SCIENTIFIC AMERICAN

FORGING

$2.00

August 1981

Pioneer announces a new c
HIGH F
FOR HU

For the past 40 years, Pioneer has built high fidelity components with technical specifications that have impressed even the most discriminating audiophiles.

While these new Pioneer components are consistent with that tradition, they go on to do something of perhaps even greater import.

They introduce a revolutionary concept in component design and engineering: *High Fidelity for Humans.* A concept so far reaching, that for the first time, components are as pleasant to live with as they are to listen to.

For instance, our receivers memorize the precise locations of your favorite stations. So you can instantly tune in any station at the touch of a button.

New Pioneer turntables have such superb suspension our Polymer Graphite tonearms won't

skip when people dance.

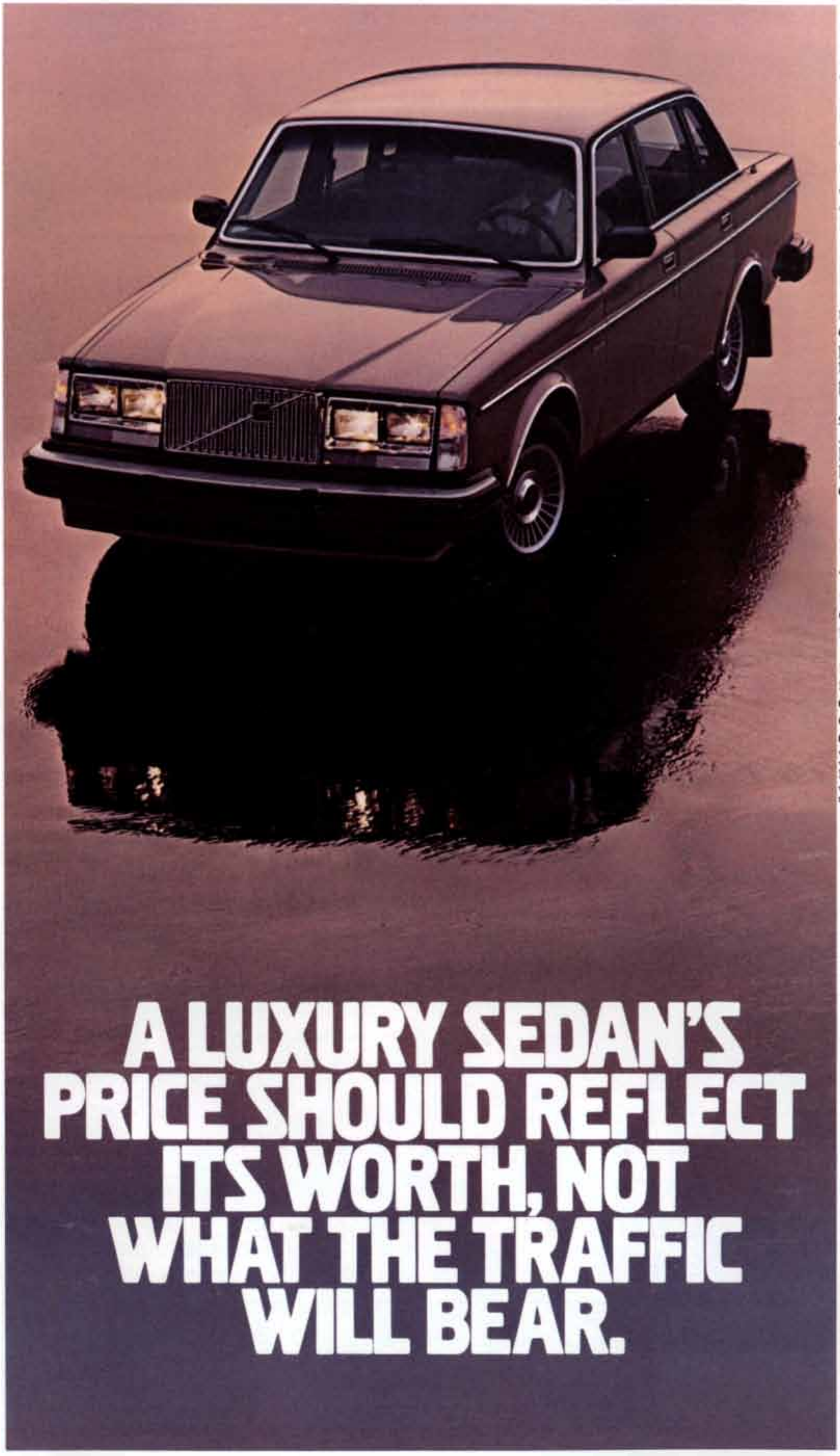
Pioneer's new cassette decks will allow you to preview a tape faster and easier than you can a record. At your command, our Index Scan will breeze through a tape, automatically stopping to play the first five seconds of each selection.

If features like these sound downright incredible, it's because they are. Which is why we suggest you go to your nearest Pioneer dealer and check out the first line of audio components that not only make major electronic advances, but enter a whole new territory: addressing human needs.

Pioneer high fidelity components are #1 with people who care about music. So you may wonder why it took us so long to get here.

We, too, are only human.

PIONEER®
We bring it back alive.

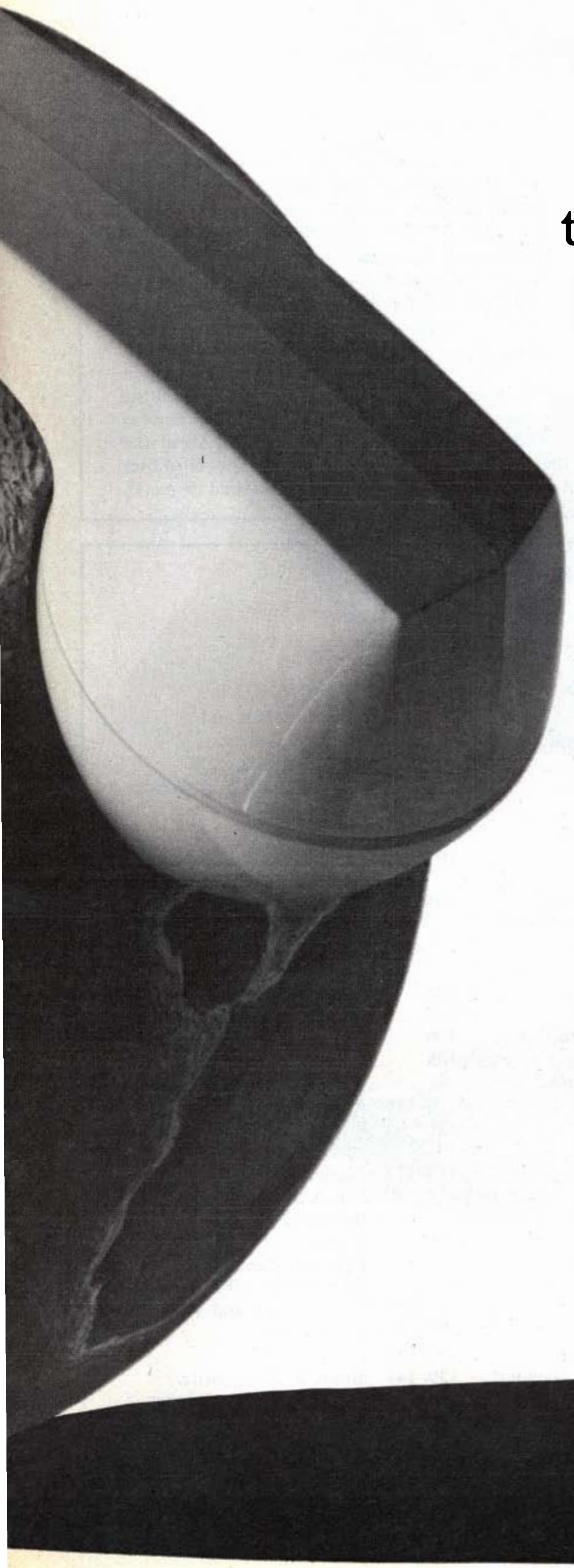

A LUXURY SEDAN'S
PRICE SHOULD REFLECT
ITS WORTH, NOT
WHAT THE TRAFFIC
WILL BEAR.

Established 1845 SCIENTIFIC AMERICAN August 1981 Volume 245 Number 2

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y 10017 COPYRIGHT © 1981 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. 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: $21 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $27 PER YEAR, ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y 10017



THE COVER

The painting on the cover portrays in a somewhat schematic way a power-assisted forging hammer (see "The Forging of Metals," by S. L. Semiatin and G. D. Lahoti, page 98). The heavy steel ram at the top, marked with safety stripes, carries a die and moves downward. The hot workpiece (the metal being forged) rests on a fixed bottom die. The heavy springs help to absorb the shock of the blows delivered by the powerful ram. Such drop hammers, some falling by gravity alone and some assisted by steam or air pressure, are the oldest, fastest and most versatile forging machines. A hammer works by striking the workpiece repeatedly. Mechanical and hydraulic presses usually make a forging with one stroke. In a mechanical press the energy is stored in a flywheel; the ram is actuated by a crank or a cam. In a hydraulic press the ram is driven by a fluid under high pressure. Although forging is an ancient technology, it has been improved in recent years by new findings about the way metals deform and by the application of computers in the design of dies and other forging equipment and in the analysis of the deformation of metal.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
36	Department of Defense
38–39	Andrew Christie
40–41	Vought Corporation
42–45	Andrew Christie
47	Harvard College Observatory
48	Kitt Peak National Observatory (*left*), Nicholas Z. Scoville, F. Peter Schloerb and Paul F. Goldsmith (*right*)
49	Lick Observatory
50–55	Allen Beechel
57–59	Tom Prentiss
60	Ruud J. Scheffer, Willie Comelin Scholten Laboratory
61	Doekle M. Elgersma, Willie Comelin Scholten Laboratory
62	Tom Prentiss
63	Tom Prentiss (*top*), Gerald N. Lanier (*bottom*)
64	Tom Prentiss
65	Gerald N. Lanier (*top*), Bruce Hemming (*bottom*)
66	Gary A. Strobel
84–85	George V. Kelvin (*left*), James A. Lake (*right*)
86	James A. Lake (*top*), George V. Kelvin (*bottom*)
87	George V. Kelvin
88	James A. Lake and Lawrence Kahan
89–95	George V. Kelvin
96	Michael Clark and James A. Lake
99	A. A. Badawy, C. F. Billhardt and W. J. Weider
100–101	Gabor Kiss
102	W. J. Weider
103	Gabor Kiss (*top*), C. R. Thompson (*bottom*)
104	W. W. Sunderland and Gabor Kiss (*top*), C. R. Thompson and Gabor Kiss (*bottom*)
105	W. W. Sunderland and Gabor Kiss
106	Mesta Machine Company
108	National Oceanic and Atmospheric Administration
110–118	Alan D. Iselin
121	P. C. Woodman, Ulster Museum
123–132	Patricia J. Wynne
135	J. L. Charmet, Collection Particulière, Paris
136	Science and Technology Research Center, New York Public Library, Astor, Lenox and Tilden foundations
137–138	Ilil Arbel
139–144	Stuart L. Rabinowitz, from the Burndy Library
147	Jearl Walker
148–152	Michael Goodman

LETTERS

Sirs:

Authors Gray and von Hippel ["The Fuel Economy of Light Vehicles," by Charles L. Gray, Jr., and Frank von Hippel; SCIENTIFIC AMERICAN, May] will no doubt be delighted to learn about the existence of a class of two-passenger vehicles that are inexpensive, fuel-efficient, lightweight, economical to operate, highly maneuverable, easy to park, relatively nonpolluting, meet the 1995 standards for fuel economy and best of all can be obtained today. They are called motorcycles.

D. E. ASPNES

Watchung, N.J.

Sirs:

...While the authors have correctly pointed out that current engines cannot deliver peak power at all vehicle speeds, they propose to rectify this by using a continuously variable transmission (CVT). Unfortunately a CVT offers about the same ratio coverage as current gearboxes, and it should not be confused with an infinitely variable transmission (IVT). In such a transmission engine speed can be varied independently of vehicle speed, and hence the engine can deliver peak power at any vehicle speed. A study by General Motors showed that a theoretically efficient CVT has very little fuel-economy advantage over a five-speed gearbox offering the same ratio coverage.

K. G. DULEEP

Arlington, Va.

Sirs:

...As a senior engineer with the Chrysler Corporation I should like to share with your readers the benefit of my experience in dealing with the practicalities of achieving better fuel economy....

The hypothetical "high-fuel-economy vehicle" does not possess the power to climb 30 percent grades, which according to the article are encountered on only a small percentage of roads. What happens to automobile users who live in, work in or visit sloping areas?...

The article estimates that a high-fuel-economy vehicle needs .5 horsepower for accessories. Today's vehicle may have to operate power steering, power brakes, radio, heater motor, defroster, windshield wiper and lights simultaneously. Five horsepower is a more accurate accessory-power requirement.

The total estimated horsepower of the hypothetical high-fuel-economy vehicle is about half what is actually required by a four-passenger vehicle. Although even a 70-horsepower engine can be made more efficient than today's engines by resorting to turbocharging and diesel technology, additional emission-control devices will be required to meet Government standards. This means higher cost to the automobile buyer....

The proposed "gas guzzler" tax would not deter the affluent customer. Similarly, a fuel tax would only serve to burden the majority of Americans who have basic transportation requirements. Mandating a 60-miles-per-gallon fuel economy is more wishful thinking than practical reality. The $10-billion cost of retooling for 60-m.p.g. vehicles, even if such a goal were achievable, would be more wisely spent developing new sources of energy for a fleet of 30-m.p.g. vehicles....

C.D. DYRDA

Troy, Mich.

Sirs:

Mr. Aspnes' letter underlines the fact that efficiency improvements in automobiles should be discussed in a larger context including not only motorcycles but also bicycles, van pools, mass transit and more rational land-use planning.

In response to Mr. Duleep's points about continuously variable transmissions, we assumed CVT's in our work for analytical simplicity. The General Motors study cited by Duleep gives us some reassurance that the fuel-economy numbers we quote could also be approached by vehicles equipped with more conventional transmissions. Our article specifies a CVT "in combination with some simple gearing" to expand its range enough so that "for most road speeds any power output up to the full rated output of the engine can be made available." For our application this is the functional equivalent of an infinitely variable transmission.

Mr. Dyrda is incorrect in concluding that a car whose engine horsepower is chosen so that it will be able to climb a 5 percent grade at 55 miles per hour cannot climb a steeper grade. Its top speed will simply decrease as the angle of climb increases. The .5 horsepower quoted by Dyrda is the average accessory-power requirement, not the peak power requirement we assumed in our analysis. We chose an engine size for four-passenger cars, for example, that could supply three horsepower to the accessories and simultaneously meet our performance requirements for acceleration or climb.

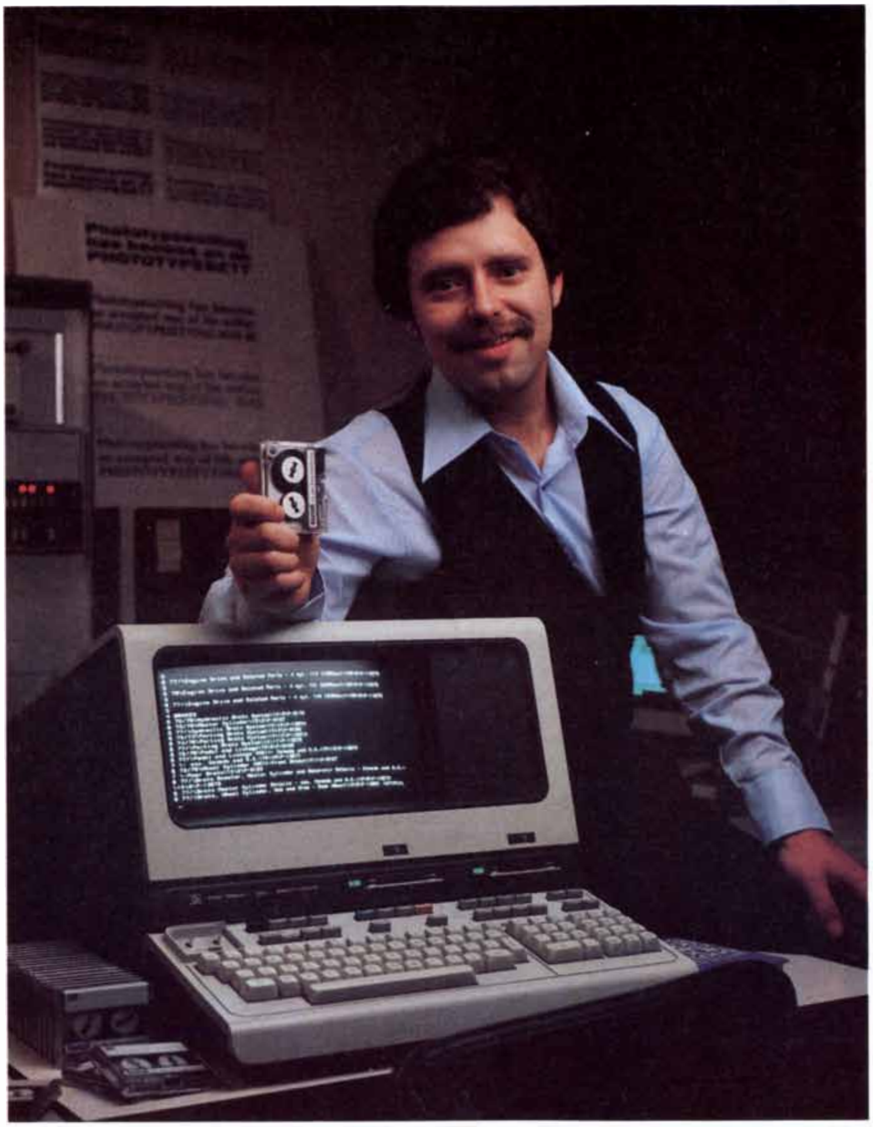
Because of the concern expressed by Dyrda and others about the low peak engine horsepower of our hypothetical vehicles, however, we have somewhat further explored the trade-off between performance and fuel economy. A typical result is that adding 20 horsepower to the 31-horsepower direct-injection diesel engine of our "current best technology" four-passenger car reduces its 0–50 m.p.h. acceleration time from 13.4 to 8.1 seconds but reduces its gasoline-equivalent fuel economy by only 4 m.p.g.: from 70 to 66 m.p.g. This result probably could not be achieved without a CVT.

Contrary to Dyrda's assertion, it is not yet clear that with "turbocharging and diesel technology, additional emission-control devices will be required to meet Government standards." Indeed, the 60-m.p.g. turbocharged Volkswagen diesel Rabbit prototype tested by the EPA in 1979 met all proposed 1985 emission standards without such devices. Although direct-injection diesel-powered prototypes have not yet met the 1985 standards, MAN, a West German company that manufactures buses and trucks, has developed a direct-injection spark-assisted diesel that burns methanol with both higher thermal efficiency and much lower particulate production than the usual diesel fuel. This is encouraging in view of the fact that in the next century methanol derived from coal and plant matter may replace petroleum as our leading liquid fuel.

Some expert readers have questioned the cost-effectiveness of the turbochargers we assumed in our analysis. We have therefore further examined the fuel economies that could be achieved with direct-injection diesel engines coupled to continuously variable transmissions in the absence of turbocharging. To our considerable surprise we find such vehicles could achieve fuel economies similar to those we had calculated for turbocharged engines. This suggests that fuel economies similar to the "current best technology" vehicles described in our article might be achieved at costs comparable to those incurred in the production of today's diesel-powered light vehicles.

Dyrda and many others question both the need and the desirability of governmental intervention in the "free market" for automotive fuel economy. There are, however, at least two interests in the improvement of fuel economy that go beyond the consumer's interest in reducing expenditures on gasoline and the manufacturer's interest in minimizing retooling investments: (1) there is an interest in reducing the vulnerability of all the industrialized democracies (not just the U.S.) to events in the Persian Gulf, and (2) because of the potential impact on the world's climate of the buildup of carbon dioxide in the atmosphere there is a global interest in shifting away from fossil fuels more rapidly than simple resource availability would require. Cur-

“Our reputation rests on digits, decimal points, and details. We wouldn’t trust them to anything less than Scotch® Brand Data Cartridges.”



Bill Birkett, Vice President,
Trade Graphics, Inc.,
Livonia, Michigan

The unique design of a data cartridge provides great reliability, high storage capacity and long tape life. And where could you possibly get better data cartridges than Scotch Brand, made by 3M, the people who invented the data cartridge system itself?

3M controls every step in manufacturing. Top quality magnetic tape and precision components are part of every Scotch Data Cartridge. Over twenty-five years of service to the computer industry assure you of the utmost reliability.

Scotch Data Cartridges are available in miniature DC 100A, the standard-size DC 300A and now, an extra-length DC 300XL with 50% more storage capacity. They are compatible with most cartridge systems including Hewlett-Packard, IBM, NCR, Tektronix and TI.

To find out where you can find Scotch Data Cartridges or virtually any other data recording medium, call toll-free: 800-328-1300. (In Minnesota, call collect: 612-736-9625.) Ask for the Data Recording Products Division.

If it’s worth remembering, it’s worth Scotch Data Recording Products.

3M

rently it appears that such a shift will not be possible at an acceptable cost unless the efficiency of our energy usage is increased greatly.

Dyrda is probably correct in assuming that without governmental tax stimuli or standards we cannot depend on average U.S. passenger-car fuel economy to rise about 30 m.p.g. At $1.50 per gallon of gasoline, for example, the payoff in reduced fuel costs associated with a switch from a 30-m.p.g. to a 60-m.p.g. car would be only 2.5 cents per mile, and even those savings would be partially offset by the higher first cost of the 60-m.p.g. car. Thus the pressure of the "invisible hand" of the market for higher fuel economy would be very weak, and manufacturers could be expected to resist successfully, for some years at least, investing the necessary billions in retooling.

For the nation, however, the difference between an average fleet fuel economy of 30 m.p.g. and one of 60 m.p.g. would amount to a difference in petroleum consumption approximately equivalent to the flow through the Alaska pipeline: 1.6 million barrels per day. If Dyrda is correct in saying that the cost to Detroit of retooling to capture these enormous fuel savings would be only $10 billion, the corresponding cost would be less than 10 cents per gallon of gasoline saved. The cost of retooling may well be considerably higher than Dyrda's estimate, but it would still be much less than the cost of the alternative: investing in the equivalent amount of new fuel supply.

Finally, we should like to withdraw the statement in our article that "the drivers of subcompact cars appear to have between 10 and 30 percent fewer accidents than the drivers of larger cars." That is the conclusion of a study of accidents conducted in the state of Michigan. Since our article was published, Dr. Leonard Evans of General Motors has brought to our attention a study of accidents in the state of North Carolina that finds small cars have higher accident rates than large cars per mile driven. Hence it appears that the last word on the question is not in.

CHARLES L. GRAY, JR.

Ann Arbor, Mich.

FRANK VON HIPPEL

Princeton, N.J.

Sirs:

I am happy to see that there has been no response to Dr. Hofstadter's column on self-reference ["Letters," SCIENTIFIC AMERICAN, May]. Those who look at self-reference perch themselves precariously on Ockham's razor, tottering between hopeless confusion and fantastic insight. One might as well open a dictionary and find, "recursion: see RECURSION." He who is able to see recursion gains fresh insight into it.

I am reminded of a conversation between an ancient philosopher, Hee Hoo (whose epitaph reads, "Who is buried in Hoo's tomb?"), and W. V. Hee:

Hoo: What is a question that is also its own answer?

Hee: Why not, "Why not?"?

Hoo: "Why not?"? Why not!

THOMAS MUNNECKE

Riverside, Calif.

Sirs:

Scientific American omitted the first reference I had listed in the bibliography for my article "The Decay of the Proton" in its June issue. The reference was "Status of Conserved Quantum Numbers," by M. Goldhaber, in *Unification of Elementary Forces and Gauge Theories*, edited by David B. Cline and Frederick E. Mills (Harwood Academic Publishers, London, 1978). I was particularly sorry to see this omission, as I had relied on the reference for much of the history of the idea of baryon conservation as presented in my article.

STEVEN WEINBERG

Harvard University
Cambridge, Mass.

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 60 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 ____________

____

____

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1931: "How did our planetary system get here? It is obviously no accidental conglomeration of bodies: the planets all go around the sun in the same direction and almost in the same plane, most of them in nearly circular orbits; the largest planets are near the middle in order of the planets' distance from the sun, whereas the innermost and outermost planets are small; most of the planets are in rapid rotation, and the rotation is in nearly the same direction as the orbital motion, except for Uranus, where it is at right angles to it; and six of the nine large planets are the center of satellite systems that reproduce the solar system in miniature with interesting but not radical variations in plan. The hypothesis accepted by practically all students of the subject is the one put forward by professors T. C. Chamberlin and F. R. Moulton of the University of Chicago, namely that the planets owe their existence to a catastrophe. Long ago, three or four billion years ago as we now have good reason to believe, our sun was an isolated star, until some other wandering star passed near it and was drawn in by the stars' mutual attraction so that the two bodies passed in a hyperbolic orbit within a few million miles of each other. The enormous tidal forces caused the ejection of vast masses of the solar material. Some of these masses fell back, others may have flown off into space, but a good part was set into motion sidewise by the attraction of the now receding star and condensed into a system of planets all moving around the sun in the same general direction. There is still, however, a great deal to be found out before we have a really satisfactory theory of the origin of our planetary system."

"Airplanes flying at heights of more than 30,000 feet, which can reach much higher speeds than existing machines, are being constructed at the famous Junkerswerke at Dessau in Germany. The chief aim of the new machine is to reach high altitudes and to find paths that can be used as regular airplane trade routes. The airplane is a Junkers metal deep-deck, single-motored machine of 60-foot wing breadth and 9,000 pounds weight. A small compressor keeps the air pressure normal for the men within the cabin, which is double-walled and air-tight. Control of the motor and steering is done by levers in the cabin working in air-tight shafts. The motor itself is of a special type and has an air pump to supply enough air from the thin air at those heights."

"The patent covering high-vacuum radio tubes, the type universally used in radio-receiving sets, has been declared invalid by the Supreme Court of the United States. The Supreme Court reversed the decision of the Circuit Court of Appeals for the Third Circuit. The validity of the Langmuir patent No. 1558436, owned by the General Electric Company, had been contested by the De Forest Radio Company, which had been charged with infringement of the patent. The court concluded in an opinion by Mr. Justice Stone that the production of the high-vacuum tube, in view of the prior art that included contributions of Dr. Lee De Forest, resulted only from the skill of those practiced in the art and did not constitute invention, and therefore was not patentable."

SCIENTIFIC AMERICAN

AUGUST, 1881: "When M. Pasteur discovered a method of so cultivating the specific germs of chicken cholera as to give them the power of preventing the disease when inoculated, he asserted it to be possible to treat other specific organisms in the same way. This he has recently done for the virus of splenic fever, more widely known as anthrax or the Siberian plague. He was furnished with an unusually good opportunity of testing the power of his modified virus. The Agricultural Society of Melun put at his disposal 60 sheep, to be experimented with as he chose. Ten of these were separated from the rest and kept subject to inspection. Twenty-five others were inoculated twice with the modified anthracoid virus. After a sufficient time these and another 25 sheep were inoculated with the pure virus of the disease. The 50 sheep were then allowed to mingle together in the same enclosure. The 25 'vaccinated' did not take the disease, while the others did. M. Pasteur made similar experiments with 10 animals of the bovine species, and he asserts that the same protective power was secured to them."

"The proposals that have recently been put forward for the construction of a fleet of high-speed Atlantic steamers, the launch of the *City of Rome,* the largest ship in the world after the *Great Eastern,* the building of gigantic ships by the Guion and Cunard companies—all point the same way and indicate the existence of a desire to place England and America virtually yet closer together than they are now. The *City of Rome* is of the following dimensions: length over all 610 ft., extreme breadth 52 ft. 3 in., tonnage 8,826 tons, indicated horse power 10,000. The distinctive type of the Inman Line has not been departed from in the perhaps old-fashioned but still handsome profile, with clipper bow, figurehead and bowsprit. The vessel is to be rigged with four masts. She will have three funnels, each painted with the company's white band."

"Some exceedingly interesting experiments lately took place at the Naval Torpedo Station with the Weeks rocket torpedo. This torpedo is a most peculiar structure. It consists of a float made of tin and sheet iron, braced internally with wood. It has two rudders similar to the tails of a sky rocket. The float portion is some 11 feet long, with the rudders being of the same length. In the forward part or head is placed some 50 pounds of dynamite, and this, coming in contact with an object, explodes by concussion. The entire structure is propelled by a rocket some six inches in diameter, three and a half feet in length and 100 pounds in weight. It moves on the surface of the water and has attained the wonderful speed of about 150 feet per second, which is kept up for some 600 yards. Captain Thomas O. Selfridge, the commandant of the Torpedo Station, witnessed a recent trial. A whizzing noise was heard and the torpedo went on its way. The velocity was something frightful, as may be judged when it is stated that the torpedo passed along and over (for it jumped occasionally) a distance of water not less than 1,375 feet in about *nine seconds.* It was impossible to time it precisely, for the smoke behind was very dense. Captain Selfridge said that the trial was a success."

"The apparatus that proved most satisfactory in cooling the bed chamber of the wounded President Garfield was furnished by Mr. Jennings of Baltimore. It was devised for use in a new process of refining lard. The apparatus consists of a cast-iron chamber, about 10 feet long and three wide and three high, filled with vertical iron frames covered with cotton terry or Turkish toweling. These screens are placed half an inch apart and represent some 3,000 feet of cooling surface. Immediately over the screens is a coil of iron pipe, the lower side of which is perforated with fine holes. Into a galvanized iron tank holding 100 gallons of water is put finely granulated or shaved ice. This water is constantly sprayed upon the sheets in the chamber. At the outer end of the chamber is a pipe connecting with an outdoor air conductor. An exhaust fan forces cold and dry air directly into the President's room through a flue. Air at as much as 99 degrees temperature has been drawn into the chamber at the rate of 22,000 cubic feet per hour. The temperature at the President's bed is maintained steadily at 75 degrees day and night."

THE AUTHORS

PAUL F. WALKER ("Precision-guided Weapons") is director of research on arms-control issues for the Union of Concerned Scientists. A graduate of the College of the Holy Cross, he joined the U.S. Army in 1969 and was trained as a Russian translator and intelligence specialist before serving a tour of duty in Germany with the Army Security Agency. He returned to the U.S. to pursue his graduate studies, obtaining an M.A. in international relations at the Johns Hopkins University School of Advanced International Studies in 1973 and a Ph.D. in political science from the Massachusetts Institute of Technology in 1978. He worked as a postdoctoral research fellow at the Center for Science and International Affairs of Harvard University's John F. Kennedy School of Government before taking up his present position in 1979. A writer, lecturer and consultant in the fields of national security, arms control and Russian-American relations, Walker was one of six coauthors of the recent book *The Price of Defense: A New Strategy for Military Spending,* published by Times Books, a division of Quadrangle/The New York Times Book Company, Inc.

GARETH WYNN-WILLIAMS ("The Newest Stars in Orion") is associate professor of astronomy and chairman of the graduate program in astronomy at the University of Hawaii at Manoa. Born in London, he was graduated with a B.A. in the natural sciences from the University of Cambridge in 1966. He taught physics for a year at University College in Nairobi before returning to Cambridge to do graduate work in radio astronomy under Martin Ryle. He was made a fellow of Trinity College in 1969 and received his Ph.D. in 1971. From 1971 to 1973 he did postdoctoral work at the California Institute of Technology, and from 1973 to 1978 he was a member of the radio-astronomy group at the Cavendish Laboratory in Cambridge. Wynn-Williams moved to Hawaii three years ago.

GARY A. STROBEL and GERALD N. LANIER ("Dutch Elm Disease") are long-distance collaborators, specializing in different aspects of the biology and control of Dutch elm disease. Strobel is R. G. Gray Professor of Plant Pathology at Montana State University. He obtained his doctorate in plant pathology from the University of California at Davis in 1963. His research tends to focus on the biochemistry of the interactions between plant parasites and the host plant. Strobel became interested in working on Dutch elm disease in 1974 after a short stint at the Connecticut Agricultural Experiment Station in New Haven. Lanier is professor of forest entomology at the State University of New York College of Environmental Science and Forestry in Syracuse. His graduate work was done at the University of California at Berkeley, where he earned a Ph.D. in 1967; the subject of his doctoral thesis was the chemical ecology of bark beetles. Before moving to New York, Lanier was an investigator at the Canadian Forestry Laboratory in Calgary, Alberta.

JAMES A. LAKE ("The Ribosome") is professor of molecular biology at the University of California at Los Angeles. His degrees are in physics: a B.A. from the University of Colorado at Boulder in 1963 and a Ph.D. from the University of Wisconsin at Madison in 1967. He writes: "It was not until I had completed my course work and examinations in theoretical physics that I decided to focus my efforts on molecular biology. I thought that understanding the molecular mechanisms of life would provide more satisfaction than pursuing physics. Since then my research has centered on exploring the structure and function of the ribosome." Before joining the faculty at U.C.L.A. in 1976 he was associated with the Massachusetts Institute of Technology, the Harvard Medical School, Rockefeller University and the New York University School of Medicine. In 1975 Lake was the recipient of the first Burton Award of the Electron Microscopy Society of America for his work on the structure of the ribosome.

S. L. SEMIATIN and G. D. LAHOTI ("The Forging of Metals") do research on the technology of metalworking at the Battelle Memorial Institute in Columbus, Ohio. Semiatin received degrees from Johns Hopkins University and Carnegie-Mellon University. He worked for a short time at Armco, Inc., before joining Battelle, where he investigates the forming properties of metals; his work is sponsored by a variety of private companies and Government agencies. Lahoti, a graduate of the University of Burdewan in India and the University of California at Berkeley, specializes in the mathematical modeling of metalworking processes. Since joining Battelle he has worked on a number of projects aimed at making such computer-assisted simulation techniques more available to industry.

PETER J. WEBSTER ("Monsoons") is principal research scientist in the Division of Atmospheric Physics of the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia. He is also an adjunct faculty member in the department of mathematics at Monash University in Melbourne. A native of England, Webster began working as a meteorologist for the Commonwealth Bureau of Meteorology in Australia in 1961. He studied physics for a year at the Royal Melbourne Institute of Technology before enrolling as a graduate student at the Massachusetts Institute of Technology, where he received a Ph.D. in meteorology in 1971. He taught atmospheric sciences at the University of California at Los Angeles and the University of Washington until 1977, when he returned to Australia. A specialist in tropical meteorology, he has also done research on planetary atmospheres and ancient climates. As a member of the National Academy of Science's Monsoon Experiment Committee, he organized and led the first phase of that multistage field project in 1978 and 1979. Earlier this year Webster served as a visiting professor at the Tata Institute of Fundamental Research in India; next year he has a one-year visiting appointment as G. J. Haltiner Research Professor at the U.S. Naval Postgraduate School in Monterey, Calif.

P. C. WOODMAN ("A Mesolithic Camp in Ireland") is assistant keeper of antiquities at the Ulster Museum in Belfast. His Ph.D. is from the Queen's University of Belfast. He writes: "Although my main interest has been the problems associated with the postglacial colonization of Ireland, I have a broader concern for how prehistoric man used and altered the landscape in Ireland. To this end I have been involved in the direction of a regional project in Glencloy, one of the Antrim glens. Besides the excavation of a series of sites ranging in date from Mesolithic to Late Bronze Age, this project involves a detailed study of how prehistoric man used the resources of this small area and how the landscape changed as a result."

S. S. WILSON ("Sadi Carnot") is professor of engineering at the University of Oxford. He was graduated from Oxford in 1944 and joined the faculty there in 1946. An expert on thermodynamics, he was responsible for the design and equipment of the present heat-engine laboratory at Oxford. His research, he reports, is concerned with the use of heavy working fluids in vapor-based power cycles, a development he sees as an alternative to steam in small power plants. In addition he is actively engaged in the design of pedal-powered machines for transportation and stationary applications in both developing and developed countries. One of his designs is an improved rickshaw chassis called the Oxtrike. Wilson was the author of the article "Bicycle Technology," which appeared in the March 1973 issue of SCIENTIFIC AMERICAN.

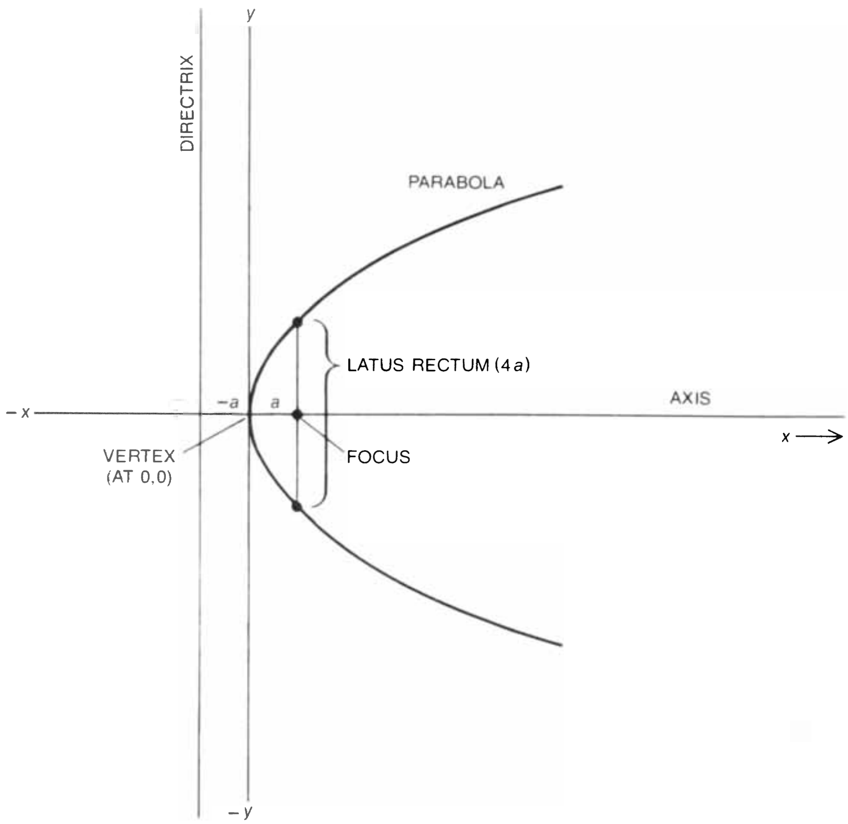
MATHEMATICAL GAMES

The abstract parabola fits the concrete world

by Martin Gardner

Mathematicians are constantly constructing and exploring the properties of abstract objects only because they find them beautiful and interesting. Later, sometimes centuries later, the objects may turn out to be enormously useful when they are applied to the physical world. There are no more elegant examples of this than the work done in ancient Greece on the four conic-section curves. Earlier columns of mine have dealt with three of them: circles, ellipses and hyperbolas. This time we take a look at parabolas.

If a right circular cone is sliced by a plane parallel to its base, the cross section is a circle. Tip the plane ever so slightly and the section becomes an ellipse, the locus of all points with distances from two fixed points (foci) that have a constant sum. Think of the circle as an ellipse with foci that have merged to become the center of the circle. As the cutting plane tips at progressively steeper angles, the two foci move progressively farther apart and the ellipses become progressively more "eccentric." When the plane is exactly parallel to the side of the cone, the cross section is a parabola. It is a limit curve, like the circle, only now one focus has vanished by moving off to infinity. It is an ellipse, as Henri Fabre once put it, that "seeks in vain for its second, lost center."



A parabola with its vertex at its origin has the type equation $y^2 = 4ax$

As you follow the parabola's arms toward infinity, they get progressively closer to being parallel without ever making it except at infinity. Here is how Johannes Kepler put it in a discussion of conic sections:

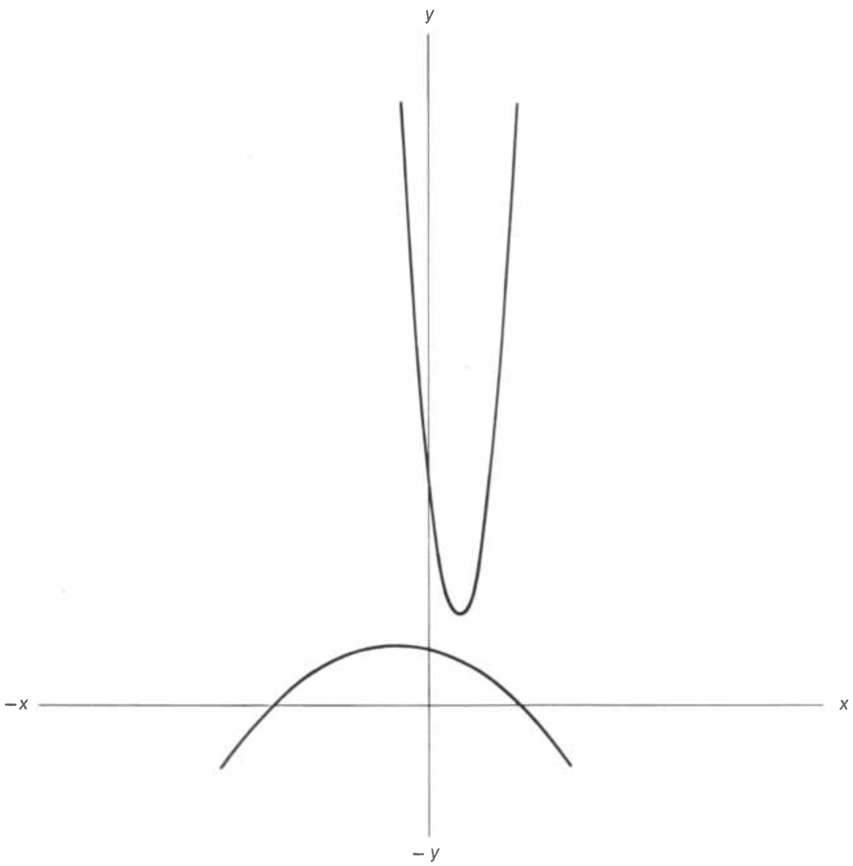
"Because of its intermediate nature the parabola occupies a middle position [between the ellipse and the hyperbola]. As it is produced it does not spread out its arms like the hyperbola but contracts them and brings them nearer to parallel, always encompassing more, yet always striving for less—whereas the hyperbola, the more it encompasses, the more it tries to obtain."

A parabola is the locus of all points in a plane whose distance from a fixed line (the directrix) is equal to its distance from a fixed point (focus) not on the line. The illustration on this page shows the traditional way of graphing a parabola so that its Cartesian-coordinate equation is extremely simple. Note that the axis of the parabola passes through the focus at right angles to the directrix, and that the tip of the curve, called the vertex, is at the 0,0 point of origin. The chord passing through the focus, perpendicular to the axis, is the parabola's latus rectum, or focal length. Let a be the distance from the focus to the vertex. Obviously a is also the distance from the vertex to the directrix. It is not hard to prove that the latus rectum must be $4a$. The parabola can now be described as the locus of all points on the Cartesian plane given by the parabola's type equation: $y^2 = 4ax$. If $y^2 = x$, the vertex of the parabola is on the directrix. More generally, any quadratic equation of the form $x = ay^2 + by + c$, where a is not zero, graphs as a parabola, although not necessarily a parabola positioned like the typical one that is shown in the illustration.

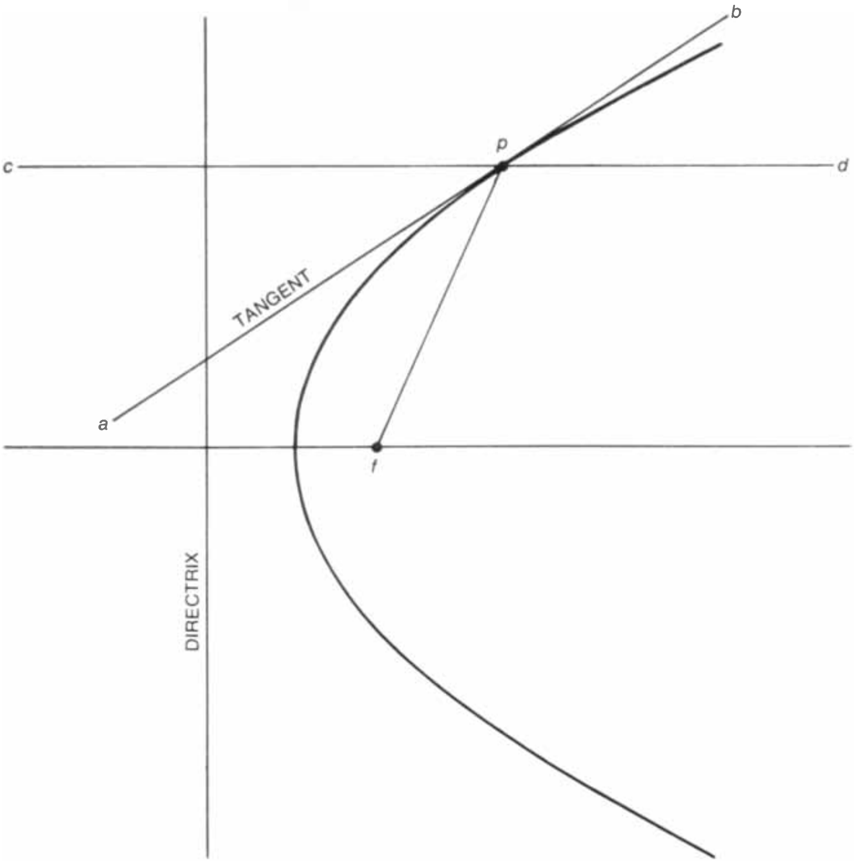
A surprising property of the parabola is that all parabolas have the same shape. To be sure, pieces of parabolas, like the two shown in the top illustration on page 18, have different shapes. If you think of either segment as being extended to infinity, however, you can take the other one, make a suitable change of scale and then place it somewhere on the infinite curve where it will exactly fit.

This property of varying only in size is one the parabola shares with the circle, although not with ellipses and hyperbolas. All circles are similar because single points are similar. All parabolas are similar because all pairs of a line and a point not on the line are similar. To put it another way, any directrix-focus pair will coincide with any other by a suitable dilation, translation and rotation. Any parabola, drawn on graph paper of the right size and properly positioned, can be given any desired quadratic formula of the form $x = ay^2 + by + c$.

If you throw a stone horizontally, it follows a path close to a parabola, as



Parabolas $y = 5(x - 1)^2 + 3$ (top) and $y = -\frac{1}{8}(x + 1)^2 + 2$ are identical in shape when extended



The parabola's reflection property

Galileo proved in 1609, although he did not publish his proof until 30 years later. You can repeat one of Galileo's experiments by rolling a marble dipped in ink sideways across an inclined plane. If the plane is covered with graph paper, the path recorded by the marble will enable you to calculate the curve's parabolic formula. In actual practice the parabolic path of a projectile is slightly distorted by the earth's roundness and much more by air resistance.

In Galileo's *Dialogues concerning Two New Sciences* he discusses at length the distorting influence of both the earth's not being flat and the air's viscosity. It is amusing to note that he discounts the deviation caused by the earth's roundness as being negligible because the range of military projectiles "will never exceed four miles."

The resistance of air to the flight of a bullet gives rise to a trajectory so resembling the curve of a breast that one of Norman Mailer's army officers, near the end of his novel *The Naked and the Dead,* sketches several pictures of the curve and muses:

"That form . . . is the fundamental curve of love, I suppose. It is the curve of all human powers (disregarding the plateau of learning, the checks upon decline) and it seems to be the curve of sexual excitement and discharge, which is after all the physical core of life. What is this curve? *It is the fundamental path of any projectile,* of a ball, a stone, an arrow (Nietzsche's arrow of longing) or of an artillery shell. It is the curve of the death missile as well as an abstraction of the life-love impulse; it demonstrates the form of existence, and life and death are merely different points of observation on the same trajectory. The life viewpoint is what we see and feel astride the shell, it is the present, seeing, feeling, sensing. The death viewpoint sees the shell as a whole, knows its inexorable end, the point toward which it has been destined by inevitable physical laws from the moment of its primary impulse when it was catapulted into the air.

"To carry this a step further, there are two forces constraining the projectile to its path. If not for them, the missile would forever rise on the same straight line. ⁄ These forces are gravity and wind resistance and their effect is proportional to the square of the time; they become greater and greater, feeding upon themselves in a sense. The projectile wants to go this way ⁄ and gravity goes down ↓ and wind resistance goes ←. These parasite forces grow greater and greater as time elapses, hastening the decline, shortening the range. If only gravity were working, the path would be symmetrical

it is the wind resistance that produces

the tragic curve

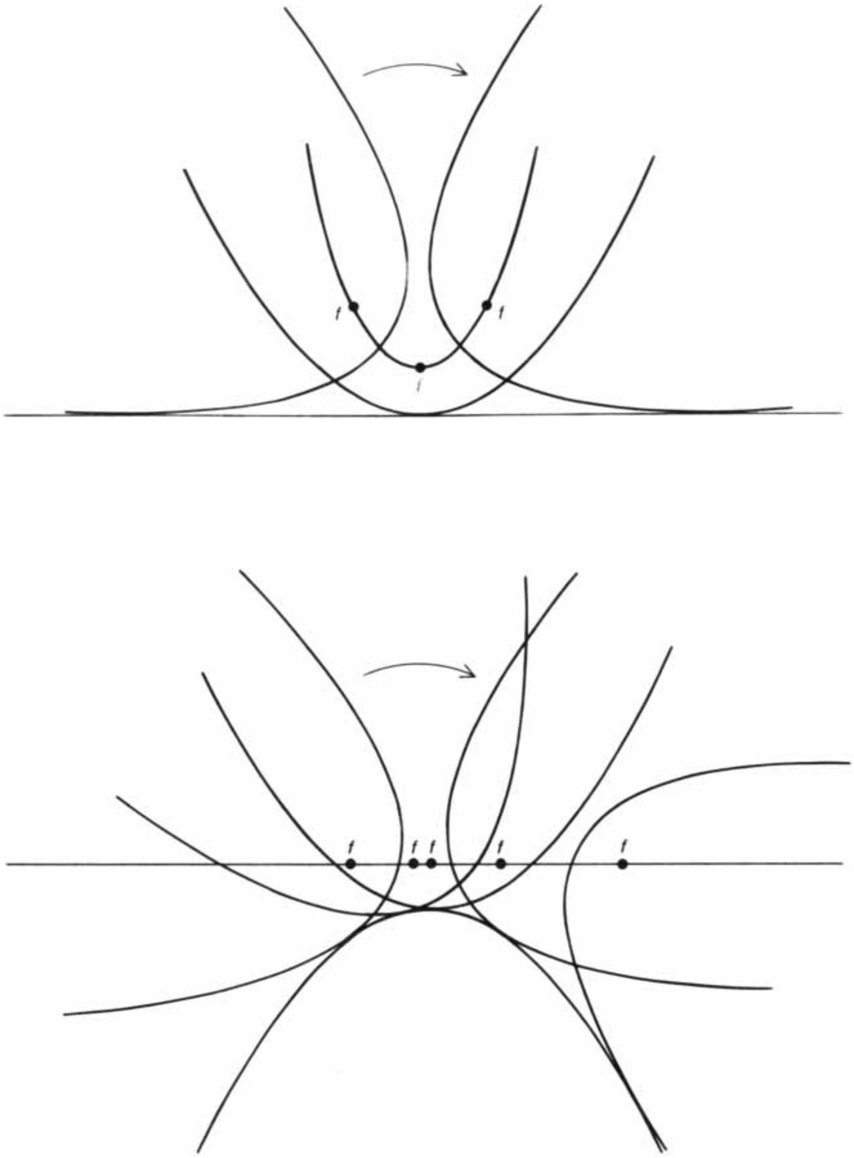
"In the larger meanings of the curve, gravity would occupy the place of mortality (what goes up must come down) and wind resistance would be the resistance of the medium . . . the mass inertia or the inertia of the masses through which the vision, the upward leap of a culture is blunted, slowed, brought to its early doom."

A jet of water from a hose also follows an almost perfect parabola. If when you water a lawn you slowly lower the angle of the hose jet from the near vertical to the near horizontal, the highest point along the series of parabolic curves as it moves away from you in a plane will also trace a parabola through the air.

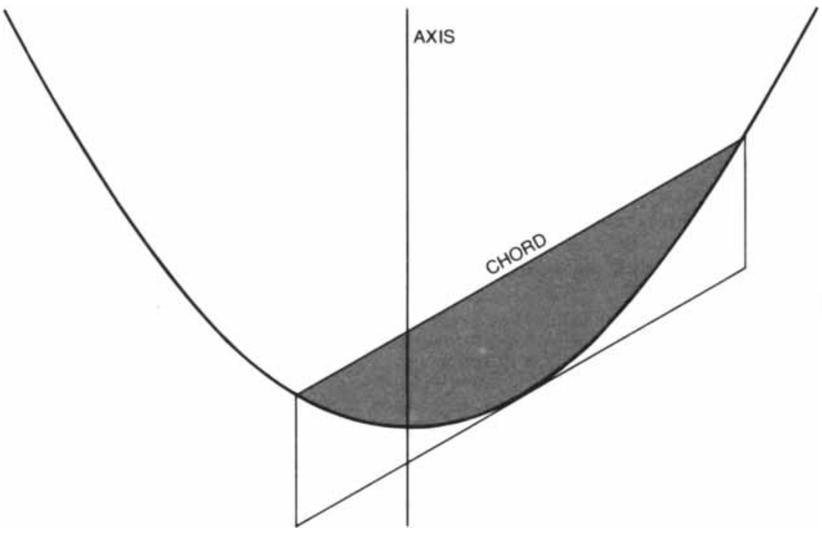
Some comets may follow parabolic paths. Comets that return periodically to the solar system move along extremely eccentric elliptical paths, but (as we have seen) the more eccentric an ellipse is, the closer it resembles a parabola. Since the parabola is a limit between the ellipse and the hyperbola, it is almost impossible to tell from observing a comet near the sun whether it is following an extremely eccentric ellipse (in which case it will return) or a parabola or hyperbola (in which case it will never return). If the path is parabolic, the comet's velocity will equal its escape velocity from the solar system. If the velocity is less, the path is elliptical; if it is more, the path is hyperbolic.

The parabola's outstanding applications in technology are based on the reflection property displayed in the bottom illustration on the opposite page. Draw a line from the focus *f* to any point *p*, and draw tangent *ab* to the curve at *p*. A line *cd*, drawn through *p* so that angle *apf* equals angle *bpd*, will be perpendicular to the directrix. It follows that if the parabola is viewed as a reflecting line, any ray of light from the focus to the curve will rebound along a path parallel to the curve's axis.

Imagine now that the parabola is rotated about its axis to generate the surface called a paraboloid. If light rays originate at the focus, the paraboloid will reflect the rays in a beam parallel to the axis. That is the principle behind the searchlight. Of course the principle also works the other way. Parallel rays of light, shining into a concave mirror with a paraboloid surface, will all be directed toward the focus. That is the secret of reflecting telescopes, solar-energy concentrators and microwave receiving dishes. Because large paraboloid mirrors are easier to build than transparent lenses of comparable size all giant telescopes are now of the reflecting type. Other optical devices serve to bring the image from the focus to an eyepiece or a photographic plate. As a child you may



***Rolling parabolas generate a catenary* (top) *and a straight line* (bottom)**



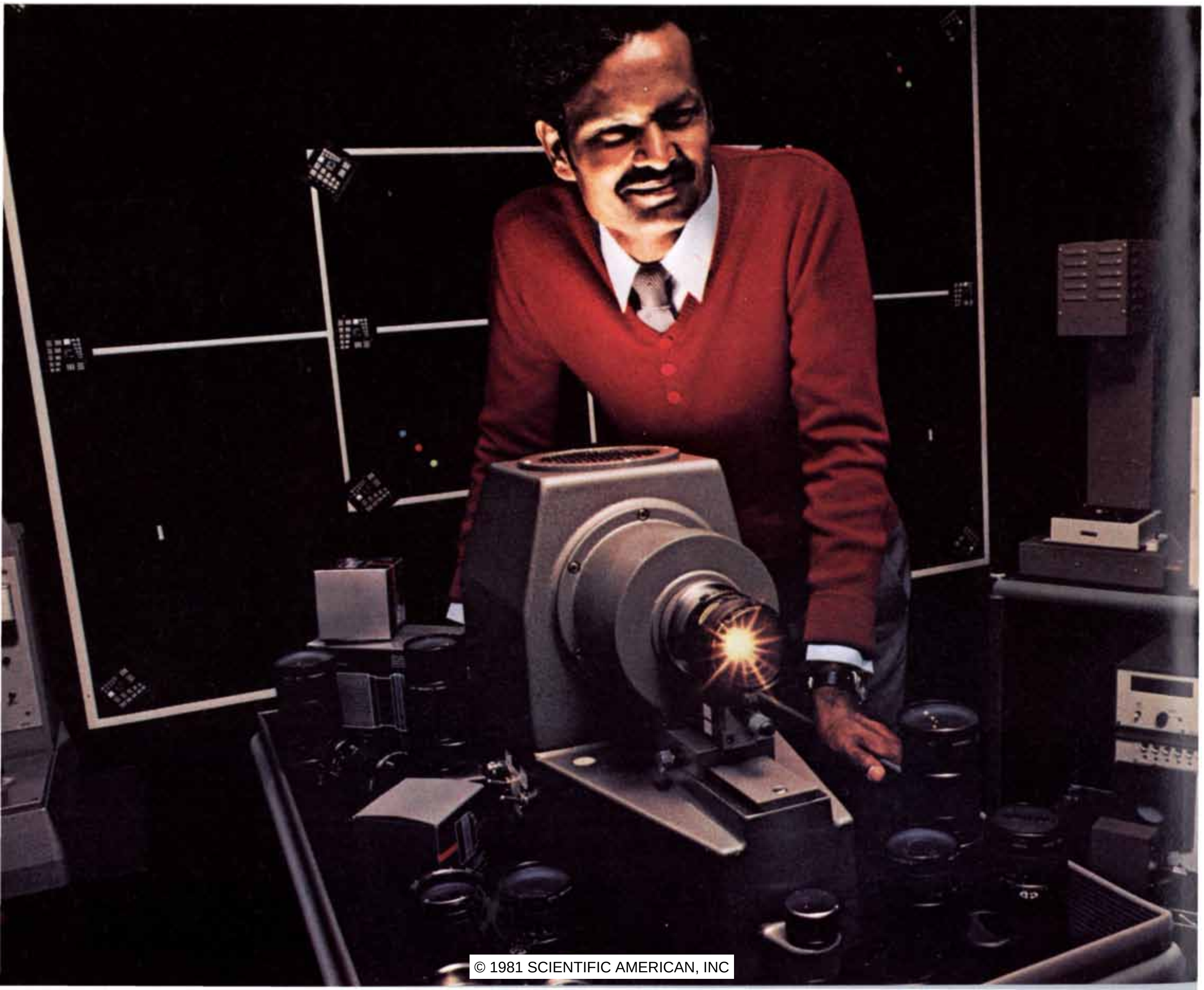
Area of the shaded segment is 2/3 that of the parallelogram

Reddy Chirra improves his vision with an Apple.

