ramp. This painting, which is in the Kunsthistorisches Museum in Vienna, is called the "large *Babel*" to distinguish it from a similar painting, the "small *Babel*," Bruegel did at about the same time (1563). The small *Babel* is now in Rotterdam.



DETAILS OF "BABEL" PAINTING include a drumlike lifting device (*left*) that is on the first level of the ramp, below the crane portrayed on the cover of this issue, and two arches (*right*) that are under construction near the center of the painting. The lifting device, like the crane, was powered by men in a treadmill and also relied on a block-and-tackle system. It had less lifting capacity than the crane.

eight laborers are working within the wheels. In the painting the crane is lifting a large slab of stone.

Could such cranes have raised all the massive stones of the tower? A healthy adult male laborer is able to deliver not much more than .1 horsepower, or roughly 75 watts, over an extended period of time such as several hours. In shorter bursts, however, he might produce as much as .3 horsepower, or 225 watts. If six workers were giving their best within a treadmill, they might deliver from 1.5 to two horsepower during a limited period of time, say a few minutes followed by a rest period.

Even if 20 percent is deducted to allow for friction in the bearings and the block-and-tackle rigging of the cranes, a net of from 1.2 to 1.6 horsepower remains. Since one horsepower is 550 foot-pounds per second, 1.2 horsepower is 660. To lift a one-ton stone slab 50 feet requires 100,000 foot-pounds of work, or energy. That amount could be supplied within about 151.5 seconds, or two and a half minutes, under the conditions I have described.

At least half a dozen man-powered lifting machines are shown rather clearly in the two Babel masterpieces; a few others are suggested less distinctly. One of the clearer devices stands one ramp below the crane. It is merely a big drum; ropes wind around the drum as men turn it. Aided by block-and-tackle rigging, the device could handle objects somewhat lighter than those with which the other cranes are designed to deal.

In 1556 Bruegel began the series of designs now widely known as Bruegel's Seven Deadly Sins. *Avarice* was engraved and published in 1556. In 1557 engravings of his designs for *Pride, Envy, Anger, Gluttony, Lechery* and *Sloth* were published. The great success of the Sins led inevitably to a series of seven Virtues. These far more sober and somber designs were drawn by Bruegel in 1559 and 1560. The virtues depicted in his designs are faith, hope, charity, justice, prudence, fortitude and temperance.

The *Avarice* engraving discloses what might be called the curious case of the crossbow arrows. In one part of the engraving several sinners are displaying their avarice by shooting with crossbows at a huge purse suspended tormentingly above them. The arrows ascend uncurvingly and then turn and drop straight down.

Bruegel, an artist exceptionally faithful in all physical and natural details, has here indicated trajectories varying from the paths that were later demonstrated by ballistic science. The flight paths of Bruegel's arrows do agree, however, with the conceptions (misconceptions, rather) that prevailed in his day. Bruegel was drawing his representation of the subject long before Galileo's discovery and publication of the fact that projected cannonballs, stones and arrows follow parabolic trajectories, at least to the extent allowed by air drag.

In the cycle on Sins, Bruegel approached not only problems of motion but also aspects of time. The subject appears specifically in the engraving *Sloth*. There one sees Queen Sloth lying in sodden stupor, surrounded by snails, slugs and other symbols suited to the theme.

A gigantic clockwork hangs in a tree at the upper right. A human or humanoid figure in front of the clock is hammering on a bell (an alarm reminding us and the slothful sinners that it is later than we think). Several large clockwork gears can be seen in this strange timepiece. Visible also are the twin weights that supplied power for such clocks in Bruegel's time. Another essential component, which appears just under the foot of the bell ringer, is a partly hidden iron bar from which hangs a metal weight that can be shifted to any of several positions by means of notches in the bar. In spite of its impossible position in the tree this device can be recognized as a foliot, an oscillating crossbar that controlled the verge escapement mechanism of the clock. Such a crossbar, mounted on a vertical axis, held two movable weights, one weight on each side of the axis. Shifting the weights closer to the axis reduced the period of oscillation, thus tending to make the clock run faster. Moving them out tended to make the clock run slower.

"THE PARABLE OF THE BLIND," painted by Bruegel in 1568, portrays a New Testament story. It also reveals Bruegel's attention to detail, since physicians have inferred five different forms of blindness from his depiction of the five faces that are clearly visible here.

"Tended to" is the appropriate phrase here, since the rate of the clock's motion was not determined solely by either the position or the size of the weights. The axle of the bar carried metal pallets, or projections, that alternately halted and then pushed forward a notched wheel, which was customarily called the crown wheel. Bruegel shows it as a wheel with hooklike projections among the clock parts in the tree. The rate at which a clock ran was heavily influenced by the shape of the hooks on the crown wheel, the form of the pallets and the size of the weights. A typical clock based on such mechanisms might gain or lose as much as half an hour in 24 hours. (The advance represented by the pendulum was still to come when Bruegel designed the engraving.)

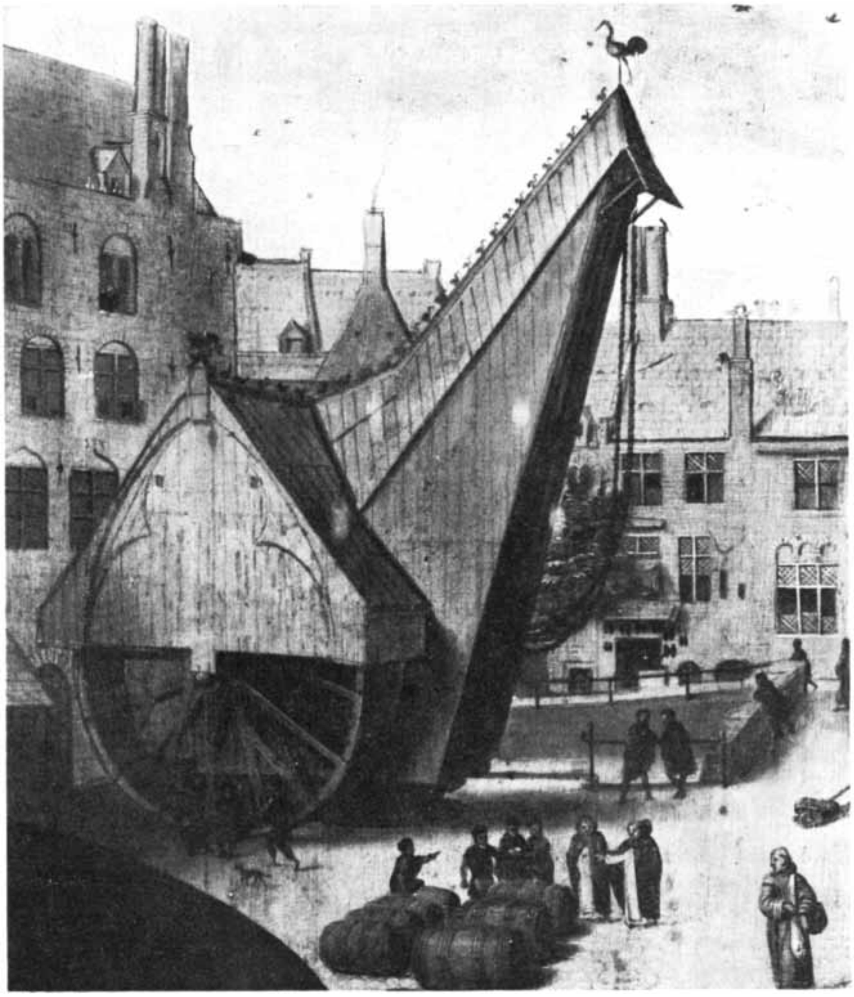
The crudeness of the timepieces is emphasized on the monstrous symbolic clock dial at the upper left in the engraving. It has only one hand (literally a hand, which is at the end of a human arm). The clocks Bruegel knew best had only an hour hand. Clocks with two hands had been made by that time, but they were rare and costly. A clock with a minute hand was of limited use anyway when a timepiece could be expected to run quite fast or quite slow during a 24-hour period.

Other symbols that combine mechanical and human characteristics appear in the engraving of Bruegel's design for *Gluttony*. A huge, horrible head is also a windmill. Into its distended jaws minuscule laborers haul sacks of grain and other materials to be ground (or masticated and swallowed). The compulsion to devour mechanically, even without need, is manifest.

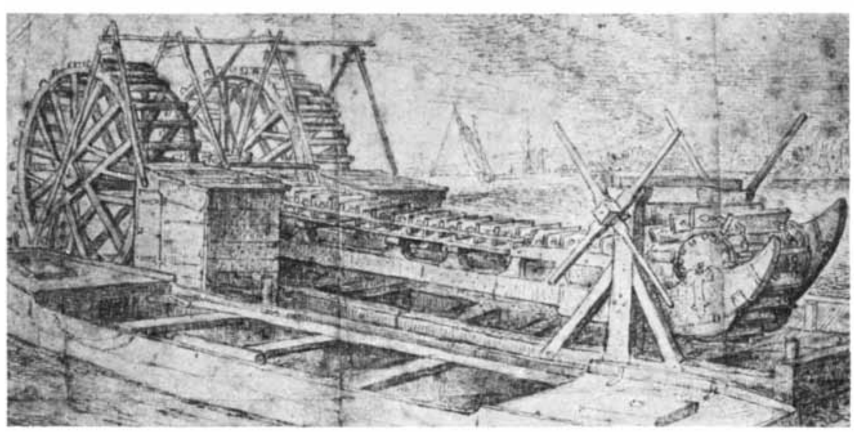
In the engraving *Lechery* one finds a relatively sophisticated technical device, an ornate fountain with twin jets. It feeds a pool in which nude couples wade and embrace. The stream meandering through this landscape of lechery originates beyond the fountain and turns an undershot waterwheel, around which slaves of lust can be seen.

In several of Bruegel's designs demons are waging symbolic wars. In *Fortitude*, a design in the Virtue series, hosts of slaughtering soldiers are in action, both mounted and on foot. Two large wood battle shields can be seen. One of these early forms of mobile armor carries a centrally mounted hinged hammer, which is poised for crashing down on approaching attackers. Near the center of the design a similar war machine has been overturned; its hammer hangs useless, and it is about to be overrun by approaching cavalry.

The engraving (probably by Philipp Galle) of Bruegel's design for *Temperance* seems to reveal most fully the artist's appreciation of the technology, sci-



MAN-POWERED CRANE painted in the 16th century by the Flemish artist Pieter Pourbus the Elder (1523–1584) shows in somewhat more detail the kind of lifting device depicted by Bruegel in the large *Babel*. The painting occupies a corner of a portrait of Jan Fernaguut of Bruges and is intended to show an aspect of the longshore trade that made him wealthy. The dark curving line at the lower left is Fernaguut's shoulder. The treadmill of the crane (presumably there was another treadmill on the other side) is covered by a rooflike structure to shelter the men who provided the muscle power for operating the crane. One also sees at the base of the crane a pivot whereby the structure could be swung toward a ship in the adjacent waterway.



MAN-POWERED DREDGER, depicted by the Flemish artist Roeland Savery (1576–1639), indicates the accuracy of Bruegel's portrayal of man-powered lifting devices of the 16th century. The two treadmill wheels at the left operated a chain-linked belt of scoops that were lowered into the water. Drawing is now in the Devonshire Collection at Chatsworth in England.

ence and art of the time. Temperantia, a tall and queenly figure, wears on her head as a crown a clock, the most sophisticated mechanical device of the era. Her right foot presses the blade of a windmill, which was the most advanced power source. In her left hand she holds a pair of spectacles, representing a major optical device in a time when the telescope and the compound microscope were still in the future.

Around Temperantia are groups engaged in activities that one can assume were esteemed by the artist and the patrons who bought his prints. Reading is being taught in a school at the bottom right. Reckoning, bookkeeping and accounting are under way at the bottom left. Behind that scene an artist works on a painting. Higher still are performances of music by instrumental and choral groups. Drama is performed (with a prompter) on a stage at the upper left. Five earnest men, probably theologians, engage in a vigorous discussion at the right center, above the school. They represent rhetoric, the art of effective presentation and persuasion through language.

Symbols of ballistics and the study of trajectories are found at the far right. A crossbowman aims at a target atop a mast. So does the user of an early long musket. Beyond them project the muzzles of two large, wheeled cannons; piles of cannonballs are nearby.

On a freestanding pillar two men make measurements. The man on the top of the pillar lowers a plumb line to ensure that the pillar is vertical. The man on the plank seat at the middle of the column employs dividers to ascertain distances. Far beyond the column, in an open field, the measurement of land areas seems to be in progress. Somewhat to the left of the column a man sights along what looks like a pole with a wheel at its far end. The device provided a simple but quite accurate means of measuring angles by sight. Behind this man is an engraver-printer, using a square to verify angles on a picture.

Beyond Temperantia's head is a globe, with the continents and oceans vaguely suggested. To its right stands the geographer-cartographer, who measures off distances on the surface with dividers. On top of the globe stands an astronomer, who wields his dividers as if to find the distance to the moon.

Seeking a suitable frontispiece for my book *The World of Measurements,* I

SIN OF SLOTH was drawn by Bruegel in 1557 as part of a series of designs for engravings on the seven deadly sins. Several of the components of a clock appear at the upper right, including the foliot under the foot of the horizontal bell ringer. At the upper left is the face of a one-handed (literally one-armed) clock. Most of the clocks of the 16th century had only an hour hand, since the clocks were too inaccurate for a minute hand to have been useful. Near the center of the engraving is an undershot waterwheel. Presumably Bruegel employed these devices to symbolize that time is wasting for the slothful. At the bottom right the publisher, H. Cock, identified himself. Cock, who was an important dealer in graphics, was proprietor of Four Winds art shop in Antwerp, where Bruegel lived for a number of years.

found nothing more appropriate than segments of the print of *Temperance.* Three segments serve the purpose, epitomizing the measurements, units and concepts by which science has advanced.

Bruegel also showed in several works that he was attentive to the principal ailments afflicting his fellow men and to the methods whereby physicians or quacks treated either the diseases or the psyches of the sick. An engraving made in 1559 of his design for *The Witch of Mallaghem* betokens the mass craziness of gullible people. Working in a chaotic crowd of distraught Flemings, the "healer" is removing stonelike tumors from the heads of the afflicted. Bruegel also probed popular gullibility and hysteria in a picture titled *Pilgrimage of the Epileptics to the Church at Molenbeek* (a Flemish town). The work survives today in the form of two engravings made after his death and a drawing (probably also a copy made after his death) in the Albertina collection in Vienna. In the engravings women who appear to be suffering more from hysteria than from epilepsy are being half-supported, half-carried by attending men. They are being taken to a bridge where they are to dance to bagpipe music until their seizures have been sweated out or worn out.

Two other paintings by Bruegel have notable medical implications. *The Beggars,* painted in 1568, shows four crippled and distorted figures propped up by crutches and sticks and strangely dressed in garments trimmed with foxtails. Some commentators have suggested that the men exhibit various symptoms and stages of leprosy.

In *The Parable of the Blind* Bruegel depicts six blind men who are leading one another for reasons of safety. The first man in the group has tumbled into a ditch and the others are certain to follow. Physicians have inferred a different kind of blindness in each of the five men whose faces can be seen. From the left they are blindness resulting from the disease known as pemphigus, atrophy of the eyeball, corneal leucoma, amaurosis and enucleation.

In 1565 a canal between Antwerp and Brussels was completed. The city fathers of Brussels chose Bruegel to portray in a series of pictures the construction of that canal. The pictures were never finished, because Bruegel died in

VIRTUE OF TEMPERANCE was portrayed by Bruegel in a design for this engraving. The drawing includes a number of representations of science and technology. Astronomy and measurement of the earth are symbolized at the top center. On the column at the right one man drops a plumb line and another calculates distances. Other types of land measurement go on in the region behind the column. The device consisting of a wheel on a rod provided a means of measuring angles. Temperantia is associated with a number of symbols of technology, including a clock, a pair of eyeglasses and the blade of a windmill. Reckoning and reading are symbolized below her. The favorable public reception of the engravings on the seven sins led Bruegel to execute designs for seven drawings on virtue. He did them in 1559 and 1560.

1569, aged not more than 44. (The date of his birth is uncertain but is usually placed between 1525 and 1530.) The choice of Bruegel for the commission suggests that the burghers of Brussels, proud of the canal and of the technology represented in the busy harbor of Antwerp, valued his faithful portrayals of people and the artifacts of their world. This ability makes him today a rewarding guide not only to important technologies of his time but also to significant scientific concepts (insights as well as errors) of that distant era.

SIN OF LUST was another of Bruegel's designs for engravings on the seven deadly sins. This detail from the engraving of the design shows a pool fed by a twin-jet fountain. It is one of several places in the drawing where lustfulness or lechery (another name sometimes given to the engraving) is evident. Presumably the water was raised to the jets by a waterwheel that received its power from the stream. A waterwheel that is apparently associated with a mill is also visible.

CONCEPT OF FORTITUDE, one of the subjects depicted by Bruegel in his series of designs for engravings on seven virtues, was conveyed by putting the winged symbol of fortitude in the midst of battles and slaughters. The plan gave Bruegel an opportunity to display a number of items of military hardware, including this mobile battle shield. The virtues depicted by Bruegel included faith, hope, charity, justice and prudence in addition to temperance and fortitude.

THE AMATEUR SCIENTIST

Visual illusions that can be achieved by putting a dark filter over one eye

by Jearl Walker

Under certain circumstances a pendulum swinging in a plane perpendicular to your line of sight can be made to appear to be moving in an ellipse around a central vertical axis. The effect was first explained in 1922 by the German physicist Carl Pulfrich and is known as the Pulfrich illusion. You can easily set up a demonstration of the illusion.