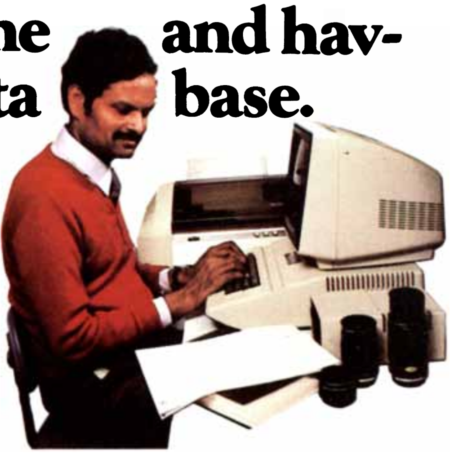
Reddy is an optical engineer who's used to working for big companies and using big mainframes.

But when he started his own consulting business, he soon learned how costly mainframe time can be. So he bought himself a 48K Apple II Personal Computer.

And, like thousands of other engineers



and scientists, quickly learned the pleasures of cutting down on shared time and having his own tamper-proof data base.



His Apple can handle formulas with up to 80 variables and test parameters on 250 different optical glasses.

He can even use BASIC, FORTRAN, Pascal and Assembly languages.

And Apple's HI-RES graphics come in handy for design.

Reddy looked at other microcomputers, but chose Apple for its in-depth documentation, reliability and expandability.

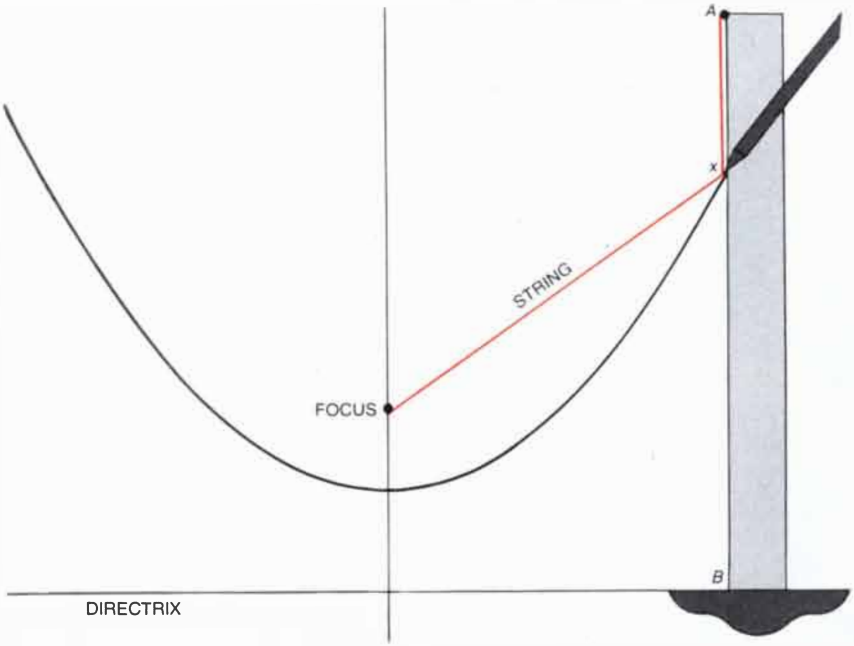
You can get up to 64K RAM in an Apple II. Up to 128K RAM in our new Apple III. And there's a whole family of compatible peripherals, including an IEEE-488 bus for laboratory instrument control.

Visit your authorized Apple dealer to find out how far an Apple can go with scientific/technical applications.

It'll change the way you see things.

The personal computer. apple®

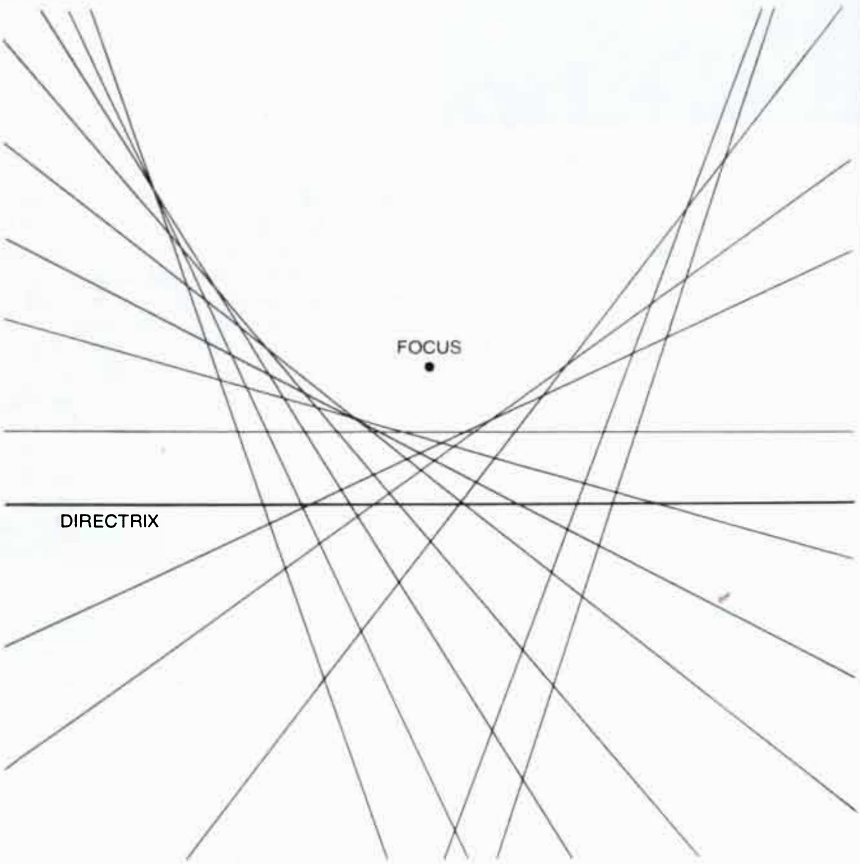
For the dealer nearest you, call (800) 538-9696. In California, call (800) 662-9238. Or write: Apple Computer Inc., 10260 Bandley Dr., Cupertino, CA 95014.



Johannes Kepler's string method of constructing a parabola

have learned how to set fire to a piece of paper by focusing the sun's rays with a glass lens. It can be done just as efficiently with a paraboloid mirror, with the paper held at the surface's focus.

If a pan of water is rotated, the surface of the water forms a paraboloid. This suggested to the physicist R. W. Wood that perhaps a reflecting telescope could be made by rotating a large dish of mercury and exploiting the paraboloid surface as a mirror. He actually built such a telescope, but there were so many difficulties in making the surface suffi-



How to fold a parabola

ciently smooth that the idea had to be abandoned.

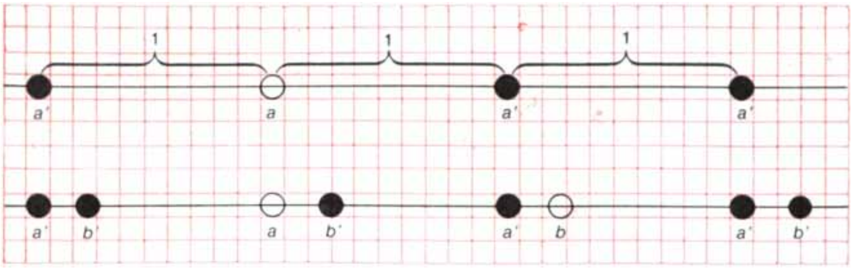
Assume that a paraboloid has a flat base perpendicular to its axis, so that it looks like a rounded hill. How do you calculate its volume? Archimedes found the amazingly simple formula. The volume is precisely 1.5 times that of a cone with the same circular base and the same axis.

A parabola is closely approximated by the cables that support a suspension bridge. The curve is distorted if the weight of the bridge is not uniform or if the weight of the cables is great in relation to that of the bridge. In the latter case the curve is hard to distinguish from the one known as a catenary (from the Latin *catena,* chain). Galileo mistakenly thought the curve formed by a chain suspended at the ends was a parabola. Decades later it was shown to be a catenary, a curve that is not even algebraic because its equation contains the transcendental number *e.*

There is a curious relation between the parabola and the catenary that is not well known. If you roll a parabola along a straight line, as is shown at the top of the top illustration on page 19, the "locus of the focus" is a perfect catenary. Perhaps more surprising (although it is easier to prove) is what happens when two parabolas of the same size are placed with their vertexes touching and one is rolled on the other as is shown at the bottom of the top illustration on page 19. The focus of the rolling parabola moves along the directrix of the fixed parabola, and its vertex traces a cissoid curve!

One of the earliest problems concerning parabolas was that of "squaring" the area of a section of the curve bounded by a chord, such as the shaded region in the bottom illustration on page 19. The problem was first solved by Archimedes in his famous treatise *Quadrature of the Parabola.* By an ingenious method of limits that anticipated the integral calculus, he was able to prove that if you circumscribe a parallelogram as is shown in the illustration, with its sides parallel to the parabola's axis, the area of the parabolic segment is 2/3 that of the parallelogram. (Archimedes first guessed this by comparing the weight of the parallelogram with the weight of the segment.) Archimedes also used the parabola for an elegant way to construct the regular heptagon. Earlier geometers had exploited parabolas for the classic task of doubling the cube: constructing a cube with twice the volume of a given cube.

There are many techniques for drawing parabolas without having to plot myriads of points on a sheet of paper. Perhaps the simplest relies on a T square and a piece of string. One end of the string is attached to a corner of the T square's arm as is shown in the top illustration on the preceding page, and the



Impossibility proof of straight-line conjecture

other end is attached to the parabola's focus. The string must have a length *AB*. A pencil point at *x*, pressing against the arm of the T square, keeps the string taut. As the T square slides along the directrix, the pencil moves up the side of the T square to trace the parabola's right arm. Reflecting the arrangement to the other side draws the left arm. This method was invented, or possibly reinvented, by Kepler. You can work with a right triangle or a rectangle instead of a T square and slide it along the edge of a ruler. It is easy to see that the string's constant length ensures that any point on the curve is equally distant from the focus and the directrix.

Lovely parabolas can be produced more easily by paper folding. Just mark a focus point anywhere on a sheet of translucent paper, rule a directrix and fold the sheet many times so that the line goes through the point each time. Each crease will be tangent to the same parabola, outlining the curve shown in the bottom illustration on page 23. If the paper is opaque, use one edge of the sheet for the directrix, folding it over to meet the point.

Familiarity with parabolas can often provide quick answers to algebraic questions. Consider, for example, this pair of simultaneous equations involving the two lucky numbers of craps:

$$x^2 + y = 7$$
$$x + y^2 = 11$$

It does not take long to discover that if x equals 2 and y equals 3, both equations are satisfied. Let us ask two questions:

1. Are there other integral solutions (where x and y are whole numbers, positive or negative)?

2. How many solutions are there altogether?

A more difficult problem, for which a parabola gives an answer, comes from Ronald L. Graham of Bell Laboratories, who both framed and solved it. It is published here for the first time. Imagine that you have an infinite supply of identical disks with a diameter less than 1/2 of some unit distance, say 1/10. Is it possible to put all of them in a plane, without their overlapping, so that no distance between any pair of points in the disks is an integer?

Since each disk is only 1/10 in diameter, no two points on the same disk can be an integral distance apart, but it is conceivable that by a clever arrangement of disks, with their centers on a straight line, no point in one disk will be an integral distance from a point in any other disk. It is not difficult to prove that it is impossible. In fact, any arrangement of an infinity of disks on a straight line will create an infinity of pairs of disks in which an infinity of pairs of points (all exactly on the line) will be separated by integral distances.

To see how it works suppose you have placed a disk of diameter 1/10 on the line as is shown by the unshaded circle ***a*** at the top of the illustration above. The shaded circles (marked ***a'***) are spaced with their centers a unit distance apart, extending to infinity in both directions. Clearly no second disk can be put on the line where it overlaps or touches a shaded circle, otherwise the disk will contain a point on the line that is an integral distance from a point on the line in disk ***a***. (We assume that points on the circumference of a disk are in the disk.)

It is possible, of course, to put a second disk on the line between any pair of adjacent circles provided it does not touch or overlap either circle. For example, a second disk ***b*** can be placed as is shown at the bottom of the illustration above. At once the line acquires another infinite set of circles (shaded and labeled ***b'***) at regular unit spacings, indicating that no third disk can be put where it overlaps or touches them. The same holds for additional disks. Since no more than eight disks will go without touching or overlapping in the finite space separating any pair of the first set of circles, it follows that no more than nine disks can be put on the line. A 10th disk, added anywhere, will contain an infinity of points that are integral distances from points in one of the nine previously positioned disks.

The proof generalizes in an obvious way to all disks smaller than 1 in diameter. If the denominator of the diameter is an integer, subtract 1 to get the maximum number of disks that can be positioned. If the denominator is not an integer, round it down to the nearest integer. Thus no disk of diameter 1 can be used. Only one disk of diameter 1/2 or diameter $1/\sqrt{2}$ can be put on the line, only two disks of diameter 1/3, only three of diameter 1/4 or diameter $1/\pi$, only four of diameter 1/4.5 and so on.

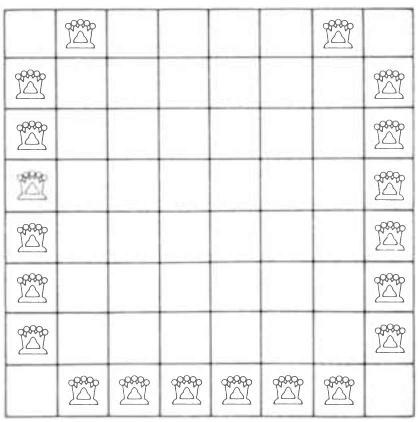
Although the problem cannot be solved by a straight line, it can be solved by a parabola. In my next column I shall give Graham's solution and also the answers to the preceding questions.

In June readers were asked to place 20 queens on a chessboard so that each queen attacks exactly four others. A solution is given in the illustration below. It is not known whether more than 20 queens can be put on the board to meet the same condition.

A letter from Dennis Weeks called attention to an error in my February column on modulo arithmetic. I misquoted Edward Waring as saying Wilson's theorem would never be proved. The correct translation of his Latin remark is that the theorem would be very difficult to prove. My statement that in 1582 the vernal equinox was creeping closer to winter meant that the traditional date for the equinox was doing this, not the equinox itself.

And when I said that April 19 was "the most frequent" Easter date, I was relying on reported results for a particular short span of time. It is not that simple. Several readers used computers for checks of other time spans, with variable results. James LaBossiere, Richard Morrish and others, independently checking the period covered by Thomas H. O'Beirne's algorithm (1900 to 2099), found that March 31, April 12 and April 15 tied for the most likely date, with March 22 not applying at all during the period.

Thomas L. Lincoln, L. V. Larsen and others investigated additional intervals, with conflicting results. Lincoln believes that in the long run, extending the present rules for Easter into the future, April 19 will become the commonest date, but this remains unconfirmed. I shall try to report any new developments on this apparently unsolved problem.



Best-known solution to chess-queens problem

Finally, high-fidelity television with stereo sound.

Once every generation it seems, something comes along that changes the entire nature of home entertainment. That time has come again.

Pioneer introduces LaserDisc,™ a whole new technology that makes your television and stereo do things they've never done before. Now you can watch a movie or hear and *see* a concert as easily as playing a record.

And when you hook up the LaserDisc player to your TV and hi-fi, for the first time, along with a television picture you get true stereo.

And a television picture that's actually a 40% sharper picture than home videotape.

And because the disc is read by a light beam rather than a needle or video head, nothing touches the disc. So, with normal use, it doesn't wear out from play. You can enjoy the disc forever.

But Pioneer LaserDisc offers you more than superb fidelity, it's truly the first form of personal entertainment. Your television viewing is no longer limited to what's on television. Because you can see what you want to see when you want to see it. A whole library of Pioneer Artists and MCA/Discovision discs with movies like Jaws, Animal House, The Blues Brothers; concerts like Paul Simon, Liza Minelli, Loretta Lynn, and Abba; sports like NFL football, and title fights.

You can study the standard-play discs in fast motion, slow motion, stop motion. Even create your own instant replay.

But there's something even more remarkable. A micro-computer built into the Pioneer LaserDisc player lets you access any individual frame on the disc at the touch of a button. You can go right to your favorite scene in a movie, or a song in concert in seconds.

And because you can view the up to 108,000 frames on the disc one frame at a time, it just may be the future of education.

The cost of Pioneer LaserDisc is surprising as well. Just $749* (with $50* more for optional remote control). And a full-length movie on LaserDisc can cost less than taking your family out to the movies.

The only way to even begin to experience the magic of this remarkable LaserDisc system is to see it. And we've arranged for that. Just call at **800-621-5199**** for the store nearest you. **(In Illinois 800-972-5855).

*Suggested retail price. Actual price set by dealers.

Simulated TV picture from Paul Simon.

"There are still huge reserves of American oil left behind in oil fields that were tapped thirty or forty years ago," says Gulf Area Engineer John Shamburger. "In those days, crude was only two or three dollars a barrel, so we would operate an oil well only as long as the oil could be pumped up easily and inexpensively. It made more sense to drill new wells than to squeeze every drop out of old ones."

"In many of these old wells, the underground pressure is so low that the oil won't come up without some extra help. The extra help is called tertiary recovery, and it's a whole technology in itself. Here at Heidelberg, Mississippi, we started experimenting with it in 1971.

"One way to build up underground pressure is to set fire to some of the oil deep in the earth. That's called fire-flooding. It drives the rest of the oil toward the well, where it can be pumped to the surface. The improvement can be pretty dramatic; in one case, production increased from one hundred to 2300 barrels per day.

"In other parts of the country with different geology, we might use water, steam, or chemicals to get the oil flowing."

All over America, Gulf is using one or another of these technologies to get more oil out of old wells. It isn't cheap and it isn't easy, but it is a certain source of made-in-America energy for tomorrow.

Gulf

Gulf people: energy for tomorrow.

"Gulf's underground heat treatment gets more oil from old oil wells."

BOOKS

Gossamer technology, the polywater episode, the Aleuts, Australian and African mammals

by Philip Morrison

GOSSAMER ODYSSEY: THE TRIUMPH OF HUMAN-POWERED FLIGHT, by Morton Grosser. Houghton Mifflin Company ($14.95). THE GOSSAMER CONDOR AND ALBATROSS: A CASE STUDY IN AIRCRAFT DESIGN, by J. D. Burke. American Institute of Aeronautics and Astronautics, 1290 Avenue of the Americas, New York, N.Y. 10104 ($15). Flight demands power. The little kitelike hang glider, sinking slowly and gracefully down from a hilltop with its intrepid pilot, has about one kilowatt of gravity power. The Wright Flyer I at Kitty Hawk in 1903 had almost 10 times as much engine power as that. But even a champion athlete can sustain for a few minutes or more a muscular output of only .3 or .4 kilowatt; the kilowatt range is possible for human beings only over a few seconds. To fly by human muscle power alone over mile-long routes demands a flying machine with an ultralow power requirement. Naive and intuitive design hardly suffices; not since Daedalus' escape has there been a well-witnessed report of sustained human-powered flight in any aircraft other than one designed by professionals cognizant of the state of the art.

That art has now matured. Perhaps 500 men and women have taken off to fly under their own physical power, mainly through the twilight-still air at dawn or dusk. A few are adept, and one young man, Bryan Allen, a superb pilot-engine, has flown the English Channel. His bicycle-racing experience was tuned to a high power-to-weight ratio by a disciplined training program under the direction of a Long Beach, Calif., professor of exercise physiology. That Channel crossing consumed two and a half hours of Allen's output, fueled by a breakfast of apples, bananas and a large sweet roll, probably not the equivalent of the nine standard megajoule jelly doughnuts his energy budget calls for. Allen is not only a high-performance engine with stamina but also a smooth and perceptive pilot of the aircraft so aptly named Gossamer.

Without underestimating the pilot-engine, success in human-powered flight demands much more. Those flights were the triumph of incentive, of enthusiasm, of a team of devoted craftsmen, a coherent organization of almost 50 men and women from the southern California center of high technology and its know-how. A common background in model aircraft, soaring and high-speed bicycles united the team as much as vocation, leadership and the hope for a rich prize. It is Paul MacCready, Jr., of Pasadena who led the Gossamer Squadron, as the conceptual designer of the aircraft and the inspired organizer of the entire effort. He himself was a national champion of indoor gram-weight model aircraft in his high school years, an international soaring champion at 30 (then with a fresh California Institute of Technology doctorate on atmospheric turbulence) and a convert to hang gliding of a decade's standing.

It was in mid-1976 on a family vacation to Florida that MacCready, watching the circling birds, came to realize that the right way to human-powered flight was through a giant hang glider: very slow flight in a plane of extreme wingspan and minimal weight. For almost two decades in a dozen countries around the world university engineering departments and designers of aircraft, moved both by the old dream of unaided flight and by the £50,000 Henry Kremer Prize, brilliantly defined since 1960 by the Royal Aeronautical Society, had been at work. Almost unanimously they chose to construct elegant cantilever-winged craft, looking like efficient sailplanes with a small propeller. The British or Japanese aircraft resembles a winged cello; the MacCready versions look like huge stringed, transparent kites. "It is a safe bet," says one of these authors, that if the designers had been offered the winner's plans, they would have rejected them with "amusement and disdain—that primitive thing with all the wires?"

Structure is the key. Low speed and very large wingspan alone allow flight with so little power; the wingspan of these aircraft matches that of a twin-jet airliner, yet the *Gossamer Albatross* weighed just 55 pounds empty, its dawn dew load excepted. The form is that of a standard hang glider, or even of the one-gram model of 40 years ago; main elements extend along three axes like a child's jack, strong steel tension wires bind and support the huge wings and a big longitudinal boom from one central vertical king post. The slogan of the structural design is: "If it doesn't break, it's too heavy."

The Channel plane was almost all modern plastic; it has strong carbon-fiber reinforced epoxy spars, made by wrapping the du Pont–supplied stuff around aluminum tubes, curing it in an oven and dissolving away the metal forms. Only the main tension wires of stainless steel and some bicycle driving parts such as sprockets and the propeller shaft are metallic; polystyrene foam, strong Kevlar lashings and weavings, and a taut Mylar skin make the aircraft reliable, repairable, beautiful. It is even easy to take it down and reassemble it for practical transport.

Structure alone, however, is not flight. Aerodynamics is the second key: the competitors never managed to make a usable turn until the seventh Japanese design, just before the Gossamer success. It took the Californians one year to win the first Kremer Figure-of-Eight Prize. After much gloomy experience MacCready realized—the equations and the computer programs quickly confirmed it—that control of the Gossamer craft lay along a surprising new path. A craft so large and light must accelerate whenever it turns a mass of air much greater than its own physical mass. It is no bird but a dragonfly; the simplest model is a glider flying underwater. The end result was a smooth and easy turn, but one made by twisting the wing in a direction exactly opposite to the conventional direction known since the Wrights. Besides MacCready, three gifted professional aerodynamic designers played a part. One produced the special low-speed airfoil form, another (at M.I.T.) the efficient propeller for the Channel aircraft and a third the incisive control study.

The volume by Morton Grosser is a personal narrative of the entire story, background and context included. Dr. Grosser is a proved and professional historian of science; his early involvement is one more happy sign of the rule of right reason that this enterprise celebrates. The second book is by another professional on the team, a veteran pilot-engineer and NASA planner, "an aeronautical Shackleton... you could trust to get you home from Elephant Island or Phobos." It is a briefer design account, centered on technical progress, although without mathematics. Both books offer the wonderful line drawings made by the British technical artist Pat Lloyd, exploded details and clever views of the aircraft; the reproductions are larger in the pages of the paperbound Burke book, although the many photographs in the Grosser book much enrich its popularly written, detailed and intimate chronicle.

It is hard to dissemble a glowing admiration for this victorious engineering. The Channel flight, for a new Kremer prize of £100,000, was fairly won by the *Gossamer Albatross* just eight months after the prize announcement. The craft's first overwater flight was the Channel crossing! That is mastery. To be sure, the importance of human-powered flight is not to be found in the marketplace or in the world of production or in the world of war. Even the sport is not likely to spread; these machines are too large, too fragile, too much at the mercy of the wind. A sport is indeed emerging around ultralight aircraft, but those craft will be smaller ones, with some power assistance for muscle whether from a battery, from silent sunlight on a square yard of photovoltaic cells or from a tiny model airplane engine, noisily drinking its pint of nitrated fuel. This tale is nonetheless important for the spirit. In our epoch many see rational technology chiefly as threat: a threat to life, a threat to work, a threat to the air itself. Playful engineering offers a healing exception; MacCready and his team are pioneers of this new form. We must laud also the distinguished engineers of the Royal Aeronautical Society whose farseeing proposals to encourage such impractical design were ringingly supported by the sporting engineer-industrialist Henry Kremer. This magnifico himself began his career with the design and fabrication of novel materials, including the plywood of the famous R.A.F. *Mosquito* of World War II.

POLYWATER, by Felix Franks. The MIT Press ($15). Almost 20 years ago Kurt Vonnegut published a brilliant ironic fantasy, *Cat's Cradle.* One of the flood of imaginative elements in the book is a polymorph of ice called ice-nine, "a crystal as hard as this desk." Ice-nine is imagined as being stabler than liquid water at room temperature. Once nucleated, all natural waters in physical contact would swiftly go over to ice-nine. "'And the oceans the frozen rivers fed?' 'They'd freeze, of course.'... 'And the rain?' 'When it fell, it would freeze into hard little hobnails of ice-nine and that would be the end of the world.'" So far we follow the novelist, who in fact ends his world in just this way.

The sober pages of *Nature,* the venerable London scientific weekly, offered a similar warning in 1969, over the signature of a competent and sincere physical chemist from Pennsylvania: "The polymerization of Earth's water would turn her into a reasonable facsimile of Venus.... Even as I write there are undoubtedly scores of groups preparing polywater.... Treat it as the most deadly virus until its safety is established." True, the commonsense rebuttal appeared not many weeks later. But the echo of ice-nine was very loud, and interest in anomalous water spread from surface-chemistry specialists and the "water club" to the world of the media, and to energetic theorists and experimenters from London to Sydney.

An epidemic of papers appeared; their unusual course in time is carefully graphed here. By 1972 some 400 authors had published on the exciting new substance, mostly in short letters in the weeklies, quick studies, quickly communicated and soon outmatched. The roar soon fell silent; its chief originator, the leading Russian surface chemist Boris V. Deryagin, a man of established experimental ability and boundless self-assurance, in whose laboratory no fewer than two dozen investigators became seized of the problem, conceded tersely in *Nature* in mid-1973: the anomalous properties reported and often confirmed over a near decade of claims "should be attributed to impurities rather than to the existence of polymeric water molecules."

Professor Franks is a weighty insider of the water club, long the director of an industrial research laboratory and the editor of a standard six-volume "comprehensive treatise" on water, completed in 1979. He never took part in the polywater quest, by happenstance of his career during those years, but he followed it all, "knowing many of those who wrote." In a lively, candid, plainly fair-minded way he chronicles the entire tale, from the prehistory of "bound water" on to the appraisals of the events as part of the sociology of science, perhaps of its pathology. The polywater paper count grew exponentially to a sharp peak, and it declined in the same way; among the 100,000 or so susceptible scientists—water has a universal niche in the world and so too in the laboratory—the fever infected its share. Immunity began to appear in about 1970.

There are a few key events. Water can indeed interact intimately with surfaces, and there it is somehow structurally modified. Far from fully understood, the phenomenon is well recognized in the mineral and the living kingdoms. When a worker in a provincial Russian laboratory reported in 1962 the growth of strangely dense liquid columns in fine capillary tubes, a tide began to rise. The Moscow Laboratory of Surface Forces took over massively. The Moscow group published a dozen papers; scientists in other countries were slow to notice them, but finally the extraordinary nature of the claims was understood. In a private London conversation with Deryagin, preserved on tape since 1968, the brilliant J. D. Bernal volunteered (perhaps with some sense of hospitality to a guest): "In my opinion this is the most important physical-chemical discovery of this century."

Even though the remark remained behind the scenes, the sense of marvel burst out in 1968. What was involved is a kind of Lamarckian process. It is no surprise to find water molecules marshaled by the forces from a solid surface nearby, a few molecule layers away. But to see that such modified water remains a new substance even remote from the solid, out within the macroscopic volume of a capillary, was remarkable. In the spring of 1969 an American spectroscopist of high repute published an infrared Raman spectrum of polywater and coined the supple name. (Polywater might not have filled all those columns of print under the name of anomalous water.) That paper was firm in tone, undoubting.

Then came the ice-nine comparison, and not long afterward the theorists of the molecular bond with their uncertain models of rings and cages, and equivocal quantum computations tailored unknowingly to fit the facts. The facts began to change, however, under the probes of the analytical chemists; meanwhile no one had managed to produce even an elfin teardrop of the rare new stuff. When it was maintained in an interview that polywater made from deuterium-bearing water yielded an unchanged spectrum, the bandwagon began to slow. Polywater drained away mainly before 1971. The last papers appeared in about 1975. In late 1978, as this book manuscript was being completed, *Science* carried an account of twin brothers in Kazakhstan who had "identified [a] form of water" having greater biological activity than ordinary tap water. Easily prepared in a milk pasteurizer, this substance increases the yield of cottonseeds, causes cattle to gain weight and makes up into stronger concrete. So far, though, no second plague has spread.

Professor Franks declines to draw any sharp moral from polywater. He makes it clear that water is a subject of power, that the cold war had a disturbing influence, particularly on the U.S. granting agencies because of their concerns for image. The establishment of science, with some hedging, was coolly skeptical, largely, as Franks sees it, for the wrong reasons: mere conservatism and naysaying, not superior insight or clear adherence to sound methods. It is certain from the materials cited that all skeptics did not make their excellent case according to the highest canons of criticism; some publications are mere derision and "lack of grace." Two referees' reports on a single proposal are worth sampling. Said one: "I presume [the agency] is more interested in supporting credible research than it is in having its name spelled right on incredible research." Said a second, to the same proposal, not even on polywater but by a polywater investigator: "All of chemistry... will be revolutionized by this type of work."

On balance it appears that the society of science weathered the storm of error well enough, with due allowance for excess, some waste and some personal misbehavior. The more general media

come off less well, but how could they not overstate a case that explicitly gave reality to ice-nine? It is good that not much personal animus remains; zealous questers for polywater fame do not appear to have "suffered unduly." A reader takes away rather a sense of the resilience of our net of fallible institutions. It would seem, however, that criticism by personal attack or by sheer force of rhetoric in the end has less social utility, as well as less corrective impact, than do serious efforts to account for influential but fallacious results. What spectrum was the one that was held to be from polywater? How did it arise? No consensus was reached.

The pop-cartoon dust jacket is a pleasure, surely new to the literature of physical chemistry.

ALEUTS: SURVIVORS OF THE BERING LAND BRIDGE, by William S. Laughlin. Holt, Rinehart and Winston ($5.95). Professor Laughlin is plainly and justly proud of the fact that the native village of Nikolski awarded him a small "piece of land on which to build a home." It was as a graduate student with the noted physical anthropologist Aleš Hrdlička that Laughlin first made his way to the windswept, treeless Aleutians. There he has worked on and off ever since, living among their people, the Aleuts. In 1974 he shared the work and the excitement of excavation by a joint Russian-American field party. This primer of Aleut life and its well-established lengthy history is one of a long series of brief case studies for students of anthropology. Laughlin's study is "more than an interdisciplinary *tour de force*." It reports, although very concisely, on the Aleuts as a whole, from the high incidence of three-rooted first molars to the Aleut semaphore signals invented by a few men before World War I (based on the Cyrillic-derived alphabet the Aleuts have used for a century and a half) and on to the Honda and walkie-talkie of today's Nikolski boatmen and hunters.

If that small house of Professor Laughlin's is ever finished, one will look from it out to the snowy flanks of the high volcanic cones across the bay, past a mile-long high island, transformed in August by a dense stand of deep-blue monkshood, loud with the bumblebee. On that island the excavators found acres of a stone-blade-working site, with a total inventory of some three million worked tools and cores, largely of obsidian. Dozens of fire hearths in the layered dig provided charcoal for the carbon-14 laboratories. The sequence is well dated and remarkable. People worked blades there 9,000 years ago. For 2,000 years the blade style remained unchanged: the tools were prismatic, blades and burins struck from a core of obsidian. The edges were very sharp but were flaked on one surface only. After that time both surfaces were worked, and bifaced tools show up along with stone lamps, dishes and grinding stones.

A couple of dozen house pits adjoin the blade site on higher ground. They represent a transitional culture, which can be followed for 2,000 years more. The same objects reappear on the Nikolski shore, in an oval green mound just south of the present village. That mound is the midden of Chaluka. Within this century it still had some residents; now they have all moved down to the flatter ground adjoining to the north. The midden is an ancient village; 4,000 years more can be traced in the debris, including the buttons and rings and skeletal remains of an identifiable Russian party, massacred in thriving old Chaluka in the spring of 1764. The continuous occupation of the shores of Nikolski Bay by the Aleuts is thus witnessed by their own bones, their particular artifacts slowly changing in style, the circumstantial stories of their tradition. Theirs is a freehold without a break over nine millenniums, the settlements moving as the sea level changed.

Why should Nikolski, apparently only one island a couple of hundred miles out along the chain from the long mainland finger of the Alaska Peninsula, be such a key site? The casual map reader overlooks the work of time. When the old Aleuts came to Nikolski, it was at the point of the mainland. Perhaps a narrow seawater channel or two, or a couple of glacier tongues, made the outward trip something less than dryshod. But Samalga Pass, the arm of the ocean that flows past the headland a few miles south of Nikolski, is the first permanent deep-ocean pass out along the line of islands. It was there that the Aleuts ended their passage along the shores of the Bering Sea, when the sea level was 100 meters lower than it is today, the great ice sheet holding the waters.

These people turned seaward along the peninsula as they came south down the coast; the Eskimo peoples instead diverged northward to follow the Arctic coast, in the end all the way to Greenland. The Samalga Pass channeled the marine life of the entire Bering Sea past the Aleut hunters, and the exposed reefs provided a sea-plenty; indeed, the permanent upwelling in the pass enriches the local fauna. The Aleuts have long used octopuses, sea urchins, shellfish and seaweed, and the lakes and freshwater streams of the islands have yielded up salmon for a long time. Aleut culture adapted to the marine wealth; less than 5 percent of their diet came from plants, although shorebirds and their eggs were an important contribution of dry land. Here the Aleut culture centered; it slowly spread both ways, out to the lonelier islands westward, and eastward back toward Alaska.

The Aleuts are superb kayak hunters of the open sea, able to take the wary sea otter regularly even during foul weather, taking the abundant whales with a somewhat greater dependence on good fortune. They knew they were "better than anyone else" at kayak hunting. Their kayaks are distinct; the designs show a specialized variety, a three-hatch model becoming the standard transport of the clerics and the watchful administrators during Russian colonial rule. The Aleuts had a monopoly on sea-otter hunting; outsiders have never learned well enough. The historical experience of the Aleuts can be evoked by noting that over the years Aleuts have been sent to display their kayak skills both on the River Neva in St. Petersburg and in a pond at the St. Louis World's Fair.

These people are admirable, and Professor Laughlin admires them. They have a traditional school for boys, to teach kayak hunting. Its curriculum includes specific exercises, such as controlled finger hanging, arm twisting and other carefully constructed schemes to prepare the muscles for the demands of the kayak and the spear-throwing board. There were recognized Strong Men, with a lifelong regime of special training, thought to lead to premature death as the price of their great power, which was both physical and spiritual. One such man broke a hand-grip dynamometer in 1910 by too hard a squeeze. Another led the uprising against the Russian party at Nikolski. They know human anatomy well, and they conduct human autopsy along with comparative anatomy in dissecting sea mammals. Their language holds a specific anatomical vocabulary, well beyond common practice in English. The Aleuts used such skills in the frequent preparation of mummies of some of their dead, stored in caves with the tools of the hunt, the relics magically consulted and employed. The Aleut midwives could skillfully cope with breech delivery; they knew how to secure the afterbirth if it did not come naturally. The Aleut healers used acupuncture, bloodletting, massage, even sutures in surgery. The Aleut primary counting words went easily to five powers of 10. The Aleuts enjoy and elaborate word games and wordplay; storytelling was the great art.

They knew warfare and even slave raids for women and children, but these episodes were never on a large scale. There is an old suit of rod armor from a mummy cave, vertical wood rods held by sinew. The Russian fur traders encountered the Aleuts in war, guns against stone weapons, until they crushed Aleut resistance after a major uprising in the 1760's. The Aleut population was sharply reduced from its peak of some 16,000 by a third or more within a decade or two. The Aleuts have not yet recovered; their well-attested longevity, known from the data yielded by their graves, is only now reappearing, more than a century after Seward's purchase. Before the Honda they had taken up Russian steam baths, the Russian Orthodox Church and literacy, all with enthusiasm. The enthusiasms remain strong, but now the people are learning to look to Washington and its processes as carefully as they watch the sea. "With their demonstrated skill in human adaptability, they may survive another 9,000 years," to see their lands once again extend as the ice cap grows.

Australia's Animals Discovered, by Peter Stanbury and Graeme Phipps. Pergamon Press ($23.40). East African Mammals: An Atlas of Evolution in Africa, Volume iii Part b (Large Mammals), by Jonathan Kingdon. Academic Press ($126.50). "To vulgar ears a black swan has the sound of a miracle." Of course, the proverbial whiteness is merely conjoined to the swan and not connected of necessity, as David Hume might have said. Doubtless someone has checked up on just what gene and what enzyme are lacking in the European birds that those of Western Australia retain, along with "the most striking resemblance to the common swan, and ... all those gracefully varying attitudes." With this remark of 1792, George Shaw, whose natural history in color "Described Immediately from Nature" was an annual triumph of deluxe publishing of the time, included an elegant engraving. The bird was so ideally figured in profile that it resembled more the chinaware decoration it soon became than anything to be seen alive, even down under. *Australia's Animals Discovered* is an attractive new bestiary, a couple of pages devoted to each of about 60 species of all kinds of Australian animals. It is also a study of the growth of knowledge. Each form that appears is represented by some period illustrations and a selection of the old accounts, whether in explorers' journals, the popular press or the scientific literature.

The time surveyed begins with the earliest Dutch explorers in New Holland before 1700 and continues to quite recent years. The visits of James Cook and Joseph Banks in 1770, the First Fleet of settlers at Sydney Cove in 1788 and the Victorian works of the famous artist-naturalist John Gould produced perhaps the most frequent sources. The illustrations include an entire span of techniques, from early woodcuts to fine color lithographs, even a few reproductions of manuscript watercolors. The subject remains fresh over the centuries. Robert Cruikshank caricatures a fine lean leaping "Escaped Kangaroo at Regent's Park," done in about 1840, an event unexpected by elegant London visitors. Shaw (who was a zoologist at the British Museum) was the first to describe the platypus in print; he expressed his strong suspicion of a hoax, "the beak of a Duck engrafted on the head of a quadruped." But after "the most minute and rigid examination" of three specimens his suspicions were completely dispelled.

A full-page black-and-white drawing from Gould shows a male great bowerbird in his shell-decorated bower. It was Captain J. L. Stokes of H.M.S. *Beagle* (that famous barque voyaged to the Antipodes with a new captain and crew after Charles Darwin and Captain Fitzroy had returned) who first among outsiders came to know that the stylish bowers were not "some Australian mother's toy to amuse her child" but artifacts intrinsic to the life of bowerbirds. It would have been appropriate to include as well one or two of the striking figures by the first artist-naturalists to enter the island continent, the aboriginal Australians; the volume, by two leading scholars of the Australian museum community, is otherwise complete, with careful citations, maps and a list of sources that includes all the later explorations of consequence.

The second of these two bestiaries is the big, handsome, expensive and fascinating continuation of Mr. Kingdon's tour de force. It treats of elephants, rhinos, horses of all species, pigs, hippos, camels, the chevrotain and giraffes, including the hidden okapi ("relatively common in some localities" within the great forest of the Ubangi watershed; "those who have eaten it have assured me of the excellence of its meat"). Once again we enjoy this talented artist-zoologist at work; once again we see impressive formal drawings, even anatomical ones, together with a lively sketchbook, careful maps and diagrams, and a thoughtful review of the literature in the context of field experience, both anecdotal and experimental.

Black rhinos show an unusually adaptive form of territoriality. Their bulls rarely fight. Male territory is well marked and firmly occupied, but any visiting male—even a displaced former owner—is readily tolerated. The Number One wanders past oblivious of the other bull, often grazing only tens of meters away. Sometimes the resident male engages his guest very briefly, only to walk off quickly. The grass-eating white rhino, the species with the remarkably long and heavy head, is found today only in two distinct pockets, one along the upper Nile and another far south in Zululand. Among the white rhinos the calf leads its mother even in flight; among the black ones the mother leads. A fine pair of thumbnail drawings make the point.

The elephant chapter is full of visual interest; it begins with a sketch of the musculature of the trunk, "typically invertebrate in its structure and mechanics." (The telling phrase is from R. M. Laws.) The built-in resiliency of the elephant skull is the result of arches and other superstructures built up of bone honeycombed with many air sinuses, which amount to a "total remodeling" of the skull. The result is a form of "powerful rhythms. These have appealed so strongly to the sculptor Henry Moore that he has embarked on entire series of sculptures and drawings based on the skull of an East African elephant."

The most striking novelty found here is in a long look at the stripes of the zebra. The author-artist fitted an enclosure holding nine Grevy's zebras with panels painted with vertical bars. One panel had wide bars, the other narrow, numerous ones. Counts showed that single animals tended strongly to stand close to the narrow bars and far from the wider ones. Field observations had suggested that the spacing adopted by neighboring animals was influenced by the spatial frequency of the stripes. A subtle discussion of behavior differences between other equines and the zebras leads to the conclusion that the stripes have evolved as a visually based bonding system, to allow larger social groups of a temporary nature; the commoner grooming and other ritual contacts tend toward a more exclusive definition of bond.

The zebra's stripes are quite constant features; they must be of importance to the species. Among these most gregarious of Serengeti grazers some interspecific signal seems a more likely adaptive function than any doubtful camouflage. A striking color plate shows 15 polymorphic equines: five are drawn from one living zebra species, five are drawn from another and five are redrawn from images on the walls at Lascaux. The sequence tends strongly away from the aberrant markings of the Lascaux horses: today's most widespread race has the best-defined striping.

What is still to come—and it is plenty—are the bovids, 43 species in East Africa. We can look forward to one or two more volumes of this rich work, now a decade in the making.

SMART MISSILE SCORES A HIT on the turret of a tank in a recent test at the White Sands Missile Range in New Mexico. The missile, called the Copperhead, has an electro-optical sensor in its nose that detects a spot of light cast on the target by a laser "designator" beam generated from a remote source, either on the ground or in the air. The missile's guidance system relies on signals from the sensor to steer the projectile to the target. The Copperhead missile, which is designed to be launched from a 155-millimeter howitzer, is scheduled to enter service with the U.S. Army later this year. It currently has a range of up to 10 miles; a longer-range version is planned.

Established 1845 SCIENTIFIC AMERICAN August 1981 Volume 245 Number 2

Precision-guided Weapons

Missiles that can home in on their target may alter the balance of forces on the battlefield. Notably vulnerable to such "smart" weapons are tanks, which the U.S. is buying in large numbers

by Paul F. Walker

Four months after the end of the 1967 Middle East war Israeli sailors on the destroyer *Elath* in the eastern Mediterranean watched a long green missile slowly approach them from the direction of Port Said in Egypt. The strange-looking object, fitted with stubby wings and tail fins and flying just above the waves, appeared to be off course; as it headed for open water away from the *Elath* the crew had time to maneuver the warship and begin firing at the missile. Suddenly, to the astonishment of the Israeli commander, the missile turned and homed in on the *Elath,* striking it amidships through a hail of machine-gun bullets. Minutes later a second missile appeared over the horizon and slammed into the engine room, disabling the ship. As crew members scrambled to extinguish the ensuing fire and help the injured, a third rocket struck the burning vessel and sank it. Finally a fourth missile splashed down among the survivors where the *Elath* had just disappeared.

An Israeli navy spokesman explained the next day that the *Elath* had been hit by Russian-built ship-to-ship missiles fired from an Egyptian 100-ton fast-attack vessel more than 10 miles away. The 2,500-ton *Elath,* originally named H.M.S. *Zealous* at its launching in Britain in 1944, was one of two such warships purchased from the British in 1956; together with its sister ship, the *Yaffo,* it had formed a major part of the Israeli navy for more than a decade. Yet it was destroyed, along with approximately a fourth of its 202-man crew, by a vessel a fraction of its size and cost. The instrument of its destruction, a 20-foot delta-winged rocket-propelled missile designed by the Russians in the late 1950's and known to Western intelligence agencies as the Styx, was equipped with a radar-based terminal-guidance system and a 1,000-pound high-explosive warhead.

This incident, the first sinking of a ship by an unmanned homing missile, was a milestone in naval warfare. It was also a startling and costly demonstration of the potential impact on military operations of "smart" weapons: conventional (that is, non-nuclear) munitions that can be precisely guided to their target after being launched. I shall review here the evolution of precision-guided munitions, their current status in the arsenals of the leading military powers and the controversy over their probable future role, particularly on the land battlefield. Wherever it is possible I shall cite the results of actual combat experience and field tests. It must be kept in mind, however, that a great deal of information about these weapons remains secret.

In the *Elath* incident a small, fast and highly maneuverable vessel armed with only two antiship-missile launchers managed to win a battle with a much larger ship fitted out with 10 antiship and antiaircraft guns and eight torpedo tubes. A powerful Goliath, formidable against opponents within range of its guns, was overcome by a puny David, deadly at longer range, in a surprise attack. This incident and many subsequent ones, both on land and at sea, have confirmed that weapons on the tactical battlefield are increasing not only in range but also in destructive potential. Although the Styx missiles of a decade or more ago were fairly crude in design compared with the smart weapons of today, they were able to sink a warship whose cost was thousands of times greater than their own.

The cost-effectiveness of this mode of attack has not gone unnoticed. The major navies of the world are now investing heavily in ships designed to fire precision-guided missiles. The most immediate effect of the *Elath* incident is evident in the modernization of the Israeli and the Egyptian navies. Israel has retired its surviving destroyer, the *Yaffo,* and is building a fleet of 44 fast-attack vessels armed with a precision-guided antiship missile made in Israel: the Gabriel. The average displacement of the vessels in the Israeli surface fleet has decreased from 460 tons in 1967 to about 200 tons at present. Egypt has acquired 54 missile-armed fast-attack craft, has retired most of its larger warships and has reduced the average displacement of its surface fleet from 415 to 365 tons.

The advent of smart weapons has also influenced American and Russian military plans, particularly with respect to armaments for central Europe, the most heavily armed potential battlefield in the world. According to William J. Perry, former Under Secretary of Defense for Research and Engineering, precision-guided munitions are "creating a revolution of major proportions" that "will change the face of battle." Perry points out that although the forces of the Warsaw Pact nations have a two-to-one advantage in tanks over those of the North Atlantic Treaty Organization (and possibly as much as a three-to-one advantage after two weeks' mobilization), the acquisition of precision-guided antitank weapons by the NATO forces "should offset [those] disadvantages" and help to maintain the balance of power in central Europe. In a similar vein Marshall Andrei Grechko, Minister of Defense of the U.S.S.R., has written that "combat actions in the Middle East... have put anew the question of the relationship of offense and defense.... Tanks have become more vulnerable and the use of them on the battlefield more complicated."

The concept of a missile that could

be remotely guided to its target after being launched goes back a long way in military history. In World War I the U.S. Department of War experimented with a flying time bomb called the Bug, which never saw combat service. Critics of the new weapon warned that unreliable systems for guidance and control could cause it to endanger friendly forces. Some of this concern was justified. On a test flight in 1923 of an early remote-control seaplane called the Wild Goose the guidance system failed, and a "safety officer" had to give chase in an old de Havilland biplane. The pursuing pilot was to toss bricks into the propeller of the Wild Goose if the errant drone threatened to crash into a populated area.

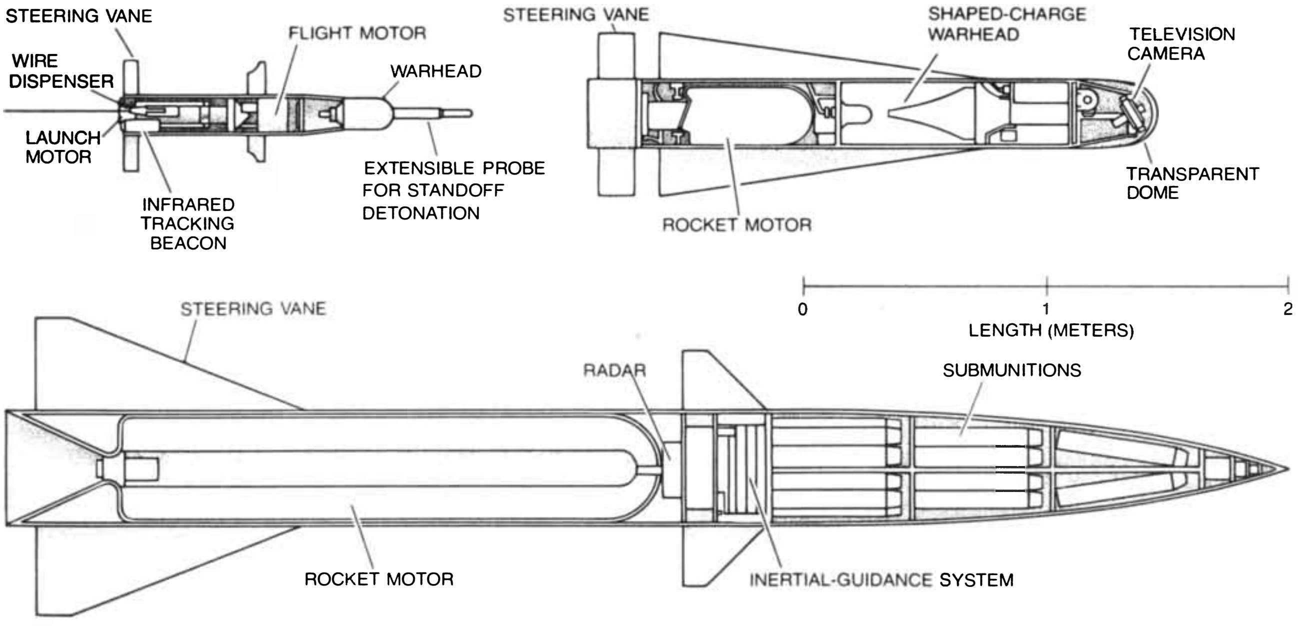
A more sophisticated approach, based on the notion of guiding a missile along a searchlight beam, was discussed as early as 1925. The missile was to be steered by means of a control system incorporating four light-sensitive selenium cells on the missile's tail fins. It was not until World War II, however, that Germany and the U.S. became actively engaged in developing advanced guided missiles. By the end of the war the Germans had at least 140 guided-missile programs under way.

One of the first American guided missiles, called the Weary Willie, consisted of a B-17 bomber loaded with 10 tons of explosives. In Operation Kastor in late 1944 and early 1945 aircraft of this type were involved in six missions against the German submarine pens in Helgoland. The pilots bailed out 20 miles short of the target and the airplanes were remotely guided the rest of the way to a crash landing. A one-ton glide bomb, the GB-1, was dropped from B-17's over Cologne in 1944. The first American homing missile, a winged glide bomb named the Bat, was fitted with a radar-based guidance system. It was launched against Japanese ships late in the war, in one instance sinking a destroyer 20 miles away. In this period the U.S. also developed the Azon bomb, which could be steered in azimuth only (hence its name). The Azon bomb was launched against various targets, including locks on the Danube, viaducts near the Brenner Pass, bridges over the Seine and lines of communication and transportation in Burma. The more refined Razon bomb, which was radio-controlled in both range and azimuth, never entered combat service.

The Japanese developed their own distinctive guided bombs—kamikaze airplanes—in World War II, but they met with only limited success. In the 1945 Okinawa campaign 93 percent of these piloted, explosives-packed aircraft either missed their target or were shot down. German military technologists meanwhile proposed the piloted Natter, a vertically launched rocket that the pilot would abandon by parachute only after homing in on the target. They also designed one of the first wire-guided missiles, the X-4, an air-to-air missile that could be controlled by means of a fine wire unwound from its tail section. In addition both the Germans and the Americans experimented with television-guided bombs long before television programs were being regularly broadcast.

These early guided weapons had a negligible impact on the outcome of World War II. In the course of the war in Europe, Allied aircraft dropped a total of 5.4 billion pounds of munitions during more than four million bomber and fighter sorties. According to the U.S. Strategic Bombing Survey, a postwar evaluation of the effectiveness of aerial bombing, "only about 20 percent of the bombs aimed at precision targets fell within the target area." (Precision targets were defined then as ones having a radius of 1,000 feet or less, measured from the aiming point.)

It was not until the mid-1950's that the first true precision-guided weapons were introduced, following a series of advances in the design of small rocket motors and guidance systems. A decade before the dramatic debut of the Styx the land battlefield had witnessed the introduction of the French SS10 antitank missile in the Arab-Israeli war of 1956. Wire-guided and manually controlled by the firer, the SS10 proved effective against older tanks with light armor. Its



THREE GENERATIONS of U.S. precision-guided missiles are represented in these cutaway views. At the top left is an improved version of the short-range TOW (tube-launched, optically tracked, wire-guided) antitank missile, which was originally developed in the mid-1960's. The semiautomatic TOW missile is designed to be fired from the ground, from a light vehicle or from a helicopter. After firing the missile the gunner has only to keep the cross hairs of a sight on the target; an infrared sensor mounted alongside the sight tracks the missile with the aid of a small infrared beacon in its tail. The guiding wire unwinds from the tail as the missile speeds away, transmitting course corrections to the missile's steering system. At the top right is an intermediate-range Maverick air-to-surface missile equipped with a television-based guidance system. In this model the pilot of the attacking aircraft selects a target in the televised image and locks the missile onto it; the missile then homes in on the target automatically. (Maverick also has infrared- and laser-guided models.) At bottom is a long-range rocket currently under development for the Assault Breaker program, which is designed to attack the second-echelon support areas of an army. The rocket has a dispenser of submunitions in its warhead that can release a variety of minimissiles over the target.

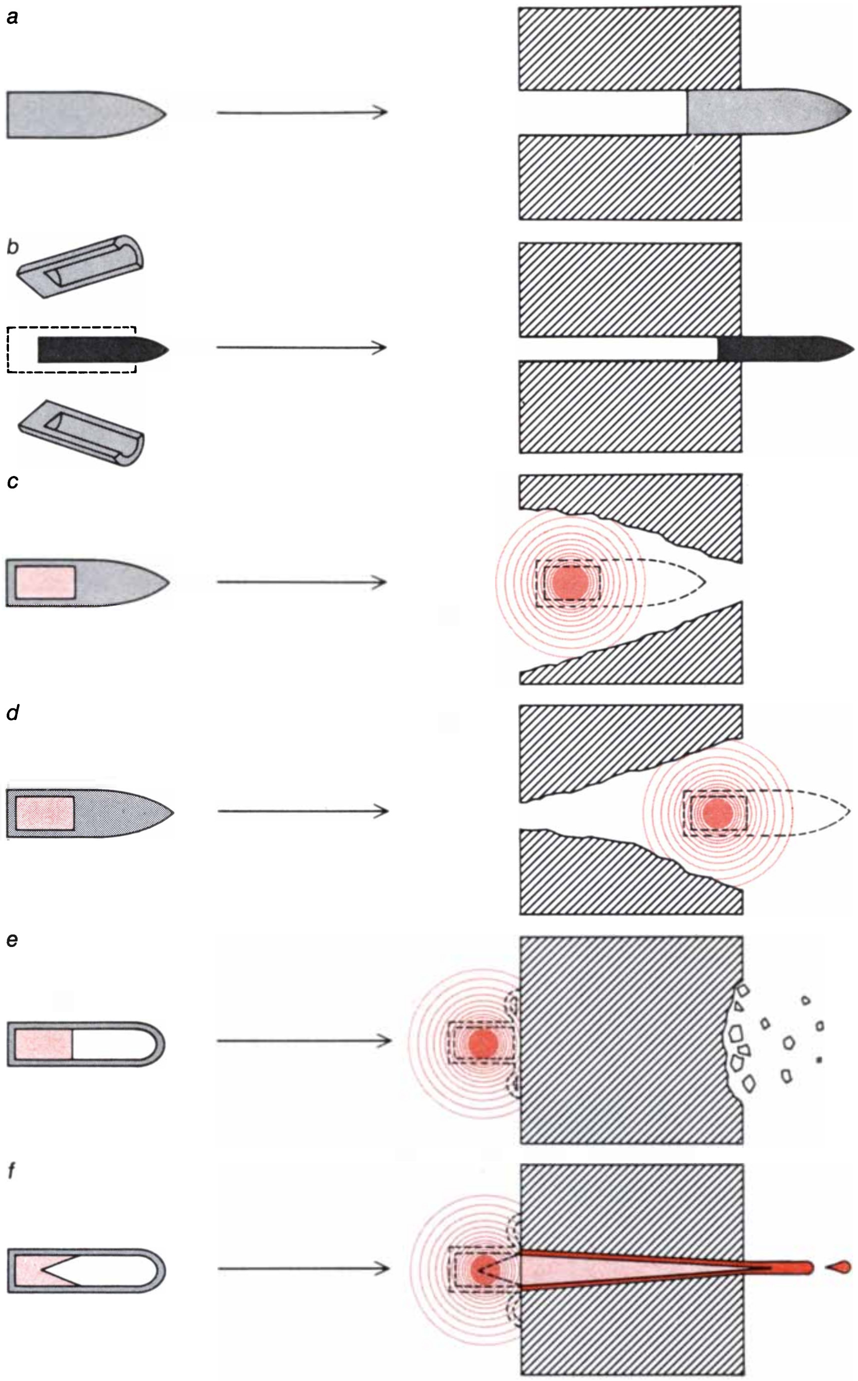
successor, the SS11, could be fired from a helicopter.