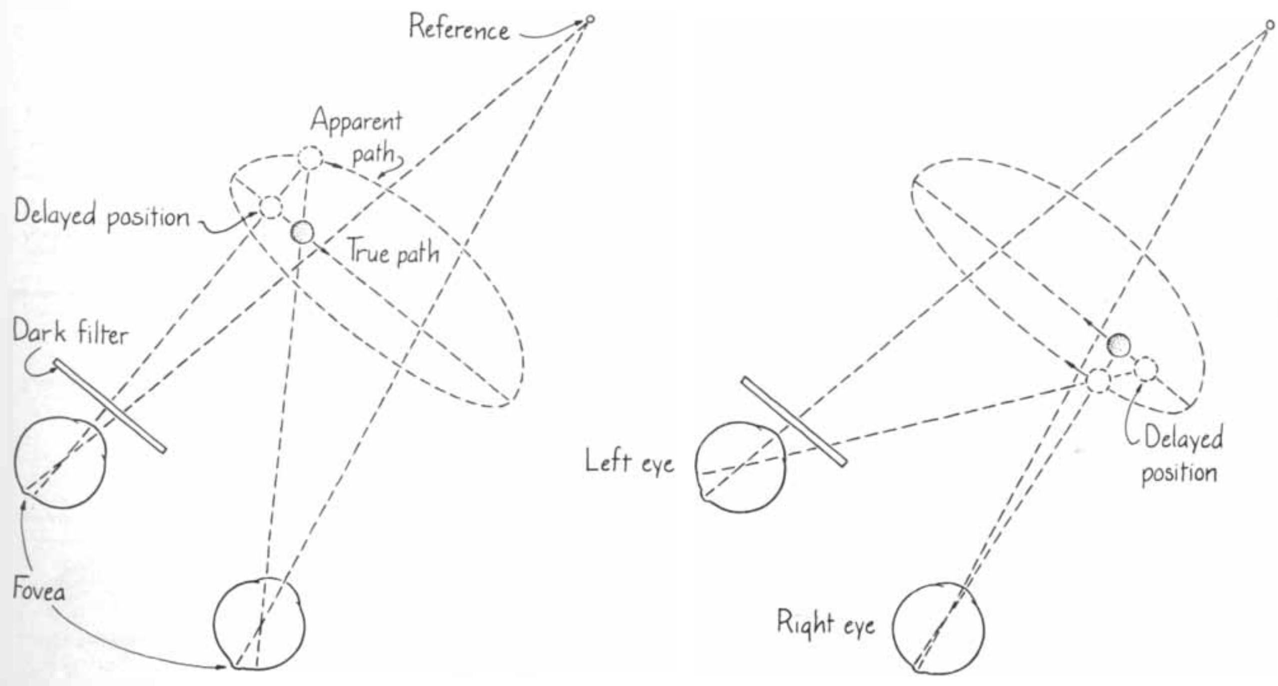
For the purpose you need three things: a simple pendulum consisting of a weight on the end of a string, a partially dark filter such as one glass from a pair of sunglasses and a thin vertical object. Set the pendulum swinging in a plane perpendicular to your line of sight and place the thin object behind it. Put the dark filter over one eye and look with both eyes at the background object. The pendulum will seem to be moving elliptically. With the filter over the left eye the rotation is clockwise from an overhead sense; with the filter over the right eye it is counterclockwise. Increasing the darkness of the filter increases the apparent depth of the elliptical path. Somehow the dark filter causes you to perceive a motion with depth when there is no depth.

Pulfrich explained the illusion after an associate suggested that the dark filter delays the image the eye affected by the filter sends to the brain. (Pulfrich himself was blind in one eye and was never able to see the illusion.) The suggestion turned out to be correct. When the illumination to one of your eyes is decreased, your perception of the scene from that eye is delayed by some tens of milliseconds from the perception you receive from the uncovered eye.

Suppose your left eye is covered by a dark filter and a pendulum is beginning to swing from left to right. Your eyes are fixed on a background object, with the result that its image falls on the fovea (the small depression in the retina that provides the sharpest vision). The images of the pendulum bob lie off center from the fovea, but because of the visual delay from the left eye the image of the bob in the left eye lies more off center. From that eye you perceive the bob where it was some tens of milliseconds earlier, hence to the left of its true position. The image on the retina is therefore more off center from the fovea than it would have been without a filter. What you apparently sense is the angle between the bob and the background reference object. The angle for the left eye is greater than it is for the right eye, and to make sense of that difference you interpret the bob as being farther away from you than it actually is. The visual delay therefore gives the impression of depth in the motion of the bob.

As the bob speeds up in approaching the center of its swing the discrepancy in the angles between the reference object and the bob increases and the apparent depth of the bob increases. As the bob slows down in reaching the right-hand extreme of its swing the discrepancy in the angles decreases and the bob seems to move closer to you. During the swing back the visual delay from the left eye makes the bob appear closer to you than it really is. The "closest" point is near the center of the pendulum's swing, where the speed of the bob is greatest. The overall motion of the bob is an apparent rotation in an elliptical path around a vertical axis.

Increasing the darkness of the filter (its optical density) increases the visual delay from the filtered eye and therefore the depth of the elliptical motion. You can most easily control the darkness by using two crossed polarizing filters over



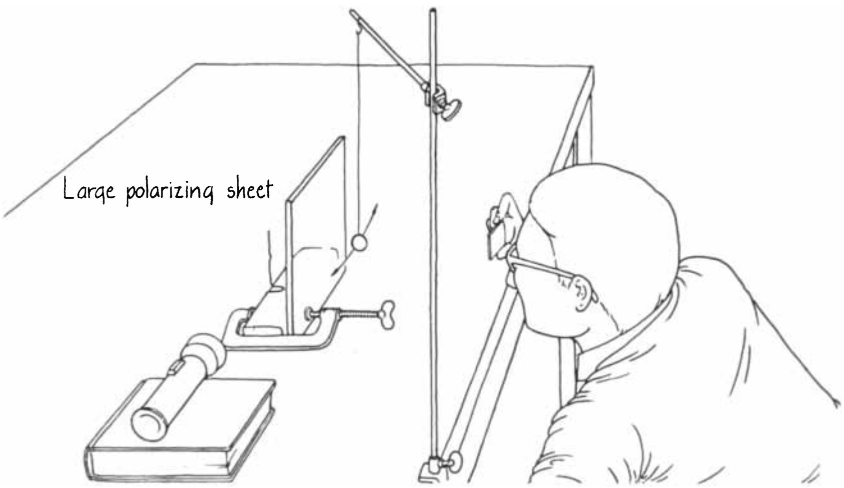
Apparent motion of a pendulum bob in the Pulfrich illusion one eye. By altering the angle between the senses of polarization of the filters you can select any degree of darkness. With enough darkness the path becomes nearly circular. Finally the filter becomes so dark that the illusion is lost because you see the bob only through the unfiltered eye. Then the perceived motion is the actual linear motion.

The physical processes resulting in the delay in visual signals to the brain are not completely understood. The visual latency is caused either by the decreased intensity of the light reaching the retina or by a change in the adaptation of the eye. Perhaps both factors are involved. (The eye adapts to changes in illumination by varying the size of the pupil and by producing more or less of the pigment involved in the visual process.) One model of the latency regards the visual pathway as being a series of four identical electrical delay lines, each stage consisting of a resistor, a capacitor and an amplifying unit. Albert M. Prestrude of the Virginia Polytechnic Institute and State University suggested that the four stages may correspond to the layers of horizontal, bipolar, amacrine and ganglion cells in the visual pathway, but this correlation is not certain.

In the analogy the sensitivity of the visual process (that is, how small an input signal can produce a sufficient output signal) depends on the values of the resistors and the strength of the amplifying units. The values of the resistors are not fixed, however, but are controlled by a feedback signal from the final output of the visual system. The response time of the system also depends on the resistor values: the greater the resistance, the longer the response time. The combined effect is that when the stimulus level to the eye is decreased, the feedback signal from the output of the visual process increases the resistance and the amplifying gain in each stage in order to increase the sensitivity. The increased resistance in turn means the system takes longer to respond. Therefore decreasing the intensity to the eye results in greater sensitivity and a longer response time.

Since the work of Pulfrich the illusion has been studied in a variety of settings. Some investigators employed two rotating disks, one for each eye. By looking through a system of prisms and lenses an observer could fuse the images of the two disks to perceive a single rotating disk.

Each disk bore a radial line. If the disks rotated synchronously, the observer saw a single radial line. Placing a dark filter over one eye, however, caused the synchronization of the lines to be destroyed. By measuring how much the phase between the two disks then had to be altered in order to bring the radial lines back to an apparently synchronized rotation the experimenter could determine the delay introduced by the filter.



Arrangement for creating a pendulum illusion

A particularly interesting technique was employed by Bela Julesz of Bell Laboratories and Benjamin White of San Francisco State University. They showed an observer motion pictures of computer-generated random dots in which each frame was statistically independent of the others. Through separate viewing systems each eye saw a motion picture. In the center of each frame were randomly selected dots that formed part of a square.

If only one eye saw the presentation, the sensation of a central square was lost in the apparent random display of dots. If both eyes saw their respective motion pictures simultaneously, the central square was visible and stood out from the background of other random dots. The stereoscopic sensation was due to the slight shift of the left-eye dots in relation to the dots seen by the right eye. The same technique is employed in conventional stereoscopic pictures, in which one eye sees a scene slightly shifted from the other eye's view, creating an illusion of depth.

In the experiment of Julesz and White one eye was shown its motion picture one frame later than the other eye. The central square was therefore lost because the selection of dots from frame to frame was randomly different. An adjustable filter was then placed over the eye receiving the earlier frame. The observer adjusted the darkness of the filter until the delay in perception of the frame by that eye was sufficient for the brain to receive the perceptions of the correlated left and right frames simultaneously. The central square became visible again.

In the pendulum illusion the apparent depth of the pendulum's motion can be produced by adaptation of your eyes as well as by a dark filter. Shine a dim light into one eye for a few minutes and then watch the swinging pendulum. The illusion of depth appears, just as if the other eye had been looking through a filter. The illuminated eye was adapted for brighter conditions and thus responded faster than usual.