This first generation of smart antitank weapons, which included in addition to the French models the British Vigilant and the German Cobra missiles, enabled the firer to guide the rocket all the way to the target, a major improvement over the earlier systems. Now the infantryman could follow the target visually and take into account changes in its direction or speed. The lack of solid-state electronics in these early weapons, however, made them unreliable. Moreover, the accuracy of the weapon depended entirely on the skill of the operator, who had to guide it over a range of a mile or two, keeping both the missile and the target in sight. On a battlefield full of dust and smoke the task was difficult even for the most determined soldier, and it could be made even more so by bad weather or other distractions, such as artillery counterfire. A small flare installed in the tail of the rocket helped to keep it in sight as it sped away, but many military analysts remained skeptical about the efficacy of such weapons.

One of the first of the early, wire-guided, manually directed antitank missiles was the Russian-built Sagger, a weapon still widely deployed in the U.S.S.R., Eastern Europe and the Middle East. A 25-pound rocket with a high-explosive warhead, the Sagger can be carried by an infantryman or mounted on an armored vehicle. Ordinarily it is launched by a team of three soldiers. It is said to have a range of almost two miles and to be capable of penetrating as much as 16 inches of steel plate.

The first time the Sagger appeared in Vietnam was on April 23, 1972, when it destroyed several American-made tanks manned by South Vietnamese soldiers. A U.S. report on the incident described the attack as follows: "At first the South Vietnamese Army tankers seemed fascinated by the missile's slow and erratic flight. Through trial and error, however, the troops soon learned to engage the launch site... with tank main gun fire and to move their vehicles in evasive maneuvers. [Nevertheless,] by midmorning all officers of the 3rd Squadron had been killed or wounded, and three M48A3 tanks had been destroyed by Sagger missiles."

The most commonly cited demonstration of the effectiveness of the Sagger missile dates from the 1973 Arab-Israeli war, when antitank weapons took a heavy toll on both the Syrian and the Egyptian fronts. Retired Major General Chaim Herzog of the Israeli army recounts the Sagger's success against Israeli armored forces in his book *The War of Atonement: October, 1973,* describing battlefields strewn with burned-out tanks and spiderwebs of missile-guiding wires. On the Golan Heights more than 1,500 Syrian, Israeli, Jordani-



SIX KINDS OF ARMOR-PIERCING PROJECTILE are depicted. The traditional solid shot (*a*) relies on high velocity, shape and hardness to penetrate the wall of an armored vehicle. A variant of this basic approach, known as the APDS (armor-piercing discarding sabot) round (*b*), consists of a heavy, small-caliber core surrounded by a much lighter sabot, or casing, which breaks up as the projectile leaves the gun barrel. The core therefore needs no rifling grooves, so that it is aerodynamically "clean." Moreover, because the core is heavy (in some instances it is made of uranium) it acquires most of the kinetic energy of the projectile, giving it a greater armor-piercing ability than a round of the same size fired from a smaller-caliber gun with the same chamber pressure. HEAP (high-explosive armor piercing) rounds have a hardened case and a chemical-explosive charge that is detonated by means of a delayed contact fuze; the projectile either embeds itself in the armor (*c*) or passes through it (*d*) and then explodes with greater effect than an ordinary high-explosive round. A HESH (high-explosive squash head) round (*e*) crumbles on impact to form a large contact area before the explosive is detonated; the ensuing explosion sets up damaging shock waves in the armor and may also cause fragments to spall off the inner surface at high velocity. The commonest armor-piercing round made for infantry weapons today is the HEAT (high-energy antitank) round (*f*), which has a shaped charge that is ignited from the rear at the instant the conically hollowed front of the charge comes in contact with the armor. The explosion is thereby focused into a jet of hot gas, which can penetrate a considerable thickness of armor, squirting molten metal into the interior of the tank.

an and Iraqi tanks were destroyed, many of them by antitank missiles. On the Egyptian front one Israeli tank brigade lost 77 tanks, three-fourths of the brigade's entire complement, to antitank ambushes in 16 hours. Egyptian General Sa'ad Al Shazli concluded that "it was impossible to ensure the success of any attack—whether of tanks or of armored infantry—without destroying or silencing in advance the antitank missiles." Herzog writes that up until the war Israeli military planners showed "a complete lack of appreciation of the new antitank capability within the Arab forces."

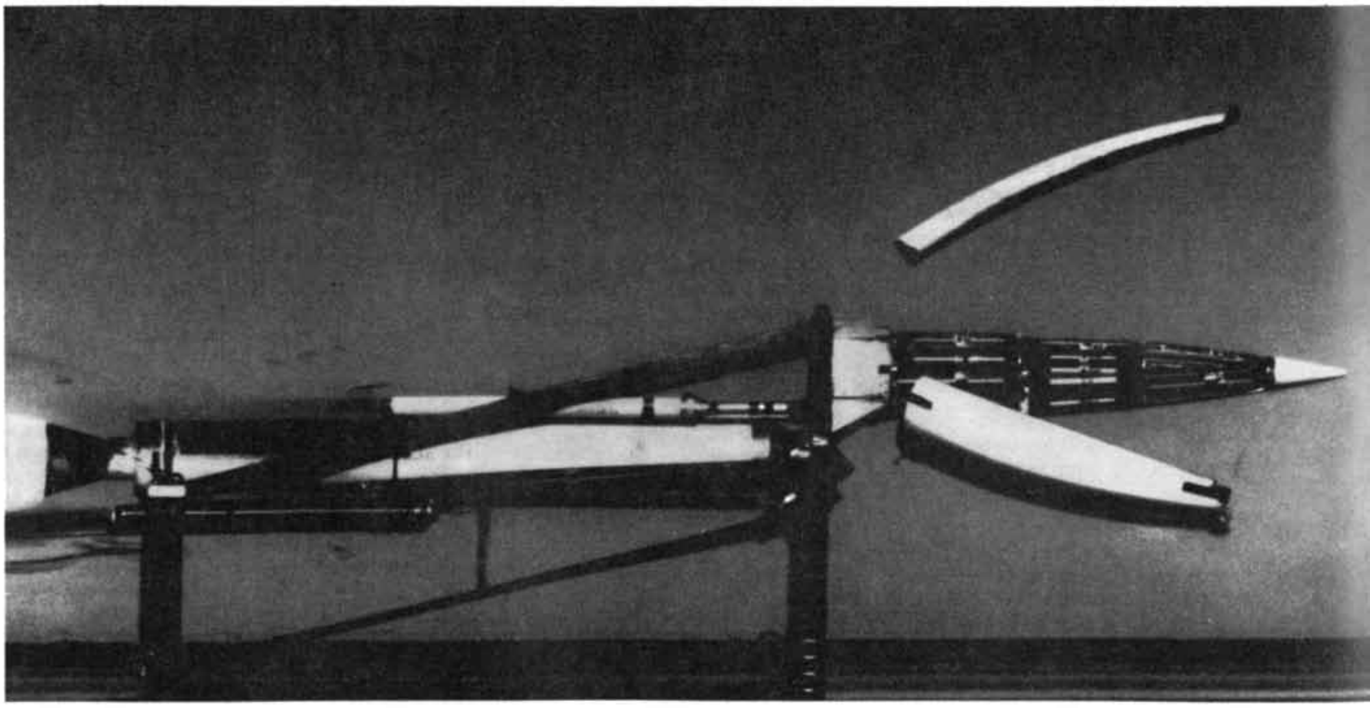
Meanwhile a major advance in smart antitank weaponry was already under way: the introduction of semiautomatic guidance. New weapons such as the American TOW (tube-launched, optically tracked, wire-guided) missile, developed in the mid-1960's, no longer need to be steered to the target by means of a manually operated "joystick." The gunner has only to keep the cross hairs of a sight on the target. An infrared sensor mounted alongside the sight tracks the missile, which has a small infrared beacon in its tail. The integrated sight unit in turn generates electrical signals that are transmitted through the wire, automatically correcting any deviation between the sight path and the track of the missile.

The TOW missile gives the infantryman a much better chance of destroying a tank with a single shot, particularly at long range. First tested in battle in Vietnam on April 24, 1972, the day after the Sagger was introduced, the TOW missile has proved to be effective and reliable. More than 275,000 of the missiles have been bought by 33 countries, including 107,000 bought by the U.S. Army and Marine Corps. The army is procuring them at a rate of 12,000 per year and is deploying them both for firing by infantrymen and for mounting on helicopters. The missile is now being upgraded with an improved armor-piercing warhead and a more accurate guidance system based on a microprocessor.

Small, inexpensive, infantry-launched antitank weapons such as the TOW missile still retain a major drawback of earlier designs: the wire "umbilical cord," which makes them vulnerable to counterfire, particularly from high-velocity tank guns. If an alert tank crewman notices the launch blast of an antitank missile or sees the missile in flight, a shot in the general direction of the launch site may be enough to disrupt the uncoiling wire. To avoid this problem most of the newer antitank missiles are radio-controlled, commonly operating on several frequencies in order to resist jamming or other electronic countermeasures. The Russian Swatter, a 60-pound antitank missile designed to be launched from an armored vehicle, is typical of the current generation of radio-controlled missiles; it is said to be deployed in Eastern Europe and the Middle East.

Eliminating the trailing wire does not entirely solve the problem of vulnerability to counterfire. Even a hasty and inaccurate shot toward the soldier steering the missile can interfere with his vision or concentration. To avoid this problem the next generation of antitank weapons will consist predominantly of automatically guided "fire and forget" missiles rather than the present "fire and follow" versions. Guidance systems based on television, on infrared sensors that detect the heat of the target and on laser "designator" beams have already been adopted by the U.S. Because of the larger size of the missiles equipped with such systems they are usually designed to be launched from ships, airplanes or helicopters. The effectiveness of these weapons against hardened targets was demonstrated in 1972, when the U.S. resumed bombing North Vietnam after a prolonged moratorium. Armed with new television- and laser-guided bombs, U.S. pilots succeeded in destroying 106 bridges in three months, including several that had withstood repeated attack during the earlier phase of bombing without the smart weapons.

One of the most widely deployed U.S. guided weapons is the Maverick, an air-to-surface missile that comes with a choice of three kinds of guidance system. In one model a television camera mounted in the nose of the missile transmits an image to a television screen in the cockpit of the attacking aircraft. The pilot selects a target on the screen, moves a set of cross hairs to coincide with it and locks the missile onto the target. From this point on the missile proceeds automatically to the target, freeing the pilot to proceed to other targets or to evade counterfire. Another model works in much the same way except that it employs an infrared sensor to detect slight differences in the amount



HIGH-SPEED-SLED TEST of the submunitions-dispensing system for the Assault Breaker development program was carried out on August 20, 1980, at the Naval Weapon Center in China Lake, Calif. The photograph at the left shows outer skins of the warhead

of heat radiated by different objects. The third method, currently preferred by the Marine Corps, employs an electro-optical system that tracks a spot of laser light reflected from the target. The laser designator can be operated from the attacking aircraft, from another aircraft or from the ground.

Homing systems of this kind have performed well in tests in the air, on land and at sea. The Maverick is reported to have scored a direct hit in more than 88 percent of all its test firings, and other systems have achieved comparable ratings. Nevertheless, there are important constraints on the expected battlefield operation of the new smart weapons that will have to be overcome if such weapons are indeed to revolutionize land warfare. The first and most obvious problem is one of tactics: it is still possible for an opponent to outmaneuver precision-guided missiles, particularly the slower antitank kind. A wire-guided missile can take as long as 15 seconds to reach its target, whereas a high-velocity tank round can cover the same distance in a second or two. Tank crews are now trained to recognize the probable positions of antitank teams and to put such positions under fire themselves or to call in artillery fire before the tank formation moves within range of the missiles.

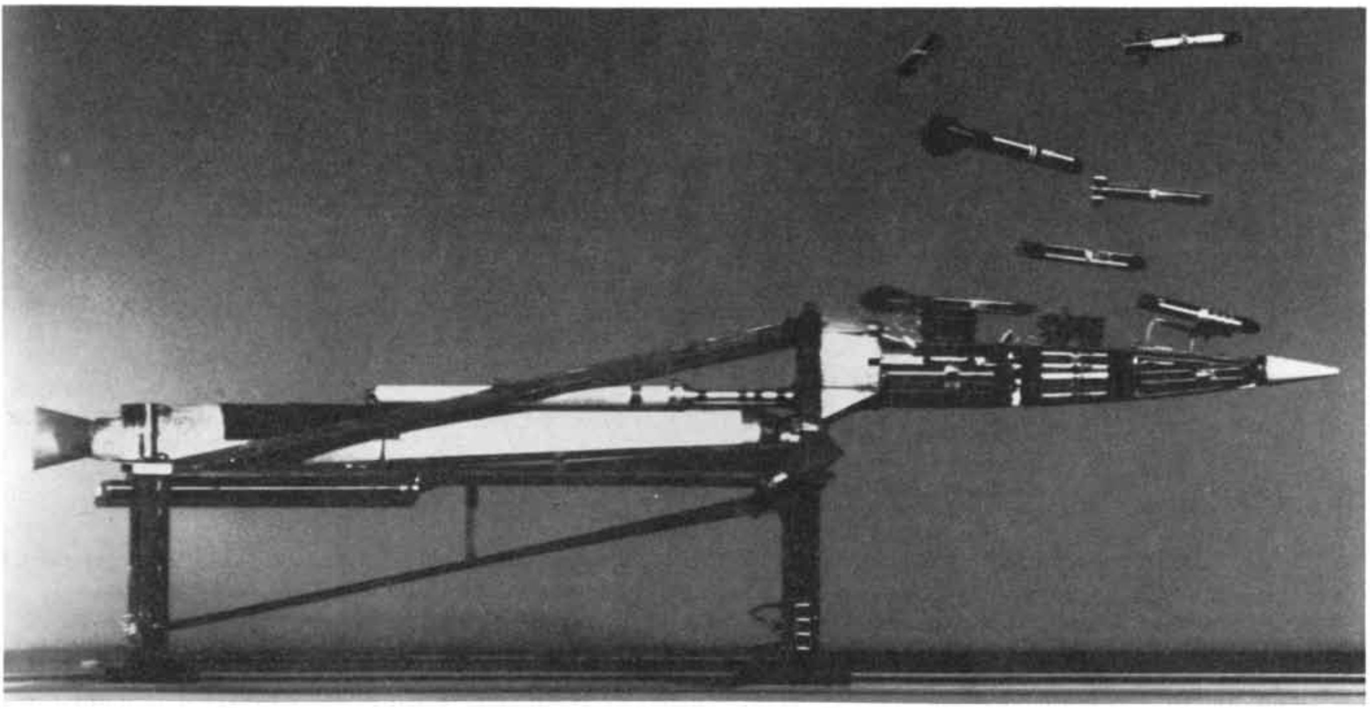
Another tactical measure available to the tank commander is camouflage: smoke to obscure the view of the antitank forces, flares to mislead infrared homing devices or bright lights to confuse laser-based guidance systems. Most tanks and armored personnel carriers are now equipped with smoke-grenade launchers capable of laying down a dense cloud of smoking phosphorus that lasts for several minutes, enough time in many instances to get within firing range of the antitank sites.

Another impediment to smart weapons can be the weather. Television systems work best in daylight and in a clear atmosphere; fog, rain and snow (which are not uncommon in central Europe) can seriously degrade the performance of most present guidance systems of this kind. Such considerations have tended to favor the laser- and infrared-guided versions of the Maverick, for example, over the television-guided one.

Another optional countermeasure, particularly for tanks, is to increase the amount of self-protection in the form of armor. Modern tanks are wrapped in heavy rolled-steel plating and advanced composite materials to safeguard them against the superior penetrating ability of new armor-piercing weapons. In recent years, however, as tanks have approached 50 or 60 tons in weight, the cost of added armor, particularly in terms of reduced speed and mobility, has become prohibitive. Accordingly it seems that armor has been losing the contest with increasingly accurate penetrating warheads. New shaped-charge, high-explosive warheads, by creating a narrow jet of extremely hot gas on detonation, are reported to be capable of penetrating steel armor to a depth of five times their diameter; in other words, a four-inch shell of this kind could go through as much as 20 inches of armor.

For almost two decades the British have been working at the Military Vehicles and Engineering Establishment at Chobham to improve armor without adding weight. On June 17, 1976, the British Secretary of State for Defense announced a breakthrough that he asserted was the most significant achievement in tank design since World War II. The new armor, called Chobham, is described as affording better protection against all forms of antitank weaponry. It consists of laminated steel and aluminum with woven fabric and ceramic materials between the layers of metal. It is designed to dissipate kinetic energy and heat much better than steel by itself does. Chobham armor has been incorporated into the British Shir tank, and a similar special armor was the basic design objective of the new American M-1, or Abrams, main battle tank. The French are introducing armor called appliqué, which consists of plates that can be added or removed depending on the threat, thereby affording minimum weight and maximum mobility under all circumstances. In addition most new main battle tanks have lightly armored "skirts" to cover vulnerable treads and wheels and a more streamlined profile to minimize vulnerability.

All these countermeasures, which at present are thought to provide a reasonably effective defense for armored vehicles under certain conditions, appear doomed to failure in view of the latest advances in smart weapons. Continuing improvements in accuracy, reliability, versatility, mobility and armor-piercing capability, combined with a lower cost



separating and clearing the dispenser. The photograph at the right shows the five kinds of submunition being ejected from the warhead. The dispensing system was built for the Advanced Research Projects Agency of the Department of Defense by the Vought Corporation.

per shot, suggest that precision-guided weaponry will sooner or later overcome most if not all countermeasures and increasingly dominate the battlefield of the future.

Among antitank weapons with the shortest range (less than two miles) the U.S. Dragon, a semiautomatic wire-guided light-infantry weapon first deployed five years ago, is scheduled to be replaced by several new systems. The Dragon has proved highly successful in test firings against heavy-tank targets, but it has operational limitations such as overly complex electronics and difficult maintenance. It will be replaced by a missile that relies on a laser designator for guidance, thereby allowing the firer to "shoot and scoot" more quickly than is now feasible.

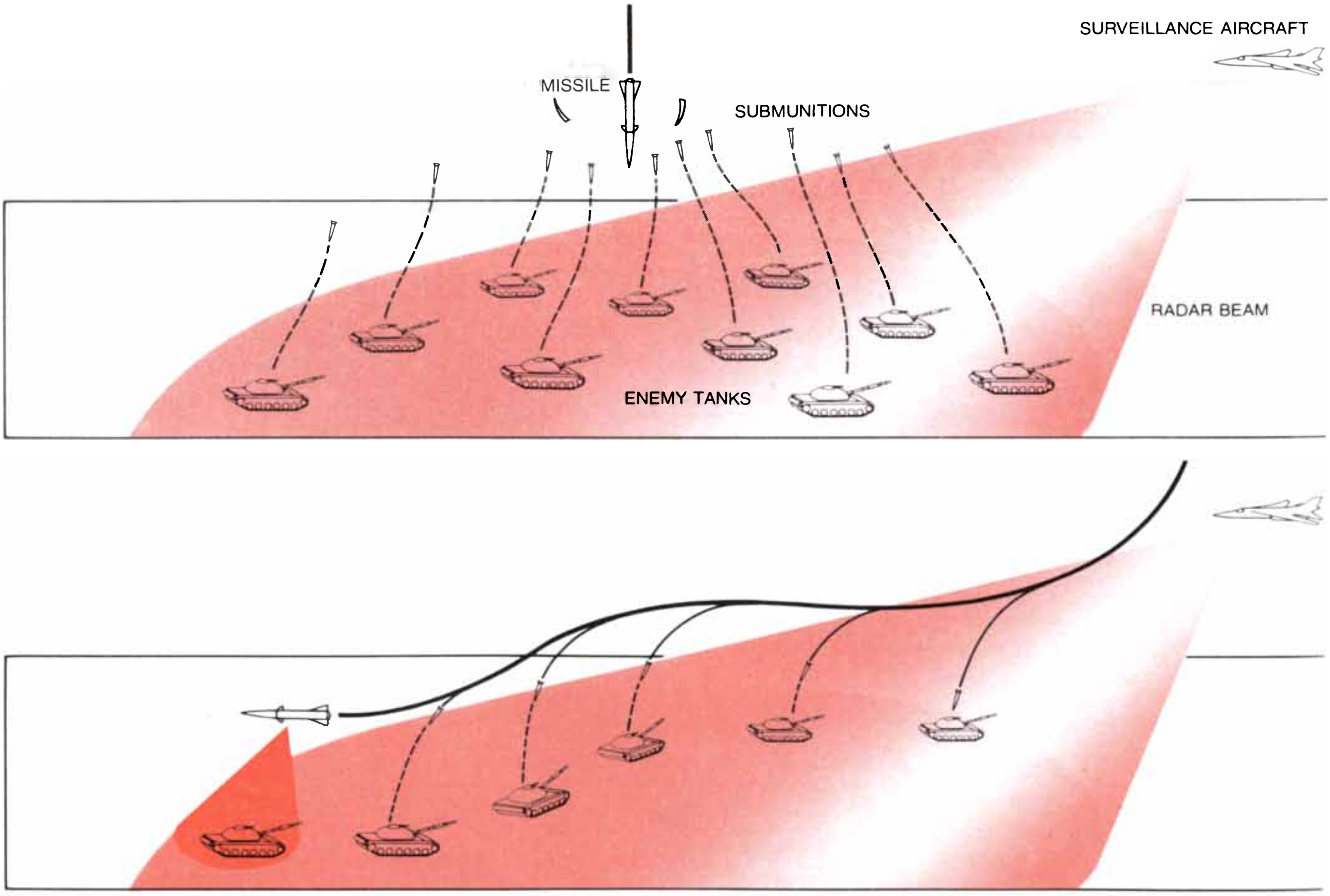
A second possible successor now under development is known as the Tank Breaker; it is an automatic homing missile with infrared guidance. A third program would rely on a tube-launched shell of 155-millimeter caliber with a sensor of millimeter-wavelength radiation for detecting armored vehicles. The projectile would be directed to fly over a tank at an altitude of 100 feet or less, detonating as it sensed the vehicle below and driving an armor-piercing fragment into the lightly armored turret of the tank at an angle of roughly 90 degrees. Currently called the STAFF (smart target-activated fire and forget) missile, it would be fired either from the shoulder or from a vehicle and would have a range of a mile or two.

An improved TOW missile will soon be able to cover the two- or three-mile range typical of attacks on tanks. It will be operable at night and will incorporate a new five-inch warhead for better armor penetration against the newer Russian T72 and T80 tanks. A planned further improvement calls for a six-inch warhead and a new tracking beacon so that the missile will be more visible in smoke.

Smart weapons of intermediate range (between five and 40 miles) and longer range (up to 125 miles) promise even more dramatic changes for the battlefield of the future. The Copperhead, a projectile launched from a 155-millimeter howitzer, homes in on a laser-designated target; it will become operational later this year. The Copperhead has already shown itself to be deadly in tests at ranges of up to 10 miles, and there is a program to extend the range to about 25 miles. The army plans to buy more than 2,000 Copperhead rockets.

A second intermediate-range antitank system under development is called SADARM (sense and destroy armor). Scheduled for deployment in five years, it would be launched from an eight-inch howitzer, travel twice the distance of the Copperhead and disperse three submunitions, each with a miniature terminal-guidance system based on a passive microwave sensor. The submunitions would descend by parachute, searching the ground for microwave-emitting targets. According to the manufacturer of the sensor, it is designed to be "effective against all known countermeasures."

The Multiple Launcher Rocket System (MLRS), which was developed jointly by the U.S. and its NATO allies and is scheduled to be deployed by 1982, will saturate an area with small submunitions at a range of about 20 miles. The design allows 12 rockets to be fired in a "ripple," or rapid volley. Each rocket contains 600 submunitions; one MLRS volley will reportedly cover an area the



BATTLEFIELD PANORAMAS illustrate two proposed applications of the Assault Breaker. Long-range targets, such as tank formations deep in the territory of the enemy, would be located by means of side-looking or forward-looking "standoff" radar operated from airplanes, helicopters or remotely piloted vehicles. The position coordinates of a target would then be transmitted to a mobile strike-control center, which would order the launching of a missile in the general direction of the target. The submunitions would be released

size of six football fields with 7,200 grenade-size fragments capable of piercing light armor. The army plans to buy approximately 360,000 MLRS rockets in the next few years. Later designs may incorporate homing submunitions and a laser seeker similar to that of the Copperhead. There are also plans for still more sophisticated rockets able to lock onto a target after launch, to discriminate between tanks and other vehicles (or decoys) and to track a target more reliably.

The most ambitious U.S. Army smart-weapons program, designated Assault Breaker, is an effort to create a long-range battlefield missile system to strike the support areas of an enemy army. New side-looking and forward-looking radars in airplanes, helicopters or remotely piloted vehicles would locate tank columns deep in the territory of the enemy; the radars would supply the position coordinates to a tactical missile battery, which would launch a warhead filled with homing submunitions. The radars would give position updates as the missile flew toward its target; in the last stages of the flight the submunitions would take over the task of guidance. Scanning with infrared or millimeter-wave sensors, the submunitions would explode and propel a slug at high speed into the turret of a tank.

Several technological developments are apparent in all these programs. First, improved armor-piercing warheads are evidently able to destroy or at least disable most armored vehicles and their crews today. It is possible that armor of the Chobham type will provide some enhanced probability of survival against such warheads, but it will be decades before current inventories of tanks with steel armor are replaced. In the meantime warhead development will most likely retain its advantage over composite armor.

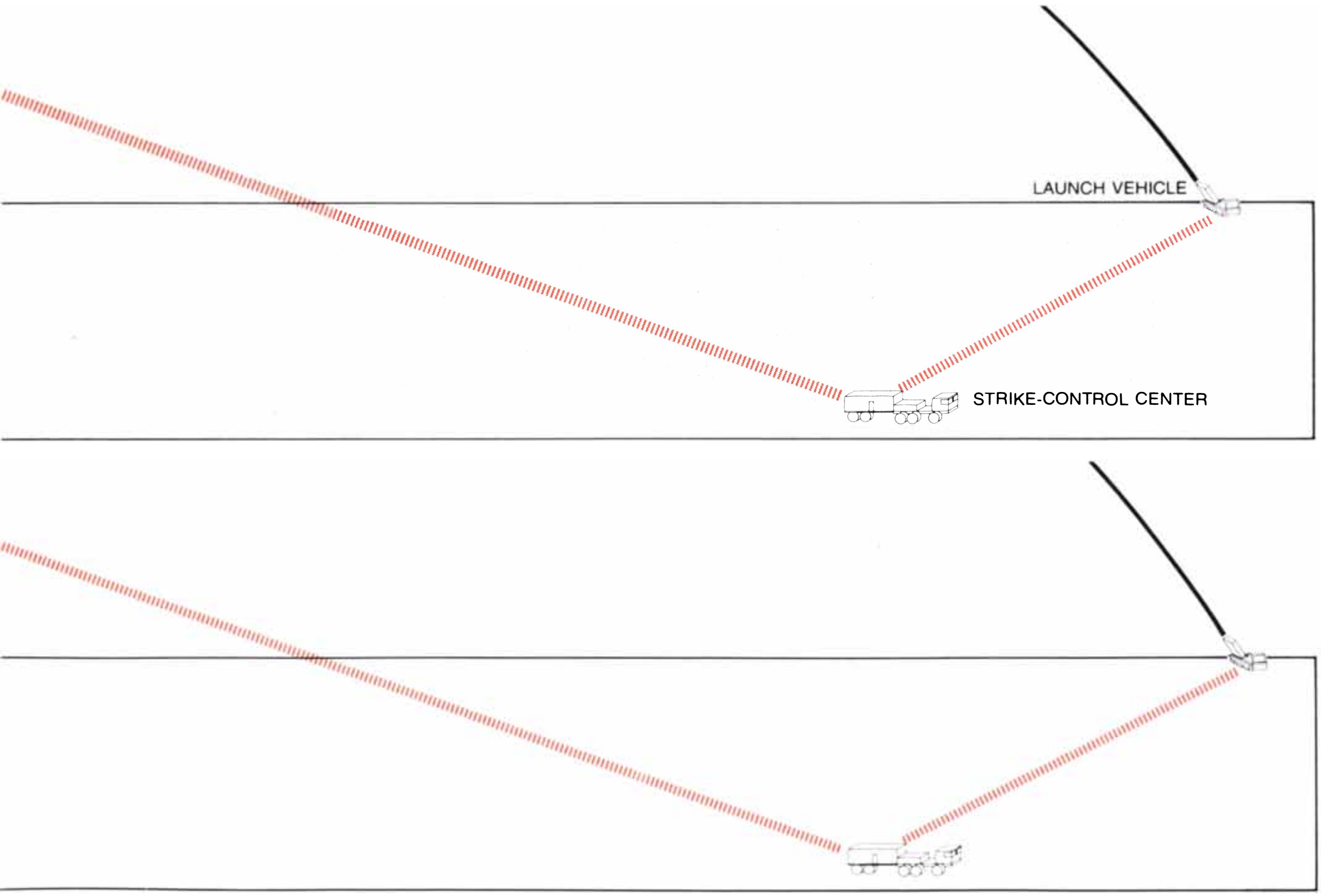
Second, weapons capable of attacking a tank from directions other than the front (where armored vehicles are most protected) will compel further increases in armor even though at 60 tons tanks are already prohibitively heavy.

Third, the range of antitank missiles is increasing far beyond the range of the tank gun; combined with improved target surveillance, this factor greatly enlarges the vulnerable zone for armored forces. Tanks will have to rely more on supporting forces, primarily long-range artillery, for successful operation.

Fourth, advances in terminal-guidance systems, chiefly in infrared and millimeter-wave sensors, are making missiles more reliable, more portable, less susceptible to jamming and more versatile under a variety of battlefield conditions. Millimeter-wave technology, utilizing a "window" in the atmosphere transparent to wavelengths between the microwave and the far-infrared regions (that is, between 30 and 300 gigahertz), is being exploited to reduce attenuation by haze, fog, clouds or rain. The new millimeter-wave components, promising small size, a high signal-carrying capacity and a narrow beam, will enable antitank forces in remote areas to make direct hits in all weather.

Fifth, as missiles become subdivided into dozens of minimissiles or submunitions, each with its own guidance system for homing in on the target, the cost per antitank round will be reduced. The cost of an antitank missile is already much less than the cost of a tank, and the disparity will become greater in the coming decades.

What does all of this mean for the



when the missile arrived over the target. Two modes of attack are envisioned. For a target spread over a wide area, such as a dispersed array of tanks (*upper diagram*), all the submunitions would be released simultaneously and their terminal-guidance systems, scanning with infrared or millimeter-wave sensors, would home in on individual tanks. For a linear target, such as a column of tanks (*lower diagram*), the missile would rely on its radar to follow the shape of the target and would release its submunitions in an appropriate sequence.

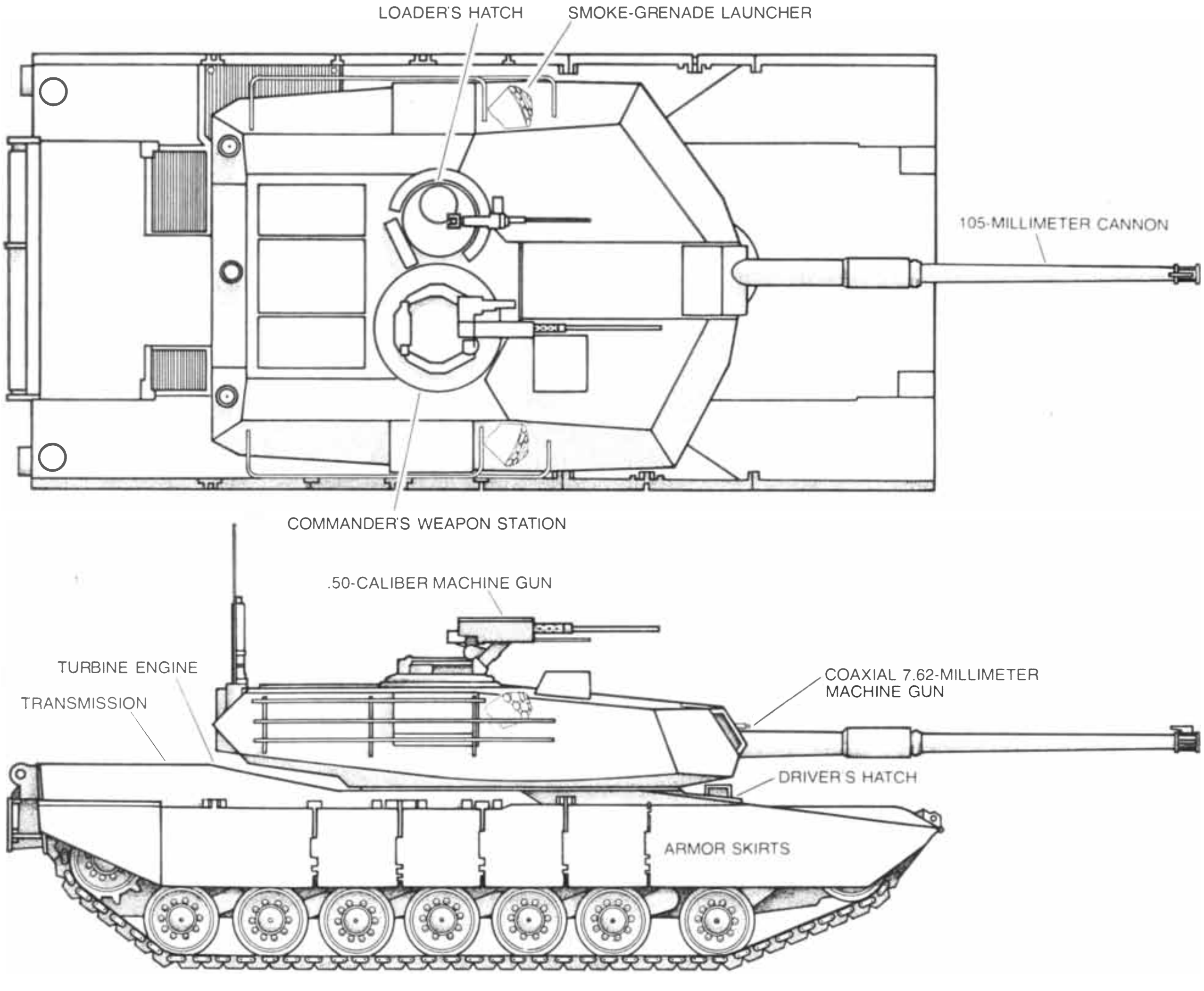
future of land warfare? Several conclusions stand out. In the trade-off between armor and mobility the emphasis in the future will surely be more on mobility than on armor. On a heavily armed front, such as central Europe or the Middle East, the side that strikes first, thereby giving away its position, will be the more vulnerable side. What can be seen, be it by eye on a clear day or by radar and homing guidance on a smoky and rainy field at night, will be hit and will probably be put out of action.

This development should come as no surprise to some military tacticians, who have practiced "hide and seek" maneuvers for some time now, certainly since the past two wars in the Middle East. During war games in Germany in 1972 U.S. Cobra helicopters armed with TOW missiles "killed" an average of 18 German Leopard tanks for each helicopter "lost." Three years later the U.S. Army Combat Development Experimentation Command conducted tests of the visibility of tanks as a function of how they are moved. The army learned that a tank column could reduce its vulnerability by as much as a third by following the contours of the terrain rather than moving in formation according to current field manuals.

Operation Bright Star, the exercises of the new U.S. Rapid Deployment Force held last December in the Egyptian desert, showed among other things that mobility is the key to survival in modern battlefield conditions. The crews of antitank helicopters reported that by flying low, hugging contours and folds in the sandy, barren terrain, "we could shoot at 3,000 meters, then move back and shoot again. We could eat up enemy tanks as long as we had room."

In short, it seems the day of the blitzkrieg is over. It could be argued that appropriate combined arms, including infantry, artillery and armor, can still prevail against antitank defenses. The technical evidence, however, does not substantiate this view. Even on a dirty and chaotic battlefield antitank weapons will be (if they are not already) too capable and available in numbers too large for them to be defeated by artillery barrages and tank thrusts. Blitzkrieg attacks will become suicidal, and as a result major defensive positions will probably be stabilized throughout the world.

The emphasis on mobility applies not only to battlefield forces but also to second-echelon supply and logistics forces. Given the increasing range of highly accurate missiles, the dispersal of troops, vehicles, aircraft, ships and supplies becomes a necessity in order to discourage preemptive attacks with smart weaponry. If depots within striking range of an enemy are not widely dispersed and somewhat protected, a first strike would become a possible option in a crisis. A premium will also be placed on command, control and communications, to ensure continuous monitoring of enemy

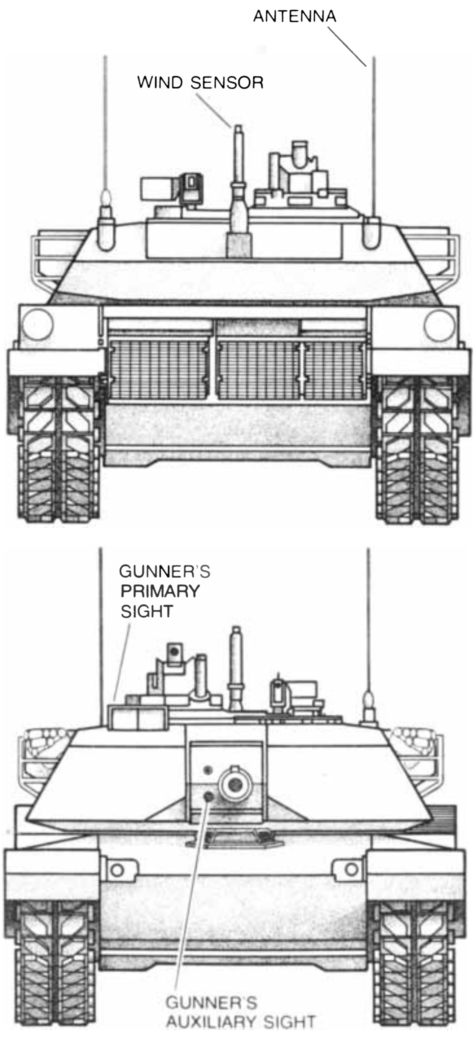


NEW U.S. MAIN BATTLE TANK, designated the M-1, or Abrams, is portrayed from several viewpoints. The 60-ton, four-man tank incorporates a number of features designed to reduce its vulnerability to precision-guided missiles: composite armor, a lower and more streamlined profile, an armored "skirt" to protect the treads and the wheels and a top speed of about 45 miles per hour. The cost of the tank, however, has increased dramatically over the past three years. Originally estimated to cost about $500,000, it is now reported to cost almost $3 million, or roughly 400 times the cost of a single antitank missile. More than 900 M-1's have been produced by the Chrys-

action and adequate coordination of dispersed forces.

Some tank commanders may still believe the best antitank weapon is a tank. Yet the heavy tank today, weighing almost 60 tons and capable of a maximum speed of 50 miles per hour, is being outranged and underpriced by accurate missiles, both land- and air-based. The future of the land battle, it seems, lies not with such ponderous, costly and easily detected armored vehicles but rather with light, mobile, cheap and therefore more numerous missile-armed jeeps or reconnaissance vehicles.

The same developments that have made heavily armored vehicles vulnerable on the land battlefield can also be applied to warfare at sea and in the air. The antiaircraft and antiship missile will soon dominate, so that an aircraft-carrier task force, for example, or an amphibious landing group will be forced to remain far behind the battlefronts out of range of missile attack. The increasing vulnerability of aircraft is attested to by numerous incidents. Between 1965 and 1973 in Vietnam the U.S. Air Force shot down 137 MIG fighters, 90 of which were destroyed by smart homing missiles. The helicopter in Vietnam was found to be extremely vulnerable even to small-arms fire. In the recent fighting between Iran and Iraq it was reported that fighter aircraft were effectively countered by older radar-directed antiaircraft guns, as they were in the 1973 Middle East war. Surface-to-air missiles also took a toll of Israeli aircraft in 1973, bringing down some 40 aircraft over the Suez Canal in the first two days of fighting.



ler Corporation at the Lima Army Tank Plant in Ohio at an estimated total cost of $2.1 billion. A decision is expected soon on whether or not to double the current production rate of the M-1 to 60 tanks per month by 1982.

Closely related to the evolution in tactics toward a greater emphasis on mobility will be a shift in procurement policies. As the tactical battlefield becomes deadlier, comparatively inexpensive, reliable equipment that can be supplied in multiple copies and quickly replaced becomes increasingly cost-effective. Large, costly items carrying high-value weapons (and people) will be increasingly wasteful. This is not to say that smart weapons are cheap. A TOW antitank missile costs about $7,000, a Stinger surface-to-air missile some $50,000 and a Harpoon antiship missile more than $700,000. In each case, however, the target attacked by the missile costs many times as much. The M-1 main battle tank, including the new composite battle armor, has risen in cost to almost $3 million, more than 400 times the price of a TOW missile; the A-10 close-support aircraft has a price of $8 million, 160 times the cost of a Stinger; a nuclear aircraft carrier costs more than $2 billion, 3,000 times the cost of a Harpoon missile.

Proponents of continued heavy procurement of tanks, aircraft and ships point out that it took several Styx missiles to sink the *Elath*, that every missile requires some launch platform, adding to its overall cost, and that large firing platforms such as tanks give a gunner 50 shots or more. Yet the cost disparity between these major capital items, which have long been dominant in the military budget, and the new precision-guided weaponry is too great for it to be ignored. In the past three fiscal years the army has bought more than 900 M-1 tanks at an estimated total cost of $2.1 billion; the same amount could have bought more than 300,000 antitank missiles. Present plans call for the Chrysler Corporation to expand its production capacity from 30 M-1's per month to 90, with a minimum purchase of 7,058 tanks. At the same time 3,596 M60A3 tanks, the predecessor to the M-1, will also be bought.

I would not argue that no tanks should be included in future budgets. The additional 9,500 main battle tanks currently scheduled for purchase, however, are clearly unwarranted in view of the capabilities of the new smart-weapons technology. A small fraction of the projected expenditures for tanks could buy a much more effective defense for the central European front in the form of vast numbers of smart antitank weapons. When procurement policies catch up with technological innovation, fewer main battle tanks will be procured and more smart antitank weapons will take their place. In the meantime matching the Russians tank for tank in either production or European deployment will be both unnecessary and undesirable.

As small missiles become more accurate and deadlier, the major capital pieces of equipment—the tank, the combat-support aircraft and the large surface ship—will ultimately become obsolete for virtually all war-fighting tasks. They will then be relegated to a largely ceremonial role, being put on display in Russian May Day and American Memorial Day celebrations.

From the point of view of arms control, smart weaponry may turn out to be a mixed blessing. The battlefield will probably become more static, neither side daring to take the initiative when being visible means being vulnerable. Political boundaries, such as those in central Europe between the NATO and the Warsaw Pact countries, may indeed turn into fixed military boundaries. This does not mean the trench warfare of World War I will return, since stationary defense is just as vulnerable as visible offense; both sides will come to emphasize mobile defensive forces. Such a situation of de facto stalemate on the battlefield may encourage negotiated arms control and disarmament in conventional weapons.

With weaponry becoming both smaller and more effective, however, fighting may shift from militarized fronts to cities and to guerrilla and terrorist activities. In the past decade there have been on the average two terrorist incidents per day somewhere in the world. Terrorists already know how to make use of hand grenades, time bombs and in a few recent instances antiaircraft rockets. Small, accurate guided missiles, carried in a suitcase or a small vehicle, capable of penetrating all but the heaviest battle armor, may become the weapon of choice for the terrorist of the 1980's and 1990's. In some cases, such as an attack on a passenger aircraft or a nuclear power plant, precision-guided munitions can cause disastrous harm. Preventing the misuse of such effective arms will clearly demand high security and restricted distribution, perhaps through security procedures similar to those in force today for tactical nuclear weapons. On the other hand, for maximum effectiveness the weapons must be widely dispersed. Some safe middle way will have to be found.

The Newest Stars in Orion

Clouds of dust and gas in the familiar constellation emit radiation at infrared and radio wavelengths. Some of the denser clouds appear to conceal new stars that are expelling fast-moving streams of gas

by Gareth Wynn-Williams

According to legend, the stars that outline the magnificent winter constellation Orion were put there by Artemis, goddess of the hunt, to commemorate the death of her companion Orion. For present-day astronomers Orion symbolizes birth rather than death; the constellation embraces one of the richest and nearest stellar nurseries in our galaxy. Within a volume a few hundred light-years across and 1,600 light-years away vast cool clouds of hydrogen are slowly collapsing under the influence of their own gravity to give rise to new stars (and presumably new planetary systems). A few of the stars that were formed in the comparatively recent past can be seen with the unaided eye, and many more have long been accessible to optical telescopes. In the past few years, however, advances in the detection of radiation at infrared and millimeter wavelengths have enabled astronomers to study the gas clouds themselves and the very youngest stars hidden deep within them.

In some of these young objects thermonuclear fusion may already have begun at the center even as a thick surrounding cocoon of matter is still radiating at warm infrared temperatures. Astronomers are puzzled by evidence that whereas one would expect such objects to be accreting matter from their immediate environment, some of them appear to be ejecting it.

A "main sequence" star (an ordinary star on the main sequence of stars in a graph of temperature v. luminosity) shines by virtue of the thermonuclear conversion of hydrogen into helium in its core. The more massive the star is, the faster it consumes its reserves of nuclear fuel; the biggest stars therefore have the shortest lives. For example, stars whose mass exceeds the mass of the sun by a factor of 15 or more are at least 20,000 times more luminous than the sun and convert all their available hydrogen into helium in less than 10 million years, a tiny fraction of the estimated age of the galaxy: 15 billion years. Such massive stars, designated Type *O*, and the slightly less massive stars of Type *B* are easily recognized by their brightness, by distinctive lines in their visual spectrum and by their bluish color, indicative of surface temperatures in excess of 10,000 degrees Kelvin. Because of their brief lifetime Type *O* and Type *B* stars have little opportunity to stray far from their site of birth. Hence when such stars are found in a cluster, one can be sure that all of them were formed recently in response to conditions prevailing in that small region of the galaxy.

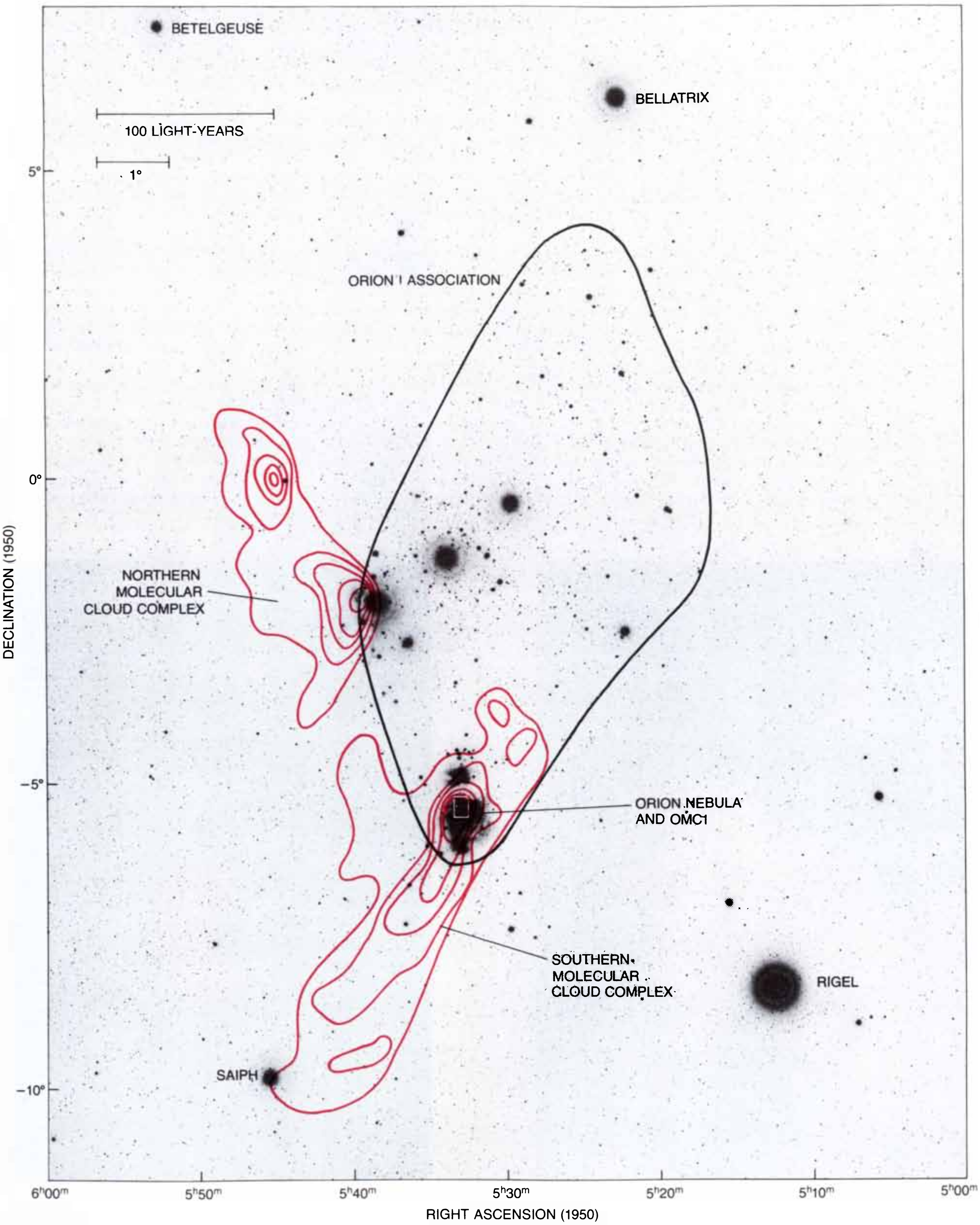
The constellation Orion abounds in stars of types *O* and *B*. A particularly prominent group is the Orion I Association, which includes the three bright stars in Orion's belt and the stars that form his sword. The stars in the association have been created at various times over the past 10 million years. The youngest of them, which are probably less than a million years old, lie in and around the glorious Orion Nebula, a cloud of glowing ionized gas heated by a group of four Type *O* stars known collectively as the Trapezium Cluster. The close connection between hot young stars and visible gas clouds is strong evidence for the hypothesis that new stars form by condensation out of the interstellar gas.

In order for an interstellar cloud to collapse, the pressure exerted by the mutual gravitation of its constituents must exceed the thermal pressure that would cause it to expand. The temperature of the ionized gas in the Orion Nebula is so high (about 10,000 degrees K.), however, that the gas must be expanding rather than contracting. It is therefore unlikely that any new stars are forming within the nebula itself. For evidence of star formation in the Orion region it is necessary to look instead at cooler clouds.

About 90 percent of the atoms in interstellar space are those of hydrogen; most of the rest are those of helium. Under the conditions that favor the gravitational collapse of an interstellar cloud (temperatures below 100 degrees K. and densities greater than 1,000 atoms per cubic centimeter) hydrogen exists mainly in the form of hydrogen molecules rather than as isolated hydrogen atoms or as hydrogen ions (that is, protons). Molecular hydrogen is difficult to detect for several reasons. It is too cool to radiate at visible wavelengths, and unlike atomic hydrogen it emits no characteristic radiation at specific wavelengths that can be detected by radio telescopes. Furthermore, clouds of molecular hydrogen are usually almost completely opaque to light because of the presence in them of tiny dust particles that absorb and scatter electromagnetic radiation of short wavelength. Only radiation with a wavelength considerably longer than that of light is capable of penetrating molecular clouds or escaping from them. Fortunately there are several processes that generate long-wavelength radiation within the clouds and thereby make them visible to modern instruments. Heated dust grains emit broadband thermal radiation in the infrared part of the spectrum, and various molecules other than those of hydrogen emit radiation at specific wavelengths (spectral line emission), mainly at wavelengths of a few millimeters.

By far the most useful of the 40-odd interstellar molecules observed so far is carbon monoxide, which was first detected in space in 1970 by Robert W. Wilson, Keith B. Jefferts and Arno A. Penzias of the Bell Telephone Laboratories with a sensitive receiver tuned to a wavelength of 2.6 millimeters. The energy levels responsible for the electromagnetic radiation from a simple molecule such as hydrogen or carbon monoxide depend on the molecule's electronic structure, on the vibration of the molecule's atoms back and forth along the molecule's axis and on the molecule's end-over-end rotation. In a molecular cloud essentially all the molecules are in the lowest electronic state, and so the only changes of state that need be taken into account are the vibrational and rotational ones.

In general, changes in the vibrational energy of a molecule are much more drastic than changes in rotational ener-



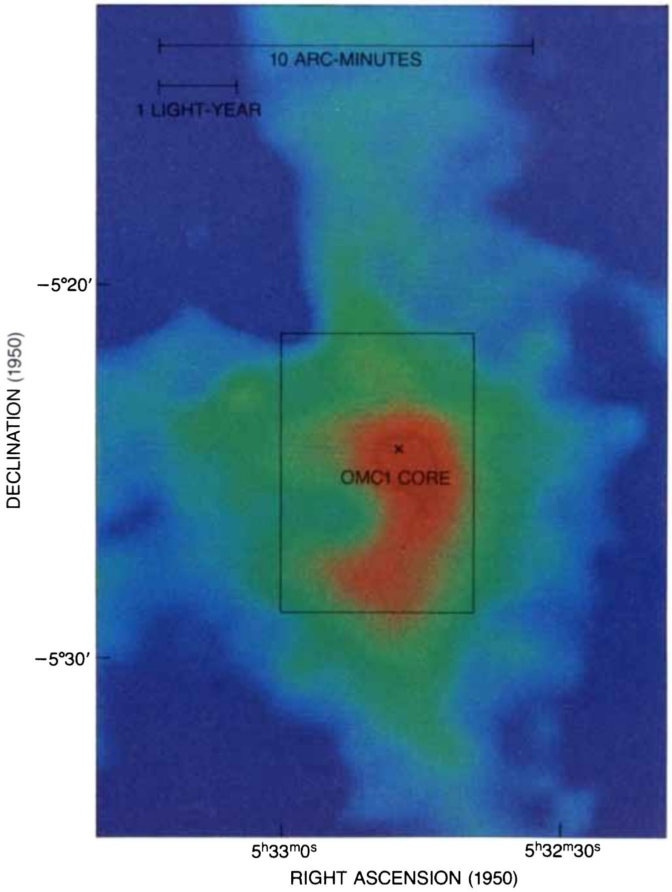
CONSTELLATION ORION, shown in a negative print, is speckled with stars that are hot, luminous and young. Many of these Type *O* and Type *B* stars, which are less than 10 million years old, are clustered in the Orion I Association. The very youngest of them are found near the Orion Nebula, which is about 1,600 light-years away from the solar system. The contour lines depict the intensity distribution of radiation at a wavelength of 2.6 millimeters from carbon monoxide present in small amounts in clouds of molecular gas that are mainly hydrogen. At the clouds' low temperature (less than 20 degrees Kelvin) there is no detectable emission from hydrogen itself. The carbon monoxide emission was observed with a 1.2-meter radio telescope at Columbia University by Patrick Thaddeus of the Goddard Institute for Space Studies and his co-workers. The concentration of gas that is coincident with the Orion Nebula but that actually lies behind it is called Orion Molecular Cloud 1 (OMC1). It is part of the southern molecular-cloud complex. The small rectangle in the illustration outlines the area in the photograph on the next page. In this photograph and others that follow it north is at the top and east is at the left.

gy. Therefore it is convenient to regard each vibrational state as being subdivided according to the rotational energy of the molecule [*see illustration on page 52*]. A given energy level can then be characterized by two quantum numbers: v for the vibrational state and J for the rotational state. These quantum numbers can have any whole-number value greater than or equal to zero. Whether or not a molecule can absorb or emit a quantum of energy and make a transition from one level to another depends on special rules that differ for hydrogen (H_2) and carbon monoxide (CO). It is rare, however, for either v or J to change in value by more than 2 in a single transition.

The energy, E, of each level above the ground level (the lowest level) is usually designated in wave numbers, computed by dividing the energy, E, by the product of the velocity of light, c, and Planck's constant, h: wave number = E/hc. The usefulness of this unit is that the wavelength of the photon (the quantum of radiation) emitted when a molecule moves from one energy level to another is given by the reciprocal of the difference between the energy levels measured in wave numbers. For example, the difference in wave numbers when the hydrogen molecule in the state $v = 1, J = 2$ emits a photon and drops to the ground state $v = 0, J = 0$ is 4,498 inverse centimeters (cm.$^{-1}$). The reciprocal of that value (1 divided by 4,498 cm.$^{-1}$) corresponds to an infrared wavelength of 2.22 micrometers.

If a single molecule in a cloud were left undisturbed, it would rapidly fall to its lowest energy level, emitting a series of photons as it did so. In a real molecular cloud, however, molecules continually collide with one another. At each collision there is a certain probability that one molecule will absorb some of the kinetic energy of impact and be excited to one of its upper levels. Having been excited in this way, it will then emit one or more photons as it reverts to the ground state, unless its energy level is again changed by collision. The relative numbers of molecules in different levels is therefore strongly dependent on the average kinetic energy of the molecules present and on the frequency of collision. Accordingly the strengths of molecular emission lines carry information about the temperatures and densities in interstellar clouds. Unfortunately, however, the interpretation of such data is rarely simple or unambiguous.

In order for spectral lines of molecular hydrogen to be observable it would be necessary at the very minimum for a significant fraction of the molecules to be excited to the level $v = 0, J = 2$. (Transitions with $\Delta J = 1$, such as the $v = 0, J = 1 \rightarrow 0$ transition, are extremely rare in a symmetrical molecule such as H_2.) In typical interstellar clouds, which are at a temperature of only about 20 degrees K., the mean translational kinetic energy (the energy of motion) corresponds to a photon with a wave number of only 21 cm.$^{-1}$. As a result the $v = 0, J = 2$ state, which is at 354 cm.$^{-1}$ above the ground state, is almost unpopulated, since collisions rare-



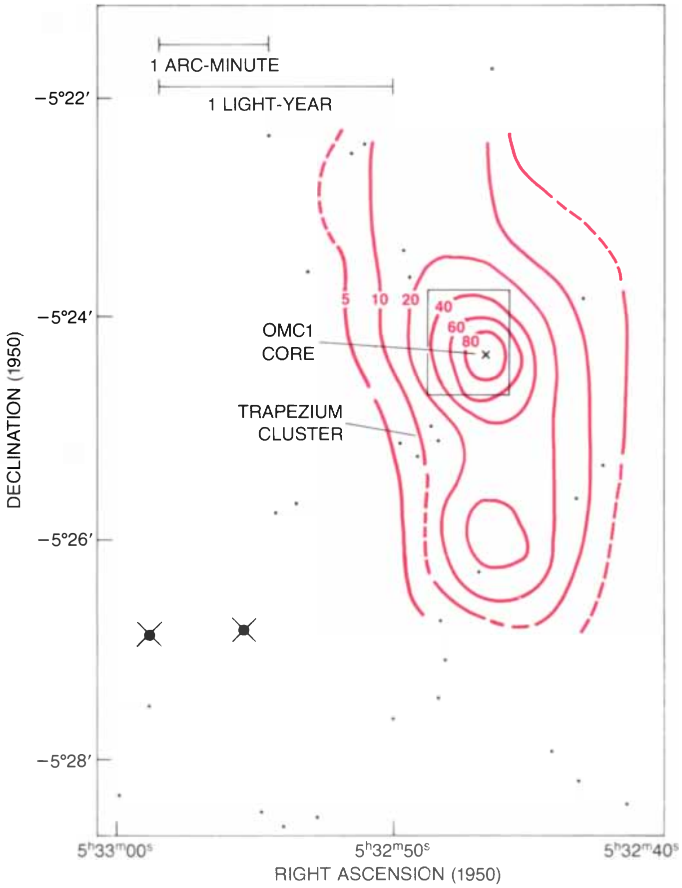
GLOWING GAS IN THE ORION NEBULA (*left*) has been heated to fluorescence, about 10,000 degrees K., by luminous Type *O* stars embedded in the nebula. Atoms in the gas are ionized (stripped of electrons) by the stars' ultraviolet radiation. The various colors in this photograph arise from different kinds of ions. The photograph was made with the 3.8-meter telescope at the Kitt Peak National Observatory. The map at the right, covering the same region, depicts the intensity of carbon monoxide emission from OMC1, the invisible cloud of warm molecular gas that lies behind the nebula. The red areas correspond to the most intense emission, the blue areas to the least intense. The map was prepared by Nicholas Z. Scoville, F. Peter Schloerb and Paul F. Goldsmith of the University of Massachusetts at Amherst. Made with a 14-meter radio telescope, the map has a resolution of about 45 seconds of arc, about 10 times higher than the map shown on the preceding page. Rectangle outlines the most luminous regions in OMC1, which are shown in greater detail on the opposite page.

ly excite molecules to levels with energies greater than three to five times their average kinetic energy. Therefore in typical interstellar clouds essentially all the hydrogen molecules are confined to the lowest two levels, and except for a few special hot regions (to which I shall return) molecular hydrogen cannot be directly observed in clouds such as those in Orion.

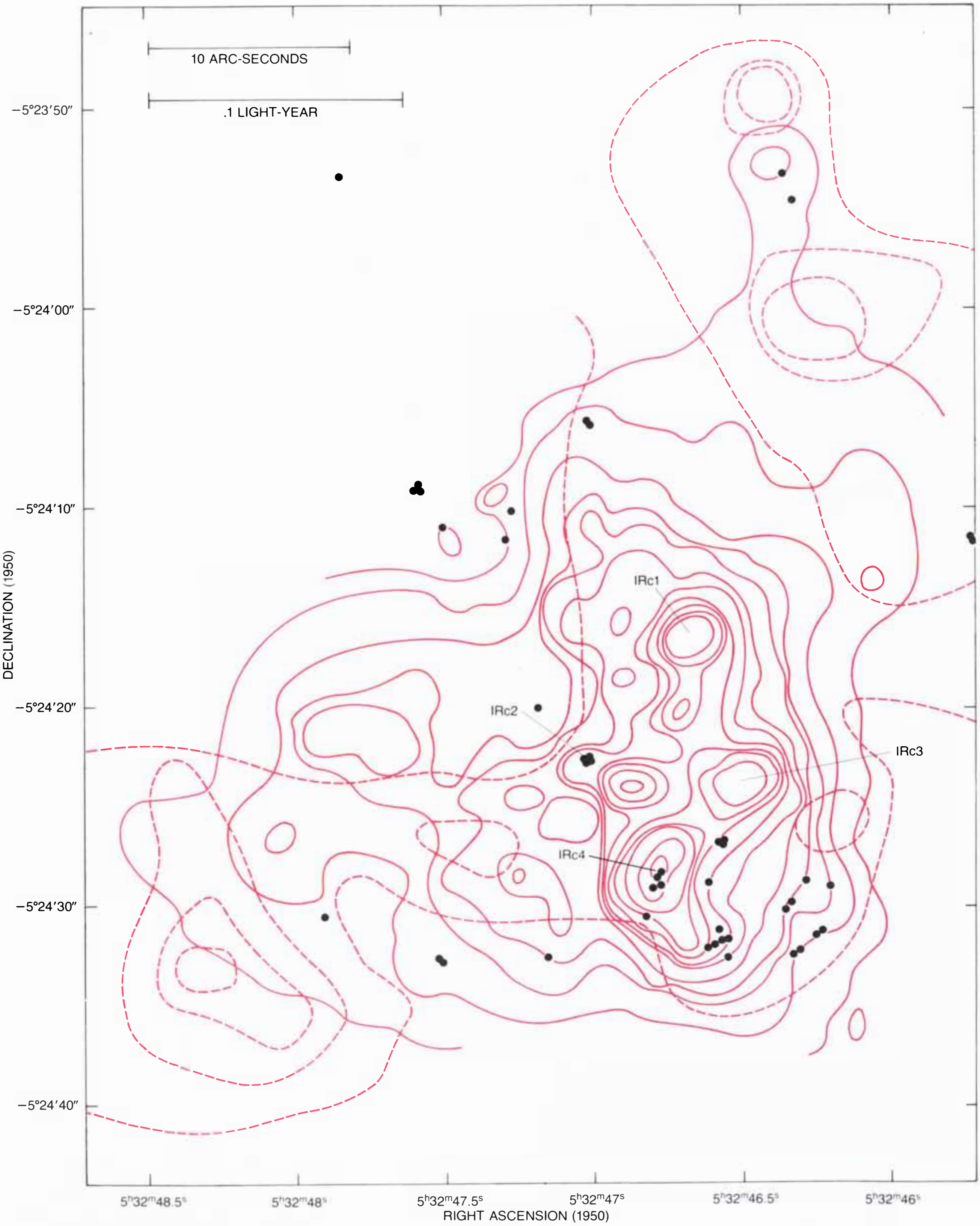
The difficulties that conceal the cool hydrogen molecule from visibility do not apply to the heavier asymmetric molecule of carbon monoxide. With this molecule transitions are possible between adjacent rotational levels that are much more closely spaced in energy than those of hydrogen. The wavelengths of those transitions form a harmonic progression starting at 2.6 millimeters when molecules in the $v = 0$ state drop from $J = 1$ to $J = 0$ and continuing with spectral lines at 1.3 millimeters, .87 and .65 millimeter for higher transitions. The two longer-wavelength transitions can be observed with high-quality radio telescopes, but the shorter wavelength transitions usually call for the more precise reflecting surfaces of optical or infrared telescopes, which are fitted with special receivers. Various different molecules have been studied by means of their transitions at millimeter wavelengths, but the carbon monoxide molecule (present in the ratio of one molecule for every 2,000 hydrogen molecules) is by far the most abundant of the molecules detected.

An extensive study of the Orion region was done some years ago by a group consisting of Marc L. Kutner of the Rensselaer Polytechnic Institute, Kenneth D. Tucker of Fordham University, Gordon Chin of Columbia University and Patrick Thaddeus of the Goddard Institute for Space Studies. A complete map of the region was made with a 1.2-meter radio telescope at Columbia tuned to detect the lowest rotational transition of carbon monoxide at 2.6 millimeters. It revealed two very large complexes of molecular clouds, each complex about 150 light-years across [*see illustration on page 47*]. Except for a few patches that are illuminated by bright stars near the edges of the clouds the material appears to be completely dark in optical photographs. Although the clouds are dense by interstellar standards, they consist of only a few hundred molecules per cubic centimeter. Nevertheless, the clouds are so large that they each hold enough matter to make 100,000 stars the mass of the sun. Measurements of the Doppler shifts in the emission lines from the clouds indicate they cannot have changed greatly in shape in the past million years.

Over most of the area mapped the clouds are cold, less than 20 degrees K.; they are warmed only by the diffuse light that filters in from the neighboring stars. The regions that show the strongest carbon monoxide emission, however, are significantly warmer: between 50 and 100 degrees. Such regions usually turn out to have a higher density than the rest of the cloud. In the southern complex, the better studied of the two, the highest peak of carbon



INFRARED EMISSION reaches a peak intensity near the visually brightest region of the Orion Nebula, as can be seen by comparing a color photograph (*left*) with an infrared map of the same area (*right*). The color photograph was made with the three-meter telescope at the Lick Observatory. The map, which shows the contours of infrared emission from warm dust grains at a wavelength of 400 micrometers, was obtained with the three-meter Infrared Telescope Facility on Mauna Kea in Hawaii by Jocelyn B. Keene, Roger H. Hildebrand and Stanley E. Whitcomb of the University of Chicago. The small group of four stars near the center of the photograph is the Trapezium Cluster, a group of Type *O* stars whose ultraviolet radiation ionizes the gas in the visible nebula and makes it glow. The warm dust grains are embedded in OMC1. The numbers on the contour lines represent the percentage of the peak flux at 400 micrometers. The hottest dust grains lie at the core of OMC1, indicating a strong source of energy. Small rectangle is the area in illustration on the next page.



HEATED DUST AND MOLECULAR HYDROGEN are concentrated in a region only a few tenths of a light-year across in the core of OMC1. The solid contours in color show the distribution of heated dust as it was observed at a wavelength of 20 micrometers with the three-meter infrared telescope on Mauna Kea by Reinhard Genzel, Dennis Downes, Eric E. Becklin and the author. The strongest infrared emission arises from about half a dozen compact sources. They include IRc1 and IRc2, which appear to be at a temperature of about 500 degrees K. The broken-line contours show the emission from molecular hydrogen at 2.12 micrometers, an energy-level transition associated with gas at a temperature of about 2,000 degrees. The hydrogen is probably heated by the collision between gas blown outward from IRc2 and stationary gas in the molecular cloud. The map of hydrogen emission was made with the five-meter telescope on Palomar Mountain by Becklin, Steven V. W. Beckwith, Gerry Neugebauer and S. Eric Persson. Becklin and Neugebauer discovered IRc1, sometimes called the BN object, in 1965. The black spots are the emission peaks of water masers, cloudlets that emit strongly at 1.35 centimeters.

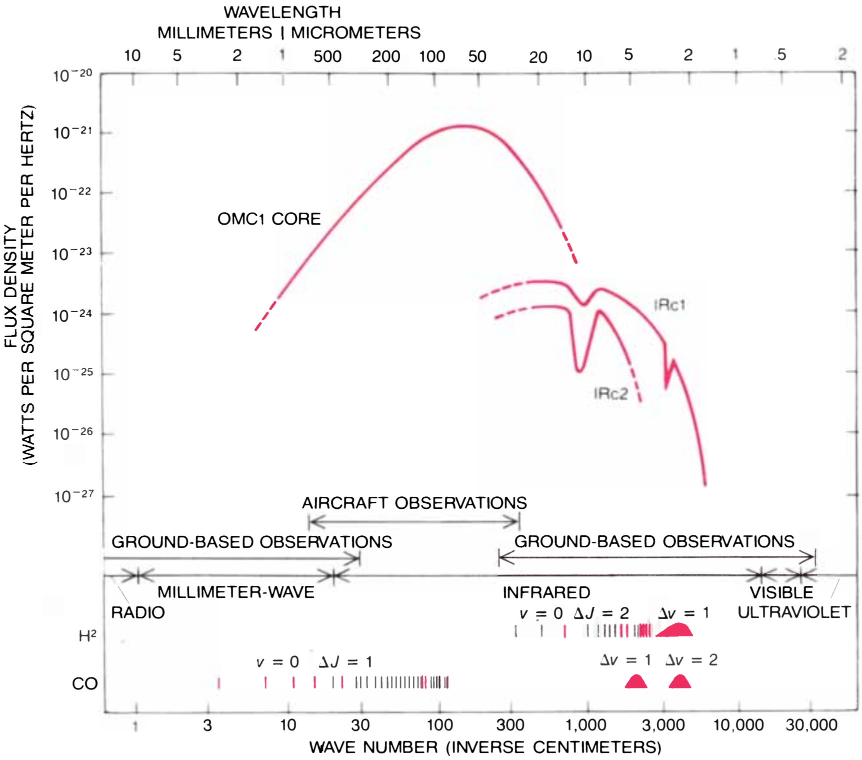
monoxide emission coincides approximately with the Orion Nebula [*see illustration on page 49*]. Although the Orion Nebula and the carbon monoxide peak appear to be superposed, the glowing gas in the visible nebula consists of fully ionized atoms at a temperature of about 10,000 degrees; the carbon monoxide emission arises from the cloud of molecular hydrogen just behind the nebula. This un-ionized cloud is usually designated OMC1, for Orion Molecular Cloud 1. The nebula is slowly eating its way into OMC1 as the hydrogen, carbon monoxide and other molecules are broken up by the ultraviolet radiation from the hot Type *O* stars within the nebula.

Although part of the heating of OMC1 can be attributed to the energy absorbed from the Type *O* stars, the most intense peak in the carbon monoxide emission demands another explanation. By examining emission lines from several different kinds of molecule, each sensitive to density and temperature in a different way, astronomers have found that OMC1 has a core only a few tenths of a light-year across in which the density exceeds 100,000 molecules per cubic centimeter. The mass of the gas in the core is somewhere between a few hundred and a few thousand times the mass of the sun. The strong concentration of gas at this point, the evidence that there must be something within the cloud heating it and the marked broadening of the emission lines (evidence of greater motion than there is elsewhere in the cloud) all suggest that the core is the site of either very recent star formation or current star formation.

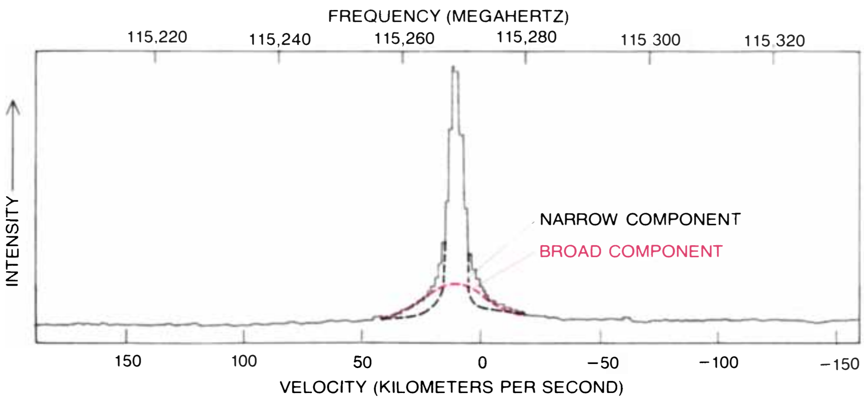
The observation of millimeter-wavelength molecular transitions in interstellar clouds has two shortcomings. First, such transitions reveal nothing about the nature of the stars or the other sources of energy that may be hidden within the cloud. Second, the method suffers from poor angular resolution; the smallest detail that can be mapped by a telescope is governed by the ratio of the wavelength of the radiation to the diameter of the telescope. The resolution that can be achieved when the 2.6-millimeter wavelength of carbon monoxide is collected by a 14-meter radio telescope is 45 seconds of arc, a factor of at least 20 poorer than can be achieved by telescopes at visible wavelengths. Observations at the infrared wavelengths between .7 micrometer and 500 micrometers (.5 millimeter) are helping to remedy both deficiencies.

The principal source of cosmic infrared emission is the thermal radiation from warm interstellar dust grains. Although the composition and size of the grains are uncertain, they probably include silicate minerals and frozen gases, and at least some of the grains must be about .1 micrometer in diameter. The grains are highly efficient at absorbing the light of stars and reemitting its energy as infrared radiation. The wavelength of the emitted radiation depends on the temperature of the grains. At 500 degrees K., for example, most of the radiation would lie at wavelengths of between five and 15 micrometers. At lower temperatures longer wavelengths would dominate, following the well-known Planck law for black-body radiation.

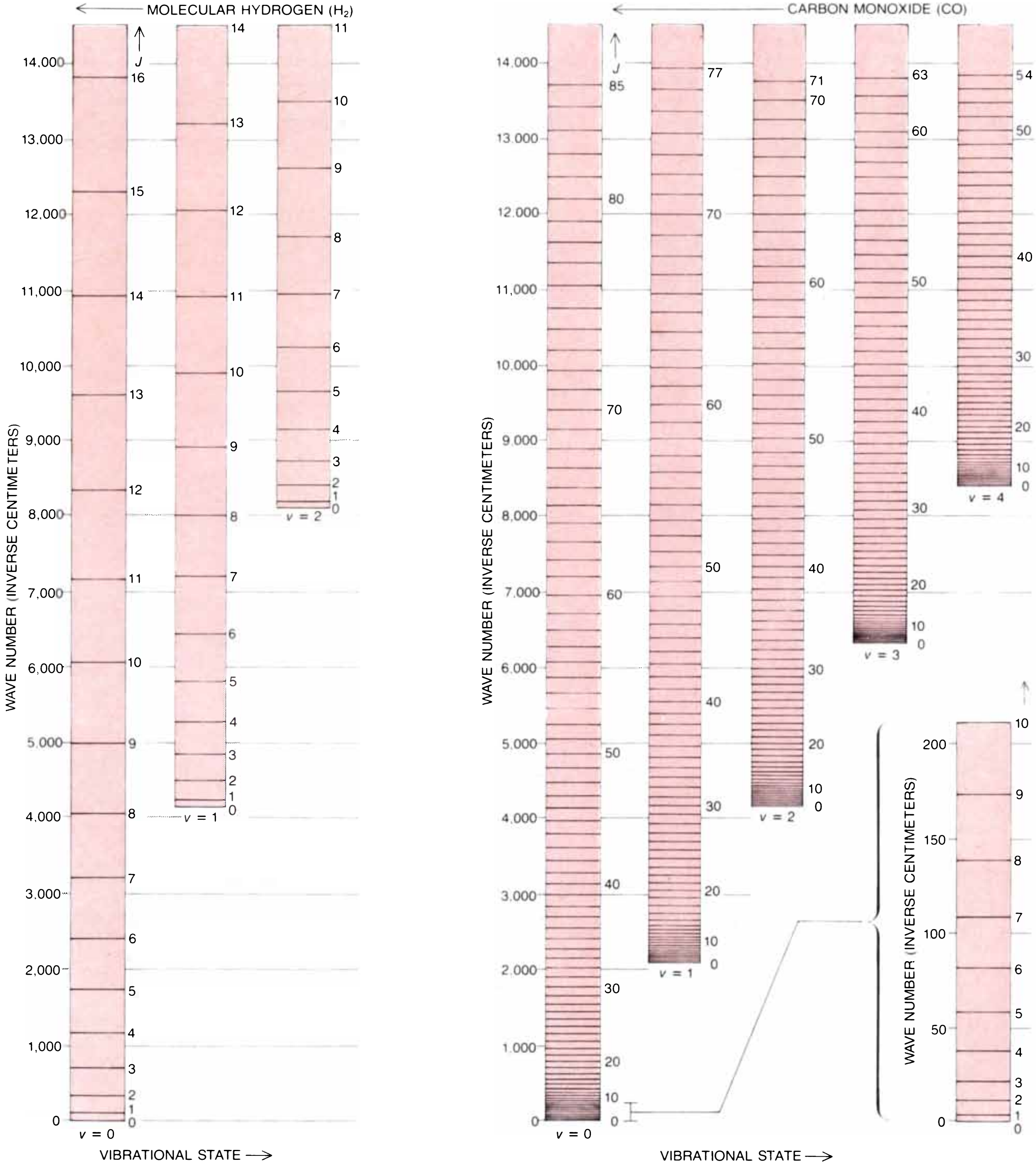
In most interstellar clouds the dust grains are well distributed and contribute about .5 percent to the mass of the



WAVELENGTH AND ENERGY DISTRIBUTIONS are compared for OMC1 and the compact infrared sources designated IRc1 and IRc2. Covering the full spectral range of these objects calls for observations from aircraft or balloons as well as observations from the ground. The wave numbers for important energy-level transitions of molecular hydrogen (H_2) and carbon monoxide (CO) are represented along the bottom of the graph. Some transitions are so closely spaced that they form bands. The three bands and the transitions indicated by the lines in color have actually been observed in Orion. The fact that IRc1 and IRc2 have emission peaks at a shorter wavelength than OMC1 has implies they are hotter and more compact. The dips in the curves for IRc1 and IRc2 at 3.1 and 9.7 micrometers are thought to result from the absorption of infrared radiation by grains of ice and silicate in the intervening interstellar dust.



GAS VELOCITIES IN OMC1 can be deduced from the magnitude of the Doppler shift in the carbon monoxide emission line at 2.6 millimeters, equivalent to a frequency of 115,270 megahertz. At most places in the gas cloud the emission is confined to a narrow line (*black*), but at the core an additional component with a broad profile is observed (*color*). The broadening indicates gas velocities are spread over the range from −50 to +50 kilometers per second. Broad component is confined to a region near IRc2 and only 40 arc-seconds, or .3 light-year, across.



TRANSITIONS BETWEEN ENERGY LEVELS of the diatomic molecules of hydrogen and carbon monoxide give rise to the radiation that has supplied much of the information about the gas clouds in the vicinity of the Orion Nebula. Diatomic molecules can exist at many different energy levels above the ground level, the minimum energy state of vibration and rotation for a molecule. In that state both the vibration quantum number, v, and the rotation quantum number, J, are zero. In a molecular cloud H_2 and CO can acquire energy and be raised to higher vibrational and rotational states by collisions with other molecules. For each vibrational state there are many rotational states. In general more energy is needed to effect a transition between vibrational states with the same J value than is needed to effect a transition between rotational states with the same v value. When a molecule falls from a higher state to a lower one, it releases a quantum of radiation in the form of a photon. The wavelength in centimeters of the emitted photon is the reciprocal of the difference in energy levels in wave numbers, shown along the vertical axis. For example, a molecule of hydrogen in the state $v = 1, J = 2$ has a wave number of 4,498 inverse centimeters. If the molecule fell to the ground state, $v = 0$, $J = 0$, where the wave number is zero, it would release a photon with a wavelength of 1/4,498 centimeter, or 2.22 micrometers. Radiation at that wavelength is emitted by hydrogen molecules in a few of the hottest regions in OMC1. For H_2, whose molecular structure is symmetrical, transitions between adjacent rotational levels are forbidden. Therefore in cool gas clouds energy is lacking to raise hydrogen molecules to levels where transitions can be observed. This does not apply to carbon monoxide molecules, which are asymmetric and free to move between adjacent rotational levels. In cool molecular clouds many transitions can be observed for carbon monoxide molecules in the state $v = 0$ and with rotational energies between $J = 0$ and $J = 6$. The lowest transition of the carbon monoxide molecule, $J = 1$ to $J = 0$, gives rise to a photon with a wavelength of 2.6 millimeters.

cloud. Jocelyn B. Keene, Roger H. Hildebrand and Stanley E. Whitcomb of the University of Chicago mapped the radiation from warm grains in OMC1 at a wavelength of 400 micrometers with the three-meter Infrared Telescope Facility of the University of Hawaii situated at an altitude of 13,600 feet on Mauna Kea. At this comparatively long wavelength and low resolution (about 35 arc-seconds) the infrared radiation appears to originate from a roughly elliptical region about two light-years in its longest dimension that embraces the core of OMC1. When the infrared source is observed at shorter wavelengths, where radiation from hotter grains becomes dominant, the radiation is found to be concentrated almost entirely within the core of OMC1, which suggests the core itself harbors the source of energy that heats the cloud. Measurements made from aircraft flying at an altitude above 12 kilometers (to reduce the absorption of infrared wavelengths in the earth's atmosphere) show that most of the radiation from the core of OMC1, amounting to about 100,000 times the power from the sun, is emitted between 20 and 300 micrometers (a range of wavelengths to which the earth's atmosphere is nearly opaque).

A map of much higher resolution than the one made at 400 micrometers (two arc-seconds compared with 35 arc-seconds) was recently obtained by Reinhard Genzel of the Center for Astrophysics of the Harvard College Observatory and the Smithsonian Astrophysical Observatory, Dennis Downes of the Max Planck Institute for Radio Astronomy in Germany, Eric E. Becklin of the University of Hawaii and me with the three-meter Infrared Telescope Facility on Mauna Kea at a wavelength of 20 micrometers [*see illustration on page 50*]. The map reveals that the infrared emission arises from a cluster of half a dozen or more objects surrounded by diffuse emission. From the variation of the infrared flux with wavelength we can estimate the apparent temperature of the objects. The hottest, designated IRc1 and IRc2 (for infrared compact source Nos. 1 and 2), radiate strongly in the range between two and 20 micrometers, indicative of a temperature of about 500 degrees K. Several other peaks, including IRc3 and IRc4, are only at about 150 degrees. All these objects have temperatures far below the temperature of the coolest normal star, and so they must represent concentrations of heated dust. The total power radiated by each concentration is hard to calculate because of the attenuating effects of dust in the foreground, but it probably exceeds the output of the sun by a factor of 10,000.

What is the nature of the compact infrared sources that lie within OMC1? Soon after their discovery it was conjectured that they might be protostars: objects intermediate in evolution between interstellar clouds and true stars. The first detailed calculations of the evolution and appearance of protostars were made by Richard B. Larson of Yale University in 1969. His theoretical model shows that a cloud collapsing under the influence of its own gravitational forces initially forms a nucleus of high-density material at its center. As more matter falls into the nucleus the gravitational energy that is released heats the dust grains in the outer shell of the protostar, causing them to emit infrared radiation. Meanwhile the density and temperature of the nucleus steadily rise until thermonuclear reactions are kindled at the center. For a protostar with the mass of the sun the entire process takes at least 10 million years; for protostars with the mass of Type *O* and Type *B* stars, however, the process is up to 100 times faster.

More recent work has shown that in high-mass protostars thermonuclear reactions start before the outer regions have had time to fall into the nucleus. The newborn star is surrounded by an infrared-emitting cocoon of gas and dust that is still collapsing. If the compact infrared sources observed in OMC1 are precursors of types *O* and *B* stars, their high luminosity is explained. According to this picture, the infrared cluster will eventually emerge as a small group of *O* and *B* stars with a size and luminosity similar to the stars of the Trapezium Cluster inside the Orion Nebula itself. The remainder of the molecular cloud will ultimately be heated to thousands of degrees and begin shining as a new ionized nebula.

Problems with this simple picture began to emerge in 1976, when the fast-moving gas in the core of OMC1 was discovered by Ben M. Zuckerman of the University of Maryland at College Park, Thomas B. H. Kuiper of the California Institute of Technology, Eva N. Rodriguez Kuiper of the University of California at Los Angeles, John Kwan of the State University of New York at Stony Brook and Nicholas Z. Scoville of the University of Massachusetts. They observed Doppler shifts in the carbon monoxide emission from the site of the infrared cluster that indicate gas velocities of plus or minus 50 kilometers per second with respect to the rest of the cloud. Such velocities are more than 10 times higher than those observed elsewhere in OMC1 and are much greater than the velocities that could be induced simply by gravitational collapse. The fast motions are confined within a small volume, about .3 light-year across, and are centered within about .1 light-year from IRc2. The distribution of Doppler velocities suggests a shell of gas expanding in all directions from one source.

Additional evidence of rapid motion in this region came with the discovery, also in 1976, of infrared emission lines from hot hydrogen molecules by T. N. Gautier III, Uwe Fink, Richard R. Treffers and Harold P. Larson of the University of Arizona. Their observations were made with the 2.29-meter telescope at the Steward Observatory. More than a dozen molecular-hydrogen transitions have since been observed, most of them between low-lying J levels of the $v = 1$ and $v = 0$ vibrational states at wavelengths near two micrometers.

Four things are remarkable about these emissions. First, they are intense, equivalent in total power to at least 100 times the output of the sun at all wavelengths. Second, the temperature of the hydrogen, as deduced from the ratio of the strengths of different emission lines, is about 2,000 degrees K., far hotter than the rest of the gas in OMC1. Third, the Doppler widths of the spectral lines are very wide, implying velocities of between 100 and 150 kilometers per second. Fourth, the hot gas is widely distributed in the OMC1 region, taking the form of clumps around the infrared cluster. These observations all suggest that the molecular-hydrogen emission is the result of a fast-moving "wind" blowing outward from the infrared cluster. As this gas, which probably includes the same material that gives rise to the broad carbon monoxide emission, collides with the surrounding parts of OMC1, it creates a shock wave that heats a thin, dense layer of OMC1 to 2,000 degrees.

A third item of evidence for expulsive gas motions in Orion comes from the observation of intense emissions from water masers. The word maser stands for microwave amplification by stimulated emission of radiation. A maser operates by raising a large number of atoms or molecules in a gas to a particular energy state and then stimulating them to tumble simultaneously to some lower state. Since the transitions to the lower state occur simultaneously, the rate at which photons can be emitted from a given volume of gas of a given density is enormously greater than the rate arising from the random kinetic processes I have been describing. In a water maser the gas consists of water molecules present in small quantities in the hydrogen gas. The first emission from a celestial water maser was observed at a wavelength of 1.35 centimeters in 1969. Since then water masers have been discovered in more than 100 different regions of the galaxy. In OMC1 at least 50 maser cloudlets are distributed over a region about the same size as the infrared cluster. The emission of a celestial water maser at 1.35 centimeters can be extremely intense. In order to emit an equal amount of energy at that wavelength by ordinary thermal processes a body would have to be at a temperature of more than 10^{12} degrees K.

The high intensity of the 1.35-centimeter line and the compactness of the water-maser sources enable radio astronomers to measure the distance between various maser cloudlets with great precision. The measurements are carried out by means of very-long-baseline interferometry. Simultaneous observations of the Orion region were made from radio telescopes in California, West Virginia, Massachusetts, Germany, Sweden and the U.S.S.R. by a group that included Downes and Genzel and Mark J. Reid and James M. Moran of the Center for Astrophysics. By observing over a period of two years the group was able to trace minute changes in the distance separating the maser cloudlets in Orion. With very-long-baseline interferometry separation angles were measured to an accuracy of 10^{-3} arc-second, which is about the angle subtended by a small coin at a distance of 4,000 kilometers.

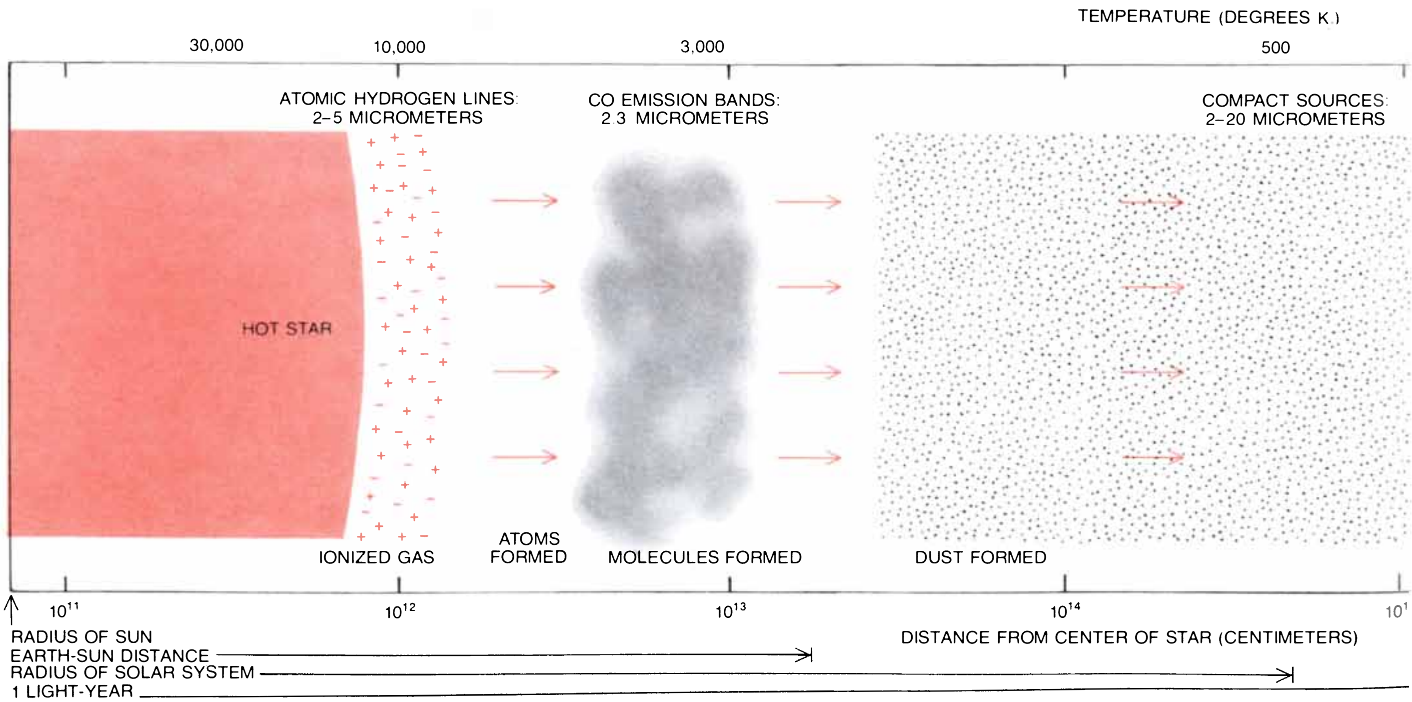
The measurements disclosed that many of the water-maser cloudlets are moving outward from a point somewhere in the vicinity of IRc2 or IRc4 with a velocity of 18 kilometers per second. The identification of IRc2 as the source of the outflow is suggested by the discovery of a second type of maser emission, one from silicon monoxide molecules, coinciding very closely with IRc2 and exhibiting an expansion velocity similar to that of the water masers. A few water-maser regions are moving away from IRc2 at speeds of up to 100 kilometers per second, but it is not yet certain whether they are parts of the same flow that propels the slower-moving water masers and the silicon monoxide maser or whether they emanate from one of the other infrared sources.

In any event the amount of matter involved in the flow from IRc2 is substantial, being equivalent to the loss of between a thousandth and a ten-thousandth the mass of the sun each year from an object unlikely to consist of more than 20 solar masses of material. So far there is no explanation for this enormous rate of mass loss, a loss about 100 times greater than the loss from mature Type *O* stars and one not predicted by any of the current theories of star formation. The possibilities that are being explored include mass loss driven by radiation pressure, by rotation, by acoustic waves, by magnetic effects and by wind-driven turbulence in the circumstellar disk of a newly formed star (perhaps even a preplanetary disk).

Whatever the mechanism of the mass loss it is clear that much of the activity observed in the core of OMC1 is directly attributable to processes taking place within one or more of the compact infrared sources. What hope is there of examining such processes in more detail and of finding out the nature and evolutionary status of these objects? A powerful tool that has recently been developed is the Fourier-transform spectrometer. This instrument is able to obtain detailed spectra of individual infrared sources at wavelengths of between one micrometer and five micrometers with a resolution that corresponds to a Doppler velocity of a few kilometers per second. Unfortunately IRc2 is faint in that range of wavelengths, but studies of the neighboring bright infrared source IRc1 have been carried out with the Kitt Peak four-meter telescope by Donald N. B. Hall, Stephen T. Ridgway and Fred C. Gillett of the Kitt Peak National Observatory, working with Susan G. Kleinmann of the Massachusetts Institute of Technology and Scoville.

Their spectrum of IRc1 exhibits both emission and absorption lines. Carbon monoxide molecules in various rotational levels of the ground vibrational state ($v = 0$) are excited up to levels in the $v = 1$ and $v = 2$ states, giving rise to two series of closely spaced absorption lines, one near 4.6 micrometers and the other near 2.3 micrometers. The absorption was expected; it is due to the gas in the part of OMC1 that lies in front of IRc1. Less expected is the finding that additional absorbing gas is traveling at about 40 kilometers per second away from IRc1 toward us, presumably part of a wind arising from the infrared source or a close neighbor.

Many emission lines are also observed at two to five micrometers in the spectrum of IRc1. Some of the lines represent transitions between the energy levels of atomic hydrogen rather than molecular hydrogen and are identical with lines emitted from interstellar



HYPOTHETICAL CONDITIONS in the vicinity of a newly formed star in OMC1 have been inferred from observations of IRc1 and IRc2. The observations are almost all explainable only if a powerful "wind" is blowing from the star, which is hidden from view in the center of each compact source. Presumably thermonuclear reactions have begun in the new star, heating its surface to about 30,000 degrees K. Ultraviolet radiation from the star ionizes the gas, chiefly hydrogen, in a surrounding shell. The shell absorbs all the ultraviolet radiation and yields emission lines from a small fraction of hydrogen atoms that are hot but un-ionized. As the outflowing gas cools, atoms combine to form molecules, including those of hydrogen and carbon monoxide. The latter emit radiation at 2.3 micrometers. Dust, heated

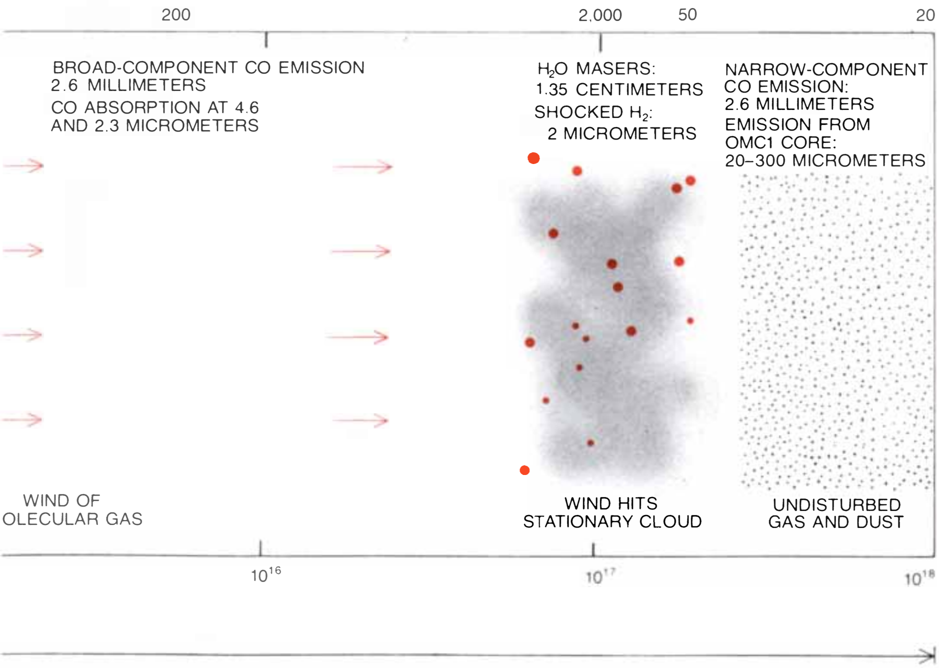
clouds that like the Orion Nebula itself are maintained in an almost completely ionized state by ultraviolet radiation from Type *O* and Type *B* stars. The fact that IRc1, unlike the Orion Nebula, does not emit continuum radiation (that is, radiation other than line radiation) at wavelengths of a few centimeters implies that the ionized region is highly compact, having a diameter of less than 10 light-hours. Continuum radiation is generated in an ionized gas by the deflection of charged particles as they are accelerated or braked when they approach other particles. The failure of such radiation to emerge from IRc1 must mean that it is being absorbed before it can escape.

The infrared spectrum of IRc1 also includes many emission lines from carbon monoxide. The strongest ones studied so far arise from transitions in which v changes by 2, namely the $v = 4 \rightarrow 2$, $v = 3 \rightarrow 1$ and $v = 2 \rightarrow 0$ transitions. Many different J levels are involved in these transitions, causing the lines to blur together. The blurring is particularly strong near $J = 50$ in each vibrational level, thus generating in the spectra bands rather than discrete lines. The surprising thing about the band emissions, which appear at wavelengths around 2.3 micrometers, is that they must originate from a region where very high energy levels of carbon monoxide (at least the $v = 4$, $J = 50$ level) are substantially populated. This can happen only if the temperature is above 3,000 degrees K. and the density is above 10^{10} molecules per cubic centimeter (which is dense for an interstellar gas but only about a billionth as dense as the earth's atmosphere at sea level). Gas under these conditions must be extremely close to its source of heat, closer even than the 500-degree dust that gives rise to the infrared emission observed at wavelengths between two and 20 micrometers.

Although the compact infrared objects IRc1 and IRc2 are by no means identical in their observed properties, they have enough in common to suggest a possible picture of what is happening in at least some of the infrared sources observed inside OMC1 and in other molecular clouds of similar appearance. Such objects have probably passed beyond the stage of accretion that would justify the use of the term protostar; nevertheless, there is little doubt that the infrared sources are associated with the very early stages of stellar evolution. Their existence as a compact group of objects, their association with dense interstellar clouds and their high rates of mass loss effectively rule out the possibility that they are actually stars in a late stage of evolution that have evolved off the main sequence.

The evidence suggests that lying in the center of the infrared objects is a hot star not unlike a main-sequence Type *O* or Type *B* star. Ordinarily such a star would emit enough ultraviolet radiation to maintain a substantial volume of gas surrounding it in a luminous ionized state, as is the case in the Orion Nebula. In the case of an infrared object, however, the star or the disk of material in orbit around it is ejecting so much matter in the form of a dense wind that all the ultraviolet photons are absorbed in a shell of gas very close to the star. The shell is almost totally ionized by the high-energy photons and gives rise to the observed emission lines of atomic hydrogen.

Beyond the range of the ultraviolet radiation the gas continues to flow outward; as its temperature falls the atoms in the gas are able to combine with one another to form molecules and dust grains. Hot molecules of carbon monoxide give rise to the band emission observed around 2.3 micrometers. The dust, which is heated to 500 degrees K. by direct radiation from the star (now a few light-hours, or about the radius of the solar system, away), emits infrared radiation at wavelengths of between two and 20 micrometers. Beyond the shell of dust the wind, now cooled to well below 500 degrees, carries molecules of carbon monoxide that yield a broad band of emission at wavelengths of a few millimeters. At a distance of about a light-month from the newborn star the wind finally collides with the relatively undisturbed part of the surrounding molecular cloud, giving rise to the infrared emission lines of hydrogen molecules heated to about 2,000 degrees. The water masers may be embedded in the wind itself or may be a phenomenon connected with the collision of the wind and the surrounding cloud.



directly by the star, is the source of infrared at wavelengths of between two and 20 micrometers. Between 10^{16} and 10^{17} centimeters from the center of the star the fast-moving carbon monoxide in the wind emits radiation with a wavelength of a few millimeters. At about 10^{17} centimeters, where the wind collides with the surrounding cloud, hydrogen molecules are reheated to 2,000 degrees K., giving rise to emissions at about two micrometers. Perhaps as part of the collision phenomenon cloudlets form water masers that radiate at 1.35 centimeters.

The discovery that at least some of the infrared sources once thought to be protostars are more probably very young, massive stars dramatically shedding mass has some important implications for the understanding of how new stars form. First of all, it means astronomers may have to search afresh for the precursors of typical main-sequence stars. Second, the wind from a large, luminous star may have a strong influence, either positive or negative, on the creation of smaller stars, such as those resembling the sun. On the one hand the wind could so badly disrupt the cloud surrounding it that further star formation would be impossible. On the other hand the pressure of wind on the neighboring parts of the cloud could promote the collapse of further fragments. Third, if a strong wind is a feature of the early evolution of all stars, not just massive ones, it could adversely influence the formation of planetary systems. Calculations of the actual effect, however, are difficult. If some astronomers confess to a slight disappointment that the search for protostars is again inconclusive, they are at least consoled that they may have discovered a new phase in the evolution of massive stars.

Dutch Elm Disease

This deadly fungal infection of elm trees may be brought under control with the aid of novel biological techniques that attack both the fungus and the beetles that spread it

by Gary A. Strobel and Gerald N. Lanier

Dutch elm disease is one of the most devastating plant diseases of the 20th century. It has killed millions of elm trees in Europe and North America, causing billions of dollars in direct economic losses and inflicting inestimable damage on the aesthetic quality of the landscape. Its cause has been known for decades: a fungus carried from tree to tree mainly by contaminated bark beetles. Only in the past few years, however, has a detailed understanding of the epidemiology and chemistry of the disease progressed to the point where there is reason to hope that the death rate of elms attributable to the fungal infection can be significantly reduced by an integrated program of pest management. The newest tools in this effort include two promising schemes for biological control: one is to manipulate the behavior of the bark beetles with the aid of chemicals essential for their communication and reproduction; the other is to inoculate individual trees with a bacterium that inhibits the growth of the fungus.

It was a Dutch botanist, Dina Spierenburg, who first called attention to the disease in the Netherlands in 1919. Soon afterward people in France and Belgium began to notice that elms there were dying too. Some observers attributed the disease to poison gas released during World War I. Others thought it was caused by attacks of bark beetles, facilitated by drought, or by a pathogenic bacterium.

Another Dutch investigator, Marie Beatrice Schwarz, announced in 1921 that she had discovered a fungus associated with diseased elm wood When she injected the putative pathogen into healthy elms, the trees developed symptoms of the disease; furthermore, she was able to reisolate the fungus from the newly infected wood. She had identified the causative agent by what pathologists call Koch's postulates (after the pioneering German bacteriologist Robert Koch).

Another part of the mystery of Dutch elm disease was how the fungus enters a tree's vascular system. Several European investigators suspected that bark beetles of the family Scolytidae, which are commonly found breeding in the phloem, or inner bark, of diseased elms, spread the infection. It was William Middleton of the U.S. Department of Agriculture's Bureau of Entomology and Plant Quarantine, however, who proved this supposition. In 1934 he reported that new infections appeared where the crotches of twigs were injured by the feeding of fungus-contaminated European elm bark beetles of the species *Scolytus multistriatus*. Later it was learned that several other species of bark beetles, both European and American, can also act as the vector, or carrier, of the Dutch elm disease fungus.

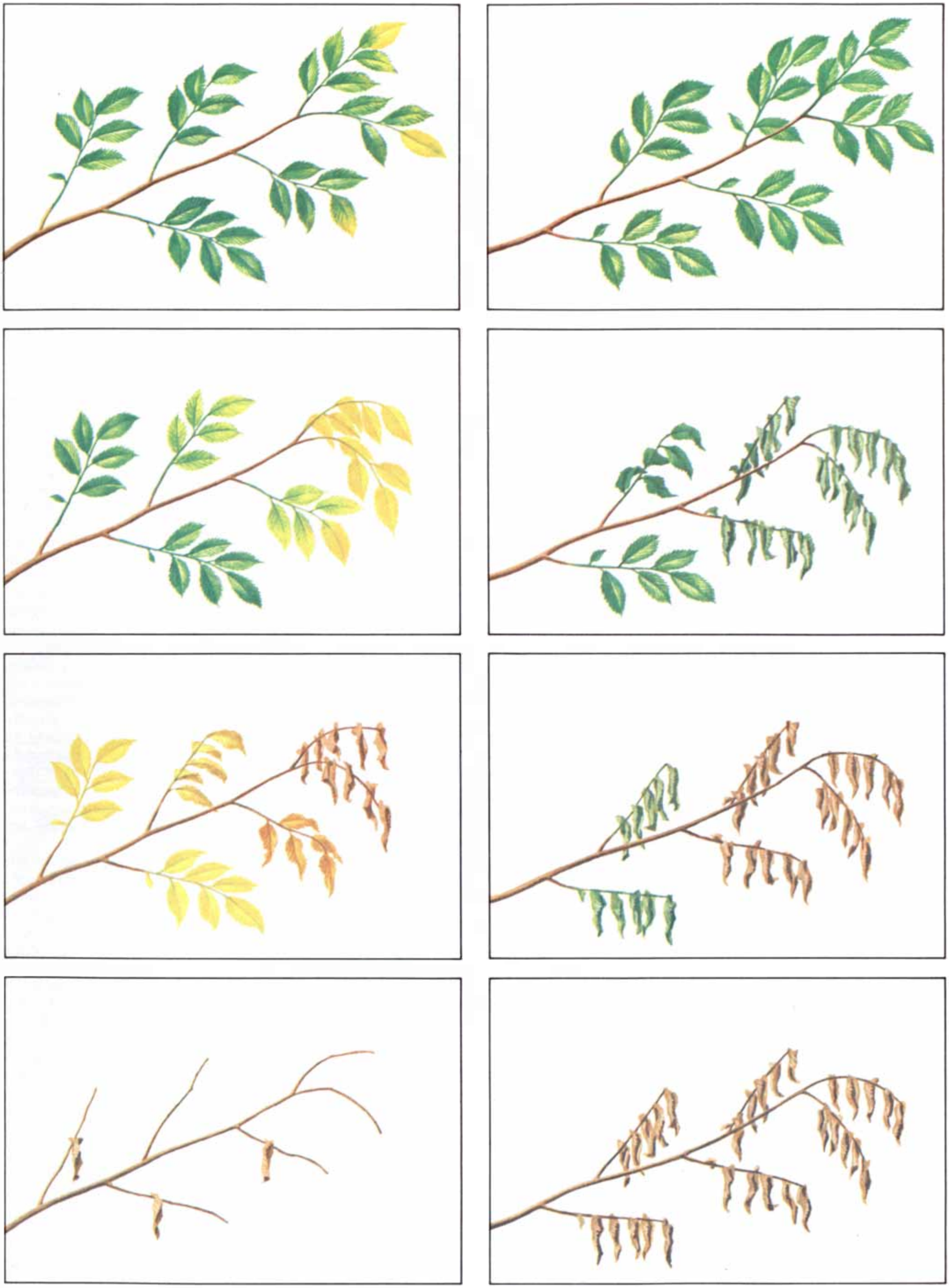
The symptoms of Dutch elm disease begin with the wilting of leaves, usually at the tips of the branches. The leaves may eventually turn yellow, curl inward and finally become brown and die. Alternatively they may dry out rapidly and turn a dull green before dying. When the bark is peeled from an infected limb, brown streaks are evident on the outer surface of the sapwood. Although other pathogens of the elm cause some of the same symptoms, they are seldom fatal. (One exception is elm phloem necrosis, or elm yellow disease, which is much less common than Dutch elm disease but is apparently always fatal; it is characterized by the uniform yellowing and death of the entire crown of the tree rather than the uneven wilting typical of Dutch elm disease.)

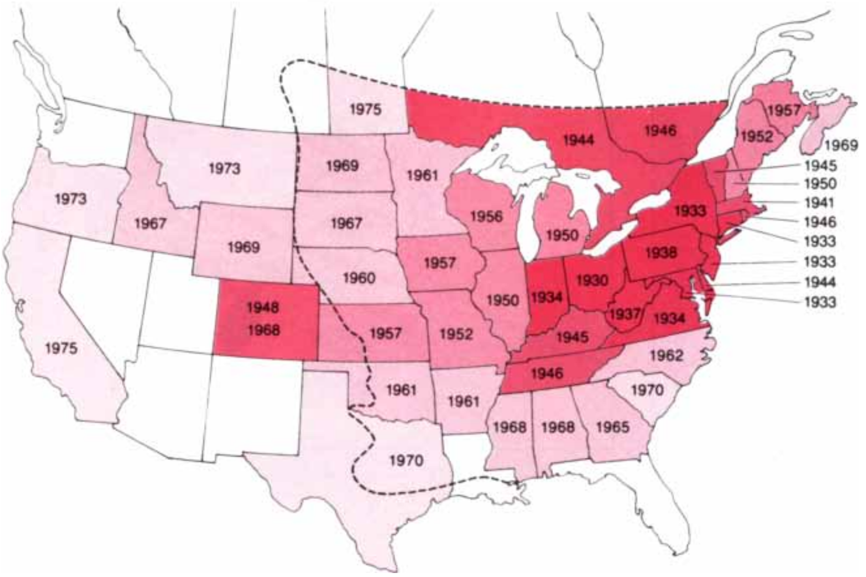
The expression of the symptoms of Dutch elm disease varies with the species of elm, the time of year, the weather, the mode of infection and the condition of the tree. The principal host in North America is the American elm (*Ulmus americana*), but the disease is also lethal to less common species, including the slippery elm (*U. rubra*), the rock elm (*U. thomasii*), the September elm (*U. serotina*) and the winged elm (*U. alata*). Most European elms, including the European white elm (*U. laevis*), the English elm (*U. procera*) and the various cultivars of the species *U. carpinifolia* are also susceptible. On the other hand, many Asian species, such as the Chinese elm (*U. parvifolia*) and the Siberian elm (*U. pumila*), are comparatively resistant to infection.

Symptoms that appear within four weeks of the emergence of leaves in the spring indicate a residual infection or an infection contracted during the dormant season through roots that have become grafted to the roots of a diseased tree. These early symptoms usually involve more than 10 percent of the crown of the tree. A new infection from fungal spores transmitted by bark beetles typically appears in a single branch 40 days or more after the leaves emerge. Hot, dry weather in the early summer enhances wilting, but extended drought seems to retard the succession of symptoms. An infection moves fastest and is most obvious in young trees and in vigorous mature ones. In old elms a new infection progresses slowly; for this reason and because an infection transmitted by bark beetles usually appears first in the top of the crown, Dutch elm disease in old trees is often not diagnosed until the second or third year after the onset of the infection.

As a rule only saplings and trees that

PROGRESSIVE SYMPTOMS of Dutch elm disease are evident in the leaves of the branch depicted in the two vertical sequences of paintings on the opposite page. In both cases the leaves begin to flag, or wilt, at the tips of the infected branch. In the sequence at the left the leaves gradually become yellow and curl inward before turning brown and dying. In the alternative sequence at the right the leaves dry out rapidly and turn a dull green before dying. In the first case most of the leaves fall from the branch soon after they die; in the second case they may remain attached to the branch for several weeks. In addition to the leaf symptoms shown here brown streaks can usually be found on the surface of the wood if the bark is peeled away.





SPREAD OF DUTCH ELM DISEASE across North America is recorded on this map. The dates mark the first appearance of the infection in each state and province. (Two dates are given for Colorado because the disease was apparently absent there for the intervening 20 years.) The graded colors show the advance of the disease across the continent by decade. The broken black outline traces the limit of the natural range of the American elm (*Ulmus americana*).

have been infected through the roots die in the first year of the infection. A tree infected in the spring by twig-feeding beetles loses part of its crown by the fall, and the symptoms progress during the following year. An elm infected in midsummer or late summer may go undiagnosed as the normal fall colors develop, but in the following spring the affected limbs die and the infection spreads through the tree.

Oceanic barriers to the spread of Dutch elm disease were overcome early in this century by the commerce in elm logs from central Europe, which were imported for making veneers. The disease was reported in England in 1927, and it turned up in Cleveland, Ohio, in 1930. Unfortunately warnings by American biologists that elm logs should not be imported went unheeded.

The Ohio infection was apparently eradicated, but independent disease centers in the New York City metropolitan area, New Jersey and southeastern Connecticut persisted and grew in spite of intensive and costly eradication efforts by the Bureau of Entomology and Plant Quarantine. In 1944 the disease was discovered in southern Quebec, and by the mid-1970's it had reached the west coast of the continent. Similar stories can be told of its advance through Europe and western Asia.

Dutch elm disease has been described as an epidemic-type disease because it spreads and intensifies rapidly in a population of susceptible hosts. The transmission of the disease is directly related to the distribution of the beetle vectors. Although the beetles by themselves can damage trees, the trees rarely die as a result. Even though the European elm bark beetle had preceded the fungus to North America by at least two decades, American elm trees were not seriously threatened until the fungus arrived. Once the beetle, the fungus and a large population of susceptible hosts came together, the epidemic began.

The main factor that determines the infection rate of Dutch elm disease is the size of the beetle population, which in turn is governed by the supply of elm tissue suitable for breeding. A healthy elm can successfully resist attempts by elm bark beetles to breed in it; the beetles tend to reproduce in trees that are either stressed by drought, injured or already diseased. In effect the fungus increases the population of beetles by enlarging the breeding habitat, whereas the growing beetle population causes a massive increase in new infections. The crucial stage of the epidemic is the early part. Unless the infection is controlled at the outset, the beetle and fungus populations multiply rapidly and the epidemic grows exponentially.

Ceratocystis ulmi, the fungus that causes Dutch elm disease, has two reproductive stages in its life cycle: one stage is asexual and the other is sexual. Infectious spores are produced at both stages. The asexual spore-bearing structures, called synnemata, are distinguished by a blue brown stalk about a millimeter long. The stalk is a bundle of filaments, called hyphae, that flares at the end to form a colorless head; there a large number of spores are suspended in a mucilaginous droplet. Synnemata form readily in the wood of a diseased elm tree, particularly in the galleries, or tunnels, carved by bark beetles, thereby facilitating contact between the vector and the pathogen.