Some workers have reported that the illusion also appears if you adapt one eye to darkness. Black out one eye for 15 or 20 minutes and then look at the pendulum with both eyes. You may find a small Pulfrich effect. (Some people do, although I do not.)

Still another example of how adaptation enters into the illusion can be seen by carefully watching the depth of the pendulum's elliptical motion as a function of time after placing a dark filter over one eye. There should be an immediate visual latency and hence depth to the motion. Both should increase over the next 20 minutes as the covered eye adapts to decreased illumination.

The elliptical motion of the pendulum in the usual Pulfrich illusion disappears if you stare at the bob instead of the background. Apparently in tracking the bob the eyes follow it in its true linear path, maintaining the images of the bob on the same areas of the retinas, eliminating the changing angles between the reference and the bob and thereby destroying the illusion of rotation. The illusion can be regained while you track the bob if you darken the background for one of your eyes. This effect has been demonstrated fairly recently by several workers using oscilloscope traces instead of pendulum bobs. The effect can also be seen with the simple pendulum.

Place a large sheet of polarizing filter behind the pendulum and a small light and the usual reference object behind the filter. (Large polarizing sheets are available from the Edmund Scientific Company, 7778 Edscorp Building, Barrington, N.J. 08007.) Over one eye place a smaller piece of polarizing filter oriented to somewhat darken the background for that eye. The polarizing filter behind the pendulum must be large

HP measurement and computer advances

New microprocessor-based quartz thermometer makes absolute accuracy to ±0.04°C (and 0.0001°C resolution) simple and affordable.

The HP 2804 digital thermometer achieves exceptional accuracy because HP calibrates each quartz temperature sensor, and stores the individual data thus derived in an electronically readable calibration module. With this module and the sensor inserted in the thermometer, calibration tables are effectively eliminated, along with manual corrections and intermediate calculations.

Temperature measurement with the HP 2804 is based on the quartz sensor's frequency output, which changes with temperature. Every quartz sensor has a unique and stable temperature-vs-frequency response. HP individually calibrates each sensor at approximately 40 different temperatures, then fits these data to a curve using a high-precision regression technique. The coefficients of each sensor are recorded into a PROM which is mounted in a calibration module shipped with each sensor. In operation, the module is inserted into the HP 2804, where its data are read and used to compute temperature from measured sensor frequency.

The HP 2804 can measure temperature with an absolute accuracy of ±0.04°C (traceable to NBS) over a range of −50°C to +150°C (or ±0.075°C over the instrument's full range of −80°C to +250°C).[1] Quartz sensors are inherently stable and can make repeatable measurements over long periods with extremely low drift. A usable resolution of 0.0001°C makes the HP 2804 an excellent choice for measuring minute temperature differences.

The HP 2804 accepts inputs from one or two sensors and can read and display temperatures from either, or can make differential measurements from both inputs. Resolution can be displayed as 0.01°, 0.001°, or 0.0001°C. Temperature can be displayed in Celsius or Fahrenheit. With optional HP-IB interface, the thermometer can easily be connected to computers, controllers, printers, or other digital peripherals. An optional analog output is provided for easy connection to strip chart recorders. A built-in self check prevents incorrect measurements.

The price of the HP 2804 starts at $3550*, including an HP 18111 sensor.

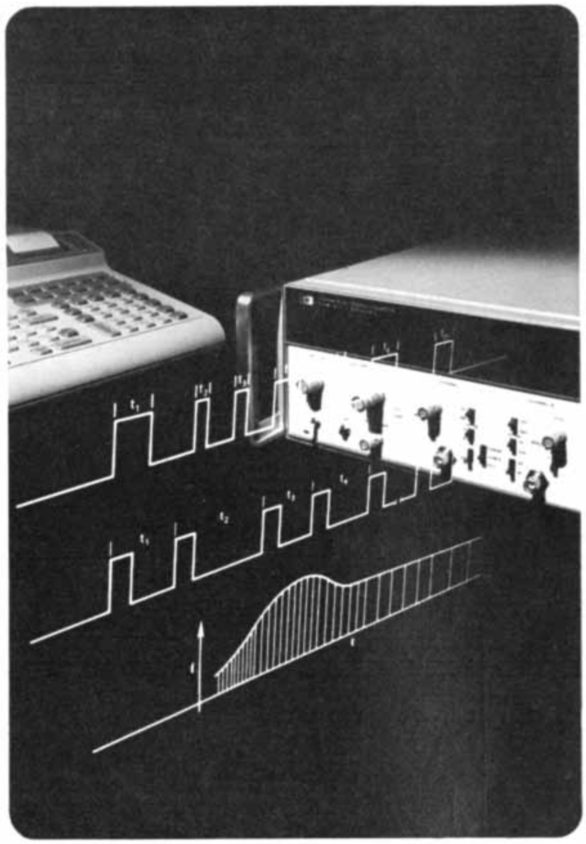
---

[1] To achieve specified absolute accuracy a reference temperature adjustment must have been made within 30 days, and the thermometer must be at the same ambient temperature during the measurement as it was when the reference temperature adjustment was made.

extend your possibilities.

If your "clocks" tick in picoseconds, we can help in some significant new ways.

Three new HP instruments have advanced the state of the art in time measurement to an unprecedented level: You can now measure time intervals with 20-picosecond single-shot resolution; generate time delays synchronized precisely to an external pulse, in steps as small as 50 ps, with less than 200 ps rms jitter; and measure successive pulses at rates up to 50,000 per second.



A new universal time interval counter, the HP 5370 employs a phase-lockable-phase-startable oscillator and digital vernier technique to all but eliminate the ±1 count quantization error that has helped to defeat picosecond single-shot resolution in previous counters. The microprocessor-based instrument thus achieves a number of important new functional advantages:

- 20-ps resolution of single events
- 1-ns accuracy
- positive or negative interval measurements from −10 to +10 seconds
- automatic characterization of pulse jitter with display of mean, max, min, and standard deviation
- pushbutton cancellation of systematic errors
- simultaneous display of trigger levels on start and stop channels
- frequency and period measurements to 100 MHz with 11-digit measurement displayed in 1 second
- fully programmable through HP-IB interface.

A new digital time synthesizer, the HP 5359 is 20 times more precise and up to 10,000 times more stable than analog pulse generators. A fit companion to the HP 5370, the HP 5359 uses the same phase-lockable-phase-startable oscillator technique to start in phase with an external pulse and thus avoid the ±1 count clock error that defeats the precision of other digital pulse generators. As a result, the HP 5359 can boast of these accomplishments:

- adjustable delays of 0-160 ms, precise to 50 ps
- jitter less than 200 ps rms
- digital selection of delay and output pulse width in steps as small as 50 ps
- pulse train output with digital selection of frequency and period to 10 MHz
- display of amplitude and offset in pos or neg, true or complement
- automatic calibration in 3 sec
- fully programmable through HP-IB interface.

If the need to make rapid measurements has you bogged down, the new HP 5391 data acquisition system for time and frequency measurements can help. It can measure successive pulse widths or periods with a resolution of 2 ns at rates up to 50,000 per sec, completely automatically—six times faster and at lower cost than is possible with a high-speed counter and computer.

The HP 5370 and 5359 are both priced at $6500*; the HP 5391 with 8K bytes of memory is $17,500*.

HEWLETT hp PACKARD

1502 Page Mill Road, Palo Alto, California 94304

For assistance call: Washington (301) 948-6370, Chicago (312) 255-9800, Atlanta (404) 955-1500, Los Angeles (213) 877-1282.

Mail to: Hewlett-Packard, 1502 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) HP 2804 quartz thermometer
( ) HP 5370, ( ) HP 5359, and ( ) HP 5391 timing products

Name ____________

Company ____________

Address ____________

City ____________ State ______ Zip ______

*Domestic U.S. prices only.