The sexual form of the fungus, called the perithecium, is often found in the same galleries. It has a flask-shaped body with a long neck that forms after the mating of the two sexual types of the fungus. Spores grow in the perithecium in groups of eight, enclosed in an envelope called the ascus, which develops after the sexual processes of nuclear fusion and cell division are completed. There are many asci in each perithecium. When the spores are released from the asci, they float freely in the spherical part of the perithecium before they are extruded through an orifice.

At the biochemical level there are currently more questions about Dutch elm disease than there are answers. For example, why does *Ceratocystis ulmi* cause disease symptoms only in a few species of trees (the susceptible American and European elms) and not in other plants? Is the growth of the fungus dependent on some nutritional substance found only in living elms? The answer here appears to be no. The fungus can grow on dead elm wood, on a culture medium prepared from plant extracts and even on a synthetic medium. And yet when the fungus is injected into another plant such as a maple tree or an ash tree, it grows very little, causes no discernible damage and soon dies.

If the growth of the fungus in the susceptible elms is not dependent on a nutritional substance, can it be related to the synthesis in such trees of a growth-promoting substance, particularly in May and June when the trees are most susceptible to infection? Perhaps there are "recognition" molecules in both the fungus and the host. If there are, what are they? These questions remain unanswered.

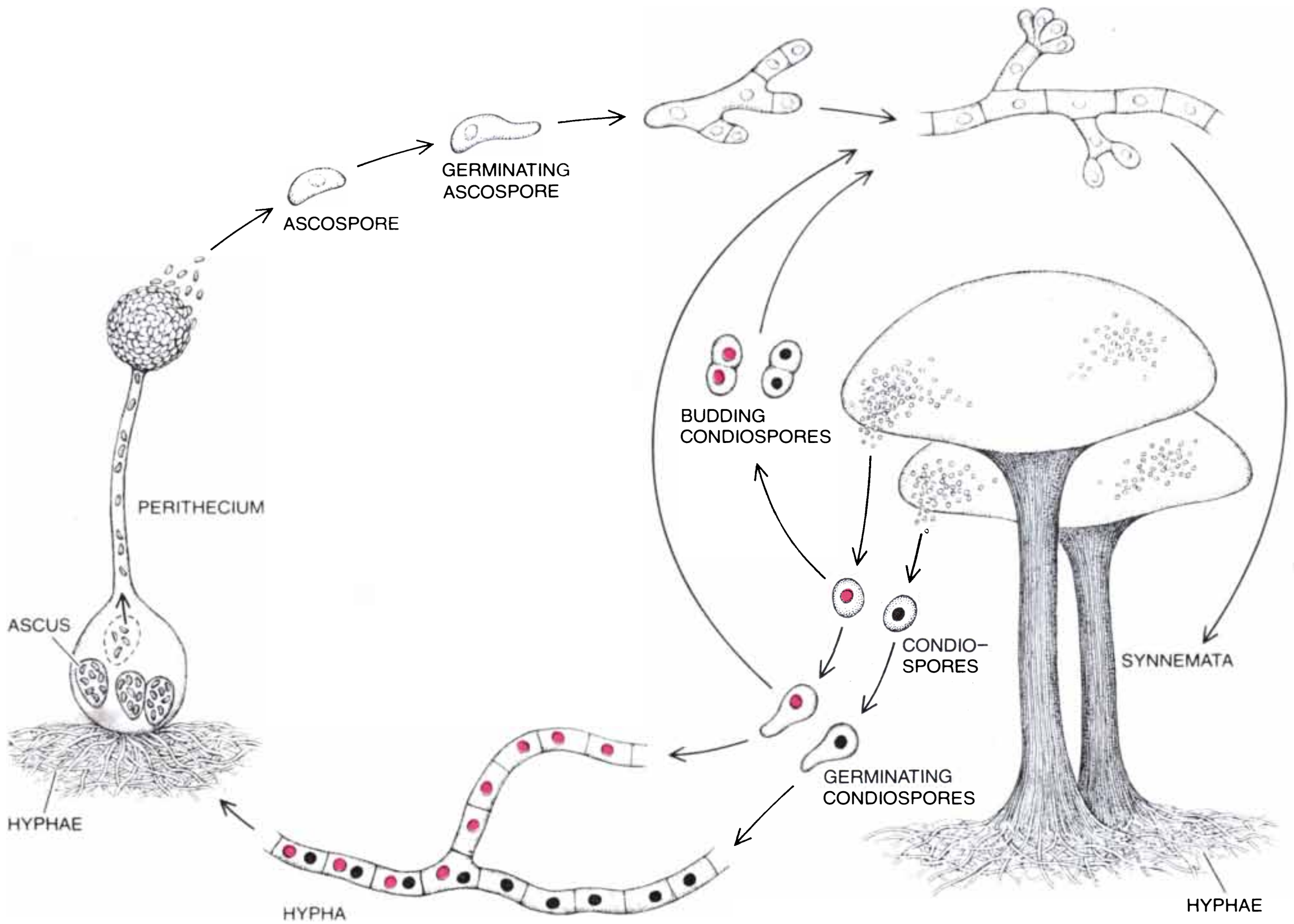
On another front some progress has been made recently toward explaining the mechanism by which the fungus damages a tree: it seems the fungus secretes a toxin. A substance that has been shown experimentally to cause wilting in elms has been isolated by a team of investigators, including Kenneth D. Hapner and one of us (Strobel) at Montana State University, Peter Albersheim of the University of Colorado at Boulder, Neal K. Van Alfen of Utah State University and John H. Nordin of the University of Massachusetts at Amherst. The existence of a toxin had been postulated previously on theoretical grounds, but it could not be characterized until it was isolated from the many other substances in the fungus cultures. We have found that the toxic substance is actually a mixture of molecules with the collective name peptidorhamnomannan. Each molecule consists of a protein backbone with at least two or three polysaccharide side chains, that is, chains of many sugar units. The protein backbone is enriched in the amino acids threonine and serine, to which the polysaccharide chains are attached [*see illustration on page 62*]. In addition shorter side chains of a few sugar units each are attached randomly to the protein. Measurements of the molecular weight of the toxin have yielded inconsistent results because of variations in the length of the polysaccharide side chains. The average molecular weight is quite high: about 130,000.

Van Alfen, working in collaboration with Neil Turner of the Connecticut Agricultural Experiment Station, has shown that the peptidorhamnomannan toxin does not directly damage the cells of an elm tree by acting as a poison. Instead it seems to interfere with the movement of water through the xylem, or woody core, of the tree. The interference appears to take place at structures called pit membranes in the wall of the xylem vessels. The supply of water to the leaves is maintained by flow from the vessels of the branch through the pit membranes to the vessels of the leaf. If water is transpiring from the surface of the leaf faster than it is being supplied by the tree, the leaves and stems begin to wilt. If the water deficit persists, cellular organelles such as the green chloroplasts are damaged and the leaf turns yellow. These effects are observed when the peptidorhamnomannan toxin is administered to elm cuttings; other large molecules have the same effect, but none of them is observed in cultures of *Ceratocystis ulmi.*

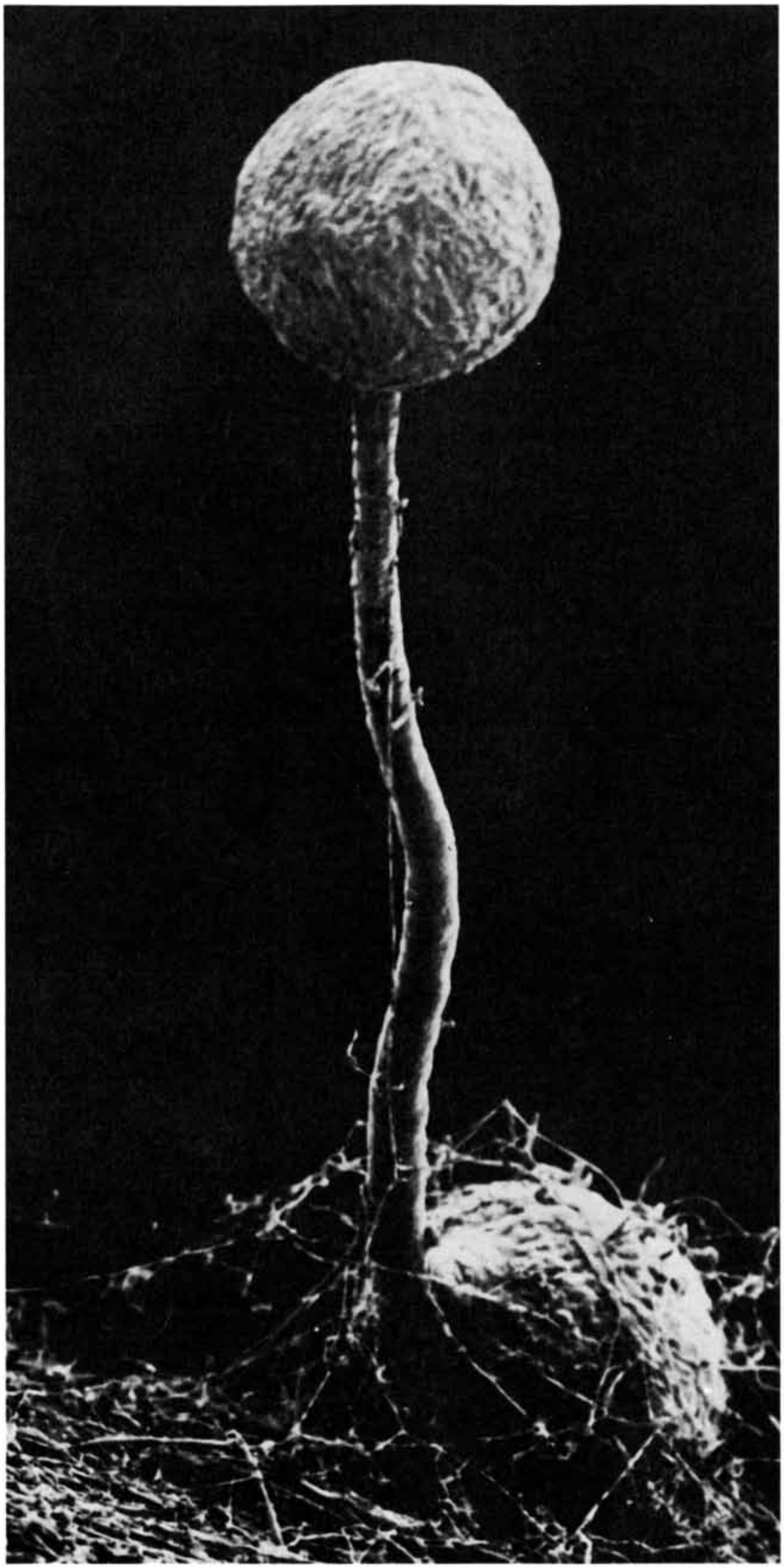
Ruud J. Scheffer and Doekle M. Elgersma of the Willie Comelin Scholten Laboratory in the Netherlands, working with a sensitive immunological technique, have demonstrated the presence of small amounts of peptidorhamnomannan in diseased elms but not in healthy ones. Since minute quantities of the toxin can measurably reduce the flow of water in elm cuttings, it appears that the toxin is present in the diseased tissues in amounts large enough to cause damage. Clearly the substance plays some role in the natural progress of the disease, although it is by no means clear that peptidorhamnomannan is what determines the specificity of the *C. ulmi* fungus for the elm tree. Other factors may be involved.

Without an insect vector Dutch elm disease would not be the problem it is. The reason is that the spores of *C. ulmi* are poorly designed for wind dispersal but are well adapted for insect transmission. Many species of insects can be contaminated with fungal spores but only bark beetles are commonly vectors of the disease fungus because only they move from diseased trees to healthy ones.

Elm bark beetles breed in the phloem of weakened, diseased, dying or newly cut trees. The female lays eggs in niches along the sides of egg galleries. In about eight days the larvae hatch and bore feeding tunnels that end in a pupal chamber. After the pupae mature the adults bore their way out of the bark and disperse, typically flying to another



LIFE CYCLE of *Ceratocystis ulmi,* the fungus that causes Dutch elm disease, is diagrammed. All the forms taken by the organism at various stages of its life cycle can be found in the galleries bored by elm bark beetles. The fungus has two fruiting bodies, one sexual, called the perithecium, and the other asexual, called the synnema (plural synnemata). The perithecium is made up of numerous structures called asci, each of which encloses eight ascospores. A perithecium is formed when the hyphae, or filaments, derived from the condiospores of two sexually compatible strains of *C. ulmi* fuse, bringing together the two kinds of cell nucleus (symbolized here by the black and colored dots). Meiosis, or sexual cell division, takes place in the perithecium, giving rise to ascospores in which the chromosomes have been reassorted. Condiospores, ascospores and the spores produced by a yeastlike budding process (*upper right*) are all infectious.



SEXUAL FRUITING BODY, or perithecium, of the Dutch elm disease fungus is seen in a scanning electron micrograph made recently by Ruud J. Scheffer of the Willie Comelin Scholten Laboratory at Baarn in the Netherlands. The spherical structure at the top of the hyphal stalk is filled with infectious ascospores. The magnification is approximately 300 diameters.

weakened elm to breed or to a healthy elm to feed. The beetles generally feed either in the crotches of twigs or in crevices in the bark of small branches. It is in the course of feeding that the beetles score the sapwood and infect healthy elm trees.

In Europe the two principal vectors of Dutch elm disease are the smaller elm bark beetle (*Scolytus multistriatus*) and the larger one (*S. scolytus*). Several other elm-breeding species of *Scolytus* and two species of the genus *Pteleobius* are also reported to transmit the disease there. In North America the main vectors are the native elm bark beetle, *Hylurgopinus rufipes,* and *S. multistriatus,* which is commonly known on this side of the Atlantic as the European bark beetle, owing to its introduction from Europe to the eastern U.S. sometime before 1909.

European and native bark beetles are similar in size (between two and 3.5 millimeters in length) but dissimilar in appearance. The European species is shiny and distinctly two-toned: the wing covers are dark reddish brown and the thorax is black. The native beetle is a drab grayish brown and has a rough texture. The immature stages of the two kinds of beetle can easily be distinguished by the gallery systems in which they are found. The female European beetle bores its egg gallery along the grain of the wood, and the tunnels made by the larvae radiate from the egg gallery, resulting in an elliptical pattern. The egg gallery of the native beetle crosses the grain, and the larvae tend to bore along the grain, creating a butterfly-shaped pattern.

The native elm bark beetle inhabits the entire natural range of elms in North America, but it is apparently absent in the Rocky Mountain states and westward, where elms were introduced in comparatively recent times. The European beetle is now found wherever elms grow in North America except for some northern areas where its overwintering larvae are destroyed by temperatures lower than about minus 25 degrees Fahrenheit.

Both beetles are common in the eastern U.S., where they can be distinguished by their different breeding habits. The more aggressive European species attacks trees at an early stage of decline, whereas the native beetle colonizes moribund tissue. In general the European beetle monopolizes trees growing in the open and the sunny upper crowns of trees that are closer together. The European beetle's avoidance of shady areas allows the native beetle to thrive in parks, woodlots and forests. Usually the European beetle begins its attack in the upper crown of a diseased tree, whereas the native beetle concentrates on the lower limbs and the trunk. Whenever the two species attack the same tree, the native beetles are soon outnumbered by the more rapidly developing European broods.

The rapid maturation of the European beetle enables it to complete two generations per year in the northern part of its range. In southern areas a third generation is begun and sometimes completed. The beetles winter over as fully grown larvae in the galleries and emerge as adults from early April to mid-June.

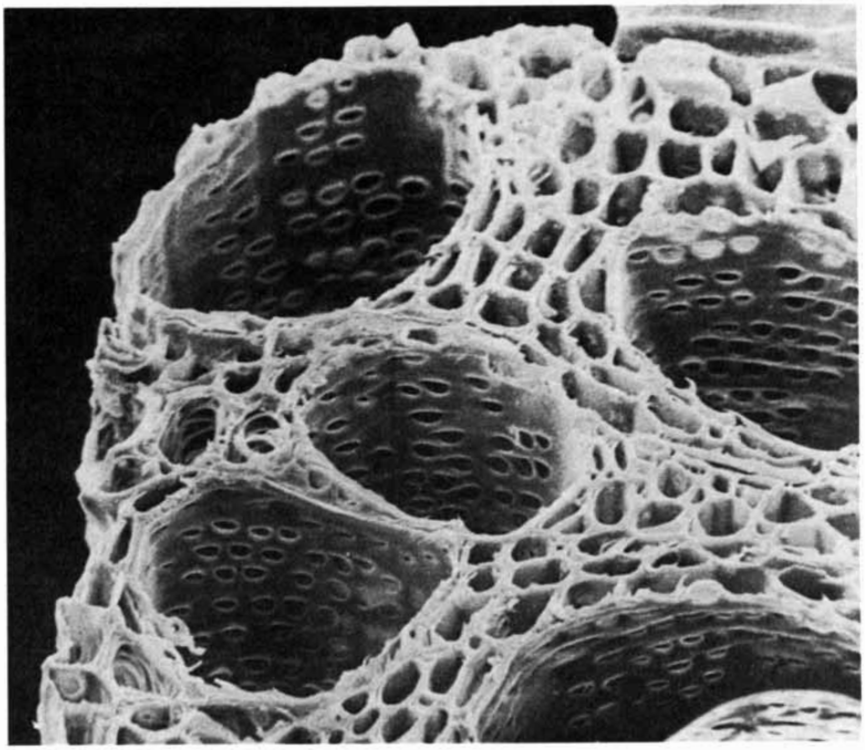
The native bark beetle completes only one full generation per year (or at most one and a half) and winters over either as a new adult or as a larva. Adults that will winter over emerge from the breeding site in late summer and fly to a living elm where they feed in the bark of limbs. As fall approaches they move downward to cut hibernation chambers in the cork, or outer bark, of the lower trunk. In early spring the tunnels are extended into feeding niches in the phloem. In April and May the overwintering adults fly to moribund trees to breed. Some of the adults leave the brood wood in May and June to produce a second brood in another tree. The adults of the new brood emerge from the tree in the summer, and the earliest of them produce broods that will spend the following winter as larvae.

A healthy elm is most susceptible to infection while it is forming early wood between late May and early July. Therefore it is the spring-maturing adults of the European beetle and the overwintering adults of the native species that are particularly effective in transmitting the fungus. The European bark beetle spreads the disease when the emerging adults fly to a healthy elm and bore into the crotches of twigs, where the fungus is deposited. The native beetle probably infects healthy elms in the spring when it extends its overwintering tunnels to the sapwood of its host or moves up the tree and feeds in the branches.

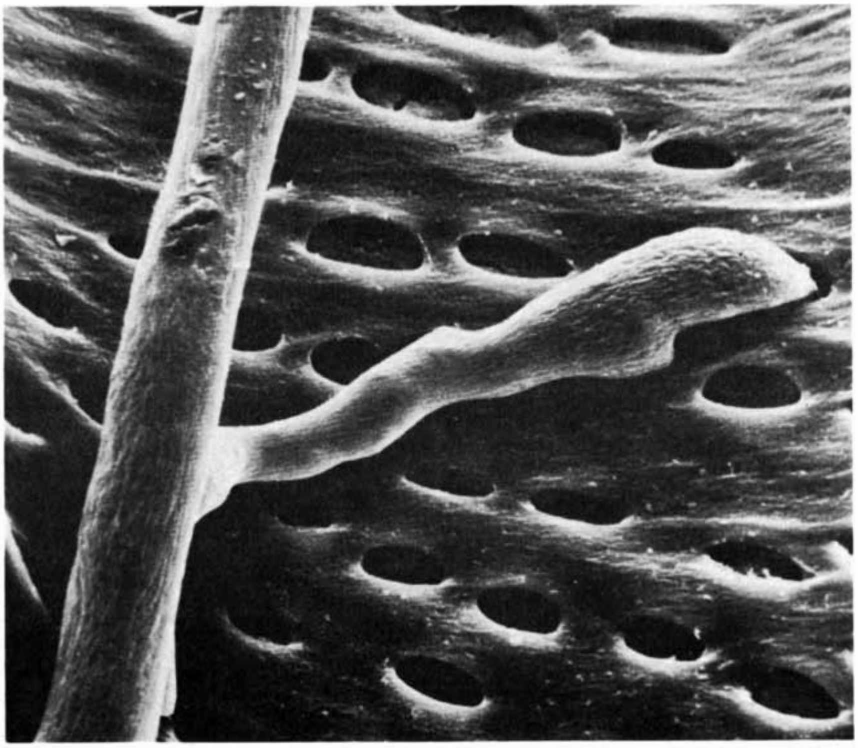
Beetles may colonize different parts of a tree for two generations or more, but the bark of one part of a tree can be utilized for breeding only once; hence each generation must find a new breeding site. The process of locating moribund bark for breeding and a healthy elm for feeding and wintering over is mediated by chemical odorants released by the host tree and by the beetles.

During the past decade one of us (Lanier) and others at the State University of New York College of Environmental Science and Forestry in Syracuse, working with a group at the U.S. Forest Service Shade Tree Laboratory in Delaware, Ohio, have studied the chemical ecology of the European bark beetle. Our ultimate objective has been to reduce the rate of Dutch elm disease by modifying the behavior of beetles with chemical odorants; the objective has been partially achieved.

Emerging European bark beetles are



CROSS SECTION OF ELM WOOD is enlarged about 700 diameters in a scanning electron micrograph made by Doekle M. Elgersma, a colleague of Scheffer's. The large vessels running longitudinally through the xylem, or woody core, of the tree are the tree's main water conduits. Each of the small pits, or openings, in the walls of the xylem vessels has a membrane that allows water to flow through it in a lateral direction. The membranes are not visible here.



HYPHAL THREAD of the Dutch elm disease fungus is seen growing along the inner wall of a xylem vessel of a piece of elm wood in a scanning electron micrograph made by Elgersma. A side branch of the hypha has grown out of the main thread and has penetrated at least two pits. The pit membranes, some of which are visible, look like cracked plates. A toxic substance secreted by *Ceratocystis ulmi* is thought to interfere in some way with the movement of water across the pit membrane. The magnification in this instance is approximately 2,500 diameters.

stimulated by ultraviolet radiation from the sun to disperse by flight. Flights of several kilometers are possible. After taking flight the beetles become responsive to the odors of elms, particularly the odors of trees that are weakened or diseased. Furthermore, when a female beetle finds a potential breeding site, it releases an odorant called an aggregation pheromone that announces the presence of wood suitable for colonization. As a result a tree found by a single female can become the target of a mass invasion.

In order to isolate and identify the pheromone, volatile substances were extracted from an airstream passing around elm logs infested with thousands of virgin female European elm bark beetles. (Males had to be excluded because our studies had shown that the attraction of the pheromone ceases soon after the females are mated.) The extract was divided into fractions by collecting parts of it at successive time intervals as they passed through a gas chromatograph. Laboratory bioassays in which beetles were presented with a choice of specimens to move toward first verified the attractiveness of the total extract and then identified fractions that could not be eliminated without reducing the response. Field tests confirmed that no one fraction was more than minimally attractive; the tests showed, however, that a combination of three particular fractions elicited a strong response.

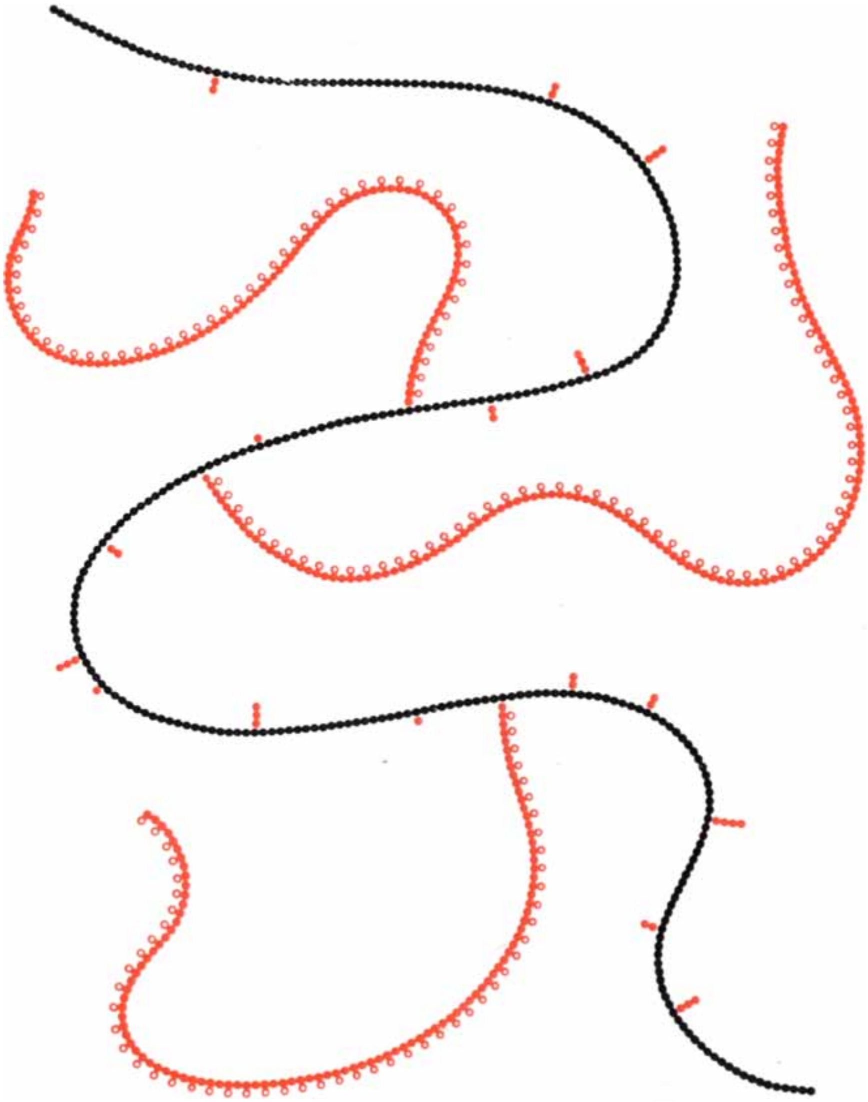
By repeated refractionation and bioassay of the three fractions, three pure compounds were obtained in microgram quantities. These were identified by spectrometric techniques and were synthesized by Robert M. Silverstein and his colleagues at Syracuse. The full chemical names of the compounds are 4-methyl-3-heptanol, $(-)\alpha$-multistriatin and $(-)\alpha$-cubebene; for convenience we refer to them respectively as compounds I, II and III [*see illustration on page 64*].

All three compounds are required for maximum attraction of the beetle. Compound III is a product of moribund or fungus-infested elm bark, whereas compounds I and II are released by virgin female beetles. Soon after mating, the release of compound I is terminated and the attraction ceases. Additional beetles that come on the aggregation center are then deflected by the excess of compound II over compound I to nearby moribund elms, where they breed, or to healthy elms, where they feed. This system of chemical communication accounts for much of the beetles' colonization behavior and for the contagious nature of Dutch elm disease.

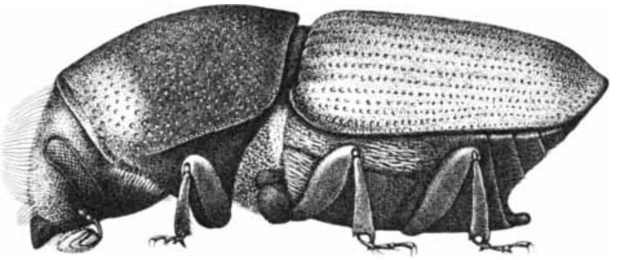
The chemical ecology of the native elm bark beetle differs from that of the European species. Volatile substances from moribund elm bark are the chief factor, and perhaps the only one, in the long-range attraction of the native beetle to its breeding material. Since the native beetle usually attacks elm wood that is in a more advanced state of decline than the wood attacked by the European beetle, a moribund potential host presumably differs in some significant way from a healthy tree and from the less stressed trees favored by the European beetle.

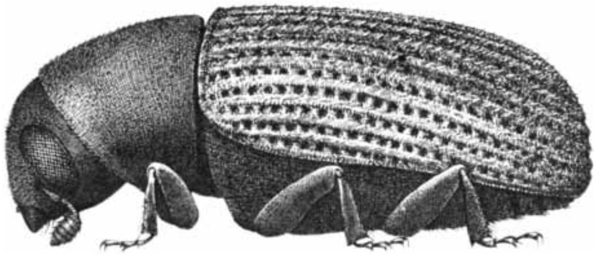
Five principal strategies have been mapped out for the control of Dutch elm disease. First, the disease can still be excluded from certain areas by quarantines. Second, it can be eradicated by the prompt removal of hopelessly diseased elms and by the elimination of elm wood that is likely to be colonized by the beetles (a combination of measures referred to collectively as sanitation). Third, the infection rate of a region can be reduced by controlling the rate of reproduction of either the fungus or the beetles. Fourth, valuable individual trees can be preserved by a variety of methods, including insecticides and fungicides. Fifth, susceptible species of elms can be replaced with resistant ones. Combinations of two or more of the strategies are also possible.

Quarantine and eradication are feasible only in those few isolated areas of



MOLECULAR STRUCTURE of peptidorhamnomannan, a toxic substance isolated from cultures of the Dutch elm disease fungus, is presented in a schematic diagram based on a model constructed by John H. Nordin of the University of Massachusetts at Amherst and one of the authors (Strobel). The molecule has a protein backbone made up of a long chain of amino acids (*black dots*). Attached to the central protein chain are several long polysaccharide chains. Each polysaccharide chain consists of linked units of the sugar mannose (*colored dots*) to which are attached numerous terminal units of the sugar rhamnose (*open circles*). Shorter side chains composed of only a few sugar units each are also attached randomly to the protein. The molecular weight of the toxin varies, depending on the length of the polysaccharide side chains.



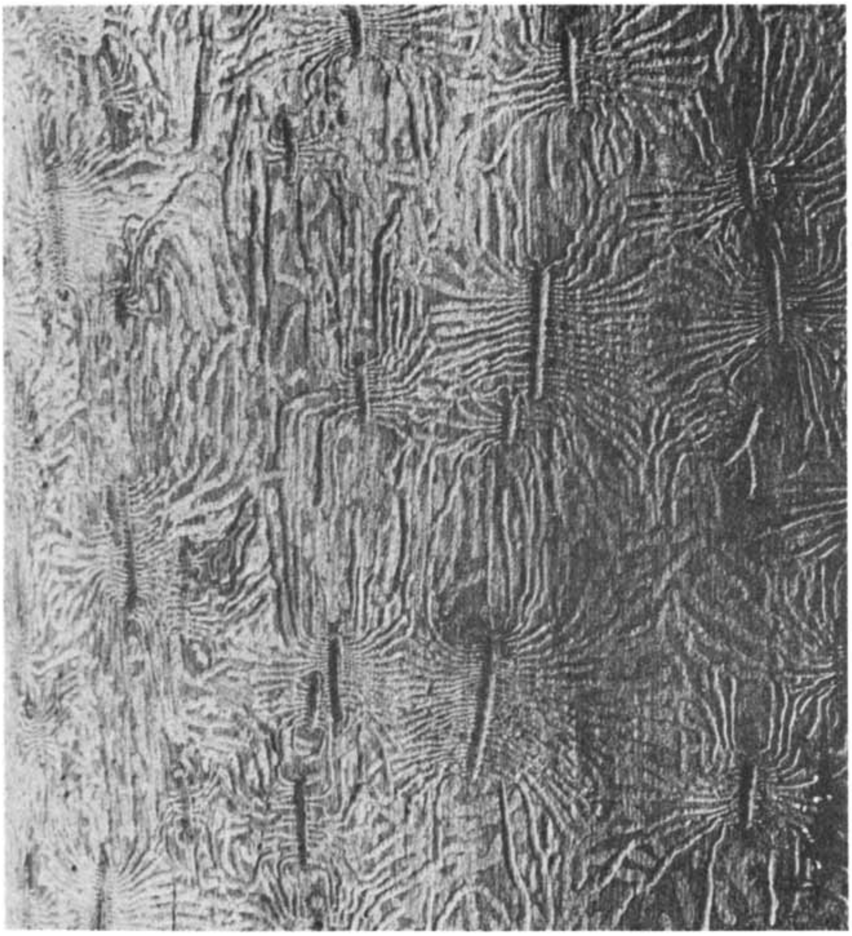


TWO SPECIES OF ELM BARK BEETLE are the principal vectors, or carriers, of the Dutch elm disease fungus in North America. The European species (*Scolytus multistriatus*) is represented by the male specimen at the left. It is shiny and two-toned: dark reddish brown on the wing covers and black on the thorax. The front of the male's head is covered with a dense brush of golden hairs, called setae. The native species (*Hylurgopinus rufipes*) is drawn at the right. It is a drab grayish brown and has a somewhat rougher texture. The two species have strikingly different silhouettes. The European beetle is concave under the back end of the abdomen, whereas the native beetle has a more domelike profile. The adult beetles of both species are approximately the same size: between two and 3.5 millimeters.

North America where the disease is not yet present or is rare. Eradication was attempted in the eastern U.S. before World War II, but the effort failed because the flying ability of the bark beetles had been underestimated, because cut elm wood was thought to be of no concern and because wild elms in which the disease could intensify without being noticed were abundant. Practical control of the disease, when it has been achieved, has been based on the control of the fungus or the beetle and to a lesser extent on the preventive and therapeutic treatment of individual trees.

The strategies that prove most cost-effective in pest management depend on a number of factors, including the size of the beetle population, the number and density of elms and the value assigned to the trees. One strategy, however, is essential to all control programs: diseased trees must be either treated or removed before they become the source of a new brood of beetles.

The protection of an individual elm in the absence of effective sanitation mea-



CHARACTERISTIC PATTERN engraved on the sapwood of a diseased elm tree by European bark beetles appears in the photograph at the left. The deep channels running parallel to the grain of the wood are the egg galleries carved by adult females. The shallower lines radiating away from each egg gallery are feeding galleries made by the hatched larvae. A closeup view of a row of larvae feeding on the phloem, or inner bark, of an elm is shown in the photograph at the right. Unlike the beetles of the European species, native elm bark beetles typically cut their egg galleries across the grain of the wood, and the larval feeding galleries generally run with the grain.

sures in the surrounding population of elms is a bleak prospect. The cost of an annual preventive treatment either by insecticide sprays or by fungicide injections can exceed $100 for a large tree. If the disease is in the area and bark beetles are abundant, even thorough annual treatments offer only limited protection. No method or combination of methods can be effective in holding Dutch elm disease at low levels if the supply of breeding material for the beetles is not controlled.

Among the newer methods for controlling Dutch elm disease are schemes for manipulating the behavior of the beetles with pheromones and host attractants and for biologically attacking the fungus with an antagonistic organism. At present plans are being tested for an integrated control strategy aimed at both the beetles and the fungus.

Since 1975 the SUNY group, led by one of us (Lanier), has deployed sticky traps baited with an artificial aggregation pheromone of the European elm bark beetle around groves of elms at 12 sites in eight states. The elms in the groves had been receiving care, but they had continued to suffer new infections of Dutch elm disease at a rate of between 5 and 20 percent of the healthy population per year. Sanitation in the groves was generally excellent. In a related program the Ohio-based Forest Service group, led by John W. Peacock, has done large-scale beetle-trapping experiments in Fort Collins, Colo., and Evanston, Ill.

The effect of the mass trapping of European elm bark beetles on the infection rates in the two city-wide programs has been equivocal. Although the losses to Dutch elm disease were generally lower in the test cities than they were in other cities in the same region, the rate of tree removal has not been decisively lower than what one would have expected without trapping. The test groves protected with pheromone-baited traps have yielded more encouraging results. With few exceptions the infection rate declined during each year of treatment. When the trapping program was discontinued in some of the groves in 1978, the infection rate increased in all but one grove, whereas it continued to decrease or remained at zero in the groves where the trapping was maintained.

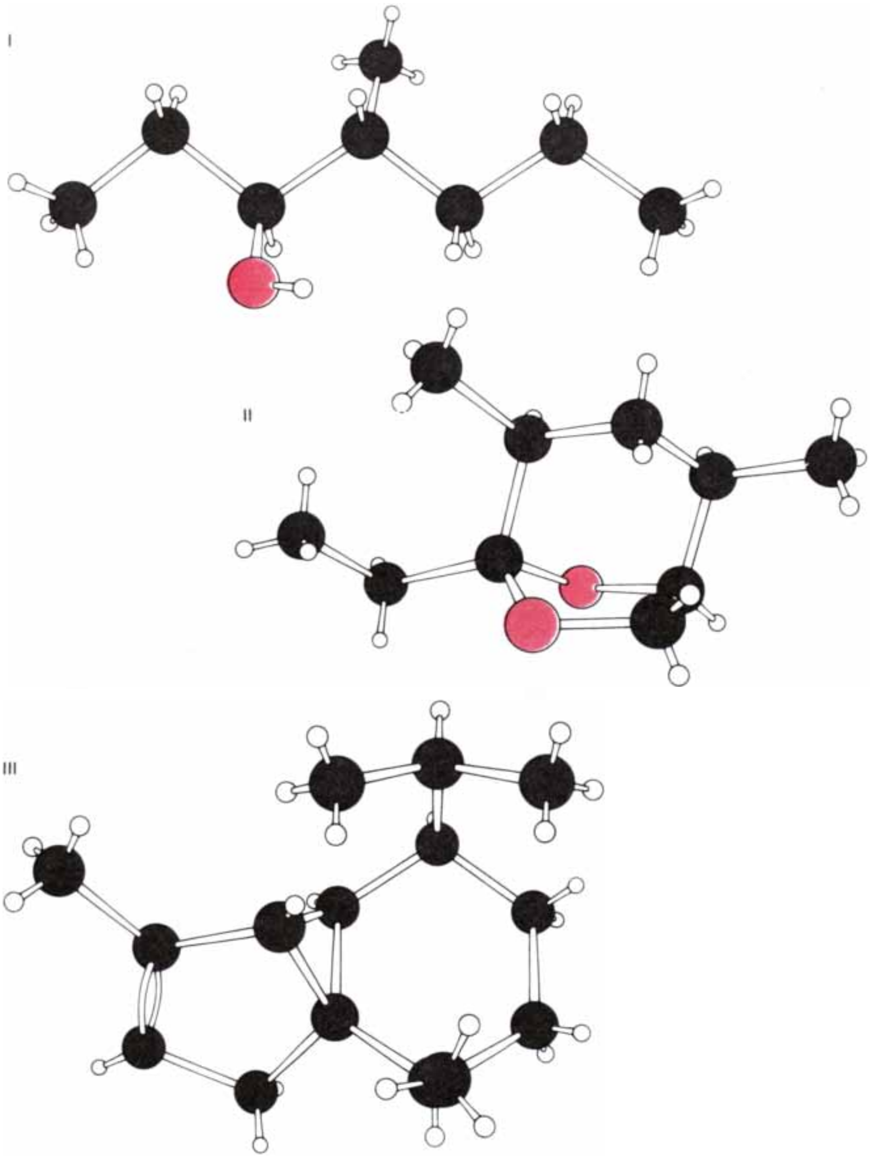
Can the destruction of potential parent beetles have a major impact on the size of the succeeding generation? Trapping experiments carried out by Martin C. Birch and his colleagues at the University of California at Davis suggest that it does not. Odorants from moribund elm wood attract some beetles even in the presence of the artificial pheromone. Once a few females bore into the bark, the weakened wood under colonization becomes competitive with the sources of the artificial pheromone. Although fewer beetles may be recruited to the attack than would be in the absence of trapping, the beetles that do participate create longer egg galleries with more eggs and their larvae have a higher survival rate because of the reduced competition for food.

Our most favorable results in protecting healthy elms by pheromone-baited sticky traps were achieved in areas that were comparatively free of brood wood and of newly infested material that acted as a competing source of the pheromone. Unfortunately such areas are rare. It seems clear that in order for pheromone traps for *Scolytus multistriatus* to be broadly effective sanitation efforts must be intensified. The fiscal and political obstacles to a wide-scale improvement in sanitation practices by traditional methods are formidable. New sanitation techniques are clearly needed.

The Dutch elm disease fungus itself is not necessary to generate elm wood that is attractive to bark beetles. Elms that are near death following exposure to the herbicide cacodylic acid, for example, are also attractive to the native elm bark beetle.

We have capitalized on the attraction generated by trees killed with cacodylic acid to manipulate the behavior of European and native elm bark beetles. After two years of developmental work on this "trap tree" method we began field

I



AGGREGATION PHEROMONES found to be attractive to the European elm bark beetle *Scolytus multistriatus* have been identified and synthesized. In these structural diagrams carbon atoms are black, hydrogen atoms are white and oxygen atoms are color. Reading from top to bottom, the compounds are 4-methyl-3-heptanol, (−)α-multistriatin and (−)α-cubebene; for convenience they are also referred to respectively as compounds I, II and III. Compounds I and II are produced by the beetle, whereas compound III emanates from a wounded elm tree.

trials in Syracuse in 1978. Hopelessly diseased and unwanted "weed" elms in green spaces were injected with cacodylic acid and baited with the pheromone. The trees were heavily attacked by both species of elm bark beetle, but the beetle broods failed substantially or entirely owing to the drying of the inner bark induced by the herbicide. By the end of the third year (1980) the infection rate in the treatment area had fallen from 7.7 percent to .6 percent. In an area set aside as an experimental control the infection rate hovered at about 5 percent from 1977 through 1979; when treatment was extended to this area in 1980, the infection rate fell to 2.2 percent.

As Dutch elm disease is brought under control in the future the trap-tree method can be phased out in favor of prompt sanitation and the mass trapping of bark beetles. At the same time other measures to protect elms can be taken when a comparison of cost and expected benefits indicates that they are warranted. In concert with the reduction in the number of beetles in a given area, an integrated program of pest management also calls for new procedures to control the pathogen.

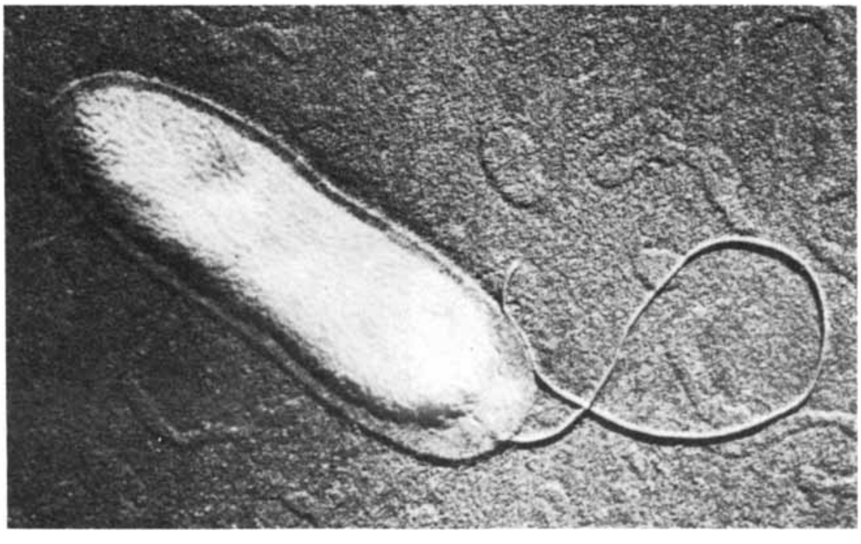
The survival of individual trees is usually important to homeowners and to others with an interest in the landscape. For the past few years fungicides based on the organic compound benzimidazole have been employed for the therapeutic and prophylactic treatment of Dutch elm disease. Their effectiveness has been limited for several reasons. First, they inhibit the growth of *Ceratocystis ulmi* but they do not kill it. Second, certain strains of the fungus are totally resistant to them. Third, the fungicides must be administered to the tree repeatedly because their effect is diluted as the tree grows. Such persistent treatment has also been shown to weaken the tree and to allow the entry of other pathogens through the treatment sites.

It was in part to circumvent such problems that work on a new biological approach to the treatment of individual elm trees was undertaken at Montana State by a group of investigators led by one of us (Strobel). Realizing that certain bacteria of the genus *Pseudomonas* synthesize antimycotics, or antifungal agents, we screened a collection of these organisms for antagonists of *Ceratocystis ulmi* and found several of them. Because some members of this bacterial genus are often found in association with plants we hoped that one of the antimycotic-producing organisms, when it was injected into an elm tree, would take up residence there and would confer on the tree some degree of resistance to *C. ulmi*. Accordingly we selected a strain of *Pseudomonas syringae* that was particularly effective as an inhibitor of *C. ulmi*. We learned that it would grow on elm

TRAPPED BEETLES ARE EXAMINED by one of the authors (Lanier) in the course of an experiment done recently at the State University of New York College of Environmental Science and Forestry in Syracuse. The beetles were attracted to the sticky paper by a synthetic aggregation pheromone. The smaller insects are European elm bark beetles of both sexes.

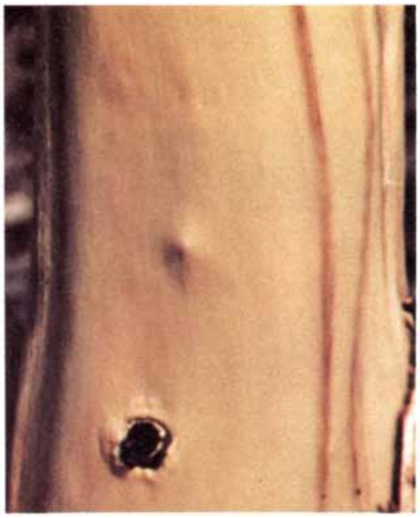
sap and would make its antimycotic on an agar medium containing sap; moreover, neither the bacterium nor its semipurified antimycotic was found to be toxic to elm trees. We also showed that the bacterium would sometimes move in the elm tree along with the regular flow of sap and would establish itself as a resident throughout the tree.

Cultures of the bacterium were administered to elm seedlings in a greenhouse in mid-May. After two weeks the trees were inoculated with a pathogenic culture of *Ceratocystis ulmi*. When the



BACTERIUM BEING TESTED for its efficacy in treating Dutch elm disease in individual trees is seen in an electron micrograph made by Bruce Hemming of Montana State University. The bacterium, a member of the species *Pseudomonas syringae*, has a whiplike appendage called a flagellum that enables it to move. Magnification is approximately 40,000 diameters.

INHIBITION **of the Dutch elm disease fungus** ***Ceratocystis ulmi*** **by the antagonistic bacterium** ***Pseudomonas syringae*** **is apparent in this laboratory demonstration. Four colonies of** ***P. syringae*** **were first established on a glass dish; they appear in the color photograph as the square array of cream-colored dots. A suspension of spores and hyphal fragments of** ***C. ulmi*** **was then sprayed on the dish. After several days of incubation the fungus became visible as a white layer covering all of the dish except the circular zone around each bacterial colony. An antimycotic agent synthesized by the bacterium effectively inhibits the growth of the fungus.**



ANTIFUNGAL EFFECT of ***Pseudomonas syringae*** **in living elm saplings was confirmed in this test. The tree at the right was treated with a culture medium bearing an antimycotic-producing strain of the bacterium. The control tree at the left was injected with an identical culture medium that had not supported the growth of the bacterium. (The injection point in both cases is the dark spot near the bottom.) Two weeks later both trees were subjected to a massive dose of spores and hyphal fragments of** ***Ceratocystis ulmi.*** **A month later the bark was peeled from the trees. The wood of the control tree was completely discolored by the fungus, whereas the area above and below the injection point in the treated tree appeared to be free of such discoloration. Cultures of** ***P. syringae*** **were recovered from the uncolored zones of the wood. In the treated tree the fungus presumably moved between the zones of wood where the bacterium had become established, giving the wood a streaked appearance resembling that of a candy cane.**

trees were examined eight weeks later, none of the trees treated with the bacteria showed symptoms of Dutch elm disease. Control trees exposed to the fungus alone or treated with bacteria that do not produce antimycotics all exhibited symptoms of the disease. The effectiveness of *Pseudomonas syringae* in disease control seems to be related to the ability of the bacterium to make antimycotics, because a mutant that does not synthesize them was completely ineffective in controlling the disease.

The biological treatment of mature diseased trees in the field has also shown some promise. In one experiment 22 diseased trees were treated with the bacteria and an equal number of diseased trees were left untreated as controls. All the control trees but one either died or declined drastically in the course of two growing seasons. In the treated group seven of the eight trees that were lightly infected and were treated early in the growing season survived with little or no sign of decline over two growing seasons. The other treated trees that either were treated later in the season or were more heavily infected died or declined. We tentatively conclude that the treatment is most successful if it is done early in the season during the time of maximum sap flow in the tree. Three of the treated trees had apparently made a complete recovery, as shown by the fact that it was not possible to reisolate *Ceratocystis ulmi* from them.

So far thousands of trees across the U.S. and Canada have been treated with *Pseudomonas syringae.* In refining the method of treatment in further tests we need to consider several factors: the ability of the bacterium to survive and spread in different kinds of elms, the remote possibility that the bacterial strain may be pathogenic to certain plants and the possibility that other strains of *Pseudomonas* may prove more effective or that the strain already tested could be genetically improved. In addition the human and environmental safety of *P. syringae* must be established.

An integrated pest-management program consisting of sanitation measures, beetle trapping and the treatment of trees with fungicides and antagonistic bacteria is now being tried in several communities in the U.S. Although the new biological techniques have shown promise, more testing is needed. In the meantime one can rest assured that substantial numbers of elms will be preserved by various control programs. Even in the countryside, where no care is provided, the American elm is not destined for extinction. Its tremendous power of reproduction will surely save it. In the longer run one can also hope that natural selection or breeding programs will someday create an American elm able to ward off *Ceratocystis ulmi* as well as its Asian relatives do.

SCIENCE AND THE CITIZEN

Mother's Milk

At the World Health Assembly last May, when the U.S. cast the only vote against a code regulating the promotion of infant formula, the U.S. position was defended in terms of market principles. The code, which was adopted with 118 affirmative votes, limits the advertising of formula as a substitute for breast-feeding and requires that mothers be informed of the advantages of breast milk. The chief aim of the code is to curb the promotion of formula in the Third World. Although compliance with the code is entirely voluntary, spokesmen for the U.S. Department of State argued that it represents unwarranted interference with the action of market forces. This focus on the restraint of trade had the effect of obscuring the underlying issue: the substantial evidence that breast milk is far superior to formula in providing nourishment and in protecting an infant's health.

In the course of evolution each species of mammal has elaborated a milk suited to the maturational needs of its young. Commercial formulas are based on cow's milk or on other sources of protein such as processed soybeans. Carbohydrates, vitamins, iron and additional protein are generally combined with the base. At best such a formulation can approximate the nutritive properties of breast milk. Work in immunology has shown, however, that formula lacks certain proteins that serve to protect the breast-fed infant from infection.

Among the proteins that cannot be supplied by formula are immunoglobulins, including antibodies to a wide range of antigenic substances. Of the various classes of immunoglobulin molecules the infant at birth has a full complement of only one type, designated IgG; it is transported across the placenta during the last phase of gestation. This ration of maternal IgG protects against viral and bacterial infections to which the mother has developed immunity. The donation is sufficient for as long as a year, during which time the child's immune system develops the capacity to make its own IgG.

Two other immunoglobulins, IgA and IgM, do not cross the placental barrier in significant amounts; their absence is compensated for by breast milk, which contains both, including a particularly high level of IgA. By binding to the antigens present on the surface of bacteria IgA prevents microorganisms from penetrating the intestinal lining and initiating infection. Alterations in the chemical structure of the IgA in milk make it more resistant to degradation by bacterial enzymes than the IgA found in blood serum. IgA in an infant's intestinal tract is therefore an excellent source of protection against gastroenteritis, a major cause of death among infants in poor countries. Commercially prepared formulas cannot supply this protection. The immunoglobulin in cow's milk is mainly the redundant IgG, and even it loses its potency when the milk is pasteurized in the process of being converted into formula.

Several other components of human milk have immunological properties. Lactoferrin, an iron-binding protein in the whey, inhibits bacterial growth; lactoferrin is particularly effective against *Escherichia coli,* a common agent of intestinal infection. Human milk has much more lactoferrin per unit volume than cow's milk does. In addition the heating needed to prepare formula denatures lactoferrin and renders it ineffective. Both macrophages and lymphocytes, the white blood cells that mediate many immunological responses, are present in breast milk. Macrophages make up 90 percent of the white cells in the milk; they are capable of engulfing and destroying both viruses and bacteria. The precise function of the lymphocytes, some of which are producers of IgA, has not been completely elucidated. It has been determined, however, that women who are orally immunized to a nonpathogenic strain of *E. coli* late in pregnancy secrete milk whose IgA-producing lymphocytes make antibodies to *E. coli.* The mother's acquired resistance may thus be transmitted to the infant by way of the white blood cells as well as by the direct transmission of antibody molecules.

It would be reasonable to expect that such differences between commercial infant formula and human milk would lead to disparities in health between groups of infants fed on one substance or the other. There is evidence of such disparities. Studies done some decades ago in Europe and North America and the analysis of more recent experience in developing countries indicate that breast-fed infants have lower death rates than infants fed on substitutes; specifically, differences have been found in death rates from intestinal and respiratory infections, in which the immunological response is critical. Because of the historical circumstances of the older studies, none is without flaw in method, and all have been roundly criticized by the makers of infant formula. Taken together, however, the studies reveal an unmistakable trend.

In a review of the older findings Joe D. Wray of Columbia University cites eight studies comparing the mortality of bottle-fed and breast-fed infants. The evaluations were made between 1900 and 1947 in eight European and North American cities. In every study the death rate was higher among bottle-fed infants than it was among breast-fed infants. In seven studies the ratio of the rates was between 2 : 1 and 10 : 1. In the eighth study the ratio was still higher. Studies done in Derby, England, in 1900–1903 and in Liverpool in 1936–42 included information on the cause of death; they showed that breast-fed infants had appreciably lower rates of death from both upper-respiratory and intestinal infections.

Infant mortality has been greatly reduced in the industrialized countries since these studies were completed. Conditions in the developing countries, however, are now comparable to those that prevailed in Europe and North America when the earlier investigations were carried out. Indeed, similar results have come from the few reliable studies done in the Third World. Surveys made by the Pan American Health Organization in El Salvador, Jamaica, Colombia and Brazil suggest that infants who were breast-fed for six months or less were from six to 14 times as likely to die in their first year as infants fed at the breast for more than six months.

In recent analyses of infant mortality among the middle class in affluent countries there is little correlation between the death rate and the duration of breast-feeding. Excellent health care and sanitation clearly compensate for the failings of commercial formula. Where sanitation and medical care are rudimentary, however, as they are in most of the developing world, the protection afforded by breast-feeding is crucial. In many poorer countries infants' bottles are used repeatedly without being washed. Moreover, a mother living in poverty must often progressively dilute portions of formula in order to extend them; malnutrition is the potential result.

Kinks in the Chain

Solitons, or solitary waves, were once a curious backwater of mathematics. They were considered charming for the serendipitous circumstances of their discovery and elegant for the properties that make them solutions to certain equations. It was not thought likely, however, that they would contribute much to the description of physical reality. Now the scientific world seems to be awash with solitons. They appear in crystal-lattice theory, nonlinear optics, plasma physics, molecular biology, elementary-particle theory, oceanography and at least a dozen other fields. They may be responsible for the Great Red Spot on Jupiter. Perhaps most significant, they cause unusual changes in the

properties of solids that have recently been observed by several groups of experimenters.

The first observation of a soliton was made in 1844 by John Scott Russell, a British naval architect and engineer. Scott Russell was watching a boat being drawn through a channel by a pair of horses; when the boat abruptly stopped, he noted that the bow wave continued to move forward. It assumed "the form of a large, solitary elevation," he wrote, "a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles per hour, preserving its original figure some 30 feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel."

Mathematical analysis shows that the localized and nondissipative character of a solitary wave is the result of a balance between two effects. One effect is the tendency of the wave to disperse because components of the wave that have different frequencies move at different speeds: the higher the frequency, the lower the velocity of propagation. As a result of such frequency dispersion most waves do not retain their shape over long distances. In a solitary wave, however, frequency dispersion is balanced by the continuous buildup of higher-frequency components. If there were no dispersion to compensate for the accumulation of higher frequencies, the wave would steepen as it moved; ultimately the wave would break, much as a whitecap does on the surface of the ocean. If the dispersion of the higher frequencies is exactly balanced by the introduction of the same frequencies, a solitary wave is formed.

The delicate balance that seems to be required for the creation of a solitary wave led most physicists to regard solitons as mere mathematical curiosities. In 1967, however, Clifford S. Gardner and his colleagues at Princeton University showed by further mathematical analysis that solitary waves are not at all difficult to generate in physical systems. Indeed, they showed that rather special initial conditions must be imposed in order to avoid creating at least one solitary wave. Since then interest in solitons has been renewed and many examples have been discovered.

In the Andaman Sea off the coast of Thailand and in the Sulu Sea between Borneo and the Philippines investigators have found internal waves at the boundary between layers of seawater that differ in density. Such waves give rise to surface turbulence that may be correlated with visual striations more than 100 kilometers long seen in satellite photographs. The waves carry enormous energy and may be responsible for strong underwater currents that have been detected aboard offshore oil-drilling platforms.

Another realm where soliton excitations have recently been proposed is in the long molecules of DNA. When a molecule of DNA is placed in a solvent, hydrogen atoms in the structure of the molecule are exchanged for hydrogen atoms in the solvent. The exchange may be promoted by a moving opening or break between the two intertwined helixes of the molecule. The break may diffuse along the length of the molecule as a wave that is localized and does not disperse.

The solitons described as topological make up a particularly intriguing class. They differ from ordinary solitons in that they can be detected merely by examining the global properties of a system. Consider a long, thin spring extended horizontally between two supports, with numerous rigid pendulums attached to it at regular intervals. In the lowest-energy state of the system the pendulums all hang straight down. If one end of the spring is then rotated 360 degrees about the horizontal axis, the rotation imparts a twist to the spring. On the twisted spring most of the pendulums still hang straight down, but in some region along the spring the pendulums must point to the side or even straight up because of the twist.

The length of the twisted region is determined by the balance of two effects. The resistance of the spring to twisting tends to minimize the degree of twist per unit length and therefore to lengthen the twisted region. For a spring without pendulums the energy would be minimized by spreading the twist over the entire length of the spring. Opposing the stiffness of the spring, however, is the force of gravity. Energy is required to raise each pendulum above the initial down position. The shorter the length of spring to which the twist is confined, the fewer pendulums must be raised. For some definite length the two forces are in equilibrium. The region of twist can be moved along the spring, but the twist retains its characteristic length and is not dissipated. The twist, which is a topological soliton, cannot be removed without cutting the spring or detaching it from its supports.

It now seems that topological solitons have been observed experimentally. They have been reported in polyacetylene, an organic semiconductor that consists of carbon atoms in a long chain, with a hydrogen atom attached to each carbon. In pure polyacetylene, as in other semiconductors, each electron is confined to a particular region of the chain. The electrons are said to occupy bound states, and every available bound state is filled. By the process called doping electrons can be added to the molecule; the added electrons move freely along the chain, greatly increasing the conductivity of the material.

The doping of a semiconductor can have other effects as well, associated with the spin, or intrinsic angular momentum, of the electron. The spin is quantized: it can have just two orientations, designated "up" and "down." Alan J. Heeger and his colleagues at the University of Pennsylvania set out to observe the effects of electron spin following the doping of polyacetylene. They found the expected change in the conductivity of the material, but they also found that when two electrons were added to the molecule, the total spin was unchanged. The spins of the added electrons always canceled, even though one would not expect pairs of electrons to always add to the molecule in a strictly complementary way.

Roman W. Jackiw of the Massachusetts Institute of Technology and J. Robert Schrieffer of the University of California at Santa Barbara have interpreted the result as compelling evidence for the existence of topological solitons. They point out that when electrons are introduced into polyacetylene, the energy of the molecular system tends toward a minimum. Since all the bound states are occupied, the added electrons create two new states in the molecule called the soliton bound state and the antisoliton bound state. In effect the soliton state introduces a twist in some region along the axis of the molecule, and the antisoliton state creates an opposite twist in another region.

Mathematically the soliton and antisoliton bound states cannot have spin, and so the added electrons must divide their spins in such a way that neither the soliton nor the antisoliton has a net spin. To ensure an even division one of the incoming electrons may have to change its spin from up to down or vice versa. Ordinarily such a spin-flip may entail an expenditure of energy, but since the soliton bound state minimizes the energy of the system, the electrons readily change spins as necessary in order to enter the state. The soliton and the antisoliton are therefore each characterized by a negative electric charge equal to the charge of an electron and by spin-up and spin-down values that are each half the value of the spin of a single electron. Hence the soliton model seems to account for the observed spin-swallowing property of polyacetylene.

Combinatorial Genetics

The most striking aspects of the immune response are its specificity and its diversity. An animal can make antibodies in response to virtually any antigen it encounters, that is, the animal can synthesize immunoglobulin molecules that recognize and bind to sites on any "nonself" molecule that enters its body. The antibodies are formed whether or


1954
The next revolution in the science of home entertainment is brought to you by the same company that brought you the last one.
1981 is the year of the RCA SelectaVision™ VideoDisc system—an innovation in electronic technology that promises to surpass the impact of an earlier RCA "first": the introduction of color television itself.
The RCA VideoDisc system has two simple components—a record-like disc and a player which reproduces sound and picture from the disc on any television set. But behind this simple, yet revolutionary, system lies a complex technology, the result of many years of RCA research in physics, chemistry, mechanics and electronics.
Created at the cost of millions of research dollars, the RCA system is a CED or Capacitance Electronic Disc system. It incorporates some of today's most sophisticated integrated circuits, including a microprocessor, and a CCD delay line filter. The disc itself is also a marvel of modern technology. Fifty of its signal tracks or grooves would fit on the edge of this page.
The RCA SelectaVision™ VideoDisc system is the product of RCA research—the kind of dedicated research that has been a tradition at RCA for over 60 years. Today, many years after Nipper started listening to "His Master's Voice," research and development are funded at the highest level in our company's history. Our commitment is, as it has always been, to the advancement of technology—to the creation of products that, like the VideoDisc, expand the human horizon.
© 1981 RCA Corporation
RCA
A tradition on the move

not the foreign substance has been encountered before and indeed even if it has only recently been synthesized by man. Moreover, the cellular machinery that makes the antibodies is not instructed by the antigen; the *B* lymphocytes, which are precursors of the antibody-secreting cells, already have among them all the genetic information necessary to make any immunoglobulin that will ever be needed.

How is the genetic capability for making antibodies that can react with innumerable antigens attained? For some time there were three competing schools of thought, each of which seemed to be supported by some evidence. The germline hypothesis held that the entire repertory of antibodies is encoded in a vast array of genes already present in the embryo. A second model proposed that a limited number of "minigenes" encoding the antigen-binding regions of antibody molecules are present in the embryo and are shuffled and recombined in the animal's lymphocytes. A third model relied on hypermutation, or unusually frequent mutations in the DNA encoding the antigen-binding sites.

In recent years powerful new techniques have been developed for breaking DNA into identifiable small pieces and for relating the pieces to the messenger RNA into which they are transcribed and the protein chains specified by the RNA. The application of these techniques in many laboratories during the past five years has revealed the genetic basis of antibody diversity. It turns out that all three of the proposed mechanisms play a role. A large number of genes encoding antigen-binding sites are present in the embryo, although the number does not approach a ratio of one gene for each antigen. The information is indeed reshuffled as *B* lymphocytes differentiate, both by the deletion and recombination of DNA segments and by the splicing of RNA transcripts in various ways. Mutation amplifies the combinatorial diversity.

An immunoglobulin molecule is a cross-linked assembly of two kinds of protein chains: light chains (of which there are two types, designated kappa and lambda) and heavy chains about twice as long. Each chain has a variable region and a constant one. The variable regions are different in each antibody; together they form the part of the immunoglobulin molecule that recognizes and binds to a particular antigen. Within the variable regions four framework regions vary only slightly from one antibody to another, and three hypervariable regions form the actual antigen-binding site. The constant region of the heavy chain establishes the class to which an antibody molecule belongs and thereby determines the function of the molecule in the immune response. The constant regions are about the same in all antibody molecules of a given class and are similar even in different animal species. The existence of variable and constant regions suggested as early as 1965 that separate genes might encode each region, so that one of many variable-region genes could combine with a given constant-region gene. That has turned out to be the case, but it is only the beginning of diversification.

Consider the kappa light chain, which accounts for almost all the light chains in the mouse (the animal in which most of the research on antibody diversification has been done) and for some 60 percent of the light chains in human immunoglobulins. Several hundred genes encoding the variable regions of the kappa light chain have been found in embryonic cells; they are designated V_κ genes, and the best guess is that there are about 300 of them. Roughly the same number of V_H genes encode the variable regions of the heavy chains. Combinations of the V_κ and V_H genes could generate some 90,000 specificities. That is a good beginning, but it is not enough.

The source of more diversity was revealed with the discovery that a *V* gene does not encode the entire variable region. In the case of the light chain the DNA for one end of the variable region is found in embryonic cells far from the *V* genes and near the gene for the constant region. This isolated fragment of variable-region DNA is called *J*, for joining, because it ultimately joins the *V* gene to a constant-region gene (*C*). Actually, there is an array of four functional *J* genes, any one of which can be combined with any *V* gene. The *J* gene is particularly important for diversification because it encodes part of a hypervariable region.

As lymphocytes differentiate, many cell lines are generated by somatic recombination: the deletion and translocation of segments of DNA. One V_κ gene, chosen at random, is thereby brought into association with one of the *J* genes; the V_κ and the *J* can be joined in several ways (at least three), providing for more diversity. The DNA of a given *B*-lymphocyte line, committed to producing a particular variable region, thus comes to include one V_κ gene linked to one *J* gene (which may be followed by other *J* genes); next comes a noncoding intervening sequence (an intron), and finally there is a *C* gene.

The genetic information is further modified when the protein-synthesizing stage begins. First the entire stretch of DNA is transcribed into RNA. Then this primary transcript is spliced: the intron and any extra *J* genes are excised. What remains after the extraneous material has been removed is a messenger RNA, comprising contiguous V_κ, *J* and *C* regions, that is translated into protein. In the kappa chain, then, one of some 300 V_κ genes combines with one of four *J* genes in one of at least three ways, for a total of approximately 3,600 combinatorial possibilities.

Still more diversity is provided by the heavy chain. Here the *V* gene is even shorter than it is in the light chain, and the third hypervariable region is encoded by genes of yet another kind, designated *D* (for diversity), of which there are perhaps five or 10; again there are four *J* genes. Somatic recombination and RNA splicing yield a messenger RNA that includes one each of some 300 *V* genes, at least five *D* genes and four *J* genes, for a total of some 6,000 distinct combinations. The association of any one of these heavy-chain coding sequences with any one of the 3,600 possible kappa light-chain genes can potentially yield more than 20 million variable-region specificities. This rough arithmetic must be modified because the variable region of the lambda light chain seems (at least in the mouse) to be encoded by only two genes, in contrast to the 300 or so genes for the kappa light chain. In the lambda chain diversity seems to be provided by mutation: something about the lambda variable-region genes makes them susceptible to a large number of point mutations, which vary the DNA sequence and so give rise to a multiplicity of functional genes.

In addition to accounting for variable-region diversity, DNA recombination and RNA splicing are thought to explain two phenomena associated with the constant region: the membrane-to-secretory switch and the class switch. There are five classes of immunoglobulins (Ig): IgM, IgD, IgG, IgA and IgE. Molecules of each class are encoded by a different set of heavy-chain constant-region genes (which are probably arranged on the DNA strand in the sequence given here). All these genes are present in the embryonic cell; they remain present in a differentiated cell committed to producing a particular variable region, and they are transcribed into RNA.

A mature lymphocyte first manufactures IgM and displays it on the cell membrane; then the cell may secrete IgM, perhaps along with IgD. These changes are apparently accomplished by RNA splicing. Either the IgM-encoding genes or the IgD genes can be preserved in the splicing process and therefore be expressed in messenger RNA. Whether the immunoglobulin molecule is secreted or remains bound to the membrane depends on whether or not the RNA retains a short segment specifying a protein chain with an affinity for the membrane.

The final commitment of the cell to the manufacture of one class of immunoglobulins is accomplished by a different process, which operates on the DNA rather than the RNA. When a given constant region gene has been selected, all constant-region genes that precede it in the order of transcription are deleted. For example, in a cell committed to making IgG the constant-region genes specifying IgM and IgD are excised. After the deletion the selected constant-region gene is hooked to the appropriate *J* gene; the resulting DNA is transcribed and the primary transcript is spliced to yield messenger RNA specifying a single immunoglobulin class.

What has been learned so far about the genetics of antibodies is enough to indicate the sources of their diversity, but the detailed mechanisms by which that diversity is attained are only beginning to come clear. The most recent evidence suggests that mutation may play a larger role than had been thought, particularly during the class switch after the first production of IgM. And painstaking analysis of DNA and RNA sequences is beginning to reveal short sequences that may serve as signals for deletion and recombination.

Bones of Contention

Among the fossil bones from Olduvai Gorge in Tanzania the remains of early hominids are not the only ones that merit scrutiny. Markings on fossil animal bones found there have now yielded unexpected information on the hominid occupants of the site. Some of the bones show signs of butchering with stone tools, which constitutes evidence of tool use by hominids almost two million years ago. What is more intriguing, a number of the bones marked in this way—in particular the limb bones of horses—would have been virtually meatless in life. Evidently the early butchers were collecting tendons and perhaps hide.

The evidence of butchering was reported by Pat Shipman of Johns Hopkins University, a specialist in the microscopic analysis of wear marks, and Richard Potts, a doctoral candidate at Harvard University engaged in the analysis of animal remains at several Olduvai sites. Their technique is to examine the bone with a scanning electron microscope and to compare the markings with those made by a known implement. They reported their recent findings at a New York conference on the research uses of anthropological collections and in *Nature.*

Shipman and Potts found a total of 85 marks on 75 fossil bones. One conclusion they reached early in their study was that the hominid butchers and the carnivores contemporaneous with them were scavengers who even scavenged from one another: both carnivore tooth marks and the marks of tools are found on the same bones, sometimes overlapping in a way that indicates which marks were made first.

Of the 85 marks 78 could be identified positively. Of the identified marks 54 were on bones that could be assigned to specific body parts of a prey animal. Carnivores were responsible for the marks on 35 bones; their teeth left grooves that varied from *U*-shaped to *V*-shaped in cross section but were invariably smooth at the bottom. Butchers' marks were found on 19 bones; the grooves were invariably *V*-shaped in cross section and rough at the bottom. Of the carnivores' marks 29 were on bones from meaty parts of the prey animal and six were on bones from meat-poor parts. The distribution of the butchers' marks was quite different: nine were on bones from meaty parts and 10 were on bones from meat-poor parts. Other identifiable marks on the fossil bones included evidence of burning, abrasive weathering, root etching by plants and digestion.

Shipman and Potts suggest that the Olduvai hominids sought out tendons and perhaps strips of hide in addition to meat. Such stringy animal products might have been useful for tying together and carrying more objects than could be carried by hand alone.

Reluctant Prophet

The concept of continental drift, an essential feature of the modern plate-tectonic theory of the earth's dynamics, is generally credited to the German meteorologist Alfred Wegener, whose views on the subject are set forth in their best-known form in *Die Entstehung der Kontinente und Ozeane* (*The Origin of Continents and Oceans*), first published in 1915. Wegener's "sailing continents" hypothesis, on which he elaborated in later editions of the book, was attacked in the following decades on various grounds. Perhaps the most telling criticism was aimed at the driving mechanism Wegener proposed for the displacement of the continents: tidal forces generated in the crust by the earth's rotation. It was not until the 1960's that new evidence, derived mainly from oceanographic and geomagnetic measurements, established that the continents do move, but that the cause is the spreading of the sea floor along a system of mid-ocean ridges where molten rock from the earth's interior wells up.

The concept of sea-floor spreading, it now turns out, was also anticipated by Wegener, although he apparently abandoned the idea before he arrived at the final formulation of his hypothesis. According to a recent reexamination of Wegener's first paper on continental drift, the inventor of the sailing-continents hypothesis described sea floor spreading in remarkably modern terms three years before the publication of his famous book.

In the early paper, published in a German journal of geography in 1912, Wegener considered the possibility of explaining differences in the depth of the ocean in terms of differences in the age of the sea floor. Freshly exposed rock would be expected to be warmer than older sections of the sea floor. The younger material would therefore be less dense and would be slightly elevated above the level of the older rock. "Such a temperature difference," Wegener noted, "would probably suffice to explain the relatively minor differences of level in the large ocean basins, although it is not sufficient to explain the weight differences between continental and oceanic material." He then added: "The depth variation appears also to suggest that the Mid-Atlantic Ridge should be regarded as the zone in which the floor of the Atlantic, as it keeps spreading, is continuously tearing open and making space for fresh, relatively fluid and hot [rock rising] from depth." Elsewhere in the same paper Wegener rejected the prevailing view of the time that the Mid-Atlantic Ridge was a relic of a land connection extending between South America and Africa.

No mention of these ideas is made in *Die Entstehung der Kontinente und Ozeane,* which is surprising because the book is essentially an expanded version of the 1912 paper. Wegener's early description of sea floor spreading was rediscovered in 1975 by Helmut Gebrande, a graduate student at the University of Munich. A critical reappraisal of Wegener's early paper, prepared by Wolfgang R. Jacoby of the University of Frankfort, appears in a recent issue of *Geology.*

Wegener never explained why he set aside some of his early ideas on sea floor spreading in favor of the sailing-continents model and the rotational driving forces. Perhaps Wegener stressed the forces related to the earth's rotation because he was a meteorologist, Jacoby writes, "but I think it was because the forces are real and can be estimated quantitatively and because their pattern (toward the Equator and westward) is similar to that of continental drift as he saw it." In any case, Jacoby concludes, Wegener "had a very clear perception of the distinction between facts and speculation.... Undeniable evidence existed only for continental drift, not for its mechanism or the driving forces. It made no difference then that some of his speculations have now been proved correct. The decisive recent evidence... was not yet available, but the ideas of sea-floor spreading and plate tectonics were there embryonically."

SINGAPORE 2000:
The engineering of rapid industrial advance

The tiny Asian island state of Singapore achieved full employment five years ago following rapid economic growth through industrialization. Premier Lee Kuan Yew is now attempting to achieve further growth through mechanization and automation, and to force the economy out of its low-wage, low-skill level into a higher-wage, higher-productivity level. This report examines the unique background and special circumstances of this industrial leap forward.

Singapore has probably the most successful 'planned' economy in the world. A major reason for its success is the willingness of government policy-makers to change plans as circumstances change; in particular, as the economy has become more developed, the country's leaders have permitted and, more recently, promoted an increasing degree of free-market operations. Indeed, with a per capita income of US $3,187 (in 1979 dollars) and an average growth rate of 7.3 per cent from 1970 to 1979, Singapore will have to depend all the more on the flexibility of a market system if its economy is to continue to grow at relatively high rates.

In the two decades since independence, the basic goals of Singaporean society have not changed at all. These have been and remain: to maintain its political independence, and to establish and maintain a high rate of economic growth as the best means of achieving increased prosperity and political independence. But the tactics for achieving these goals have changed considerably over the years.

In the early 1960s, the emphasis was on import substitution – the manufacturing of a wide range of products behind high tariff protection under the continuing guidance of paternalistic officials. But events moved too swiftly for this policy to work. Indonesia's so-called 'confrontation' policy, under the late President Sukarno, temporarily undermined Singapore's traditional role as an entrepôt.

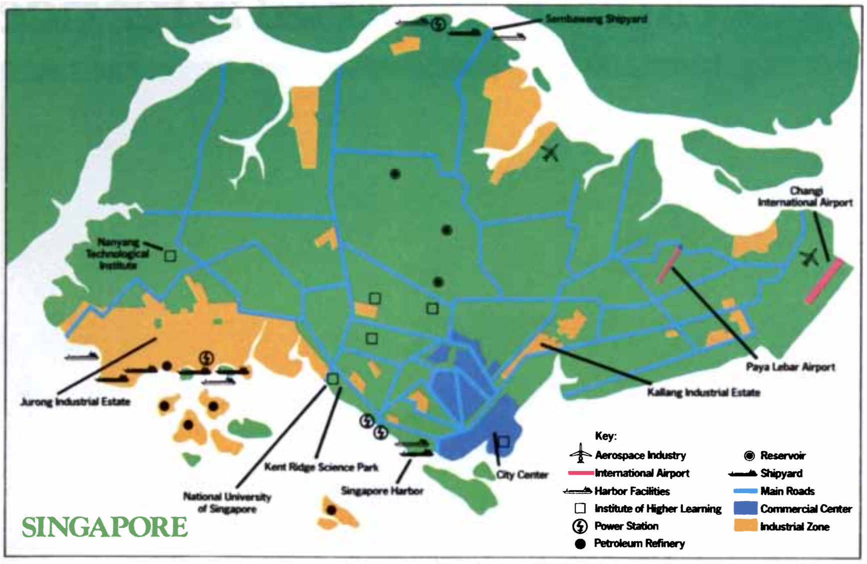
Its role as a finance and service capital for the Malay peninsula was undermined – again, temporarily – by ethnic rivalries between relatively well-off overseas Chinese and relatively poor Malays. After helping stave off the threat from Indonesia, the British forces still stationed in Singapore after independence began a permanent withdrawal, reinforcing a long term concern for the island nation's political survival and an immediate concern about how large numbers of base workers were to be employed.

As each of these challenges cropped up, political and bureaucratic leaders moved swiftly to adopt new policies to cope with these urgent problems, while also maintaining a high rate of growth in line with the long term goal of trying to ensure the political survival of a predominantly Chinese city-state in a predominantly non-Chinese region.

Import substitution was dropped in favor of export-oriented manufacturing. This meant that Singaporean products had to compete in world markets, which in turn required foreign capital and technology. This led to a change in previous attitudes about the relative merits of a market economy.

The Singapore Economic Development Board (EDB) was established in 1961 to spearhead the industrial drive by attracting foreign investors. Lee's People's Action Party's task was to convince foreign investors that an independent Singapore was politically and economically viable. New labor laws promised employers industrial peace. Tax incentives, relatively low wages and a good infrastructure were promoted, and a huge industrial complex was constructed.

Multinational corporations (MNCs) were seen as the best equipped institutions to provide modern technology and quality control – and they were encouraged to do so. A newly created tourist industry could employ thousands and raise incomes. It was set up, in spite of anti-colonial feelings against 'serving foreigners' – with air schedules, hotels, car rental agencies, shopping arcades, cultural events, courteous customs inspectors, and generally more vertical integration than can be found in most manufacturing industries. Generous investment incentives made it profitable to establish various export-oriented manufacturing industries, and these same kinds of incentives were also used to induce the world's leading



banks and financial institutions to establish a money-market center that utilizes many of the skills built up during the earlier entrepôt days.

As Prime Minister Lee explained in a speech to the International Chamber of Commerce in 1978, 'The question was how to make a living? How to survive? This was not a theoretical problem in the economics of development. It was a matter of life and death for two million people . . . The sole objective was survival.'

With this same goal in mind, the government initiated another policy switch two years ago, deliberately pushing wages higher than productivity growth in an effort to force low-wage, labor-intensive industries like textiles and electronics assembly to become more capital-intensive and to encourage the development of higher technology industries like computers and petro-chemicals.

The effort is not succeeding as well as had been hoped, though it may yet represent the boldest attempt anywhere to force through a restructuring of industries and employment in line with changes in comparative advantage. To date, because such wage increases cannot easily differentiate between manufacturing and service industries, they have had the unintended effect of putting additional, and possibly damaging, pressure on various industries that government planners certainly want to keep – notably tourism, financial services, and insurance.

Eventually, the mixture of 'planning' and market forces that emerges from this attempt at planned restructuring will offer an interesting example for the rest of the world. Perhaps the major lesson for now is to see once again that planning becomes more difficult as an economy becomes more developed.

INDUSTRIALIZATION FOR GROWTH

The route that Singapore took to achieve rapid economic growth through industrialization was by encouraging MNCs to build manufacturing plants in Singapore. In the 1960s this policy was aimed primarily at solving the unemployment problem. Given a relatively well-educated and skilled labor force, the quickest way to create jobs was to attract labor-intensive industries to Singapore. MNCs offered the advantage that they had the technical expertise to start up instantly and the international marketing network to ensure that whatever they produced in Singapore could be sold.

Growth in the 1970s

Throughout the 1970s, many MNCs saw new business opportunities and recognized that Singapore was a stable and profitable place in which to invest. Manufacturing industry grew rapidly. In 1965, the year when Singapore became an independent republic, there were around 1,000 manufacturing firms. Their combined output was only US $360 million. They provided employment for 47,000 workers, or 25 per cent of the total workforce. By 1980, the number of manufacturing establishments had grown to 3,347 with a total output of US $16,400 million. Industrial employment accounted for 287,314 or 30 per cent of the total workforce of 1,068,900. The success of the industrial development program in the 1970s not only solved the unemployment problem but also made the manufacturing sector the leading sector accounting for 23 per cent of the gross domestic product (GDP) of the nation.

Restructuring for the 1980s

Starting in 1979, a deliberate upward turn was taken in the direction of the continuing industrialization of Singapore. The aim is to restructure the economy, particularly the manufacturing sector, so that it will engage in higher technology production, and become more capital- and less labor-intensive.

In the earlier phase of industrialization, the aim was to create an industrial base which could absorb the excess labor that was available. Labor-intensive industries were therefore brought in to take advantage of the low-cost, high-quality labor that was available. However, with the rapid economic growth, full employment was reached by the mid 1970s. Labor became short in supply.

To achieve further growth, it was necessary to encourage firms to mechanize and automate their operations as much as possible, thereby releasing some of their labor into the job market for new industries to use. To break out of the low-wage, low-productivity syndrome, it was necessary for the National Wages Council (a government-employers-labor council which makes annual wage increase recommendations) to recommend substantial 'corrective' wage increases over the past three years. This influenced many firms to increase their investment in production machinery and automation to substitute for scarce and costly labor.

THE WAY FORWARD

The grand plan to restructure the economy via an increased wages policy is enshrined in the Ten Year Economic Development Plan, announced to the Singapore parliament in 1979 by Mr Goh Chok Tong, then Minister for Trade and Industry.

The Ten Year Plan takes into account both domestic and external factors. Domestic labor supply will grow more slowly in the 1980s (32,000 now, 24,000 yearly in 1981-85, 16,000 yearly in 1985-90). Rising expectations for better wages can be afforded only by the creation of many higher skilled jobs, and Singapore will have to continue to depend on foreign investment for skills, technology and markets.

Three external factors must be taken into consideration:

(1) the slower growth of the industrialized countries would prevent Singapore's higher growth unless Singapore upgraded and restructured its economy. ('We can attract the medium technology industries only if we step up training', said Mr Goh.)

(2) oil prices will continue to rise. ('Only if we upgrade can we afford to pay higher energy costs.')

(3) Singapore faces increasing competition from South Korea, Taiwan, Hong Kong. ('Unless we restructure our economy we will face graver pressure from other developing countries with abundant labor, especially China. On the other hand, as these developing countries expand and diversify they will import higher technology products which we may help to supply.')

Hence the strategy is first to break the vicious circle of low wages sustaining labor-intensive activities which lead in turn to poor productivity growth, an over-tight labor market and slower economic growth.

As Dr Goh Keng Swee, the First Deputy Prime Minister, put it, 'The purpose of the exercise is not to give more pay for the same work, neither more pay for less work. It is more pay for better work and more

skilled work.'

Reflecting Mr Goh Chok Tong's belief that the key to the long term success of economic restructuring is manpower development, the National University of Singapore (NUS) took in 40 per cent more engineering students in 1980, the two technical colleges increased their enrolment by 11 per cent, the EDB's Joint Industrial Training Centers expanded their classes by 32 per cent, and the Skills Development Fund gave employers grants to train over 10,000 workers.

'We must also provide tax incentives to encourage mechanization, automation and computerization, and research and development' said Mr Goh in his speech to parliament. He saw the Ten Year Plan's objectives, targets and strategies dependent upon five growth areas:

(a) manufacturing,
(b) trade,
(c) tourism,
(d) transport and communications, and
(e) the high-skill sector (e.g. computer, financial, medical and consultancy services).

As Mr Goh concluded, 'Our aim is modest – to step into the shoes left behind by countries like Germany and Japan as they restructure, they from skill-intensive to knowledge-intensive industries and we from labor-intensive to skill-intensive industries.'

TECHNOLOGY TRANSFER

Industrial R&D in Singapore has developed from humble beginnings to include a significant number of local companies and MNCs now participating in effective R&D in electronics, electrical, food, plastics, pharmaceutical and other areas.

More firms are beginning to engage in product engineering and design, development and testing, with the objective of improving their products, reducing costs and increasing productivity. Six examples here presented performed under different challenges, constraints, manpower resources and equipment to achieve the desired results.

Hewlett Packard: *the local R&D group, supporting 1,600 workers, designs oscilloscopes (transferred from Germany), and magnetic data cartridges for mini-computers (transferred from the USA).*

Philips: *has the biggest industrial R&D team – 150 strong – in Singapore. This team designs all the audio products manufactured locally, taking them from artist's sketch to final prototype.*

Beecham: *its process development and production support group has reduced prime costs of producing penicillin nucleus by 20 per cent, and has made unquantifiable improvements in product quality.*

Nestlé: *set up a US $12 million research center to adapt Nestlé food products to local tastes and nutritional requirements.*

Everbloom Mushrooms: *was established to exploit a unique high yield-fast harvest mushroom cultivation technology developed by the company's founder.*

Lamipak: *propelled into worldwide prominence when it successfully developed a superior version of high molecular, high density polyethylene film. Its 50 workers now export more than 90 per cent of its 24,000 tonnes output per year to Europe, USA and Japan for a variety of packaging purposes.*

SINGAPORE AS AN R&D CENTER

Research and development, especially where it is related to the development and design of products and to the improvement of productivity, quality or cost of a product, is being actively promoted. Singapore is firmly committed to the path leading towards higher technology for its industries: the whole economy is undergoing restructuring to make it more able to move up market and to compete internationally in higher technology products, and R&D is expected to play a significant role in this move.

Every means, including fiscal incentives and increasing R&D manpower supply, are being accelerated to ensure that it will be attractive and advantageous for firms to incorporate R&D into their other manufacturing operations.

Until recently, the need for MNCs and local manufacturers to carry out R&D was not high so long as their operations were relatively low in technology. But, with the push to move up market, companies will need to introduce new product lines and upgrade their manufacturing technologies to maintain their competitive edge.

Fiscal incentives for R&D

In the promotion of industrial development, the Singapore government had long adopted the basic philosophy and approach of providing a conducive environment in terms of favorable laws and regulations, tax incentives and loans, and infrastructure supports such as industrial estates and manpower training; the actual running of the enterprises and the economic outcome are left to be determined by private resources and free market forces.

The same policy will also prevail in the promotion of R&D: the government will provide the appropriate incentives and supports but does not intend to engage in extensive direct and active participation in R&D as is practised in nations such as the USA, USSR and Israel.

Science park

Another major impetus which R&D activities will receive over the next five years will be the development of a science park. The 136 hectare park will be opened to pure R&D organizations as well as choice high technology firms which will engage in industrial R&D in addition to manufacturing. The first phase will be ready for occupancy by 1983.

Central R&D support

While some of the best and most profitable industrial R&D work is and should be done by the firms themselves, there are many industrial problems which are most economically done on behalf of firms at central laboratories. The providing of technical support and testing services by the Singapore Institute of Standards and Industrial Research (SISIR) assists many firms, especially the smaller ones, to adapt and modify their capital equipment, improve their productivity, develop their quality control systems and to upgrade their products to standards acceptable by foreign markets.

THE FUTURE

When in 1979 Mr Goh Chok Tong announced the Ten Year Plan, he spoke of a calculated risk. In 1981 is there still a risk of failure? Mr Goh states: 'The risk has receded. Our economy grew by 10.2 per cent last year, despite adverse external economic conditions. Productivity growth doubled from 2.6 per cent to 5 per cent. Only a few companies closed down because of inability to adjust to the restructuring. Most are buying new machines and tightening their operational methods.' The Minister admitted that the risk of failure still existed, 'though remote', because the world economy could still go into a deep and sustained recession.

Effects of the Ten Year Plan

Mr Ng Kee Choe, senior vice-president management services, the Development Bank of Singapore, describes the impact of higher wages on the overall costs of banks as significant: wages form between 60 and 70 per cent of general and administrative charges. Banks in Singapore had already started computerizing, and higher wages meant that the need to accelerate computerization had become pre-eminent. 'Further progress', said Mr Ng, 'will depend on the availability of skilled computer personnel. This may take a few years.'

The decision to restructure the economy made very little impact on some companies. Philips Singapore, for example, has completed the transfer to Malaysia of the production of small personal radios previously manufactured in Singapore; a move which had already been planned.

'Introducing higher technology is a gradual process in our company. We believe quality can be improved with sophisticated mechanization and automation' said Mr W. G. Maeyer, managing director of Philips.

A spokesman for the chairman of the Port of Singapore Authority, Mr Lim Kim San, a former cabinet minister, said that the Authority's policy has for years been to emphasize mechanization and improve operational systems to increase labor productivity. Hence, while general cargo (excluding oil in bulk) handled across the wharves from 1970 to 1980 increased by 43 per cent to 15.2 million tonnes, PSA's overall staff strength grew at 6.5 per cent, from 10,019 to 10,669 employees.

Neville Watson, chief executive of Sembawang Shipyard, believes it is too soon yet to measure improvement in productivity: 'On a world basis, the Singapore shiprepair industry is strongly competitive. We can absorb these high wage costs.'

While Neville Watson believes the future for shiprepair in Singapore is good, he also holds the opinion that eventually '... it will become the wrong industry for Singapore.' It is a relatively low technology industry, needing a large workforce. 'In the next decades Singapore will have more sophisticated uses for the young men and women than simple shiprepair.'

Archie Gilchrist, managing director of Vosper, builders of warships and other vessels, classifies shipbuilding and ship-repairing as capital- *and* labor-intensive; 'Hence a company operating in this field must be affected by a policy which deliberately increases wages by an amount greater than the inflation rate'. Mr Gilchrist reckons that in his company the biggest improvements in productivity so far have come from better systems of management. 'Material scheduling, quality control and target setting, combined with an efficient system for monitoring progress, have all helped to improve performance', Mr Gilchrist explained.

Dr Richard T. T. Hu, chairman and chief executive of Shell, says that Shell has not been significantly affected by the decision to restructure the national economy: 'Our business has always been capital-intensive and of high technological content. Continuous upgrading of technology has also been taking place in our refinery.' Manpower numbers had always been kept in check by work rationalization, automation and mechanization, wherever possible.

On the higher wages plan, Dr Hu remarked that pushing up wage costs might have the desired result of weeding out overly labor-intensive industries; but it also affected the competitiveness of others already paying attractive wages. 'Preferential capital allowance for automation or technology-upgrading investments, on the other hand, would have had the desired effect without the unwanted drawbacks' he claimed.

The Singapore International Chamber of Commerce and Industry argues that there should be two sets of National Wages Council recommendations, one for the trading and services companies and one for the industries and manufacturers. Chamber members say that an across-the-board increase tends to defeat the objective of encouraging labor to move to companies involved in higher technology manufacturing. Besides, it is widely believed that only a relatively low percentage of Singapore companies meet the annual National Wages Council guidelines in full: the unionized MNCs are the main upholders of the guidelines. Press reports confirm that many Singapore companies do not follow them, relying upon annual bonus payments to retain staff.

The Chamber accepts the view that the needs of each company vary, and feels strongly that the service and trading companies should not be treated in exactly the same way as manufacturing and industrial enterprises. Even within this latter field, companies differ enormously, from the capital-intensive oil refining industry to labor-intensive manufacture of garments and cheap electronics.

Confidence confirmed

Many companies are ready to show their confidence in the future of Singapore in unequivocal terms: by making huge investments.

NEC Singapore, the subsidiary of Japan's largest semi-conductor company, Nippon Electric Company, plans to more than triple its investments. A new plant costing US $11 million for land and building alone will start production next year. Japan's largest producer of instant noodles, Nissin Foods, is building a factory costing US $14 million. Japan Medical Supply (Singapore) intends to invest US $18 million to manufacture a wide range of medical equipment and supplies, including artificial internal organs. It has been granted ten years' pioneer status (zero tax on profits) in order to develop the industry. Glaxo, the leading British pharmaceutical company, and one of the world's largest, is investing another US $100 million to make basic drugs primarily for export. The US $20 million factory will be ready before the end of the year.

The right position

Given its geographical location, Singapore is extremely well placed to be the industrial production, distribution and service center for the rapidly developing regions of Far East and ASEAN countries (ie those in south east Asia comprising Indonesia, Philippines, Malaysia, Thailand and Singapore itself). As these countries develop their primary and secondary industries, it will be natural for Singapore to become a source of technological expertise, hardware, spares and technological support.

In addition, it is expected that as the Peoples' Republic of China develops, its requirements for expertise, capital equipment, etc, will grow. Singapore, which has a population equally at home with the Chinese and English languages, would be well placed for business with the PRC.

Thus, Singapore's economic perspectives are well keyed-in with those of the region as well as those of the advanced, industrialized nations. By focusing on the high-growth, high technology areas which are compatible with Singapore's economic perspectives, the conditions for rapid R&D growth also seemed to be assured.

New industries sought

Minister Goh explained, when introducing the Ten Year Plan, that the prime objective is to develop Singapore into a modern industrial economy based on science, technology, skills and knowledge. 'If we succeed we shall be manufacturing high technology products like integrated circuits, computers, industrial electronic equipment, aircraft components, numerical-control machine tools, medical instruments like X-ray machines and blood counters, and speciality pharmaceutical products', he said. 'To realize our objective we require more engineers, accountants, lawyers, doctors, skilled workers and other managerial and professional personnel. We shall train them. We shall also induct talent from abroad'.

The changing face

Singapore's ruthlessness in wrecking old Chinatown has been criticized. Old Singapore is fast disappearing, leaving hardly a trace of the quaint, unhygienic old buildings of colonial days. Planners say they have no alternative: the buildings were almost falling down anyway. Much better to build afresh than to spend vast sums of money on trying to prop up the facades while rebuilding the interiors.

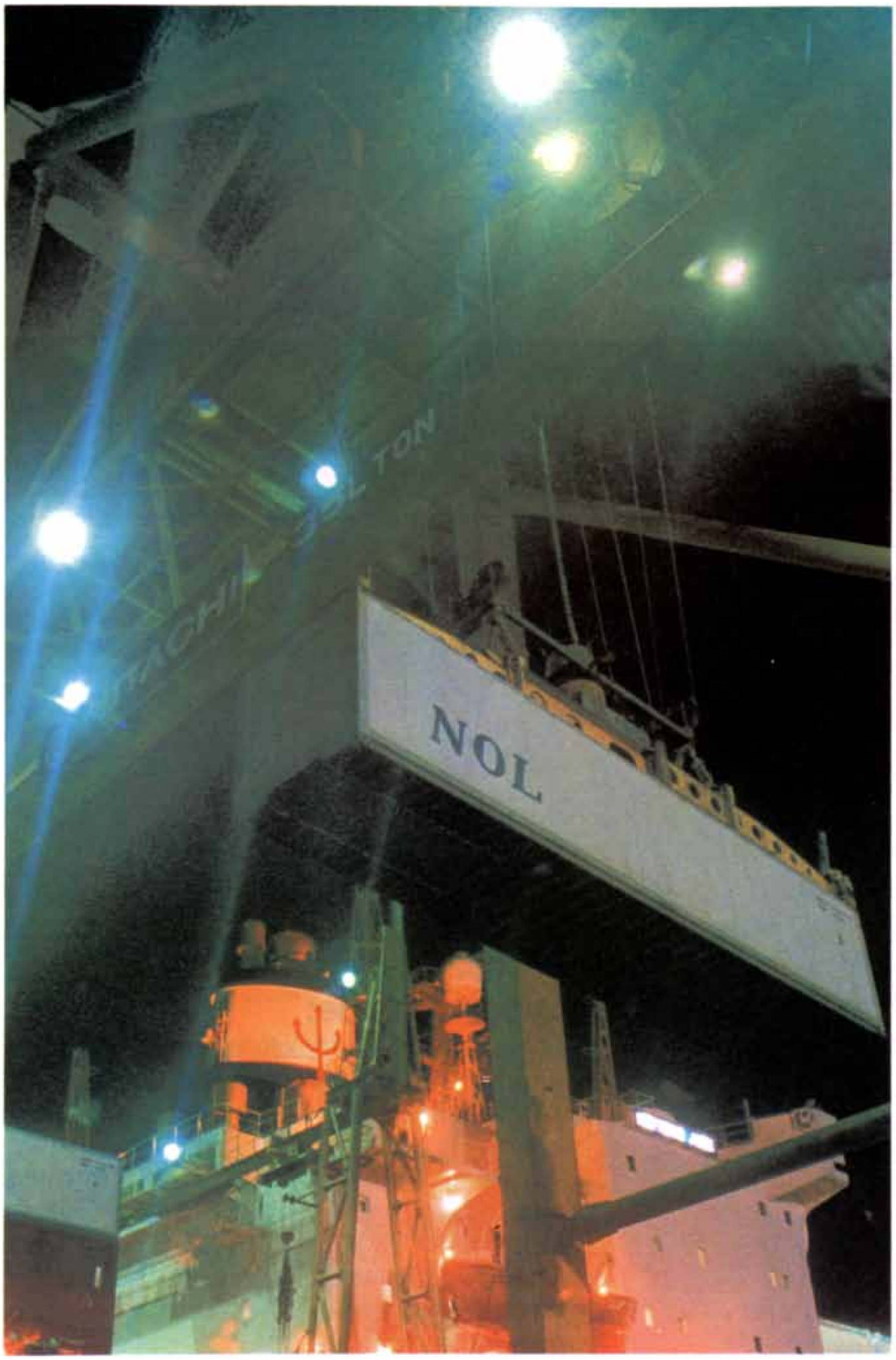
Instead, considerable sums are invested in highrise blocks of flats, offices, hotels, shops and factories. Spaces are reserved for green places, little parks and arbors. Streets and roads are lined with instant trees and carefully selected bushes and flowers. Singapore is the cleanest city in the east; it is fast becoming the greenest. Every week the face of Singapore is changed. A canal becomes a pedestrians' underpass. Narrow lanes are converted into wide motorways. Maps constantly have to be redrawn.

Singapore is on the move, into the age of higher technology, higher productivity and continuing growth: 'Our philosophy is simple', says Mr Yeo Seng Teck, director of the Economic Development Board, 'Since we have no natural resources nor a large domestic market, we have no choice but to depend on export-oriented manufacturing for our continuing growth, and, indeed, for our economic survival'.

Singapore expects difficult times in the near future if the world economic and political outlook does not improve. Even so, the government is confident of being able to attract foreign investments and to move its industry into higher technology, because of the very strong assets the nation has built up over the years – namely, political and economic stability, a sound infrastructure, competitive skills, and a capacity for hard-headed and practical adjustments to changing circumstances.

By the year 2000 all available land for industrial factories and workshops will have been allocated, the population will have reached the maximum of four million, and, if the current economic restructuring is successful, Singapore will have advanced industrially into an economy capable of providing a standard of living not far behind that of the developed nations.

We have you berthing Oakland. Chassis awaits discharge 2200 hrs precisely.



At NOL we know every move your cargo makes. At sea. Or on land. Making sure it is where it ought to be, when it ought to be.

It's this close, personalized attention that's earned us our global stripes... on the Trans-Pacific, Far East - Europe, Far East - Mediterranean, South Asia - Europe, and Straits - Australia services.

And as a committed carrier we only operate reliable and technologically advanced vessels. Including our feeder service within South East Asia.

So whenever your shipment is ready, NOL is ready. With the vessels. The equipment. The back-up. And the follow through.

We're with you right down the line.

NOL
NEPTUNE ORIENT LINES LTD

HEAD OFFICE: Neptune Orient Lines Ltd., Neptune Bldg., 13, Trafalgar St., Singapore 0207. Republic of Singapore. Tel:2223311 Cable:NEPORIENT Telex:Neptune RS24371 **REGIONAL OFFICES: USA** Regional Representative: Neptune Orient Lines Ltd., Suite 540, 300, Montgomery St., San Francisco, California 94104. Tel:(415) 433-4090 Telex:RCA 278-577 USNOL UR **UK/Continent** Regional Representative: Neptune Orient Lines Ltd., 201, Haverstock Hill, Belsize Park, London NW3 4QG. Tel: 01-794-5776 Cable: EUNOL LONDON NW3 Telex:8811656 EUNOL G **Japan** Regional Representative: Neptune Orient Lines Ltd., New Diamond Bldg., 4-4 (7th Flr.) Kasumigaseki, 1-Chome, Chiyoda-Ku, Tokyo. Tel:506-2539/2540 Cable:NEPTUNEREP TOKYO **Australia** Regional Representative: Neptune Orient Lines Ltd., Suite 31, 3rd Flr., "Kyle House", 27-31, MacQuarie Place, Sydney NSW 2000. Tel:2311975/2311644 Cable:MCARTHURS Telex: 20226/24045 **South East Asia** Regional Representative: Neptune Orient Lines Ltd., Suite 1522 (15th Flr.), International Plaza, Anson Road, Singapore 0207. Republic of Singapore Tel: 2225877 Cable: NOLSEA Telex: NEPTUNE RS 34822

GK&L/N4/USA/81

SCIENTIFIC AMERICAN

The annual single-topic issue of this magazine, published in September, will be devoted to Industrial Microbiology. The occasion for this issue is supplied by the advent of the technology of genetic engineering.

Articles written by scientists and engineers actively engaged in the work will present the full background necessary to common understanding of the new power to manipulate life processes that genetic engineering confers. These authors will review the origins of genetic engineering in the science of molecular biology and will place in perspective the industrial application and prospective economic impact of this technology.

The advent of genetic engineering was inevitable. Every industrial process is a scientific experiment scaled up to industrial dimensions and repeated over and over again or, when the engineers have perfected it, run continuously.

The cultivation of microorganisms in food processing, in the fermentation of wines and spirits, in the manufacture of pharmaceuticals and of such industrial reagents as the alcohols and acetic acid already contributes $88 billion to the gross national product. Genetic engineering will bring new sophistication to these ancient and well-established industries. It promises also to create new industries, as the technology of solid-state switching made possible the computer. In addition to the familiar bacteria, yeasts and molds of industrial microbiology, these new industries will bring novel organisms into cultivation, including human tissue cells. Beyond industrial microbiology, genetic engineering portends major changes in other branches of applied biology, such as agriculture, forestry and medicine, and in industries as remote from biology as the extraction of metals from their ores.

The engineering of life processes implies, as well, profound questions of ethics and values. Citizens of a self-governing society that would use such power must understand how it works and what it portends. For years to come there will be no more authoritative and comprehensive picture of this new technology than that presented between the covers of this issue.

INDUSTRIAL MICROBIOLOGY

September 1981

The Ribosome

Clues from electron micrographs yield a three-dimensional model of the intracellular machine that synthesizes proteins. Then the model is made to embody the steps by which a protein is produced

by James A. Lake

The function the tiny granule called the ribosome serves in the living cell is easy to summarize: it receives genetic instructions and translates them into protein. In the first aspect of this task it processes information. It binds a messenger RNA, a linear molecule whose sequence of the units called nucleotide bases is a code for amino acids. Then it captures the amino acids in the proper coded sequence. In the second aspect of its task it is in essence a catalyst: it joins the amino acids in a chain to make a protein.

In one sense the ribosome is small. Its largest dimension is 250 angstrom units, or about a millionth of an inch, which makes it about a tenth the size of a wavelength of visible light. In the light microscope ribosomes are therefore invisible except as patches in the cell that include large numbers of ribosomes and polysomes. The polysomes are clusters of ribosomes that are translating a single strand of messenger RNA.

In another sense the ribosome is large. One of the principal techniques for deducing the structure of a molecule is X-ray diffraction, in which the repeating three-dimensional arrangement of the atoms in a crystal of the substance gives rise to a diffraction pattern, or interference pattern, when a beam of X rays is directed through the crystal. The ribosome, however, has the weight of 100 to 150 average protein molecules. It is a formidable machine; most investigators consider it an intracellular organelle. For a structure as large (and as complex) as the ribosome, growing a crystal for X-ray diffraction has until recently been seen as impossible. Certainly the diffraction pattern yielded by the crystal would be extremely complicated.

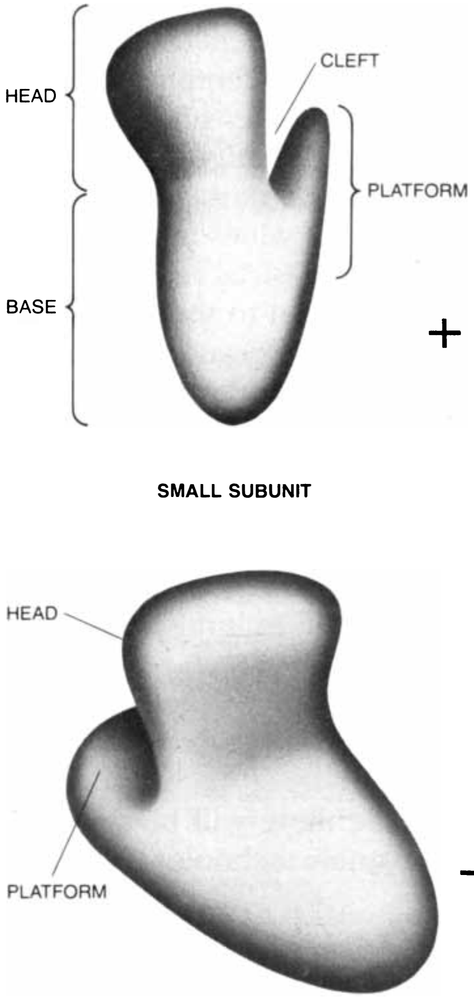
In a third sense the size of the ribosome is ideal. The electron microscope can now resolve two features of a structure if the features are roughly 15 angstroms apart. At such a resolution a protein molecule 30 angstroms in diameter appears in electron micrographs to be more or less spherical or elliptical. At the same resolution a ribosome shows up as a complex image.

The size of the ribosome means that an array consisting of several thousand items of data would be needed to specify its three-dimensional structure. Yet even a map giving the location of every atom in the ribosome would not in itself reveal how the ribosome makes protein. Techniques that I shall describe make possible the identification of functional sites on the ribosome. In this way the findings about the function of the ribosome can be mapped onto the structure. One hopes, of course, that the elucidation of the structure together with the finding of functional sites will suggest in detail the molecular mechanism by which the ribosome works. At the end of this article I shall describe the locations of several functional sites and propose the outlines of a possible mechanism. The first problem, however, was to find out what a ribosome looks like.

The Discovery of the Ribosome

Since the size of the ribosome makes it well suited for electron microscopy, the electron microscope has been an important source of information about the structure of the ribosome throughout the three decades that have passed since the ribosome was discovered. Indeed, the electron microscope made the discovery possible. In the early 1950's Albert Claude, Keith R. Porter and George Palade of the Rockefeller Institute for Medical Research observed under the electron microscope that spherical granules are plentiful in the cytoplasm of cells. A typical bacterium holds 10,000 such granules; a cell of a higher organism can hold many times more. Some of the granules float freely in the cytoplasm; others are attached to membranes inside the cell. Palade and Philip Siekevitz employed the ultracentrifuge to segregate the different constituents of cells according to their size. One of the fractions they obtained (the microsomal fraction) was rich in the spherical granules. Chemical assasy then revealed that the granules are rich in ribonucleic acid (RNA). This property led to the coining of the term ribosomes for the granules. Later it was shown that if amino acids labeled with radioactive atoms are added to the fraction, the amino acids (and their radioactivity) are incorporated into newly synthesized protein.

The quality of electron micrographs



THREE-DIMENSIONAL MODEL of the ribosome gives an asymmetric shape to the two subunits of which a ribosome consists.

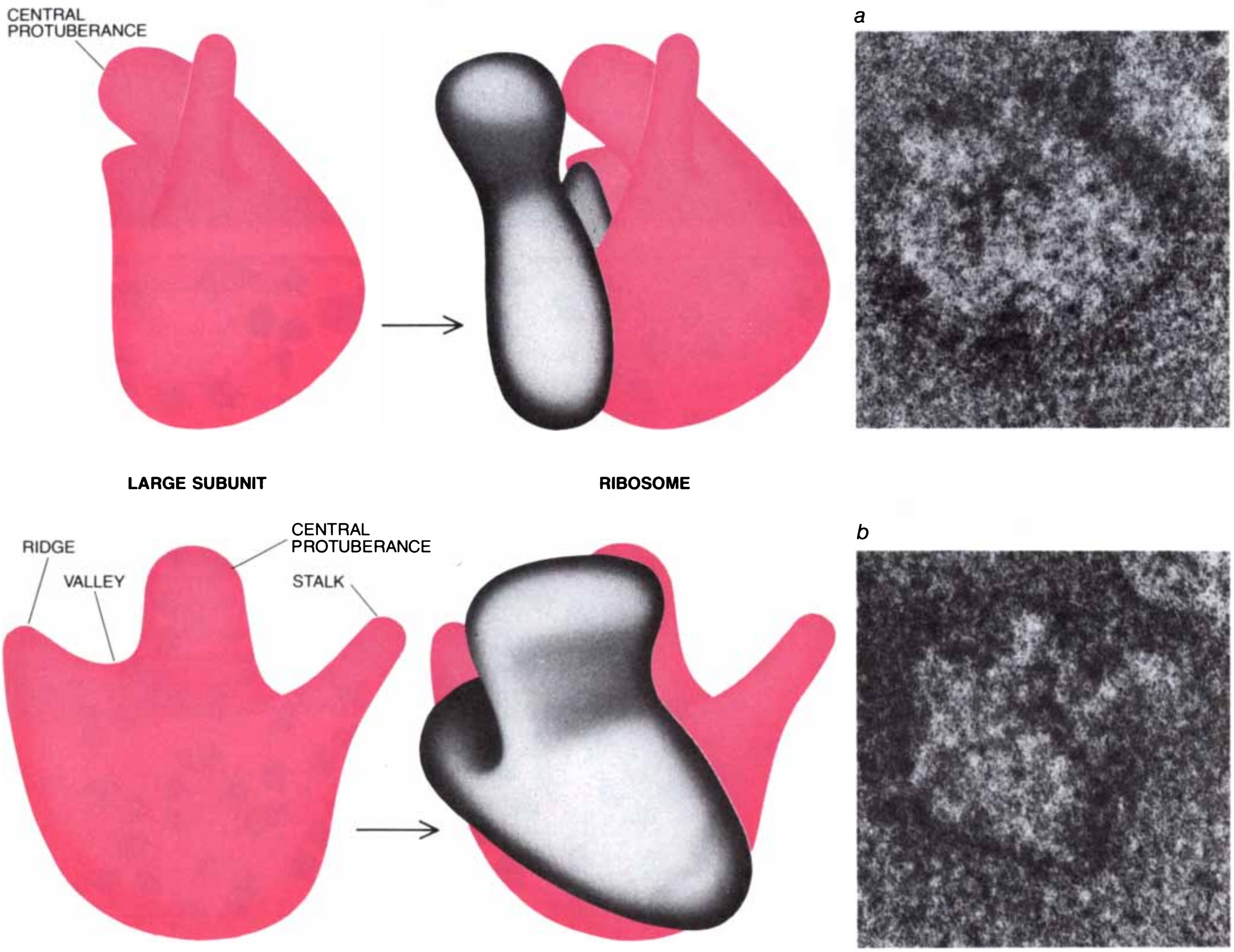
advanced through the 1950's as the techniques and the microscopes were improved. By 1959 Henry S. Slayter and Cecil E. Hall of the Massachusetts Institute of Technology and Geoffrey L. Zubay and H. E. Huxley of the Medical Research Council Laboratory of Molecular Biology in Cambridge had made micrographs showing that the ribosome consists of two subunits. One of them is about half the size of the other. At the time most investigators doubted that the electron microscope would yield much further information. Many thought the three-dimensional structure of the ribosome would be revealed only when ribosomes could be crystallized and subjected to X-ray diffraction. For a structure as large as the ribosome the project was taken to be essentially impossible.

With the path to X-ray diffraction apparently blocked David DeRosier and Aaron Klug of the Medical Research Council Laboratory of Molecular Biology reexamined the potential of electron microscopy. In 1968 they demonstrated that an electron micrograph is mathematically equivalent to an X-ray-diffraction pattern. Specifically, a diffraction pattern is related to the three-dimensional structure of a crystal by the mathematical operation known as the Fourier transform. In contrast, an electron micrograph is a map that compresses a three-dimensional structure into two dimensions. The Fourier transform of the data that make up the map, however, turns out (by what is called the projection theorem) to give the central plane of the diffraction pattern: the plane that passes through the center of the pattern and is perpendicular to the line of sight along which the electron micrograph was made. Given electron micrographs from several vantages, therefore, one can build up the data that X-ray diffraction would have provided. The Fourier transform of the built-up data then yields a map of the pattern of density throughout the three-dimensional structure.

DeRosier and Klug employed their technique to produce from electron micrographs a three-dimensional density map of part of a virus. With that impetus Slayter and I collaborated at the Harvard Medical School on an attempt to produce a density map of a ribosome. The result was quite fuzzy: the resolution of the map was only 100 angstroms. We were pleased to get an image, but we were also troubled: the map revealed nothing about how the ribosome worked. It was clear that a detailed view of the ribosome in three dimensions was needed.

Experimental Techniques

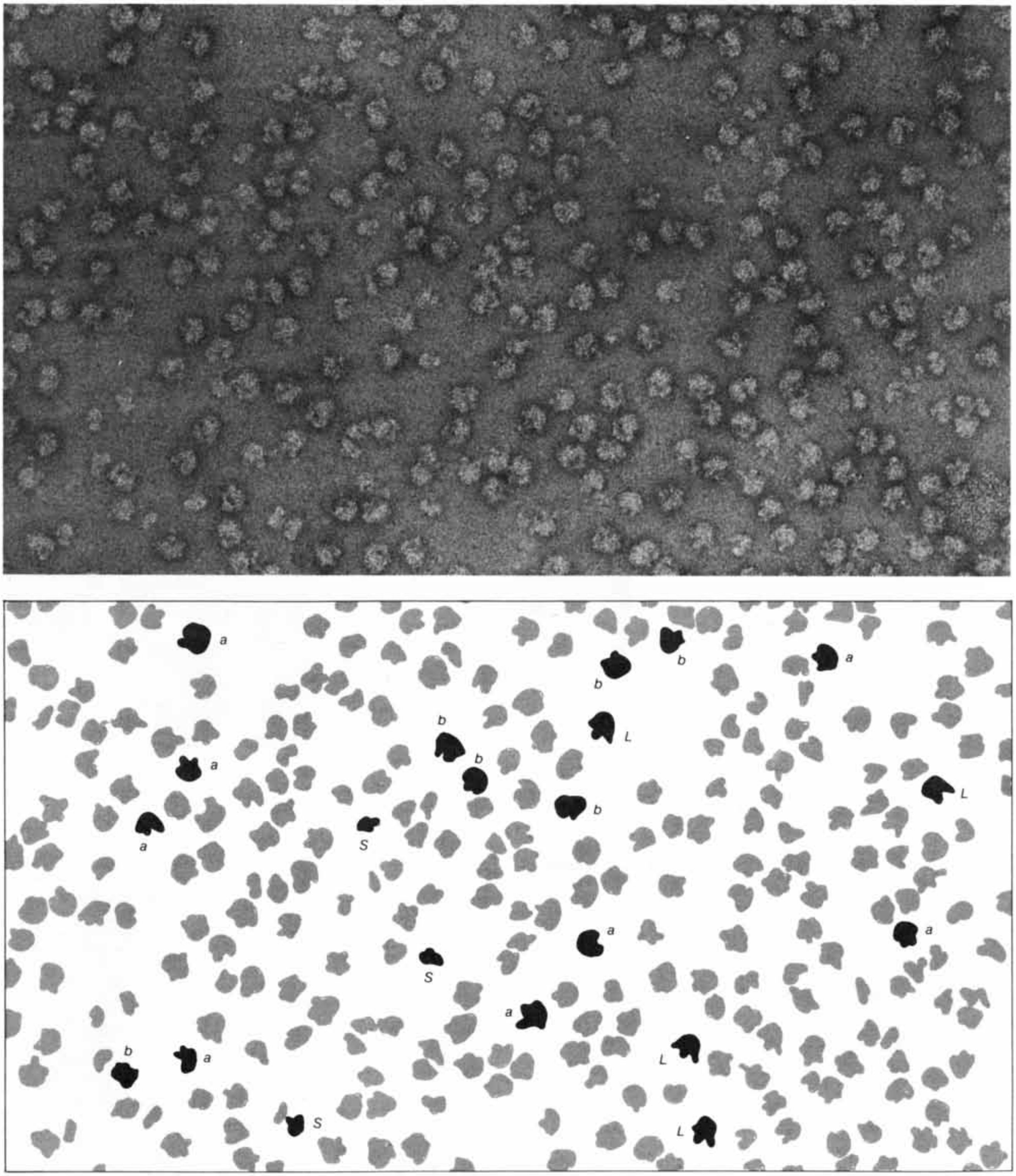
Biochemical studies of the ribosome had also been advancing. Most fundamentally the ultracentrifuge and other separative techniques had shown that the smaller of the two subunits of the ribosome is made up of a molecule of RNA whose molecular weight is about 600,000, together with one molecule each of 21 different proteins. The total



The small subunit (*left*) includes a head, a base and a platform. The large subunit (*second from left*) includes a central protuberance, flanked by a ridge on one side and a stalk on the other. Two orientations of the model are shown. They match the orientations of ribosomes from the bacterium *Escherichia coli* that appear in electron micrographs at the right. The length of a ribosome is about 250 angstrom units.

molecular weight of the small subunit is 900,000. The large subunit is made up of two RNA's and 34 proteins. Its total molecular weight is about 1.6 million, and again two-thirds of the weight is RNA.