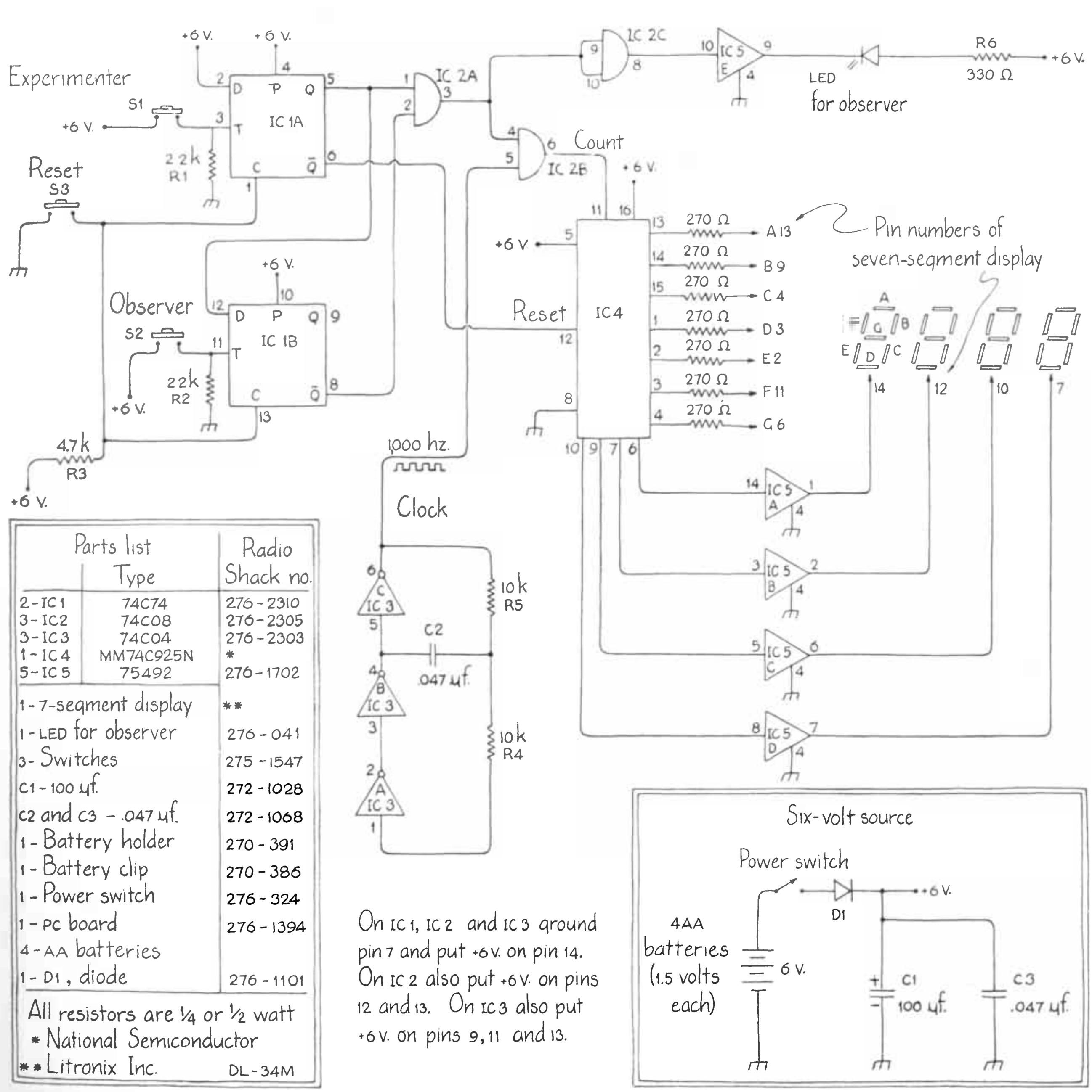
00843

enough so that when the pendulum swings, the filter remains as a background. Now track the bob. If the left eye sees a darker background than the right eye, the bob appears to rotate clockwise as before in spite of the tracking. Darkening the background by rotating the polarizing filter over the eye predictably increases the depth of the bob's elliptical path; putting the small filter over the other eye yields the expected counterclockwise motion.

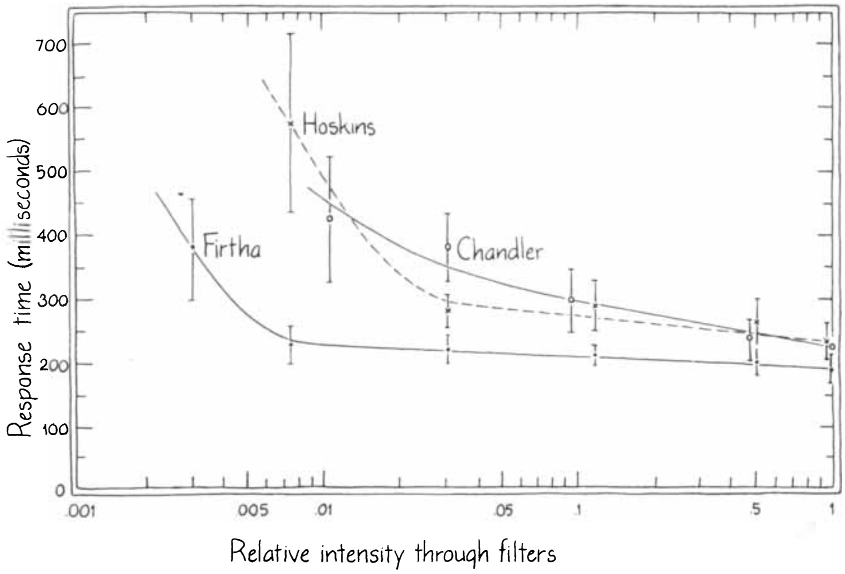
As the images of the large polarizing filter and the reference object move across the retinas the increased visual latency of the filtered eye creates the illusion that the distance between the bob and the reference object is constantly changing. Much of the changing depth is attributed to the bob, perhaps because a change in the distance between you and the bob is easier to believe than a change in the distance between you and the background and the reference object, both of which are known to be stationary. As a result the bob appears to move in an elliptical path even if it is the bob you are tracking.

The experiment with the Pulfrich pendulum illusion is a difficult way of measuring the visual latency produced by a given dark filter. One reason other workers designed different experiments was the aim of measuring the latency more precisely. In some experiments of my own I chose to measure the response time by a relatively simple arrangement in which an observer would turn off a light I turned on. The time was measured in milliseconds. The circuit that controlled the light and the digital clock for the experiment was designed and built for me by Harry Freedman of Cleveland State University.

The integrated circuit chips labeled *IC* (for integrated circuit) 1*A* and 1*B* in the illustration below were "*D*" flip-flop circuits. Initially the *Q* output from *IC* 1*A* would produce a logical-0 signal and the $\bar{Q}$ a logical-1 signal. (In a logical-0 signal no voltage is produced; a logical-1 signal is one in which a voltage greater than zero is created.) When I tripped the switch *S*1, the outputs from *Q* and $\bar{Q}$ of *IC* 1*A* changed states; the *Q* then produced a logical-1 signal. This output state remained until the run was over



Circuitry of apparatus for measuring visual latency



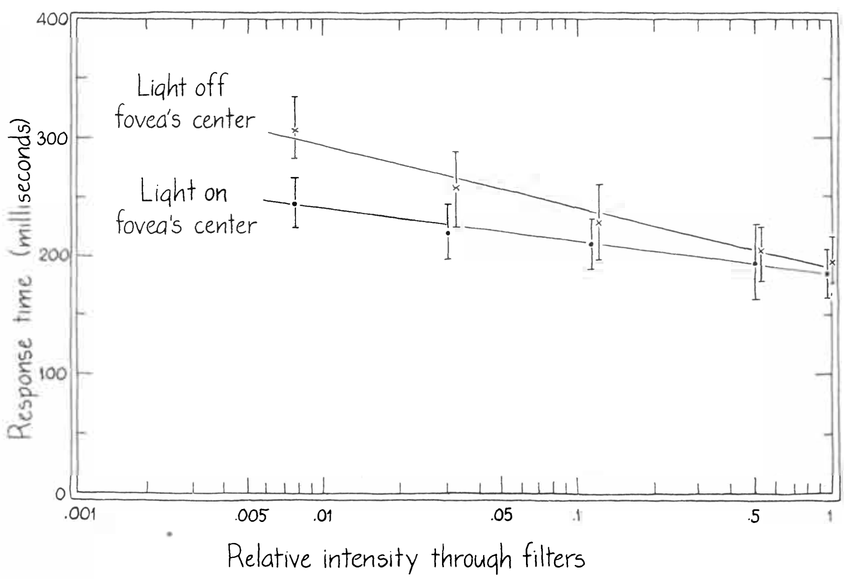
Response time of three observers as intensity of light was increased

and I tripped the reset switch $S3$, restoring Q and Q to their initial states.

On tripping $S1$ the logical-1 signal from the Q of *IC* 1*A* went to *IC* 2*A*, which was an "and" gate. Such a gate sends out a logical-1 signal only if both inputs to it are of the same logical states. In this case *IC* 1*B* was already sending a logical-1 signal to the gate; when I tripped my switch, the gate had two such signals coming in and therefore sent a logical-1 signal to *IC* 2*C* and *IC* 5*E* to turn on the light-emitting diode for the observer to see.

Another logical-1 signal was sent from the gate *IC* 2*A* to another "and" gate (*IC* 2*B*), which compared it with the signal coming from an oscillating circuit functioning as a clock. At a rate of 1,000 times per second the clock generated a logical-1 signal for *IC* 2*B*. Thus *IC* 2*B* found a match 1,000 times per second in its inputs, and so it sent at that rate a logical-1 signal of its own to *IC* 4. The *IC* 4 was a four-digit counter with seven segments of decoded outputs that were used to drive a four-digit, seven-segment display indicating the observer's response time in milliseconds.

The *IC* 4 continues to receive signals at the rate of 1,000 per second, and the display continues to flash the count until the observer trips switch $S2$. The output from Q of *IC* 1*B* is then flipped to be a logical-0 signal, and the gate *IC* 2*A* no longer makes a logical match between the flip-flops *IC* 1*A* and 1*B*. The observer's light-emitting diode goes off, and a



Responses when the stimulus light was on the center of the fovea and off it logical-1 signal is no longer sent to the gate *IC* 2*B*. The clock's signal is therefore ignored, and the display stops with its last count and hence the observer's response time in milliseconds. All remains the same until the reset switch is tripped and both *IC* 1*A* and 1*B* are put back to their original states.

Nearly all the parts are available from the Radio Shack (check a local store or write 2617 West Seventh Street, Fort Worth, Tex. 76107) for less than $15. The *IC* 4 is available from the National Semiconductor Corporation (2900 Semiconductor Drive, Santa Clara, Calif. 95051) for $12. The seven-segment display can be obtained from Litronix Inc. (19000 Homestead Road, Cupertino, Calif. 95014) for about $4.