In the late 1960's a technique was developed by which the RNA and the proteins that make up the small subunit of the ribosome could be isolated from one another and purified, so that when these various substances were mixed together, the small subunits reassembled themselves. The initial investigations had been done by Masayasu Nomura, Theophil Staehelin and Matthew Meselson, working at Harvard University. Finally Nomura, working with Shoji Mizushima at the University of Wisconsin at Madison, had conducted a series of experiments in which the ribosomal proteins were added in varying order to



FIELD OF RIBOSOMES from *E. coli* includes some whose orientation is that of the model of the ribosome on preceding two pages. The letters *a* and *b* identify them on the map. A few ribosomal subunits that also lie in the field are identified by the letters *S* and *L*, which stand for small and large. The field has been negatively stained for electron microscopy; each ribosome is defined by an outline of the salt of a heavy metal, which in this case is uranium. The electron micrograph, like the others in this article, was made by the author.

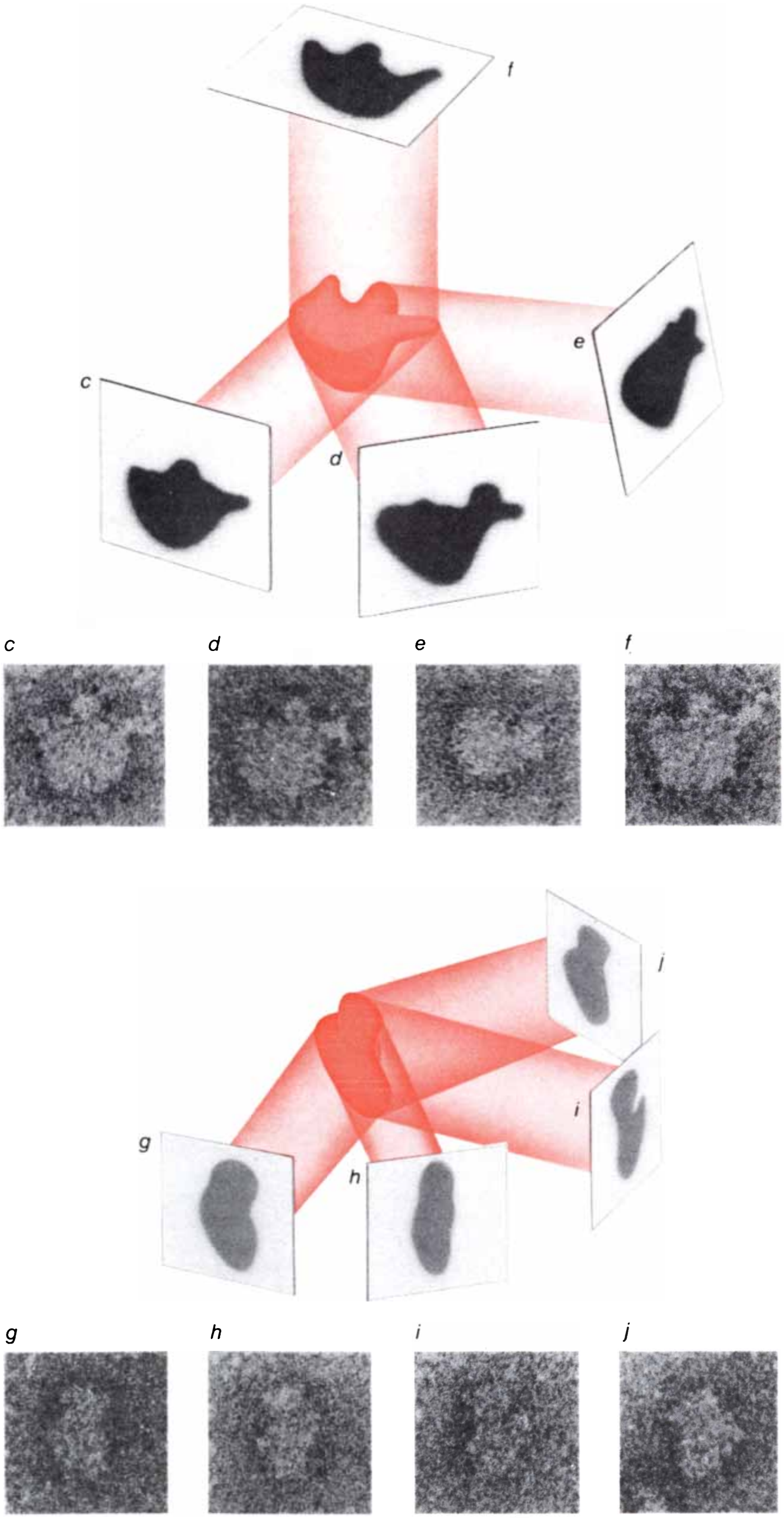
a solution containing the small-subunit RNA. Their intent was to determine which proteins must bind to the RNA in order for other proteins to do so.

The results were complex. No single protein turned out to be responsible by itself for the subsequent binding of another protein to the self-assembling small subunit. Instead the proteins designated *S*4, *S*8, *S*15, *S*17 and *S*20 bound themselves directly to the RNA. The binding of *S*20 was facilitated—that is, it proceeded with greater efficiency—if *S*4 and *S*8 were already bound. The binding of each subsequent protein to the self-assembling small subunit turned out to depend in varying degree on the presence of more than one of the earlier arrivals. The binding of protein *S*11, for example, was facilitated by no fewer than eight other proteins that had already become part of the assembly. What emerged from the work was an assembly map outlining the sequences by which the proteins self-assembled.

The assembly map provided a set of clues to which proteins are neighbors in the small subunit. Some further thought suggests that such clues may in some instances be misleading. It is conceivable, for example, that the binding of a protein to the small-subunit RNA can cause the shape of the RNA to change, so that a new loop in the RNA, perhaps 100 angstroms away from the site of the binding, becomes the binding site for a protein arriving later. Nevertheless, the publication of Nomura's assembly map renewed hope that the structure of the ribosome could be elucidated. Soon a number of laboratories devised new biochemical techniques toward that end.

One example is the cross-linking technique, in which ribosomes are mixed with bifunctional reagents: molecules having two sites that bind to other molecules. After such bonds have formed, the ribosomes are made to disassemble. The pairs of proteins that have been linked by the reagent can then be identified. The cross-linking technique is being employed in many laboratories to determine which ribosomal proteins are neighbors. Among the laboratories doing the research are those of Robert R. Traut at the University of California at Davis, Charles G. Kurland at the University of Uppsala, Donal Hayes at the Institute of Physical Chemistry in Paris and David Elson at the Weizmann Institute of Science in Israel.

A second technique, devised by Peter B. Moore and Donald M. Engelman of Yale University, requires the preparation of small-subunit proteins in which the hydrogen atoms have been replaced by atoms of deuterium (heavy hydrogen). Small subunits are then allowed to self-assemble from a mixture in which two of the proteins are the deuterated versions; the rest are normal. The small subunits are placed in a beam of neutrons from a nuclear reactor. The result-



MODEL WAS DEDUCED for the ribosome by finding distinctly different images of the large subunit (*top*) and the small subunit (*bottom*) in fields of subunits that had been negatively stained for electron microscopy. The letters *c, d, e* and *f* identify images of the large subunit; *g, h, i* and *j* identify images of the small subunit. Each image is a two-dimensional projection of a subunit that results from directing a beam of electrons through a three-dimensional structure and the uranium salt that surrounds it. Electron micrographs do not show a ribosome's surface; thus the three-dimensional shape of the ribosome is inferred from the multiple projections.

ing interference pattern is dominated by the diffraction of neutrons by deuterium atoms. Specifically, the pattern yields the distance between the two deuterated proteins. From a series of such experiments in which various pairs of proteins in the small subunit are the deuterated versions, a set of distances emerges. As the set becomes complete the position of the center of mass of each protein becomes well determined.

In the 1970's a third approach, immune electron microscopy, came into its own. It capitalizes on the preparation of antibodies that bind to specific ribosomal proteins and thereby reveal the places where the proteins are exposed at the surface of the ribosome. The basic program is to purify a given protein from the ribosomes of, say, a bacterium and inject the protein into the bloodstream of an animal, say a rabbit. The response of the rabbit's immune system to the ribosomal protein is the same as it would be to any protein from another species: the rabbit's antibodies are primed to bind to the foreign substance. The antibodies can be purified from the blood of the rabbit and mixed with whole ribosomes from the bacterium. The hope is that the antibody will bind to the parts of the protein exposed at the surface of the ribosome.

By 1972 Georg Stöffler and Heinz-Günter Wittmann of the Max Planck Institute for Molecular Genetics in Berlin, and also Lawrence Kahan, who was then a postdoctoral fellow in Nomura's laboratory at Wisconsin, had purified antibodies against many proteins from the ribosomes of the bacterium *Escherichia coli.* It was becoming increasingly likely that antibodies might serve as markers to tag specific proteins at the surface of the ribosome. Nomura, Kahan and I began to collaborate on further efforts; Kahan continued his association with us after he joined the faculty at Wisconsin, and we were joined by William A. Strycharz, another postdoctoral worker in Nomura's laboratory.

Our strategy was to induce in rabbits the formation of the antibodies called immunoglobulin G (IgG) and then allow them to react with ribosomal small subunits. Each antibody is a *Y*-shaped molecule that has two binding sites, one at the end of each of its two short prongs. Hence the antibody links small subunits into a pair. In the electron microscope the pairs can be seen. In general their appearance varies because they lie in different orientations. The location of the antibody and its binding site to the small subunit is a clue to deriving not only the position of the protein but also the three-dimensional structure of the ribosome.

Transmission Images

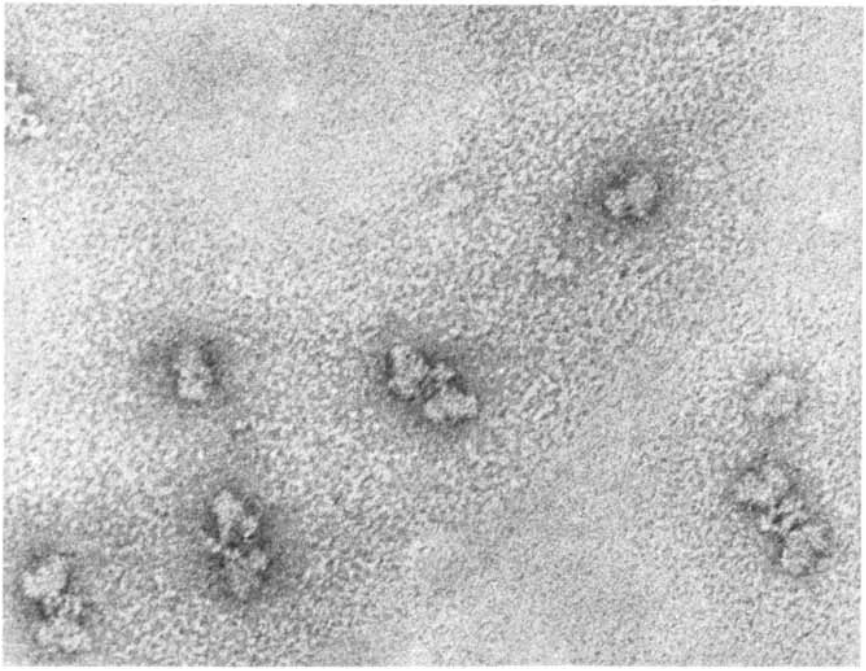
At this point it will be helpful to discuss the nature of electron microscopy in somewhat more detail. The technique we employ is called negative staining. A flat meshwork of copper wires is coated with a layer of carbon. The ribosomes (in solution) are deposited on this layer. Then a solution containing a salt of a heavy metal such as uranium is deposited. When the ribosomes dry, each one is surrounded by a shell of uranium atoms. The preparation is placed in the electron microscope. A beam of electrons passes more or less unimpeded through the ribosomes themselves. It is the uranium outlining each ribosome that deflects the electrons and shows up darkest in the image.

Fundamentally an electron micrograph (like a medical X ray) is a transmission image. It is the two-dimensional projection of a three-dimensional structure that results from directing a beam through the structure. In contrast, an ordinary photograph is a record of reflected light. In other words, an ordinary photograph shows surfaces; a transmission image does not. And in studying a transmission image one's visual experience is less help than one might wish.

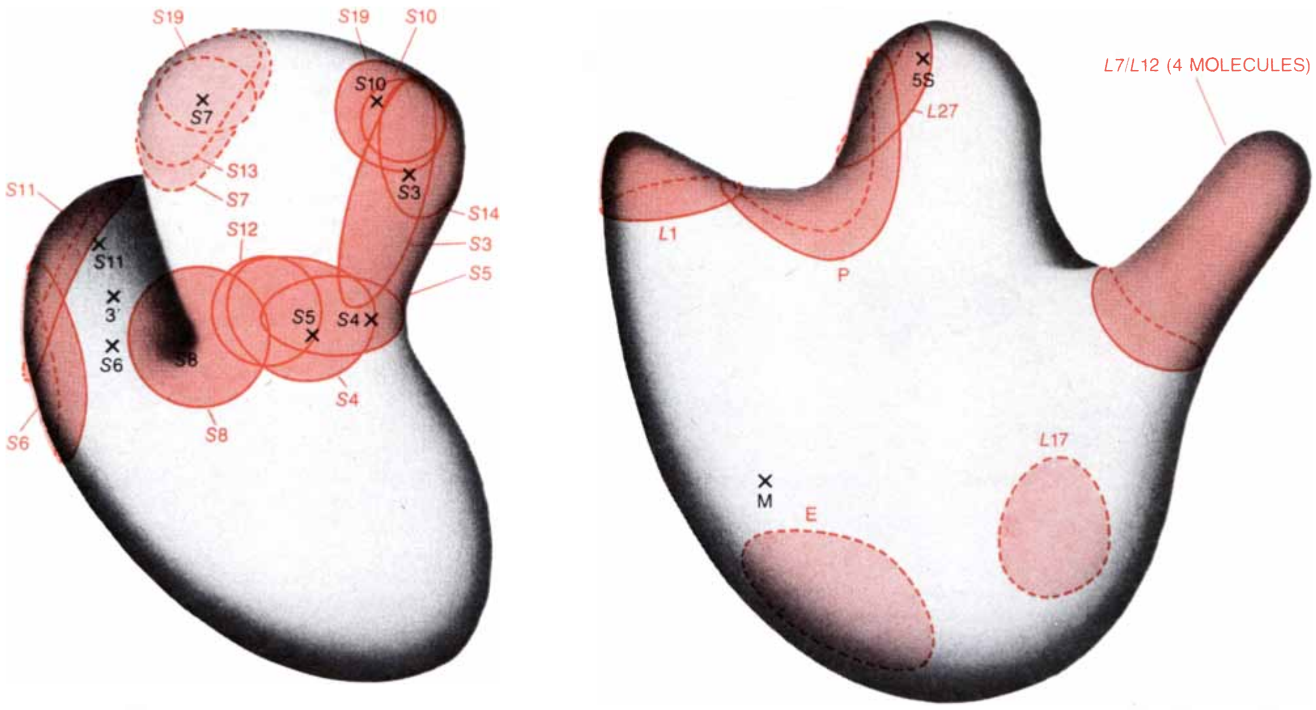
For example, in a transmission image there is no way to distinguish which part of a structure is closest and which part is farthest away. Imagine that you are presented with an X-ray plate of a hand. The fingers extend toward the top of the image and the thumb is at the right. Is the X ray a picture of a left hand or a right hand? The answer is you cannot tell. The X ray of a left hand with the palm down is indistinguishable from the X ray of a right hand with the palm up. This property of transmission images has an immediate implication for electron micrographs of a field of ribosomes oriented at random on a copper-wire grid. If a particular image of a ribosome appears, its enantiomorph, or mirror image, is likely to be somewhere in the field.

Plainly the projections of the ribosome are clues to its three-dimensional shape. In particular, one can hope to find several distinct projections of the ribosome in negatively stained preparations and deduce the shape that yields these projections. At Rockefeller University, Yoshiaki Nonomura, Günter Klaus-Joachim Blobel and David D. Sabatini had made such an effort in 1970. They identified specific profiles in electron micrographs of ribosomes from rat-liver cells and deduced a low-resolution model of a three-dimensional structure.

In making an effort to derive a high-resolution model a second problem arises. Specifically, one distinctive projection of each ribosomal subunit is more or less mirror-symmetric. Hence a three-dimensional model inferred from only two projections (the quasi-symmetric projection and a second distinctive view) will have bilateral symmetry. The fact that electron micrographs are transmission images dictates that any immune electron micrographs interpreted



IMMUNE ELECTRON MICROSCOPY reveals the site (or sites) at which a ribosomal protein comes to the surface of the ribosome. First the protein is purified from bacterial ribosomes; then it is injected into the bloodstream of an experimental animal. The immune system of the animal primes antibodies against the protein. The antibodies in turn are mixed with the ribosomes. Here the rabbit antibody immunoglobulin G has been primed against the small-subunit protein *S6*. Each antibody molecule is *Y*-shaped. It can bind to *S6* at the end of each short prong of the *Y*. Hence the antibody links the small subunits into pairs. The experiment was done by the author in collaboration with Lawrence Kahan of the University of Wisconsin at Madison.



MAP OF PROTEINS in the ribosome shows their location in the small subunit (*left*) and the large subunit (*right*). Two methods were employed. The colored patches were mapped by immune electron microscopy; they are the sites on the surface of the ribosome where antibodies bind to the protein against which they were primed. Each protein makes one such appearance at the surface of a subunit, except for protein *S*19, which evidently makes two appearances, and protein *S*4, which is under *S*5 and *S*12. The crosses were mapped by neutron diffraction, a technique in which ribosomal subunits are made to include a pair of proteins in which the hydrogen atoms are deuterium (heavy hydrogen). The interference pattern made by directing a beam of neutrons through the subunit then yields the distance between the two deuterated proteins. From a set of such distances the positions of the proteins emerge. In the illustration each cross marks a protein's center of mass, as determined by neutron diffraction. The immune electron microscopy was done by the author in collaboration with Kahan, Masayasu Nomura and William A. Strycharz at Wisconsin; the neutron-diffraction mapping was done by Peter B. Moore and Donald M. Engelman of Yale University. Five additional sites are shown. They are M, the site at which the ribosome can be anchored to intracellular membrane (determined by Nigel Unwin of Stanford University); 5S, the location of an RNA that forms part of the large subunit (determined by Vladimir Vasiliev and his collaborators at the Institute of Protein in the U.S.S.R.); 3′, an end of the ribosomal RNA denoted 16S rRNA (determined by Helen Olson and Dohn G. Glitz of the University of California at Los Angeles); P, the site at which the ribosome joins successive amino acids to make a protein chain (determined in the course of the author's work), and E, the site where the newly synthesized protein chain emerges from the ribosome (determined by Carmelo Bernabev and the author).

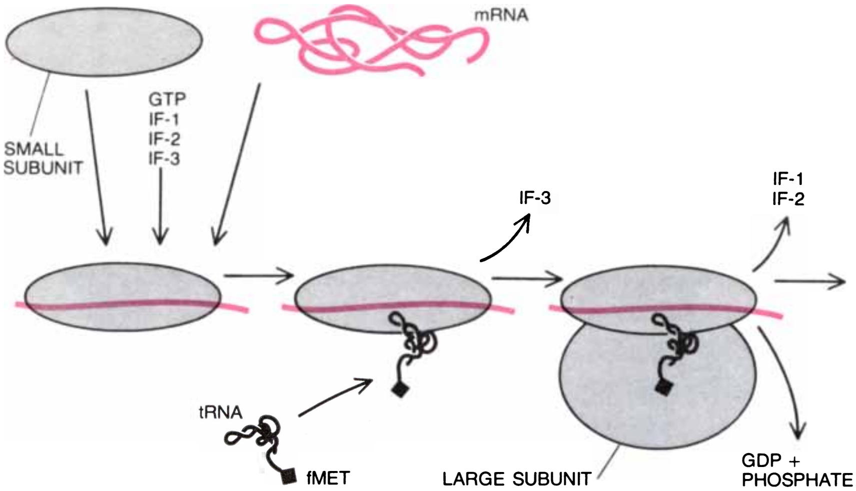
to place a protein on the left side of a symmetric model can be interpreted with equal justification to place it on the right. I myself was hopeful that I had found three distinct projections of the small subunit. One of them was the quasi-symmetric projection. In that case, however, the three-dimensional structure would be completely asymmetric. The binding sites of several small-subunit proteins in each of the projections (and in their enantiomorphs) were clues to the rotations that related the three projections. After some months I determined an asymmetric three-dimensional model.

Two other groups had also derived a model. By examining negatively stained electron micrographs of small subunits, Stöffler, working with Gilbert Tischendorf, a graduate student in his laboratory, had derived a symmetric model. Vladimir Vasiliev of the Institute of Protein in the U.S.S.R. had followed a different strategy. He had coated fields of small subunits with heavy metal by bombarding the fields with metal atoms from an angle, so that the metal deposited itself around each ribosome much as a snowdrift would. Although the electron micrographs can then be treated as reflection images, such "shadowing" yields a resolution inferior to that of negative staining. Indeed, the resolution is often too low to reveal the location of the antibodies employed in immune electron microscopy. Vasiliev had nonetheless derived an asymmetric model for the small subunit of the ribosome that much resembles ours.

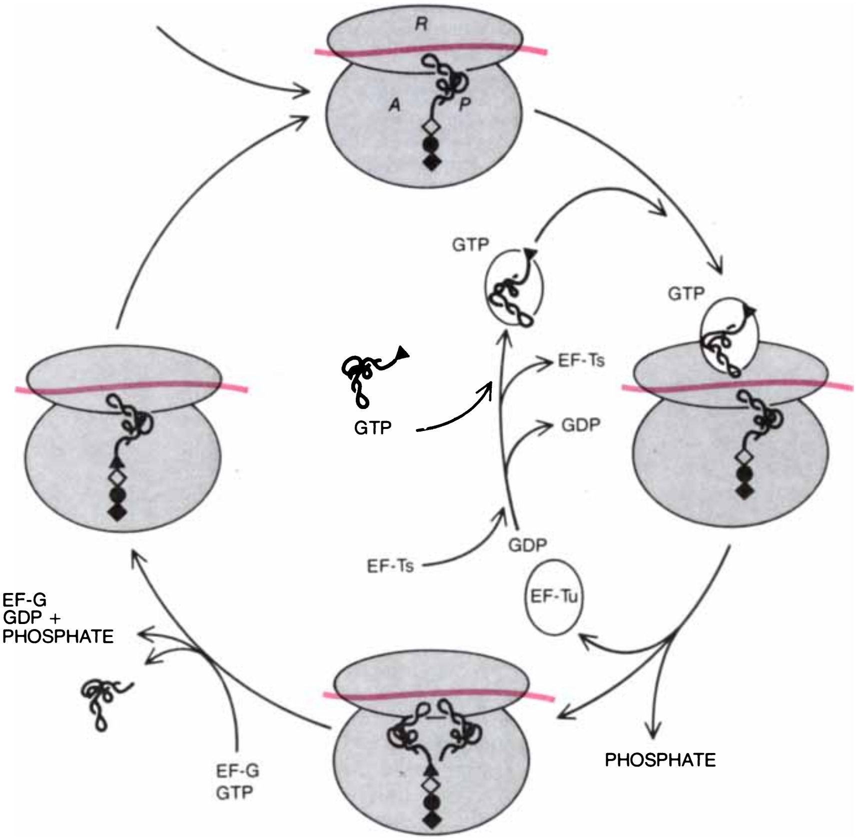
Once an asymmetric model is derived, immune electron micrographs showing two distinct views of the binding site for a given protein in the small subunit are sufficient to define unambiguously the position of the site on the small subunit's three-dimensional surface. Each view yields a line through the ribosome on which the site must lie. The intersection of the lines determines the site. Our mapping of proteins proceeded. Meanwhile by 1979 Moore and Engelman and their collaborators had mapped enough small-subunit proteins by neutron diffraction for them to be able to compare their results with ours. Their data fit the asymmetric model and not the symmetric one.

Fidelity Proteins

For ease in discussing the asymmetric structure of the ribosome we have given its features names. The model for the small subunit consists of three regions: a head, which makes up roughly a third of the small subunit; a base, which makes up the other two-thirds, and a platform, which is separated from the head by a gap called the cleft. A constriction in the small subunit demarcates the head from the base. The model for the large subunit (derived by similar methods) includes a central protuberance, flanked by a projection on each side. The more extended projection is called the stalk. The other projection is called the ridge. Between the ridge and the central protuberance lies the valley. When the large subunit and the small subunit are bound together, the stalk of the large subunit has its base near the constriction on the small subunit, and the head of the small subunit and the central protuberance



INITIATION, the first stage in the synthesis of a protein, requires that several molecules bind to the small subunit of the ribosome. Specifically, three initiation factors (IF's) bind, along with guanosine triphosphate (GTP) and messenger RNA (mRNA). The GTP is a source of energy; it will lose one of its phosphate groups and become guanosine diphosphate (GDP). The mRNA is an instruction tape; its sequence of the units called bases is a code (read three bases, or a codon, at a time) for the sequence of amino acids of which the protein will consist. Next (in bacterial initiation) the modified amino acid formylmethionine (fMet) is brought to the small subunit by transfer RNA (tRNA), a molecule specialized at one end to bind a specific amino acid and at the other end to bind to the ribosome. The result of initiation is a complex consisting of a large ribosomal subunit, a small ribosomal subunit, the mRNA and the fMet-tRNA.



ELONGATION CYCLE, the second stage in the synthesis of a protein, adds amino acids to the nascent (partially synthesized) protein chain. Each iteration of the cycle requires the participation of molecules called elongation factors (EF's) and also two molecules of GTP. At the beginning of the iteration a peptidyl-tRNA (a tRNA bearing the nascent chain) is bound to the ribosome at what is called the *P* site. An aminoacyl-tRNA (a tRNA bearing the amino acid specified by the next unread codon on the mRNA) binds to the *R* site. It binds in complex with EF-Tu and GTP. The binding is controlled by the binding of a codon on the mRNA to three bases (an "anticodon") on the tRNA. The tRNA is transferred to the *A* site, where its amino acid accepts the nascent chain. Finally, the *P*-site tRNA is ejected and the *A*-site tRNA takes its place. A cycle within a cycle prepares complexes of EF-Tu, GTP and an aminoacyl-tRNA.

of the large subunit are approximately aligned.

The grafting of the biochemical studies onto this structure is beginning to bear fruit. Consider protein *S*4. In the course of mapping small-subunit proteins by means of immune electron microscopy, Kahan, working with Donald A. Winkelmann, a graduate student at Wisconsin, discovered that the antibody they had prepared by sensitizing the immune system of the rabbit to protein *S*4 of *E. coli* bacteria could bind to *S*4 itself and could also bind to the complex made by binding *S*4 to small-subunit RNA. It could not bind to small subunits. Kahan and Winkelmann traced the reactivity of the antibody by mixing it with small subunits in partial states of assembly. Eventually they found that the antibody to protein *S*4 could bind to the small subunit only if one or both of two other proteins—*S*5 and *S*12—were missing from the assembly. It is as if *S*5 and *S*12 cover *S*4. Indeed, when we recently mapped the binding site for protein *S*4 on subunits lacking *S*5 and *S*12, we found the site for *S*4 lies where the site for *S*5 would lie if the subunit were complete.

Protein *S*4 is unusual in many ways. It binds strongly to the small-subunit RNA, it binds to the RNA in the absence of any other ribosomal protein, and once it is bound it protects a large part of the RNA from being broken down by the enzymes called nucleases. In addition protein *S*4 occupies a keystone position in Nomura's assembly map in that it facilitates the binding of four other proteins. Yet the facilitation is weak. Moreover, our protein mapping suggests no simple relation between *S*4's assembly interactions and its location. For example, *S*4 has no direct assembly interactions with *S*5 and *S*12 even though the three of them are neighbors on the side of the small subunit that faces away from the large subunit.

*S*4, *S*5 and *S*12 are linked functionally as well. In particular, it is known that the antibiotic streptomycin causes ribosomes to turn out proteins full of errors in their sequence of amino acid units. The ribosomes from certain mutant strains of bacteria resist this effect of the drug. If protein *S*12 is purified from the ribosomes of resistant bacteria and substituted for the *S*12 in the ribosomes of normal bacteria, the normal bacteria become resistant. Experiments with other drugs that cause "misreading" (and therefore mistakes) also implicate *S*4 and *S*5 as "fidelity proteins."

Biochemistry of Protein Synthesis

The location of proteins *S*4, *S*5 and *S*12 in a specific place on the small subunit, together with the biochemical findings about the function of the proteins, serves as an example of how regions on the ribosome are beginning to acquire significance. Before I go into such matters, however, there is more to be said about protein synthesis as a biochemical process and as a user of the cell's resources of energy.

The synthesis of a protein proceeds in three distinct stages. The first stage is initiation, in which a messenger RNA (mRNA) that was synthesized on a template of DNA is positioned on the ribosome in such a way that its message can be translated sequentially into a chain of amino acid units. The process begins when three protein molecules called initiation factors (IF's) bind to the small subunit of the ribosome. Two other molecules also bind. They are guanosine triphosphate (GTP) and the mRNA. The guanosine triphosphate is a source of energy. It has three high-energy bonds that link phosphate groups (PO_3) to the guanosine, and at the end of a process to which GTP contributes energy the GTP is found to have split into guanosine diphosphate (GDP) and an inorganic phosphate. The mRNA embodies genetic instructions. The instructions (the message) consist of a sequence of bases, which protrude from the RNA strand like a stack of tiles. A triplet of bases (a codon) specifies a particular amino acid.

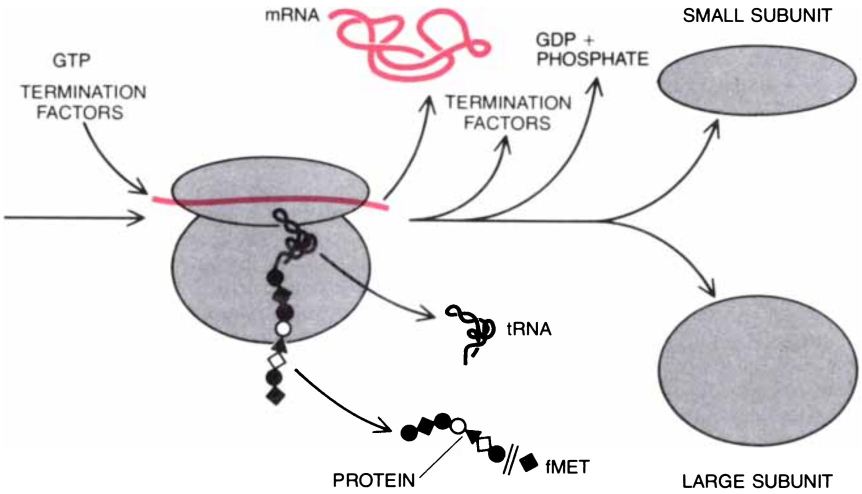
In bacteria the "start" codon (the triplet of bases at the beginning of the transcribed message) is adenine, uracil, guanine (AUG), or sometimes guanine, uracil, guanine (GUG). As a "start" codon either sequence specifies the amino acid methionine. It happens that the molecule placed by the ribosome at the beginning of a bacterial protein chain is a modified version of methionine known as formylmethionine (fMet). Initiation, which includes the translation of the "start" codon, therefore requires that fMet be bound to the small subunit. It is brought there by a transfer RNA (tRNA), an adaptor molecule specialized at one end to bind a particular amino acid and at the other end to bind to the ribosome. The binding of the tRNA forces one of the initiation factors, IF-3, off the ribosome. The remaining aggregate then binds to the large subunit. Immediately the GTP, now split into GDP and inorganic phosphate, falls off. So do IF-1 and IF-2.

The second stage in the synthesis of a protein is called elongation. It consists of a cycle that is repeated as each amino acid is added to the nascent (partially synthesized) protein chain, from the second to the last. (The second amino acid is of course linked to fMet.) Each iteration of the elongation cycle requires that two proteins called elongation factors temporarily bind to the ribosome. It also requires that two molecules of GTP bind to the ribosome and split into GDP and inorganic phosphate.

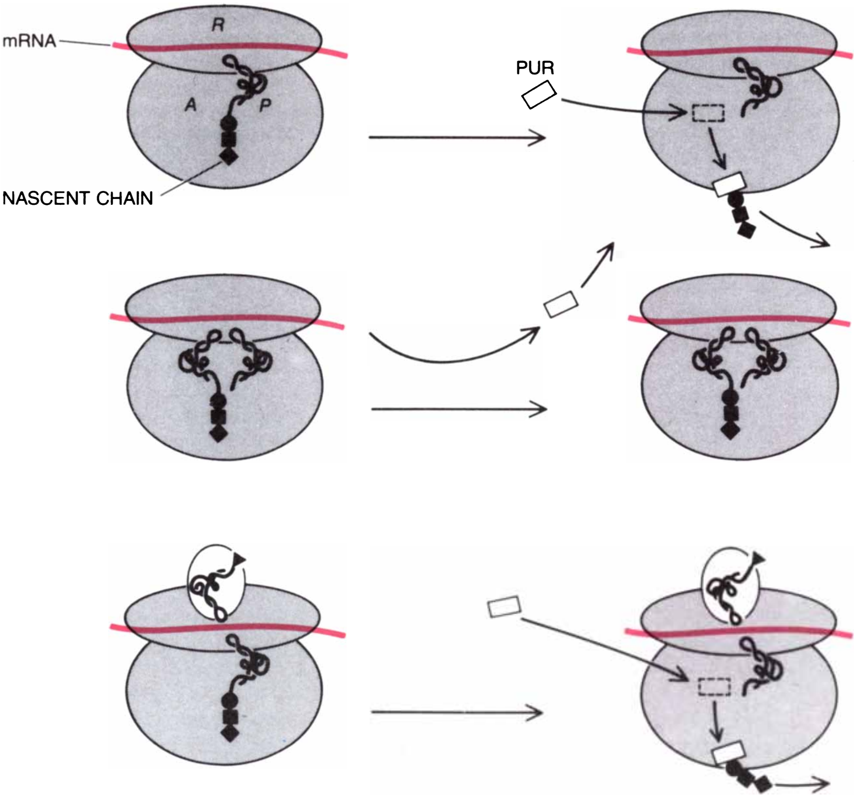
The final stage of protein synthesis is not yet as well understood as the stages that precede it. Two proteins called termination factors are involved, and it appears that UAG, UAA and UGA all serve as termination codons: triplets on the mRNA that cause the ribosome to release the messenger and the newly synthesized protein. Termination very likely requires that one molecule of GTP bind to the ribosome and split into GDP and phosphate.

The Elongation Cycle

Efforts to understand the details of the synthesis of protein have tended so



TERMINATION ends the synthesis of a protein and releases it from the ribosome. It also releases the mRNA. When the ribosome next begins to synthesize a protein, it must reassemble from subunits. Termination requires two termination factors and probably a molecule of GTP.



EFFECT OF PUROMYCIN on the ribosome tests the occupancy of the binding sites for tRNA's. At the beginning of an iteration of the elongation cycle (*top row*) a peptidyl-tRNA occupies the *P* site. The antibiotic puromycin (Pur) can occupy part of the *A* site, accept the nascent protein chain and then fall off the ribosome. At the middle of the iteration (*middle row*) the *A* site is occupied by a peptidyl-tRNA, and an arriving molecule of puromycin cannot bind. Under experimental conditions an aminoacyl-tRNA can be immobilized at the *R* site while the *P* site is occupied by a peptidyl-tRNA (*bottom row*). Puromycin then proves able to release the nascent chain. It follows that the *A* site is vacant when the *R* site is occupied.

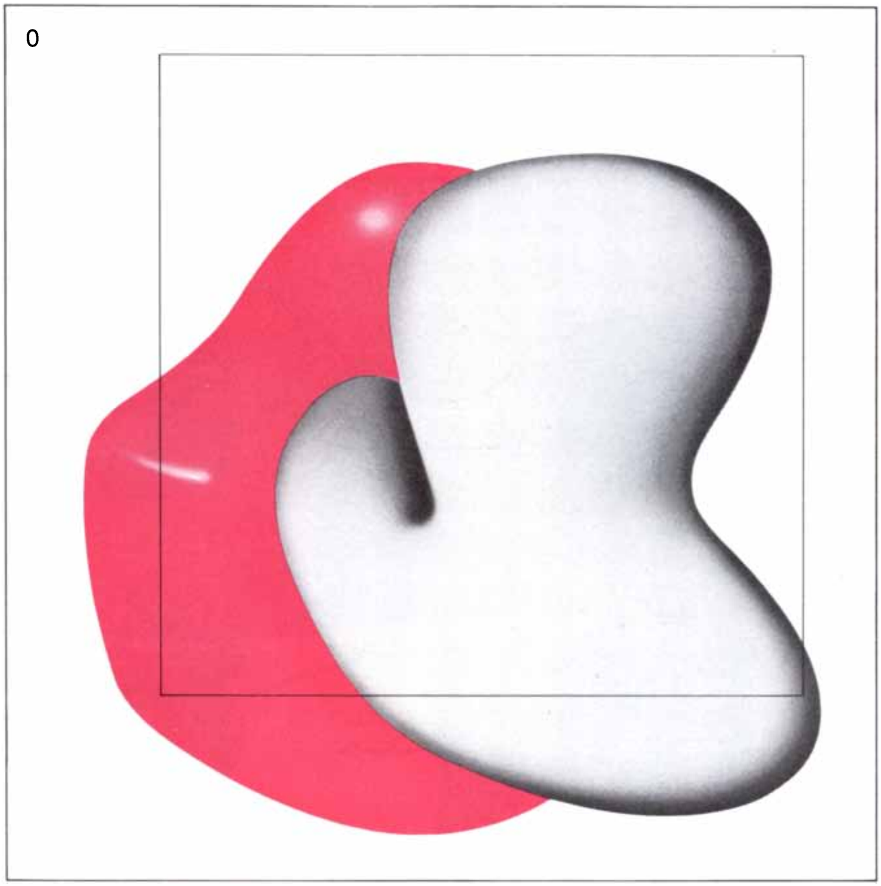
far to center on the middle stage: the elongation cycle. In this regard the antibiotic puromycin has been crucial. Puromycin is a molecule that is capable at certain times in the elongation cycle of disrupting the synthesis of protein. Its structure mimics the end of a transfer RNA that holds an amino acid. The mimicry is good enough so that when puromycin has usurped the binding site of the tRNA, the ribosome mistakenly joins to it the nascent protein chain. The binding of the puromycin to the ribosome then proves to be weaker than the binding of a tRNA would have been. The incomplete chain, now ending with puromycin, falls off the ribosome, and the elongation cycle stops.

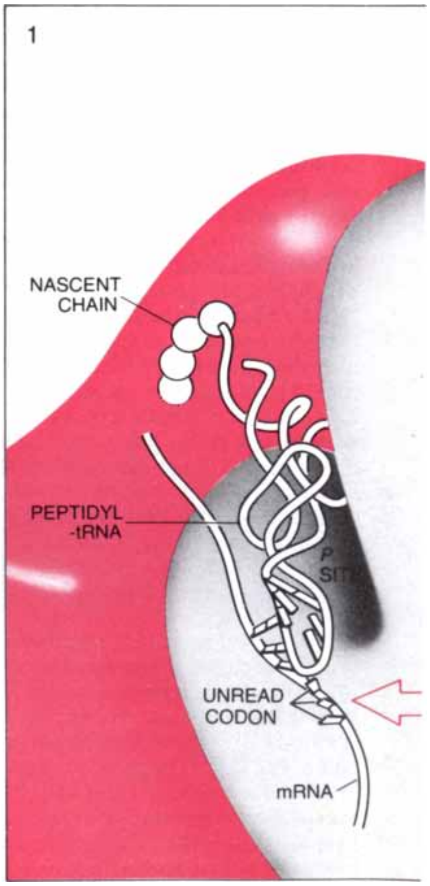
For 15 years the properties of puromycin have been taken to establish that tRNA's have two binding sites on the ribosome. Indeed, the model for the elongation cycle pictures a succession of tRNA's being shuttled between the two sites. At the beginning of each iteration of the cycle a peptidyl-tRNA (a tRNA bearing the nascent protein chain) is bound to what is called the *P* site. An aminoacyl-tRNA (a tRNA bearing an amino acid) becomes bound to what is called the *A* site. The amino acid is the one specified by the next untranslated codon on the message. The nascent chain is removed from the tRNA at the *P* site and joined to the amino acid on the newly arrived tRNA. The tRNA at the *P* site, now stripped of the nascent chain, is expelled from the ribosome; the tRNA at the *A* site takes its place, leaving the *A* site vacant for the next iteration of the cycle.

The effects of puromycin are in accord with this scheme. At the beginning of the cycle, when a tRNA bearing a nascent chain occupies the *P* site, the puromycin can occupy the *A* site and subsequently cause the nascent chain to fall off the ribosome. At the middle of the cycle, when a tRNA bearing the nascent chain occupies the *A* site, the puromycin cannot bind to the ribosome. It therefore cannot interrupt the synthesis of the protein. The introduction of puromycin constitutes a test of whether or not the *A* site is occupied.

In the early 1970's it began to appear that this model was incomplete. For one thing it was difficult to understand why each iteration of the elongation cycle should require the splitting of two molecules of GTP. The splitting of the second molecule is easy to explain. It can provide the energy that moves the peptidyl-tRNA from the *A* site to the *P* site and advances the mRNA so that the next codon can be read. The splitting of the first molecule is harder to justify, particularly because the energy released by breaking the bond of the nascent chain to the peptidyl-tRNA is more than enough to bind the nascent chain to the amino acid on the aminoacyl-tRNA.

Moreover, it can now be seen that the model had been decisively compromised by an experiment that Jean M. Lucas-Lenard of Rockefeller University had done in 1969. A few years later the result of the experiment was confirmed by Yoshito Kaziro and his colleagues at the University of Tokyo. The experiment is based on the fact that when an aminoacyl-tRNA binds to the ribosome, it binds in combination with two other molecules: the elongation factor called EF-Tu and also a molecule of GTP. In the test tube one can substitute for the GTP a substance called GMPPCP. It is identical with GTP except that a carbon atom takes the place of an oxygen atom. The tRNA can bind to the ribosome in





HYPOTHETICAL MOLECULAR MECHANISM for the synthesis of a protein by a ribosome is shown in a sequence of drawings that begins on these two pages and continues on the next two pages. Drawing *0* (*left*) depicts the entire ribosome. The black square indicates the part of the ribosome that appears in each of the following drawings. Drawing *1* depicts the events that begin an iteration of the elongation cycle. A peptidyl-tRNA is bound to the small subunit's *P* site; it bears a nascent protein chain that is shown as entering the large sub-

combination with EF-Tu and GMPPCP. The molecule of GMPPCP, however, cannot be split. The phosphate bonds therefore do not give up their energy, and the tRNA cannot move from the place where it binds.

Lucas-Lenard, then, had produced the complex consisting of an aminoacyl-tRNA, EF-Tu and GMPPCP, had allowed it to bind to a ribosome in the middle of the elongation cycle and had shown that puromycin could release the nascent chain. At such a time the *P* site was occupied by the peptidyl-tRNA and the *A* site was vacant. (This was established because the *A* site had accepted the puromycin.) Yet the aminoacyl-tRNA had found a place to bind. In the spring of 1977 I published a proposal for the existence of a third binding site for tRNA's. I call it the *R* site. Six months later Arthur Johnson, who was then at the Columbia University College of Physicians and Surgeons, independently published a similar proposal.

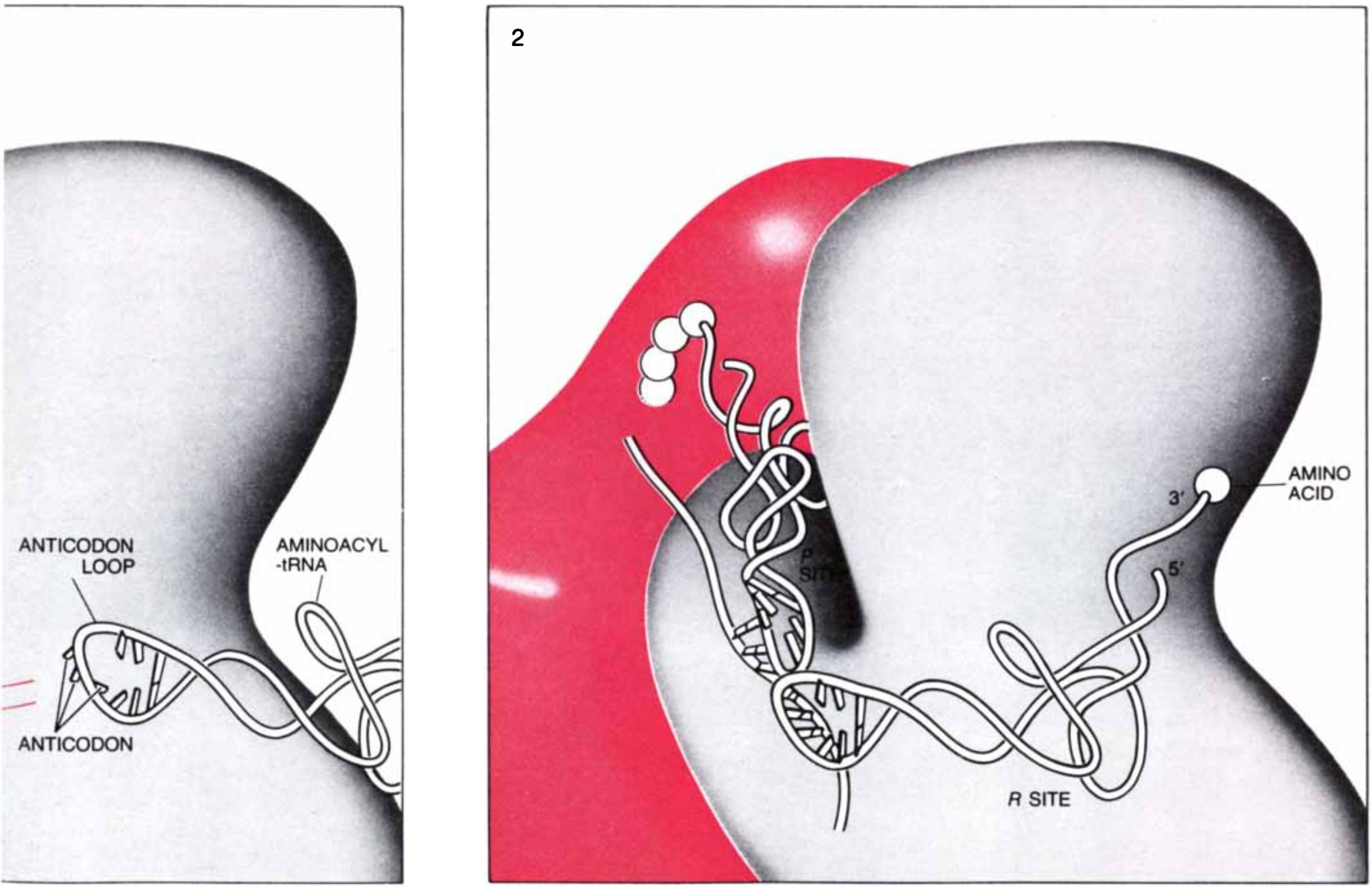
Transfer RNA

To understand the function the *R* site might serve one must know the basic structure of a transfer RNA, as revealed by X-ray diffraction. Like other molecules of biological significance, transfer RNA can be characterized by describing its primary, secondary and tertiary structure. The primary structure is simply the linear composition of the molecule: a transfer RNA is a string of about 80 of the tilelike nucleotide bases, which protrude from a backbone of sugar and phosphate. Three of the bases (guanine, cytosine and adenine) are also found in DNA. A fourth type of base is not. It is uridine, which takes the place of the thymidine one finds in DNA. In addition more than 50 modified versions of the usual bases in tRNA make an occasional appearance.

The secondary structure describes the way certain sequences of bases bind (or form base pairs) with sequences elsewhere along the strand. Normally guanine pairs with cytosine; the two are called complementary. Similarly, adenine pairs with its complement, uridine. In RNA some of the base pairing fails to fit this pattern. The important point, however, is that the tRNA becomes double-stranded in the places where the single strand binds to itself. The single-stranded regions in between become loops.

The tertiary structure describes the way the tRNA molecule with its loops and its double strands is arrayed in three dimensions. Overall a tRNA is *L*-shaped. At the end of the short limb of the *L* are the two free ends of the strand, which come to lie near each other. One end is designated 5′. By convention it is the end at which the numbering of the primary sequence of bases begins. The other end, designated 3′, is the place at which the tRNA binds a specific amino acid. At the end of the long limb of the *L* (and thus the other end of the tertiary structure of the molecule) is a loop consisting of seven bases. The central three of them (typically Nos. 34, 35 and 36) make up the anticodon: a triplet of bases forming the complement of the triplet that codes for the amino acid the tRNA is specialized to carry.

In the two-site model of the elongation cycle a succession of aminoacyl-tRNA's arrive by chance at the *A* site. When the correct one arrives at the site, the three bases that make up its anticodon bind readily to the three bases on the messenger that make up the next un-



unit. The anticodon of the peptidyl-tRNA is bound to a codon of the mRNA. The codon was translated in the preceding iteration of the cycle. An aminoacyl-tRNA draws near the small subunit. GTP and the elongation factor EF-Tu, which accompany the tRNA, are not depicted. In drawing *2* the arriving tRNA has bound itself to the *R* site; its anticodon becomes bound to the next unread codon. The anticodon is part of a loop of seven bases. Five of them are stacked. They are the five that lie nearest the end of the tRNA designated the 5′ end.

read codon. The binding of the correct tRNA to the ribosome is therefore securer than the binding of an incorrect tRNA. This constitutes the "recognition" of the correct tRNA.

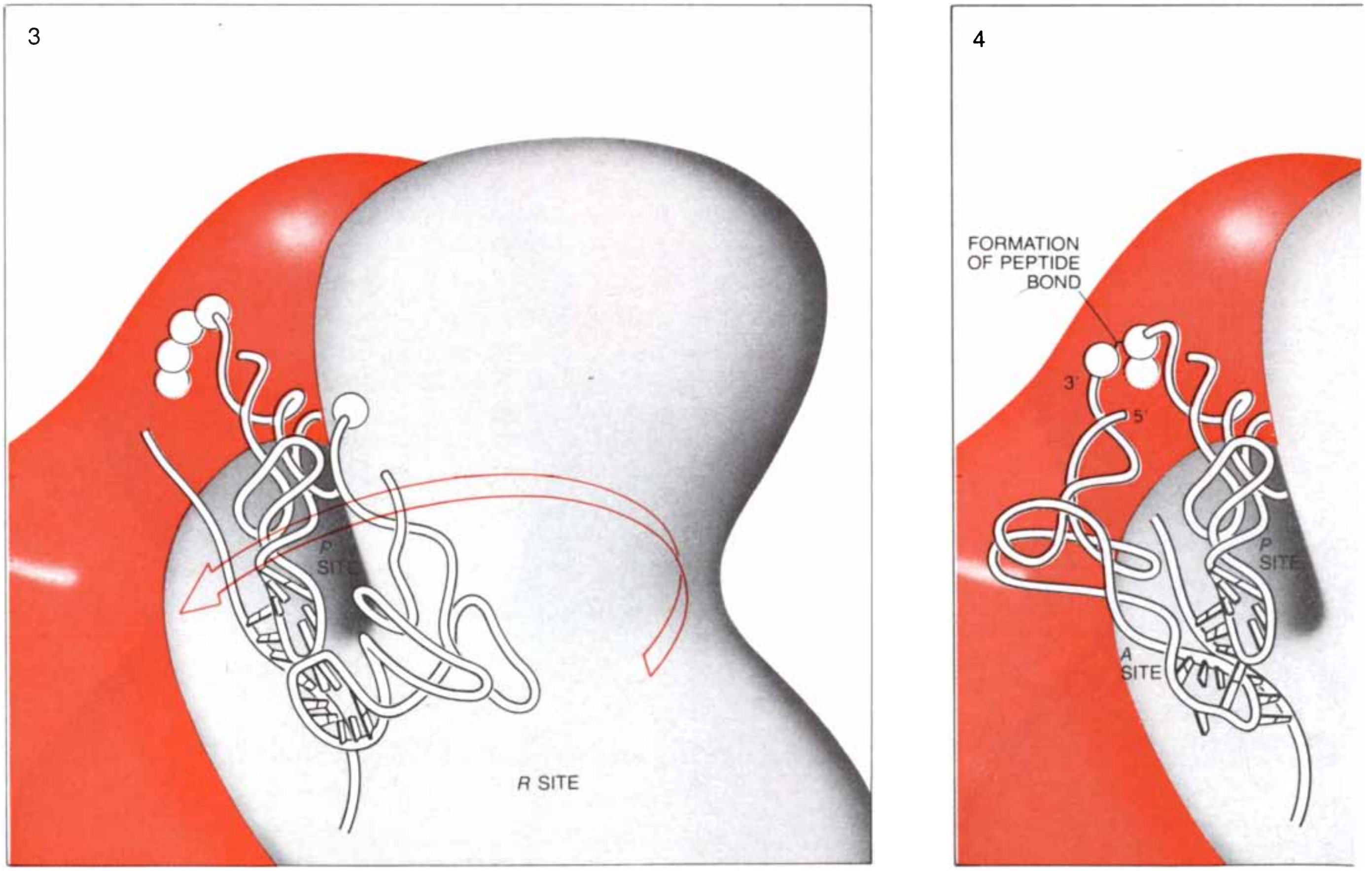
How good is the recognition? From the difference in energy between a correct and an incorrect matchup between codon and anticodon it can be calculated that an incorrect tRNA will bind (and an incorrect amino acid will be incorporated into the nascent chain) roughly once for each 100 correct incorporations. Evidence suggests, however, that the actual error rate is less than one mistake in 2,000 incorporations. For example, the longest proteins in *E. coli* are the β and β' subunits of the enzyme RNA polymerase, which each consist of roughly 1,400 amino acid units. Errors in their sequence are thought to be uncommon. It seems likely, therefore, that the ribosome must improve on the accuracy inherent in the interaction between codon and anticodon. Perhaps the improvement in accuracy is the function of the *R* site. Perhaps too the improvement is bought at the cost of a high-energy phosphate bond for each amino acid that is joined to the nascent chain.

What kind of molecular mechanism could bring about this enhancement of accuracy? Where might it lie on the ribosome? How might the *R* site be involved? A definitive answer cannot yet be given, but one can essay some guesses. After all, attempting to piece new data together is one of the pleasures of science.

Constraints on the Model

Any mechanism to improve the accuracy of protein synthesis must conform to a few basic observations. First consider the binding site for the messenger RNA. In 1974, as the three-dimensional model of the small subunit was being conceived, Kahan and I learned the results of several cross-linking experiments. Groups at the University of Vienna and at the Max Planck Institute for Molecular Genetics in Berlin had prepared analogues of mRNA in which some of the bases had been replaced by bases not normally found in RNA. Each analogue binds to the ribosome just as a natural mRNA would bind. When the analogue is bound, however, a further treatment (typically irradiation with ultraviolet light or the addition of a chemical cross-linking agent) makes strong, lasting bonds between the analogue and whatever lies nearby on the ribosome. The cross-links had established the proximity of the mRNA analogue to protein *S*11 and to a neighboring protein, *S*18. The locations of those proteins in our model allowed Kahan and me to propose that the platform and the cleft of the small subunit form the most likely binding site for the messenger RNA.