Be certain never to reverse the battery connections or wire the power-supply pins of the integrated circuits backward or they will be destroyed. The resistor *R*5 should be adjusted so that the clock counts at the rate of 1,000 hertz (cycles per second). You can check it by allowing the circuit to count for 10 seconds as measured on a stopwatch. Since the highest reading on the display is 9999, at 10 seconds the display should again show 0000. Adjust the resistor until you have achieved this correlation. For example, increasing the value of *R*5 decreases the counting rate.

My observers were three students at Cleveland State: Teresa Chandler, Sandy Hoskins and Bruce Firtha. An observer viewed the light-emitting diode through two large, crossed sheets of polarizing filters. By adjusting the angle between the polarizing senses of the filters I could select the illumination reaching the observer. To measure the filter angles I affixed a length of clear cellophane tape to each filter so that the tapes were perpendicular when the filters were perpendicular, thereby blocking the maximum amount of light from a light bulb. Other filter orientations were measured on a protractor laid between the tape lengths. Accuracy was probably no better than plus or minus two degrees.

The intensity of the light transmitted by the crossed filters was compared with the intensity (designated 1) transmitted by parallel filters. For each orientation the relative value of the transmitted intensity was the square of the cosine of the angle between the senses of polarization of the crossed filters. I estimated the relative intensity transmitted through perpendicular filters from transmission data from the Polaroid Corporation for the type of polarizing filters I was using. Although for perpendicular filters the transmission was small, it was not exactly zero. When data were taken with the room lights on, the observers cloaked their head and the filters with a black cloth so that the lights did not produce a glare on the observers' side of the filters.

Whenever the diode light came on,

the observer turned it off and I recorded the response time. For each setting of the polarizing filters I took at least 50 measurements and then plotted their mean and their standard deviation. Such an average response time includes any visual latency and also the larger motor-response time of the observer. I tried to turn on the diode light randomly so that the observer could not anticipate the event. Particularly large or small response times that occasionally resulted from neglect or anticipation by the observer were eliminated from the data.

One problem you may encounter is to find an observer sufficiently interested in the experiment to concentrate continuously on the diode. The experiment is tiring, and a lack of interest could increase the standard deviations considerably. I took data for observers in a dark room, first allowing at least 10 minutes for their eyes to adapt to the darkness. All data for an observer were taken at a single sitting. Mixing data taken on different days would probably complicate the experiment because of the possible daily changes in motor-response times.

A plot of the data on a semilog graph is shown in the top illustration on page 148. If the motor-response times were eliminated from the experiment, the remaining visual-response times would presumably follow this same type of curve, but at shorter time intervals. In other words, the visual latency would increase as the illumination of the stimulus decreased; the rise would be sharper at low levels of illumination.

I also collected data when the room lights were on. Some researchers have found that the visual latency decreases when the overall illumination surrounding the stimulus increases. Within the standard deviations in my data, however, I could find no significant difference in an observer's response curve with the room lights on and with them off.

When one of my observers stared at the diode, the image of the diode fell on the fovea, which has a dense packing of cone cells but no rod cells. Some published research indicates that the visual latency may be different if the image of the stimulus falls elsewhere, partly because the visual response would then involve rods. You can search for this effect by having the observer stare at a small object displaced to the side of the diode.

To check this possible change in visual response I first redetermined the response curve for Bruce Firtha (left eye only) in a semidark room. His right eye was covered with a cardboard patch. Firtha stared directly at the diode as before. The resulting curve was essentially the same as the previous one. Then I repeated the experiment with Firtha staring at a small dot three centimeters to the side of the diode. The diode was otherwise surrounded by a featureless cardboard background. This procedure made the image of the diode fall about four degrees off the center of the fovea and thus in a region that has some rods. The response curve was the same for the higher illumination levels but then increased significantly for lower levels.

Such an increase suggests that in conditions of low illumination the visual latency increases faster in retinal areas outside the fovea than directly in the center of the fovea. Vision in bright light is termed photopic and is dominated by cone response. Vision in dim light is termed scotopic and is dominated by rod response. The results suggest that the scotopic vision away from the center of the fovea incorporates more visual latency than it does in the center of the fovea because of the greater abundance of rods.

You can do much other work on visual latencies, either with this type of apparatus or with some other type. How does the response curve change as a function of angular distance from the center of the fovea? Is the response when the stimulus is moved toward the nose different from the response when it is moved away from the nose? You could try to map response curves for the area surrounding the fovea. If more work indicates that the rod-free center of the fovea has a response curve different from the curves of other retinal areas, you might draw tentative conclusions about the visual latency of the rods as distinguished from that of the cones.

You might also try stimulus targets of different sizes. A small target viewed directly stimulates only the center of the fovea, whereas a larger target stimulates a larger portion of the retina and thus could invoke responses from rods. The visual response in such a case is a summation over the entire retinal area that is stimulated.

I am not sure how the response curves should look as a function of the room lights. Although I did not find any change in the response curve by turning on the room lights (for as much as I measured, at least), you might try to repeat the experiment. You could also illuminate the eyes directly with a small light instead of illuminating the diode stimulus and its surroundings. As the eyes adapt to a higher illumination level, do they respond noticeably faster, corresponding to my findings in the pendulum illusion?

Remember that in any experiment you must collect a lot of data, average the results and then compare the means by examining the standard deviations. Two means that have been collected under different circumstances and that fall within one standard deviation of each other are normally regarded as being indistinguishable.

Visual latency can affect your perception of passing scenery while you are riding in a car if you watch the roadside with a pair of sunglasses from which one glass has been removed. Fix your view on distant objects, first to one side of the car and then to the other. At the same time compare the speeds at which foreground objects seem to pass you on the two sides. If your left eye is covered with the dark glass, the objects on the left seem to pass slower than those on the right. If you judge the speed of the car by watching the passing scenery, your left-hand view indicates that the car is moving slower than it actually is. On your right side the car seems to be moving faster than it actually is.

The procedure also alters the sizes and distances of foreground objects. The objects on the side of the car that gives a sensation of decreased speed tend to seem closer and smaller than they really are. The objects on the other side seem farther away and larger than they actually are.

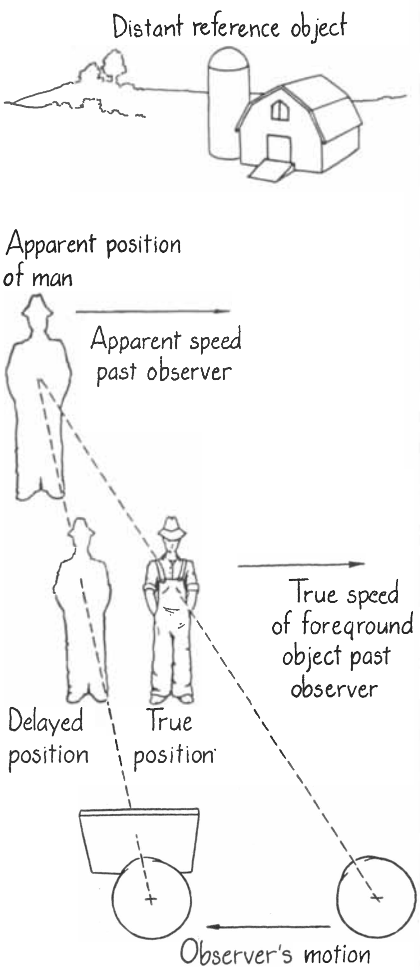
These distortions of speed, distance and size are produced by the visual latency caused by the dark filter over one eye. That eye sends a delayed response to the brain. As with the pendulum illusion, the delayed response alters the angle between a foreground object and the reference background object. If your left eye is covered and you look to the right, the angle between the foreground object and the reference object is increased because of the delay. The increase is interpreted as meaning that the foreground object is farther away than it actually is. If the object is regarded as being more distant but it still occupies the same angular size in your field of view, you are forced to conclude that it is larger than it really is. Similarly, if you judge it to be more distant but it still crosses your field of view at the same rate (that is, in a given amount of time it crosses a particular angle in your field of view regardless of how far away you judge it to be), you are forced to conclude that it is traveling faster than it actually is. The visual latency alters your perception of the object's depth, and you in turn misjudge its size and speed past you.

The distortions of speed, distance and size can be demonstrated to a small audience by employing a motion picture made from a moving car. James T. Enright of the University of California at San Diego suggests that the camera be aimed at some distant object to the side of the car while the car is traveling at about 10 miles per hour. A location that has such a distant object and also foreground objects some 100 yards away will enhance the effect. Each observer in the audience wears sunglasses with one glass removed or holds some other dark filter in front of one eye.

Writing in *Journal of the Franklin Institute* in 1940, Ronald Ives noted that at dusk the streetlights at intersections seemed to go on slightly sooner than the ones between intersections. Moreover, the second group seemed to go on in sequence down the street and away from the observer. Assuming that all the streetlights were in the same circuit, the effect had to be an illusion.

The effect was due to the visual latency I have been describing. There were more lights at the intersections than at any point between intersections, and so the observer received more light from the intersections. The visual latency for a bright stimulus is less than that for a dim one. Although all the lights went on simultaneously, the ones at intersections were perceived first because of the reduced visual latency from the stimulus. The lights along the street seemed to go on one by one, the remoter ones later, because their increasingly dim stimulus resulted in increased visual latency.

Can the visual latency resulting from sunglasses significantly affect your reaction time when responding to a hazard? It might. Suppose your sunglasses are dark enough to add 100 milliseconds to your response time (an extreme case). If you were driving at a speed of 55 miles per hour (about 25 meters per second), the increased response time would mean an additional stopping distance of about 2.5 meters.



Distortions seen by a moving observer

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

MATHEMATICAL GAMES

COMPUTABILITY AND UNSOLVABILITY. Martin Davis. McGraw-Hill Book Company, 1958.

GÖDEL'S PROOF. Ernest Nagel and James R. Newman. New York University Press, 1958.

THE UNDECIDABLE. Edited by Martin Davis. Raven Press, 1965.

WORLD OIL PRODUCTION

PETROLEUM RESOURCES: HOW MUCH OIL AND WHERE. John D. Moody and Robert E. Geiger in *Technology Review,* pages 38–45; March/April, 1975.

ENERGY: GLOBAL PROSPECTS 1985–2000. Workshop on Alternative Energy Strategies. McGraw-Hill Book Company, 1977.

ENERGY SUPPLY TO THE YEAR 2000: GLOBAL AND NATIONAL STUDIES. Workshop on Alternative Energy Strategies. The MIT Press, June, 1977.

HEAVY LEPTONS

ELEMENTARY PARTICLES: FRONTIERS OF HIGH ENERGY PHYSICS. Gerald L. Wick. Geoffrey Chapman, London, 1972.

ELECTRON-POSITRON ANNIHILATION AND THE NEW PARTICLES. Sidney D. Drell in *Scientific American,* Vol. 232, No. 6, pages 50–62; June, 1975.

WHAT IS THE WORLD MADE OF? Gerald Feinberg. Anchor Press, 1977.

FUNDAMENTAL PARTICLES WITH CHARM. Roy F. Schwitters in *Scientific American,* Vol. 237, No. 4, pages 56–70; October, 1977.

THE ELECTRONIC TELEPHONE

AN IMPROVED TELEPHONE SET. A. H. Inglis and W. L. Tuffnell in *The Bell System Technical Journal,* Vol. 30, No. 2, pages 239–270; April, 1951.

DIALOG WITH MICROPHONE AMPLIFIER AND TONE RINGING. H. Wangensteen and T. Wessel in *Ericsson Review,* Vol. 44, No. 3, pages 98–110; 1967.

NEW ELECTRONIC TELEPHONE SUBSCRIBER SET WITH DYNAMIC TRANSDUCERS. P. V. Arlev, K. Damsgaard and Sv. Aa. Jäger in *Teleteknik,* Vol. 18, No. 2, pages 62–69; 1974.

A HISTORY OF ENGINEERING IN THE BELL SYSTEM: THE EARLY YEARS (1875–1925). Edited by M. D. Fagen. Bell Telephone Laboratories, 1975.

THE SURFACE OF MARS

MARS AS VIEWED BY MARINER 9: NASA SPECIAL PUBLICATION 329. National Aeronautics and Space Administration. Government Printing Office, 1974.

THE NEW MARS: THE DISCOVERIES OF MARINER 9. William K. Hartmann and Odell Raper. NASA Special Publication 337. Government Printing Office, 1974.

MARS. James B. Pollack in *Scientific American,* Vol. 233, No. 3, pages 107–117; September, 1975.

THE GEOLOGY OF MARS. Thomas A. Mutch, Raymond E. Arvidson, James W. Head, Kenneth L. Jones and R. Stephan Saunders. Princeton University Press, 1977.

SCIENTIFIC RESULTS OF THE VIKING PROJECT. *Journal of Geophysical Research,* Vol. 82, No. 28; September 30, 1977.

THE FLOW OF ENERGY IN A FOREST ECOSYSTEM

ENERGY EXCHANGE IN THE BIOSPHERE. D. M. Gates. Harper & Row, 1962.

ECOLOGY. Robert E. Ricklefs. Chiron Press, 1973.

BIOCHEMISTRY OF A FORESTED ECOSYSTEM. Gene E. Likens, F. Herbert Bormann, R. S. Pierce, J. S. Eaton and N. M. Johnson. Springer-Verlag, 1977.

HOW CELLS MAKE ATP

ACID-BASE TRANSITIONS AND PHOSPHORYLATION BY CHLOROPLASTS. André T. Jagendorf in *Federation Proceedings,* Vol. 26, No. 5, pages 1361–1369; September–October, 1967.

PROTON-TRANSLOCATION PHOSPHORYLATION IN MITOCHONDRIA, CHLOROPLASTS AND BACTERIA: NATURAL FUEL CELLS AND SOLAR CELLS. Peter Mitchell in *Federation Proceedings,* Vol. 26, No. 5, pages 1370–1379; September–October, 1967.

KINETICS OF ADENOSINE TRIPHOSPHATE SYNTHESIS IN BOVINE HEART SUBMITOCHONDRIAL PARTICLES. William S. Thayer and Peter C. Hinkle in *The Journal of Biological Chemistry,* Vol. 250, No. 14, pages 5336–5342; July 25, 1975.

A NEW LOOK AT MECHANISMS IN BIOENERGETICS. Efraim Racker. Academic Press, Inc., 1976.

OXIDATIVE PHOSPHORYLATION AND PHOTOPHOSPHORYLATION. Paul D. Boyer, Britton Chance, Lars Ernster, Peter Mitchell, Efraim Racker and E. C. Slater in *Annual Review of Biochemistry,* Vol. 46, pages 955–1026; 1977.

THE COMBINATORIAL MATHEMATICS OF SCHEDULING

BOUNDS ON MULTIPROCESSING ANOMALIES AND RELATED PACKING ALGORITHMS. Ronald L. Graham in *AFIPS Conference Proceedings: 1972 Spring Joint Computer Conference,* Vol. 40, pages 205–217; 1972.

ON PACKING SQUARES WITH EQUAL SQUARES. Paul Erdös and Ronald L. Graham. *Journal of Combinatorial Theory (Series A),* Vol. 19, No. 1, pages 119–123; July, 1975.

COMPUTER AND JOB-SHOP SCHEDULING THEORY. Edited by Edward G. Coffman. John Wiley & Sons, Inc., 1976.

NEW DIRECTIONS IN CRYPTOGRAPHY. Whitfield Diffie and Martin E. Hellman in *IEEE Transactions on Information Theory,* Vol. IT-22, No. 6, pages 644–654; November, 1976.

PIETER BRUEGEL THE ELDER AS A GUIDE TO 16TH-CENTURY TECHNOLOGY

A HISTORY OF SCIENCE, TECHNOLOGY AND PHILOSOPHY IN THE 16TH & 17TH CENTURIES. A. Wolf. George Allen & Unwin Ltd, 1950.

GRAPHIC WORLDS OF PETER BRUEGEL THE ELDER. H. Arthur Klein. Dover Publications, Inc., 1963.

GREAT STRUCTURES OF THE WORLD. H. Arthur Klein and Mina C. Klein. The World Publishing Company, 1968.

PETER BRUEGEL THE ELDER, ARTIST OF ABUNDANCE: AN ILLUSTRATED PORTRAIT OF HIS LIFE, ERA, AND ART. H. Arthur Klein and Mina C. Klein. The Macmillan Company, 1968.

SCIENCE AND SOCIETY IN THE SIXTEENTH AND SEVENTEENTH CENTURIES. Alan G. R. Smith. Thames and Hudson, London, 1972.

THE WORLD OF MEASUREMENTS. H. Arthur Klein. Simon and Schuster, 1974.

THE AMATEUR SCIENTIST

STEREOPSIS, VISUAL LATENCY, AND THREE-DIMENSIONAL MOVING PICTURES. James T. Enright in *American Scientist,* Vol. 58, No. 5, pages 536–545; September–October, 1970.

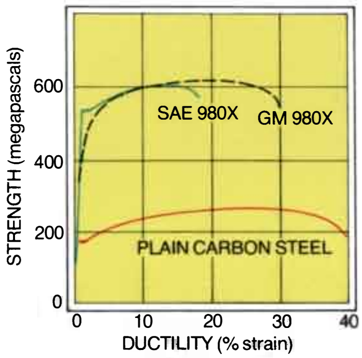
LATENCY FUNCTIONS IN HUMAN VISION. R. J. W. Mansfield in *Vision Research,* Vol. 13, pages 2219–2234; 1973.

EYE MOVEMENTS AND THE PULFRICH PHENOMENON. B. J. Rogers, J. J. Steinbach and H. Ono in *Vision Research,* Vol. 14, pages 181–185; 1974.

Why all the interest in using high-strength, low-alloy (HSLA) steel in the automobile? To help cut weight and thus conserve fuel.

Not that HSLA steel is lighter than other steels. It's not. But it can do an equivalent job in a thinner gauge.

Until recently, however, HSLA steel was limited in use to simple shapes because of its poor ductility, or formability. A discovery by our metallurgists here at the General Motors Research Laboratories changed all that. They found that contrary to traditional belief, ductility can be greatly increased without sacrificing strength.



How? By heat treating. The treatment is applied to commercial SAE 980X steel. It improves ductility as much as 50% by dissolving brittle carbide particles in the microstructure and then introducing a strong but ductile second phase (martensite) through controlled cooling.

Result #1 was a new steel called GM 980X. Although weakened by annealing, which enables it to form into complex shapes, it recoups its high strength during forming.

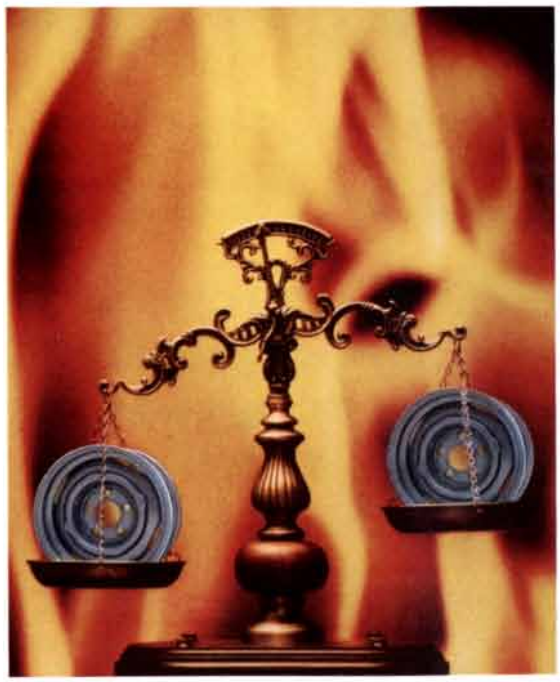
Significantly, steel companies can employ our heat treatment with their existing annealing facilities to make this more ductile HSLA steel. And they are now doing so.

Result #2: GM 980X is already being used to stamp lighter weight parts for production and prototype automobiles.

Result #3: The research behind GM 980X has stirred considerable activity in the steel industry, at universities, and in independent laboratories throughout the world to develop additional HSLA steels having better strength-ductility properties.

All in all, just part of GM's continuing effort to help America meet the growing energy challenge.

Putting the heat on high-strength steel to make it shape up.



General Motors Research Laboratories
Warren, Michigan 48090

At the Indian hospital in Sells, Arizona, Papago Indian Peter Ruiz checks the quality of the TV signal coming in from a mobile health van miles away. He's watching the screen of a TEKTRONIX waveform monitor in the STARPAHC* communications console. The TEKTRONIX automatic video corrector, just above the waveform monitor, automatically corrects distortions in both color and black and white signals, freeing Peter to concentrate on the other communication channels.

Small, isolated villages, in an area of the Sonora Desert roughly the size of Connecticut, regularly receive mobile health van visits. Two-way video,

"Getting health care to people in remote areas is a worldwide problem that telecommunications is helping to solve - whether the people are astronauts in space or Papago Indians on the Sonora Desert."

Norman Belasco, Project Officer, STARPAHC*
NASA's Lyndon B. Johnson Space Center, Houston, Texas

At NASA, Norman Belasco explains that one of NASA's early telemedicine projects involved the development of systems that would support biomedical research and provide clinical health care capabilities for long-duration, manned space flights.

"The STARPAHC* project originated when NASA was asked by the President's Domestic Council to examine ten pressing national problems — and one of them turned out to be the delivery of health care to remote areas."

The resulting program involved the cooperation of NASA, HEW and its Indian Health Service, and the Papago Indian Tribe. Lockheed Missiles and Space Company, Inc., was selected to assemble the system and support the field operations.

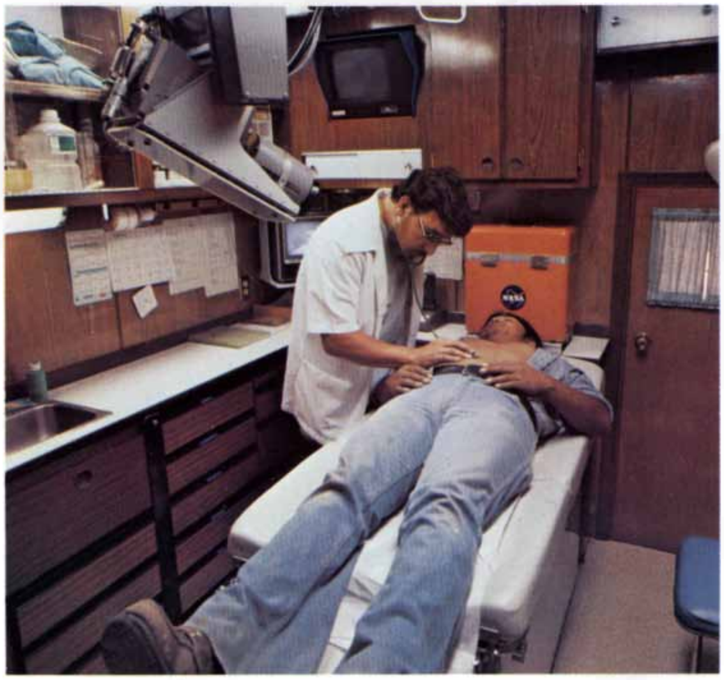
One of the requirements of NASA's remote-health-care program was that commercially available, off-the-shelf equipment — or as much as possible — be used in the project. That's how Tektronix became involved.

The communications industry uses a wide range of Tektronix electronic instrumentation. This includes sophisticated television broadcast equipment, which has found many uses beyond the commercial television broadcast industry. Noncommercial uses are as varied as NASA Control's elaborate system that enabled us to see man take his first step on the moon — to the production of video tapes that record what a physician sees through a bronchoscope.

The TV broadcasting industry around the world looks to Tektronix for instruments that help to maintain the quality of television transmission from the camera to the home receiver. The Grass Valley Group, a wholly owned Tektronix subsidiary located in California, is especially well known for its production and routing switchers, which are used in program editing and for special effects, such as the split screens, inserts, and montages often seen in applications such as sports programming.

*Space Technology Applied to Rural Papago Advanced Health Care voice, and data signals are transmitted from the van to a mountain-top, microwave relay station that is aligned with the Sells hospital and a fixed, remote clinic at Santa Rosa. Sells has telephone tie lines to consultant and referral centers in Phoenix and Tucson, and to a computerized patient-record file in Albuquerque.

Inside the mobile van Medic Ken Tiokasin examines patient Alfred Pablo, while a doctor at the physician's console miles away in Sells conducts a two-way video and voice consultation. The physician can remotely control the TV cameras on the mobile unit and at the Santa Rosa clinic — zoom in on a patient's problem, read x-rays, examine color specimen slides under a microscope, receive an ECG readout, and send or receive pertinent health-record data. Now, with communication satellite technology, it will be possible for physicians to hold two-way visual consultations with colleagues and patients anywhere in the world.



What the engineer sees. The TV camera transforms images into complex analog signals. Special test signals, not visible on a home TV set, are sent along with the television signal. At each step in the broadcast chain, engineers use these test signals to evaluate picture-signal quality.

For example, in the STARPAHC system, color intensity and hue are controlled by the insertion of a relatively new signal called a VIR (vertical interval reference). The insertion is accomplished with a TEKTRONIX signal generator just after the camera picks up the image.

Then a TEKTRONIX Automatic Video Corrector samples the VIR signal and *automatically* adjusts both color and black and white values before the picture is transmitted to the doctor's (or home viewer's) receiver. The engineer can check the color signal on the TEKTRONIX waveform monitor periodically to be sure that the picture being transmitted is the same as the picture at the source.

Many other Tektronix products are used throughout the TV industry. Picture Monitors. Vectorscopes. Demodulators. Analog-to-Digital Converters. Spectrum Analyzers. Digital Photometers. And oscilloscopes for system servicing (including the computer in the STARPAHC system).

But TV is only part of the picture.

Tektronix serves the total communications industry. For example, when we saw that the telephone industry needed a faster and more precise way to find cable faults, Tektronix engineers solved the problem with a Cable Tester. We work to anticipate the needs of people working with everything from CB radios to communication satellites.

A half billion dollars in sales. Tektronix' diversified products are used in many industries worldwide.

Progressive managers are using TEKTRONIX computer graphics to solve problems in areas as diverse as international monetary exchange to the mapping of subterranean oil and coal deposits. Almost everywhere computers are found, TEKTRONIX portable oscilloscopes and other service instruments help to maintain them. Quantity buyers and suppliers of integrated circuits use TEKTRONIX IC Test Systems to help develop, select, and maintain quality control of "chips" for their products.

Foundries save energy and man-hours with TEKTRONIX digital ultrasonic testing systems. Almost every R & D lab uses TEKTRONIX high-performance oscilloscopes to test, measure, and interpret electrical phenomena. Hospitals improve patient care with TEKTRONIX physiological monitors.

And now in microelectronics, TEKTRONIX logic analyzers and Microprocessor Labs give engineers an edge in designing new products that promise to revolutionize our business and personal lives.

These are just a few of the more than 800 Tektronix products that serve the test, measurement, and information display needs of thousands of customers, worldwide.

We'll gladly send you information about our products, or a copy of our annual report. Write to the Marketing Communications Department, 76-260, Tektronix, Inc., P.O. Box 500, Beaverton, Oregon 97077.

By phone—call 503-644-0161.

Tektronix
COMMITTED TO EXCELLENCE

Jameson Irish

"~~Scotch~~ on the rocks."

If you like Scotch, you'll love light Jameson Irish—and what better day to try it than St. Patrick's.

Try a glass of Jameson Irish the way you would your favorite Scotch.

You'll notice how much it tastes like fine Scotch—only lighter and more delicate. Not smoky tasting like Scotch.

The dedicated Scotch Whiskey drinker will instantly appreciate this flavor difference.

Though it may take a little time getting used to saying, "Jameson Irish on the rocks, please."

Jameson. World's largest-selling Irish Whiskey.

80 & 86 PROOF • CALVERT DIST. CO., N.Y.C.