An experiment performed by Kahan with Robert Traut and John W. B. Hershey of the University of California at Davis supports that conclusion. They cross-linked IF-3, an initiation factor that facilitates the binding of natural mRNA's, to three proteins. The proteins are *S*11, which our model places on the platform of the small subunit; *S*13, which our model places on the head, and *S*19, which evidently comes to the surface of the small subunit at two places on the head. The significance of these findings is plain. The anticodon of a tRNA must bind to a codon on the messenger. Hence in any proposed mechanism the anticodon of a bound



PROTEIN SYNTHESIS CONTINUES with drawing *3*. Here the aminoacyl-tRNA has begun to flip into a position that will bind it to the *A* site. The flip constitutes a test of the strength of the binding between the codon and the anticodon; hence it improves the likelihood that the arriving amino acid is the one the code requires. In drawing *4* the aminoacyl-tRNA has arrived at the *A* site. The five stacked bases in the anticodon loop are now the ones that lie closest to the 3′ end of the molecule. This change allows the tRNA to flip without

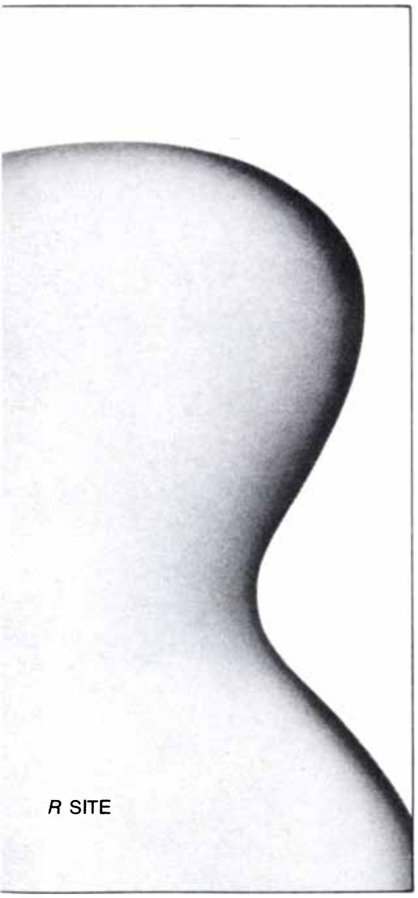
tRNA should probably be positioned in the cleft.

Next consider the binding sites for the transfer RNA's. To begin with, it can be inferred from cross-linking experiments that the initiation factor IF-2 acts at the cleft and the platform to facilitate the binding of initiator tRNA to the ribosome. Thus a credible model will probably position the *A* site and the *P* site in the cleft. A clue to the location of the *R* site is derived from locating a cluster of proteins—*S*3, *S*4, *S*5, *S*10, *S*12 and *S*14—on the side of the small subunit that faces away from the large subunit. Proteins *S*4, *S*5 and *S*12 are the ones that experiments involving streptomycin and other drugs implicate in the fidelity of the transcription of the message. The other three proteins—*S*3, *S*10 and *S*14—suggest their function another way. If any one of them is missing from a ribosome that has been reconstituted in a test tube from a mixture of the other constituents of the small subunit, the ribosome proves to be less efficient at binding tRNA's. Both of these clues implicate the external face of the small subunit in the recognition of aminoacyl-tRNA's bearing the correct anticodon. The place of such recognition presumably includes the *R* site.

Note that all the foregoing clues implicate the small subunit in the genetic (or information-processing) task of the ribosome: the binding of the instruction tape (that is, the messenger RNA) and the binding of a succession of transfer RNA's in accord with those instructions. A role for the large subunit is suggested by cross-linking experiments employing a peptidyl-tRNA in which an amino acid on the nascent chain has been modified. When the peptidyl-tRNA is bound to the ribosome, a further treatment makes the modified amino acid bind strongly to whatever is nearby. It then emerges that the nascent chain lies near protein *L*27, a constituent of the large subunit that occupies the central protuberance.

That is not to say protein *L*27 is the long-sought peptidyl-transferase: the catalyst in the ribosome that links amino acids into a chain. Experiments with partially assembled large subunits show that the lack of any one of several constituent proteins slows but does not abolish the lengthening of the chain. If peptidyl-transferase can be said to exist, it exists as a property of a region of the large subunit. In general the small subunit serves the genetic task of the ribosome and the large subunit serves the biochemical task.

A final constraint on a plausible model emerges from the research of Sigrid S. Thatch and Robert E. Thatch of Washington University in St. Louis and Peter Lengyel and his collaborators at Yale. Both groups employed enzymes that digest RNA. They found that the part of the length of the messenger RNA that is bound to the ribosome (and thereby protected from the enzymes) changes its position by the length of three bases when the peptidyl-tRNA moves from the *A* site to the *P* site. This is easy enough to accept; one has little difficulty imagining a model in which the motion of the tRNA from the *A* site to the *P* site is coupled to the repositioning of the message so that the next codon can be translated. The stringent constraint is that the messenger RNA apparently is motionless throughout the rest of the elongation cycle. In the model I shall now describe the interaction of the anticodon on the transfer RNA with the codon on the messenger RNA is unaltered by the



tugging on the mRNA. The proper alignment of the tRNA's on the mRNA may be ensured by a bond between bases that forms a bridge between the two anticodon loops. The large subunit can now transfer the nascent chain to the amino acid on the aminoacyl-tRNA. Drawing *5* completes the iteration of the cycle. The naked tRNA is expelled; the *A*-site tRNA takes its place at the *P* site and the mRNA is repositioned for the arrival of the next aminoacyl-tRNA. Neither GTP nor the elongation factor EF-G, which act in these final events, is shown.

motion of the tRNA from the *R* site to the *A* site.

The Hypothetical Mechanism

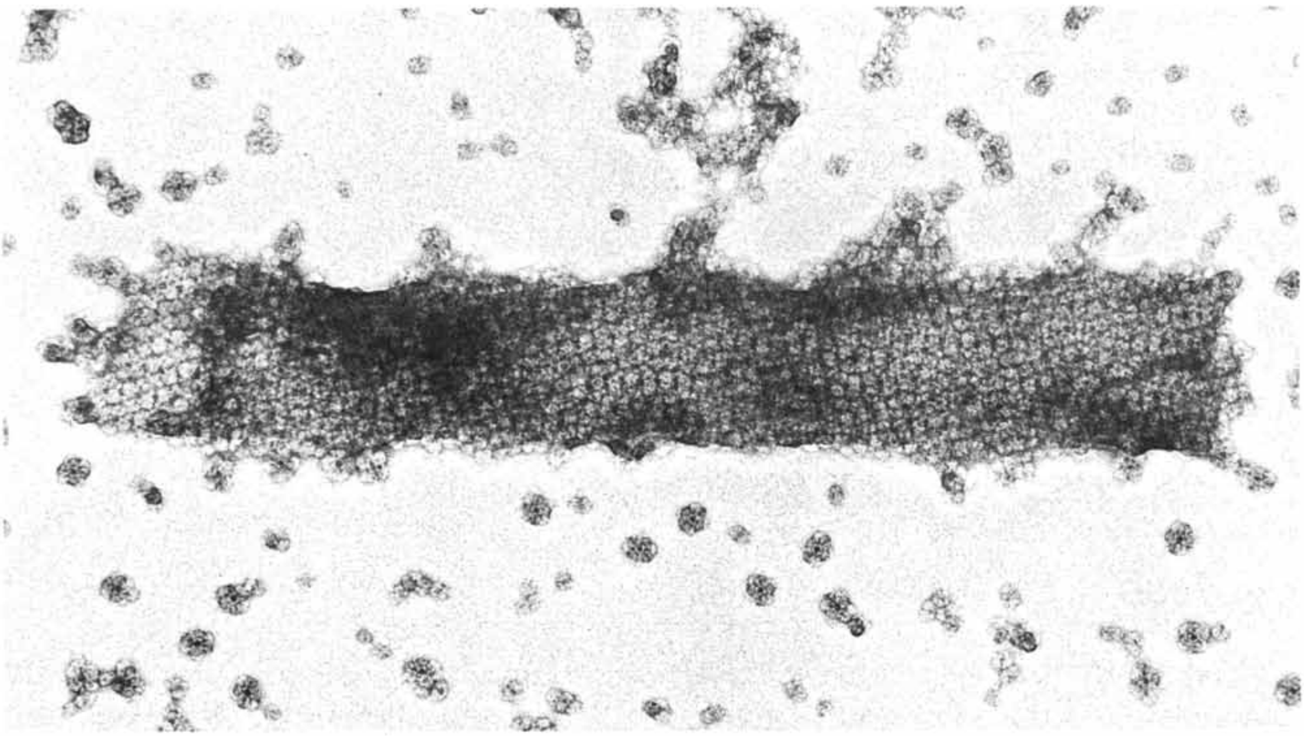
Each iteration of the elongation cycle begins (according to the model) with the peptidyl tRNA at the *P* site and with the *R* site and the *A* site vacant. A succession of aminoacyl-tRNA's are arriving at the *R* site, which lies on the external face of the small subunit. The *R* site therefore lies exposed (quite appropriately) to the cytoplasm of the cell. Each transfer RNA arrives in complex with EF-Tu and GTP, and the binding of this complex to the *R* site includes the binding of the anticodon of the tRNA to the next unread codon on the message. The chances are roughly 99 in 100 that the correct tRNA will bind. In any case the tRNA that binds is switched (or flipped) immediately from the *R* site to the *A* site, on the platform of the small subunit. The energy for the flip derives from the molecule of GTP. While the tRNA is moving it is held to the ribosome by no more than the binding of the anticodon to the codon. The motion of the tRNA therefore constitutes a test of the strength of the base pairing; it is a second reading of the anticodon. If the binding is weaker than it should be, the tRNA flies off the ribosome.

An analogy may help to make the process clearer. Suppose you have a fish tank in which some of the fish are silver and some are gold. You want to segregate the gold ones. To do this you maneuver a net around in the water and find that you can scoop out a mixture of fish in which the proportion of gold ones is larger but not large enough. Therefore the fish you trap in successive nettings you release into another tank with no fish in it. If you now apply your net with equal skill to the fish in the second tank, the proportion of gold ones you catch will improve. What you need for the project is a pair of tanks. In the ribosome the tanks are the *R* site and the *A* site.

In the course of the tRNA molecule's switch from the *R* site to the *A* site the interaction of the anticodon with the codon is unaltered. To show how that is possible I must explain a further aspect of the structure of the anticodon loop of a tRNA. How the loop is arrayed is a compromise between two constraints. On the one hand the seven bases that make up the loop tend to stack themselves one on top of another. Since it is a loop, however, the string of seven bases must turn back on itself. The compromise arranges the loop so that only five of the bases are stacked. If the five are the ones closest to the 5′ end of the tRNA, the loop is said to be in a 5′ stacked configuration; if they are the ones closest to the 3′ end, the loop is 3′ stacked. According to the model, the anticodon loop switches its conformation from 5′ stacked to 3′ stacked as the tRNA moves from the *R* site to the *A* site. The tRNA can therefore change how it is arrayed on the ribosome without yanking on the message.

One imagines that when the aminoacyl tRNA arrives at the *A* site, it might interact with the peptidyl tRNA at the *P* site. After all, the anticodon of the arriving aminoacyl-tRNA should bind to the codon just after the codon that was transcribed in the preceding iteration of the elongation cycle: the codon to which the peptidyl-tRNA remains bound at this stage of the cycle. If there were a gap between successive anticodons (a "frame-shift error"), it would mean that an incorrect codon had base-paired with an anticodon and that a tRNA bearing the wrong amino acid had therefore been bound to the ribosome. After such a mistake the synthesis would go utterly wrong: every amino acid would thereafter be incorrect unless further frame-shift errors fortuitously brought the message back into alignment.

The model provides that the "reading frame" can be checked by means of base pairing between adjacent *A*- and *P*-site tRNA's. A pairing between the 33rd base of both tRNA's seems well suited for this checking. For one thing the 33rd base lies on the anticodon loop. Indeed, it is next to the anticodon; thus its binding ensures there has been no frame-shift error. Moreover, the structure of a tRNA allows the 33rd base to stick out of the molecule and enter into the bond. Further still, the 33rd base is constant throughout every tRNA that participates in the elongation cycle. Whatever



ARRAY OF LARGE SUBUNITS from the ribosomes of *E. coli* was grown by Michael Clark and the author at the University of California at Los Angeles by allowing the subunits to precipitate from a hanging drop of solution. Arrays of small subunits have also been grown. The arrays give hope that ultimately large crystals of ribosomes might be made the subject of X-ray diffraction, the technique in which the atomic regularities in a crystal give rise to an interference pattern from which the atomic structure of the crystal is deduced.

amino acid it is specialized to carry, its 33rd base is uracil. The only exception is the special tRNA that serves the initiation of protein synthesis in eukaryotic cells (cells, such as those of organisms higher than bacteria, that have a nucleus). Here the 33rd base is cytosine. This difference may be important. The unusual nature of the base pairing in RNA dictates that the cytosine of a tRNA bound to the ribosome at the *P* site can base-pair with the uracil of a tRNA bound at the *A* site. The reverse is not the case: the uracil of a tRNA at the *P* site cannot pair with the cytosine of a tRNA at the *A* site. Perhaps this prevents initiator tRNA from entering the *A* site during the elongation cycle.

The base pairing of uracil with uracil is, however, a detail. The important aspects of the model are its inclusion of the *R* site and the flip of the aminoacyl-tRNA from the *R* site to the *A* site. After the tRNA is bound at the *A* site the elongation cycle proceeds in accord with the accepted scheme. The nascent chain is transferred to the amino acid on the tRNA at the *A* site. For that no outside source of energy is required. The EF-Tu, the GDP and the phosphate leave the ribosome and two other molecules arrive. They are the second elongation factor, called EF-G, and a fresh molecule of GTP. The deacylated tRNA is ejected from the *P* site; the *A*-site tRNA, now bearing the nascent chain, rotates into its place, and the message is advanced one codon. EF-G and the molecule of GTP (now split, like its predecessor, into GDP and phosphate) leave the ribosome, so that the ribosome is ready to begin the programmed addition of the next amino acid.

Ribosomal RNA

The model suggests the ribosome is large because it must simultaneously cradle two transfer RNA's that are interacting head to head. That is, the ribosome must be large enough to support a molecular complex about 150 angstroms long and at the same time provide binding sites for GTP and the various elongation factors.

Why is the ribosome itself composed mostly of large RNA's? Perhaps they are the evolutionary remnants of the primordial ribosome. F. H. C. Crick of the Salk Institute for Biological Studies has reasoned that since proteins are made by ribosomes, the first ribosome would have existed before any ribosomal proteins. It would thus have consisted of RNA. It counts in favor of this hypothesis that the tRNA's are made up solely of RNA.

A calculation suggests an additional reason. Assume that an intracellular organelle such as a ribosome requires for its structure a molecule whose molecular weight is a million. One possibility is that the molecule will be a nucleic acid: a strand of nucleotides. (A nucleotide consists of a base, a sugar and a phosphate group.) The molecular weight of a single nucleotide is about 330; hence about 3,000 nucleotides would be needed. The genetic instructions for their sequence would be a DNA strand of the complementary nucleotides, and "reading" the instructions to produce the nucleic acid would call for 3,000 steps.

The other possibility is a protein: a strand of amino acid units. The molecular weight of a single amino acid unit is about 110; hence about 9,000 such units would be needed. The genetic instructions for their sequence would be a strand of DNA that has a codon, or three nucleotides, for each amino acid. Reading the instructions to produce the messenger RNA that codes for the protein would require 27,000 steps. Assume that the translation of the message by a ribosome is error-free and only the transcription of DNA into RNA is imperfect. Even so, it is plain that an RNA some nine times the weight of a protein can be synthesized for any given probability of error.

Imagine, finally, that an error in the transcription of DNA into RNA places an extra nucleotide somewhere in the RNA. This addition might well produce no more than a slight bulge in the RNA's tertiary structure. On the other hand, if the nucleic acid were a messenger RNA, the extra nucleotide would lead to a frame shift, and the protein synthesized from the instructions on the messenger would almost certainly be useless. Thus if a very large molecule is needed for the structure of the ribosome, the accurate synthesis of the molecule probably dictates that it be an RNA.

The past decade of progress in understanding the structure and function of the ribosome raises hope for further progress. For example, the indications that ribosomes have only a single structure and a single shape raise new hope that they can be stacked in a crystalline array for X-ray-diffraction analysis. Indeed, Michael Clark, Jerome Langer and I recently produced ordered arrays of ribosomal subunits from *E. coli* by applying the "hanging drop" technique. The subunits (in a drop of solution) hang from a sheet of glass in a chamber containing a vapor of ethanol. As the concentration of ethanol in the drop approaches 10 percent the subunits begin to precipitate.

Perhaps X-ray diffraction can ultimately be brought to bear on an array larger and more orderly than the ones we have so far produced. Still, a view of the ribosome that locates every atom would not necessarily show which atoms belong to which molecule. The techniques employing antibodies and the electron microscope are sure to remain invaluable. The goal of all the efforts remains the complete elucidation of how the ribosome makes protein.

The Forging of Metals

This ancient art shapes metals by pounding or squeezing them. It is currently evolving with advances in knowledge of how metals deform and with the application of computers to steps in the forging process

by S. L. Semiatin and G. D. Lahoti

To the blacksmiths of earlier times forging meant shaping a piece of metal by holding it on an anvil and pounding it with a hammer. As such it is probably man's oldest means of working metal into useful shapes. Today forging is formally defined as the compressive deformation of material between dies. Often the dies are components of huge presses capable of exerting enormous force, and the part that emerges can be as elaborate as the entire wing structure of an aircraft. Although one might suppose such a mature technology would offer few opportunities for evolutionary change, the techniques of forging are in fact being much altered by new knowledge about the deformation properties of metals and by the increasing application of computers in the design of forging dies and other forging equipment and in the analysis of what happens to a piece of metal as it is being forged.

Modern forging occupies a prominent place in what is termed primary metalworking, which comes after the steps in which metal is extracted from ore, is melted to remove impurities or to add alloying elements and then is cast, sometimes into a final shape and sometimes into an ingot for further working. Forging accounts for about half of primary metalworking, which also includes such processes as extruding, drawing and rolling. In many instances a shape made by one of these processes is subjected to a secondary metalworking process such as stamping or machining.

The earliest records of forging describe the simple hammering of gold and copper in various regions of the Middle East around 8000 B.C. The forming of these metals was crude by modern standards, since the art of refining by smelting was unknown and the ability to forge the metal was limited by impurities that remained after the metal had been separated from the ore. The advent of copper smelting around 4000 B.C. provided a useful method of purifying metals through chemical reactions in the liquid state. Later in the Copper Age it was found that the hammering of metal brought about desirable increases in its strength (owing to what is now called the work hardening of the metal). The quest for strength spurred a search for alloys that were inherently strong and led to the utilization of alloys of copper and tin (the Bronze Age) and iron and carbon (the Iron Age).

The Iron Age, which can be dated as beginning around 1200 B.C., followed the beginning of the Bronze Age by some 1,300 years. The reason for the delay was the absence of methods of achieving the high temperatures needed to melt and refine iron ore. Then it was discovered that iron alloys are easily forged when the temperature is above only 60 percent of the absolute melting point expressed in degrees Kelvin, or degrees Celsius above absolute zero. Forging at such temperatures is now called hot-forging or hot-working. (One of its results, in addition to the shaping of the metal, is the development of iron oxides in the form of black scale. It was the ubiquitous presence of this material in early forging shops that gave rise to the term blacksmith.)

The development of the tilt hammer in the 13th century was a boon to blacksmiths; it relieved them of much of the labor needed to forge metals. This machine employed waterpower to raise a lever arm with a hammering tool at one end; it was called a tilt hammer because the arm tilted. After raising the hammer the smith let it fall, thus generating the forging blow. This relatively simple apparatus remained in service for some centuries, primarily for the forging of bars and plates. With the Industrial Revolution came processes for making iron and steel in larger quantities, together with a great increase in the demand for metal products. A need arose for forges with larger capacity. The need was met by the development of the high-speed steam hammer, in which the hammer is raised by steam power, and of the hydraulic press, in which the force is supplied by hydraulic pressure. From such forging machines came products ranging from firearms to locomotive parts.

The past 100 years have seen the development of new types of forging equipment and new materials with special properties and applications. The utilization of these materials has benefited greatly from new knowledge (much of it acquired over the past 25 years) on the effects of temperature and deformation speed on the basic working characteristics of metals and the effects of the design and lubrication of dies on the flow of metals and the force needed for forging. The application of this knowledge to forging work has been facilitated by the development of sophisticated mathematical analyses of forging processes, analyses that are best carried out on high-speed digital computers.

It is now commonly accepted that the explanation of the ability to deform metals lies in the nature of the metallic bond. In this type of bond the metal atoms share their outer electrons. The bond is also nondirectional, so that the atoms can pack themselves closely in regular three-dimensional crystalline arrays. The presence in these structures of dislocations, or imperfections in the three-dimensional array, makes it possible for the crystal lattice to shear on close-packed planes of atoms along close-packed directions. As one might suppose, the comparative workability of different metals depends strongly on the ease with which dislocations can be made to move without causing the metal to break or without introducing other kinds of defect into it.

The ease with which dislocations can be made to move without undesirable results is a function of structural features, the temperature of the metal and the rate of deformation. The structural features include the three-dimensional crystal structure, the presence of alloying elements and the presence of other dislocations, but probably the most important feature is the grain size. Grains are the building blocks of metal aggregates. Each grain is a separate crystal, with a crystal-lattice orientation different from that of its neighbors. When the grain size is large in relation to the overall size of the workpiece (the piece of metal that is being worked), cracks can develop and can propagate easily along the boundaries between grains.

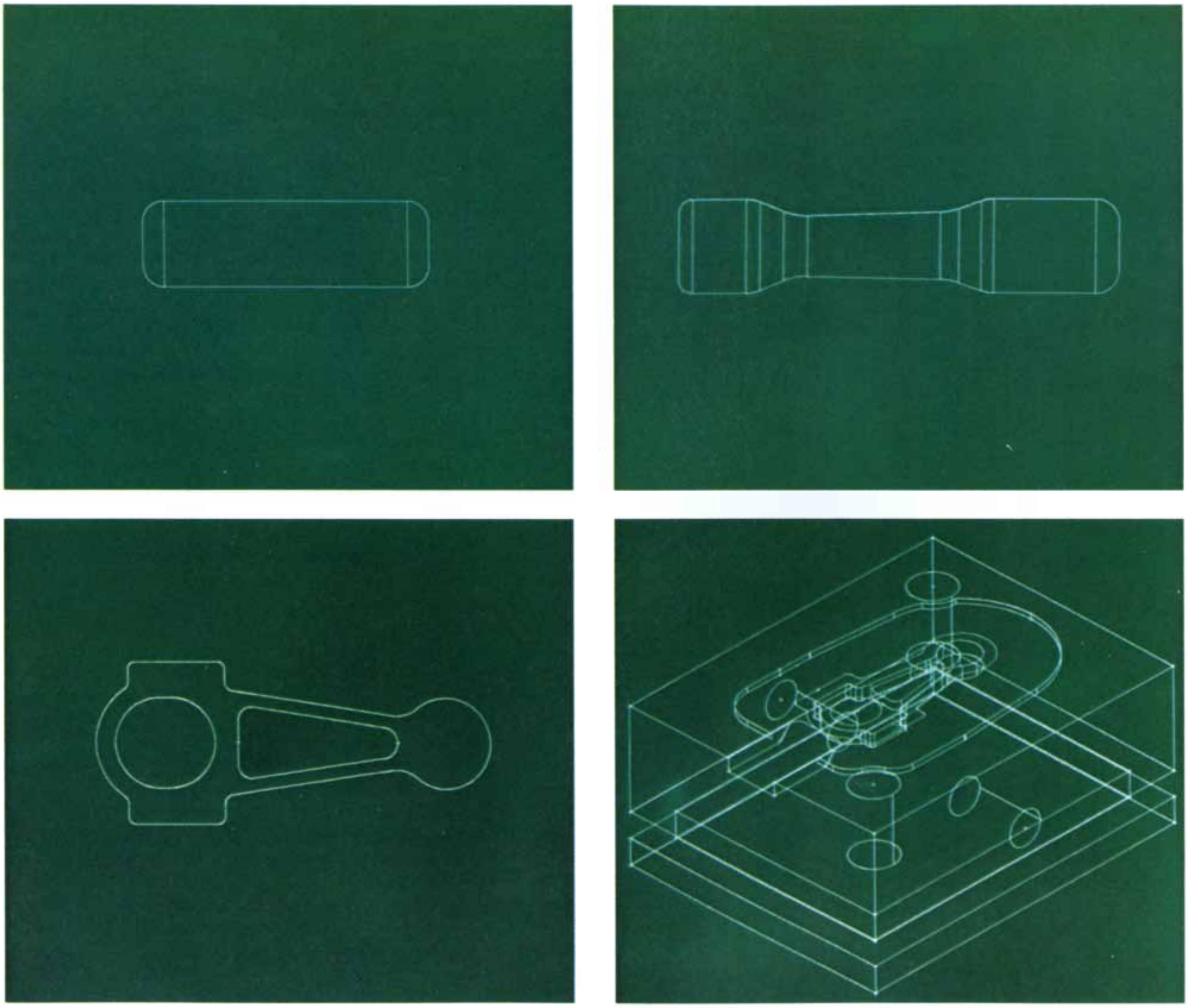
This problem is of particular concern when the structure of the metal is characteristic of a piece that has been cast, for example a piece that has solidified from the melted state in an ingot mold. As the metal solidifies it freezes first at the cold walls of the mold. Coarse, columnar grains grow inward as much as several centimeters. The metal is difficult to forge without cracking it because the interfaces between grains are planes of weakness.

As a result the forging of cast metals is often done in two steps. In the first step forging serves to break down the coarse-grained structure, giving rise to a chemically homogeneous wrought structure with a much finer grain size, usually measurable in hundredths of a millimeter, and far superior workability. It is from metals of this structure that finished forged parts are made.

Although most wrought metals can be worked over a wide range of temperatures and rates of deformation, care must nonetheless be exercised in selecting the forging temperature. Hot-forging takes a minimum of force in forging and imparts a maximum in workability, but the forger finds it difficult to control the dimensions of the product because the metal contracts nonuniformly as it cools. Therefore forging is often done at room temperature, at which most metals lie in the cold-working regime (25 percent or less of the absolute melting point of the metal). At such temperatures the workability of the metal is somewhat less than it is at hot-working temperatures but is far greater than it is at the warm-working regime (between 25 and 60 percent of the absolute melting point). This apparent paradox is a subject to which we shall return.

It goes almost without saying that the main objective in forging is to make parts of the right size. Beyond that the major considerations are the force needed to achieve the deformation (in other words, the capacity of the equipment) and the workability of the metal. Secondary considerations, mainly economic ones, are the rate of production and the durability of the dies.

For a particular metal an estimate of the force required for deformation, a force that is formally termed the forging load, is often obtained by compressing cylindrical samples between flat, parallel dies that have been well lubricat-



COMPUTER-AIDED DESIGN is a recent development in the forging of metals. Here the cathode-ray-tube display of a Computervision system shows the starting bar stock for a connecting rod in an automobile engine (*upper left*), an intermediate forging (*upper right*) and the finished forging (*lower left*). The final display is a drawing of one of the two dies employed in the final forging operation. The techniques for computer-aided design in forging are being developed by A. A. Badawy and C. F. Billhardt of the Battelle Memorial Institute.

ed. This mode of deformation simulates the simplest kind of forging operation, which is known as upsetting from an old usage of "upset" in the sense of "beat back." Through this test the flow strength, or resistance to deformation, in terms of stress (load divided by area) is plotted as a function of some measure of deformation, usually strain (the logarithm of the ratio of the initial height of the sample to the final height) or the reduction in height expressed as a percentage. The stress-strain curves thus obtained show the effect of the amount of deformation, the rate of deformation and the test temperature on the metal's resistance to deformation. They are an essential input when one wants to model a forging process mathematically.

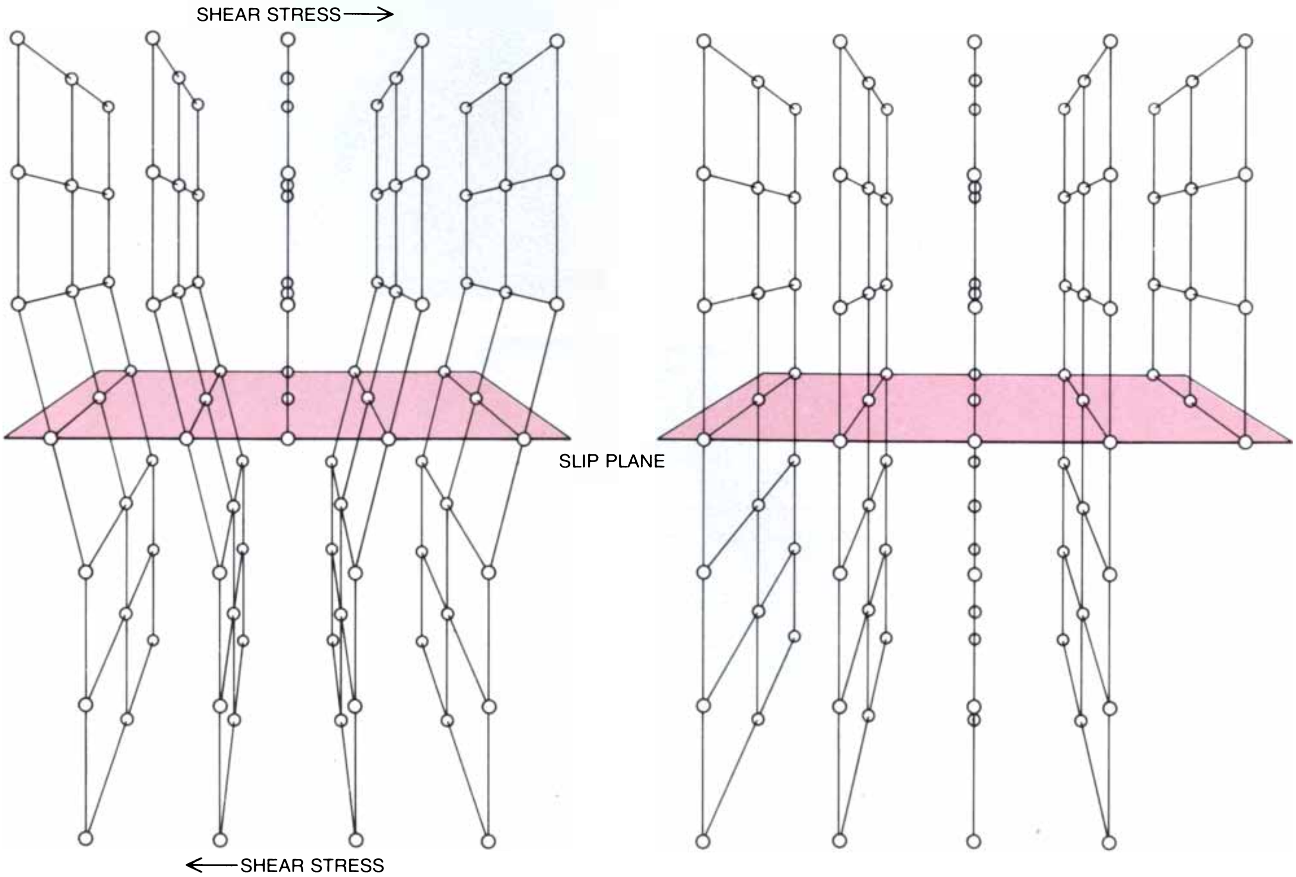
At cold-working temperatures the stresses are high, primarily because of the difficulty of moving dislocations through the lattice and because the dislocations multiply rapidly, leading to work hardening: increases of stress with strain. The workability limits are modest. At warm-working temperatures the stress levels are somewhat lower, but there is still a large amount of work hardening and the workability is low.

The stress levels are lowest at hot-working temperatures because thermal energy causes large fluctuations of the atoms in the crystal lattices of the metal grains around their equilibrium positions; as a result it is easier for dislocations to move through the lattices. Moreover, the stress-strain curves reflect no work hardening because thermal energy is effective in "annealing" out many of the dislocations generated during deformation. Indeed, the structure of a metal may change so drastically during hot-working that the result is work or flow softening (decreases of stress with strain). The microscopic mechanisms of this change have been documented by John J. Jonas of McGill University and Michael Luton of the Exxon Corporation, who also demonstrated that the softening can give rise to grossly nonuniform deformation in compression samples.

In a real forging operation the forging load and the forging pressure depend markedly on the friction at the interfaces between the dies and the workpiece as well as on the flow strength of the metal of the workpiece. Under conditions of perfect lubrication the pressure required to forge a cylindrical sample is uniform and is equal to the flow strength of the material. With friction at the interface the distribution of forging pressure takes the form of a tent: the pressure is minimum (equal to the flow strength) at the outer rim of the sample and maximum at the center. Associated with this pressure distribution are radial compressive stresses in the sample, which develop from the constraining effects of friction. Depending on the magnitude of friction and deformation, the maximum pressure in forging can be as high as 10 times the flow strength of the material.

The friction conditions at the forging interface are governed by lubrication and, in warm- and hot-forging, by the temperature gradient between the cold dies and the hot material. (This gradient can lead to a chilling of the material and a "locking" of the metal touching the dies.) Lubricants reduce the friction and control or improve the uniformity of flow in the metal. They also reduce die wear and keep the workpiece from sticking to the dies.

In the cold-forging of steels the usual lubricants are soaps. They not only provide excellent lubrication at room temperature but also can be removed easily from the forged parts without affecting



DEFORMATION OF METAL arises from the motion of dislocations through the crystal lattice of the metal. A dislocation is a defect in the stacking arrangement of atoms in the crystal. An "edge dislocation" is generated by the presence of an extra half plane of atoms in an otherwise regular lattice (*left*). The dislocation makes it easy to shear the lattice on a "slip plane," which is usually a close-packed plane, and along close-packed directions (*right*). When forging pressure is applied to a piece of metal, each crystal in the metal deforms along several sets of planes and directions in order to accommodate the deformation being imposed on the workpiece by the forging operation.

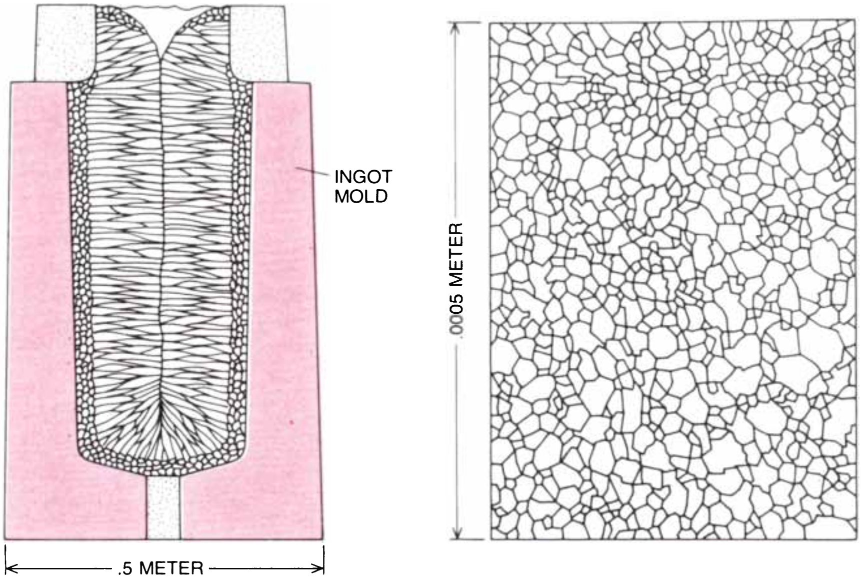
the appearance of the parts. In warm- and hot-forging the commonest lubricant is probably graphite. Oil-based graphites tend to give off fumes, and so the latest trend in the forging industry is toward the use of water-based graphites.

Forging lubricants are often evaluated by means of the "ring test," in which a ring-shaped sample of known dimensions is deformed between two lubricated flat dies to a predetermined reduction in height. The change in the inside diameter of the ring specimen is highly sensitive to the friction at the interface; the amount of change increases if the friction is small and decreases if it is large. The percentage of change in the inside diameter provides a basis for rating the effectiveness of lubricants.

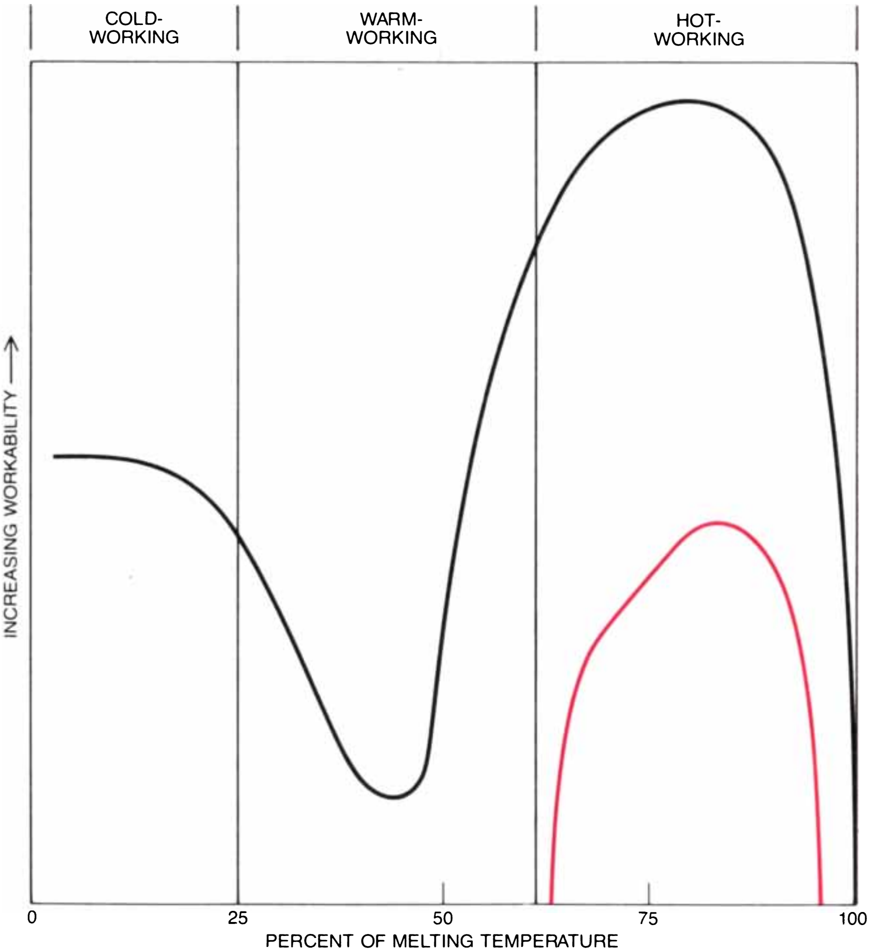
When the part to be made has a fairly simple shape, it is relatively easy to estimate the forging load by means of the measured data on flow strength and friction. Estimates of load for a complex part are difficult to obtain. They depend strongly on the preform, or starting shape, of the workpiece, on the design of the die and on the degree of contact between the die and the material. The degree of contact is small in open-die forging and high in closed-die forging.

In addition the manufacture of complex parts typically calls for several forging steps. Designing the number and configuration of intermediate shapes (known as preforms or blockers), the dies and the forging conditions for the different steps is probably the most difficult task of the forging engineer. Forging a connecting rod for an automobile engine may require three stages to obtain the finished shape. In forging shops experienced design engineers usually follow empirical guidelines to reach their decisions. Sometimes they test materials such as wax, lead and clay in small-scale dies or examine flow patterns by studying the effects of deformation on grid lines previously inscribed on test specimens. Grid-line analysis is also done at times to find regions in which undesirable states of stress may develop in the course of forging. For example, the technique reveals regions where geometrical effects may lead to tensile stresses, called secondary tensile stresses, even under the nominally compressive loading. Computer-aided techniques for designing preforms and dies and for simulating the flow of metals are now coming into the picture.

When the dies have been designed and the forging load has been determined, the die material and the equipment for forging are selected. The materials for dies range from common-alloy and high-strength tool steels to superalloys that can withstand high pressures at high temperatures. When the pressure and the temperature of forging are moderate and the contact time is short, common-alloy steel dies are adequate. Sometimes



WORKABILITY OF METAL depends strongly on the size of the grains (individual crystals) in a piece of metal being forged. A metal that has been cast in an ingot mold (*left*) has large grains, measurable in centimeters, and typically is difficult to forge. A wrought structure (*right*) has small grains, measurable in hundredths of a millimeter, and is much easier to work. Forging is therefore often done in two steps. In the first step forging is employed to break down the large-grain cast structure to achieve the wrought structure from which finished forgings are made.



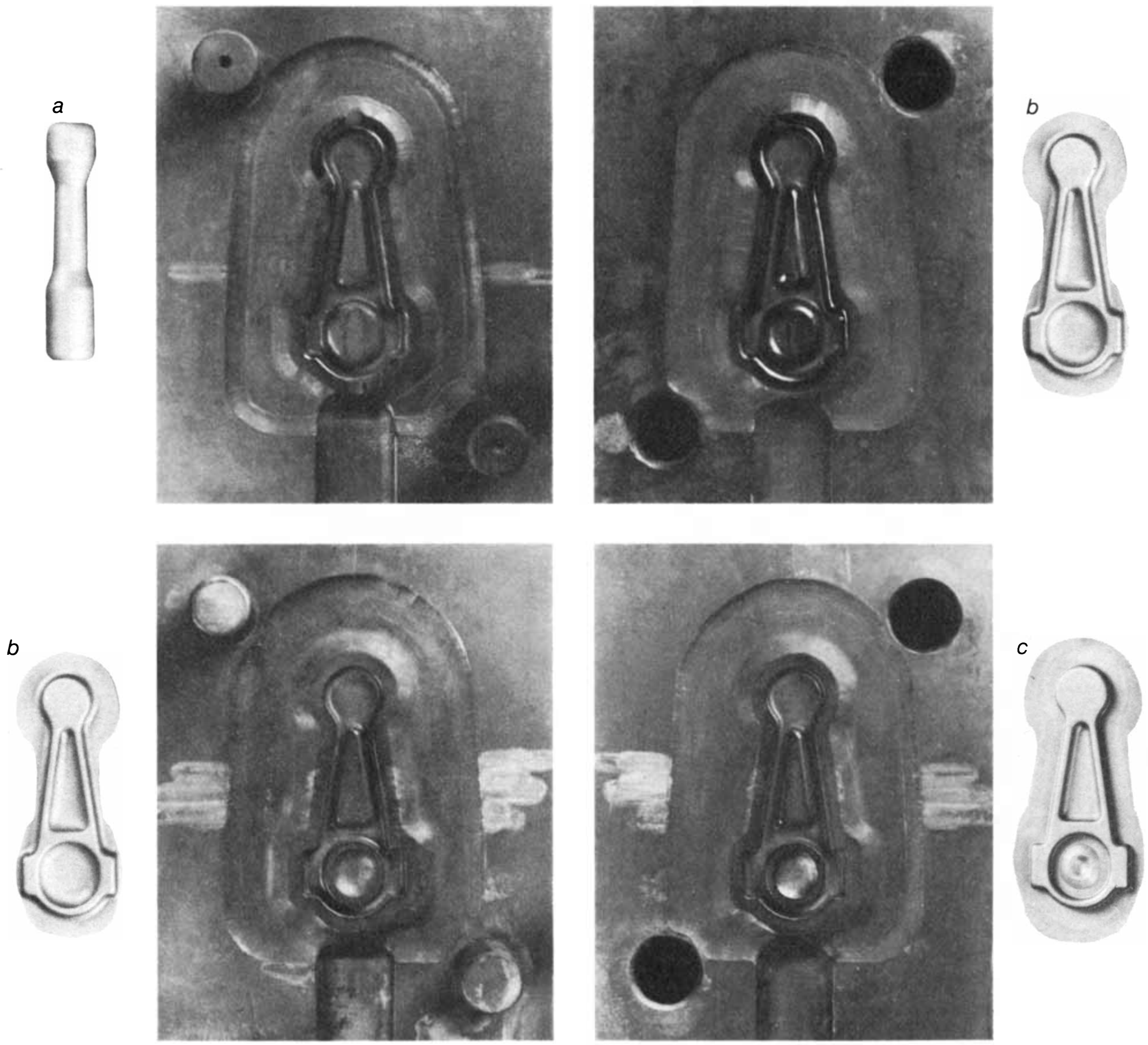
ROLE OF TEMPERATURE in the workability or forgeability of metals is charted for cast structures (*color*), which have coarse grains, and for fine-grain wrought structures (*black*).

such a die is coated with a hard, wear-resistant superalloy to increase the die life at a reasonable cost. The same objective is achieved in many cold-forging operations by means of tungsten carbide die inserts: small dies in supportive containers. At the other end of the spectrum, in exotic hot-forging operations where the temperatures of the die and the workpiece are nearly equal, the entire die assembly is made from special materials (such as IN-100, a nickel alloy, or TZM-Molybdenum) and is heated to as much as 1,000 degrees C. in a controlled atmosphere to avoid oxidation. This kind of forging operation is employed in the precise shaping of alloys of metals such as titanium, nickel and beryllium.

Finally, the forging engineer selects the equipment that is most suitable for forging a particular part. His primary consideration will be the force and the energy that the forging machine can deliver, but other characteristics, such as the machine's speed, which may affect the workability of the part, are also taken into account.

In making the choice the engineer has three basic types of forging equipment to consider: the drop hammer (dropping either by gravity alone or with power assistance), the mechanical press and the hydraulic press. Hammer forging entails repetitive blows to the workpiece; a mechanical or hydraulic press usually achieves a forging with one stroke. Otherwise the main differences among the three types of equipment are their forging speed and the way in which they store energy. Hammers are fastest, with speeds ranging from three to nine meters (120 to 360 inches) per second. They store energy in the form of the potential energy of a raised weight, which is converted into kinetic energy as the ram falls. In a power-assisted drop hammer steam or air pressure is applied to the ram during the downstroke. A hammer is rated according to the energy it can deliver; large hammers have capacities of about 680,000 Newton-meters (500,000 foot-pounds). Hammers are the least expensive and most versatile of the forging machines.

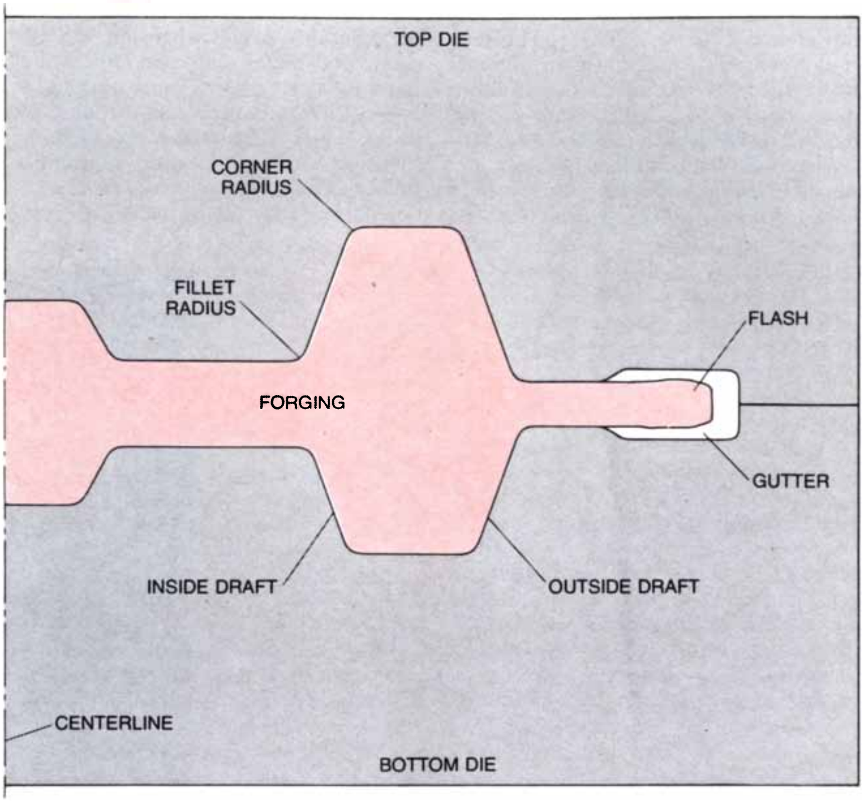
The energy in a mechanical press is stored in a flywheel. The ram is actuated by a mechanical system such as a crank



FORGING IN STAGES is often necessary to achieve a finished part. The process is shown here for a connecting rod for an automobile engine. Each stage requires a preform, or starting shape, and a set of dies. For the connecting rod a round bar is forged to make the first preform (*a*); that object is forged in a set of dies to make the second preform (*b*). The second preform is then forged in another set of dies to yield the connecting rod (*c*), which is finished except for the flash material (the excess metal) that still must be trimmed off.

or a cam. Since a mechanical press has a stroke that can be accurately controlled, it is often the choice when a forging must be made to close tolerances. The velocity of the ram is .06 meter to 1.5 meters per second. A large mechanical press can develop loads exceeding 90 meganewtons (10,000 tons).

Hydraulic presses are the slowest (.06 to .30 meter per second), because such a press is actuated by a large piston driven by a fluid under high pressure. Some of the largest forging presses in the world are hydraulic. The two biggest ones in the U.S. are rated at 445 meganewtons.



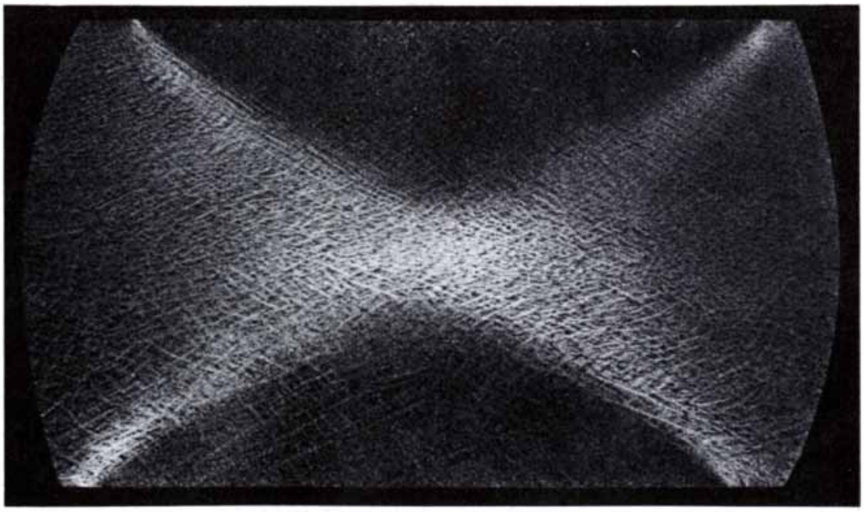
ELEMENTS OF DESIGN of dies for forging are shown in a vertical cross section through the right-hand halves of a pair of dies and the right-hand half of a forging. The design of the dies calls for the determination of the fillet and corner radii, which enhance the flow of metal; the inside and outside draft angles, which facilitate the removal of the part after it has been forged, and the gutters into which the flash can move. For a complex part the design can be difficult.

The second major consideration in determining the conditions for forging is the workability of the metal. As one might expect, the modes of fracture in forged metals are highly dependent on temperature. In cold-forging the fractures usually develop as cracking on surfaces that are not touching the dies. The cracking results from a friction-induced bulging or barreling of those surfaces, which in turn leads to the development of secondary tensile stresses there. Peter Lee of the Timken Company and Howard A. Kuhn of the University of Pittsburgh have developed criteria for the cracking that are of great use in cold-forging design.

Other kinds of fracture and defect appear in cold-forging but are either uncommon or negligible in their effect on the service life of the part and so are generally ignored. One is porosity, in which holes develop because of secondary tensile stresses generated internally in the workpiece. The holes can be microscopic or many times the grain size. Small holes usually do not affect the service properties of the part; large ones, called central bursts, generally cause the workpiece to break during forging, so that the part is never put into service.

Another defect occasionally found in cold-forging is the presence of shear bands. They are regions of localized shearing that develop between areas of widely differing degrees of deformation inside the workpiece. One of the main causes of the nonuniform internal deformation is friction, which can cause the metal touching the dies to lock by raising the load needed to deform it considerably beyond what is required for the rest of the workpiece. A locked area is known as a dead-metal zone. In cold-forging the shear bands separating the dead-metal zone from the rest of the workpiece are usually quite diffuse because of the influence of work hardening and only rarely become localized enough to give rise to a gross fracture of the part. Improvements in lubrication or minor changes in the design of the die or the preform usually suffice to alleviate a situation in which shear bands are appearing.

In warm- and hot-forging the development of internal holes or shear bands is a much more troublesome problem. The major factor limiting workability in warm-forging, to which we alluded above, is holes or cavities that arise at grain boundaries. One cause of such holes is secondary tensile stress. The holes are initiated by the action of dislocations moving through the crystal lattices and impinging on the grain boundaries. If there are any microscopic particles at the boundaries, the dislocations may induce a shearing action on them, opening weak interfaces between the

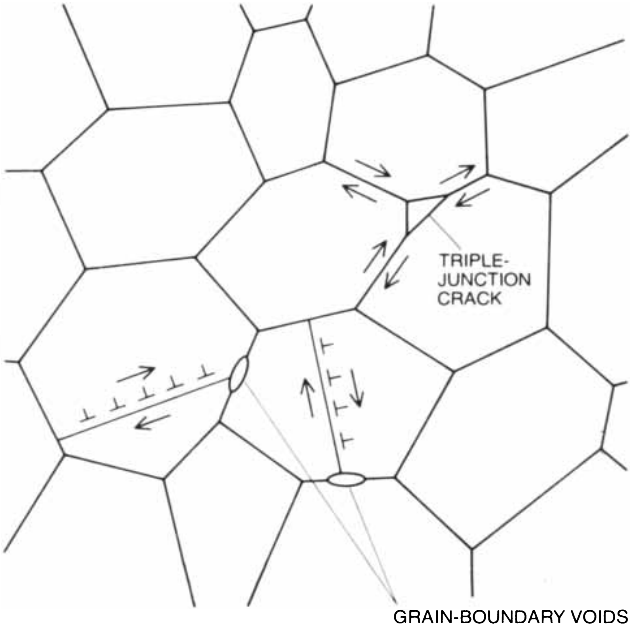


SHEAR BANDS are one type of defect that can appear in cold-forging. Here they are the lighter regions in a cross section of a cylinder of carbon steel that was compressed between unlubricated dies. The bands result when friction constrains the deformation of the metal touching the dies. Those areas become "dead metal" zones; here they are the darkest regions. The dead-metal zones are separated from the remainder of the deforming metal by regions where the extent of deformation is much higher than the average. Those regions are the shear bands.

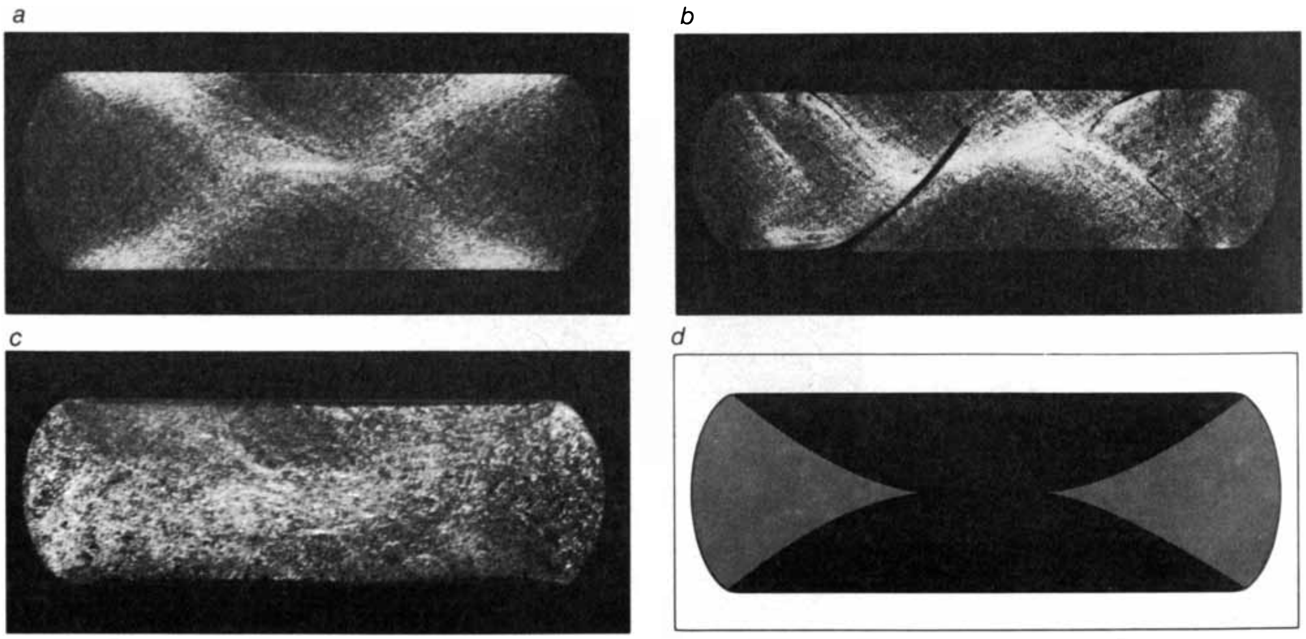
particles and the boundaries. The secondary tensile stresses cause the holes to grow. If the imposed deformation is large enough, the holes impinge on one another and give rise to a fracture.

A second type of hole, which is common at high warm-working and low hot-working temperatures, appears at triple junctions: places where three grains meet in the metal. At these temperatures shearing along the grain boundaries becomes thermally activated. Because the stress levels are moderately high the shearing can lead to stress concentrations and holes at the junctions. The lattice deformation is insufficient to relax the stress concentration, and so cracks traversing the length of a series of grain boundaries may start at the junctions, eventually causing a gross fracture of the workpiece.

Often simply increasing the temperature of forging to the hot-working regime remedies the kind of workability



TWO PROBLEMS that limit the workability of metal at warm-working and moderate hot-working temperatures are the formation of cracks at triple junctions and voids that appear at grain boundaries. The problems are diagrammed at the left and shown in a photomicrograph of a titanium alloy at the right. In the diagram the arrows denote shear stresses and the symbols resembling an inverted *T* denote dislocations. Triple-junction cracks result from a thermally activated shearing at grain boundaries. Voids are initiated by the actions of dislocations that impinge on particles found at the grain boundaries. In the photomicrograph, in which the enlargement is about 250 diameters, the boundaries of three grains in the titanium are visible as the thin white ribbons that converge at the triple-junction crack.

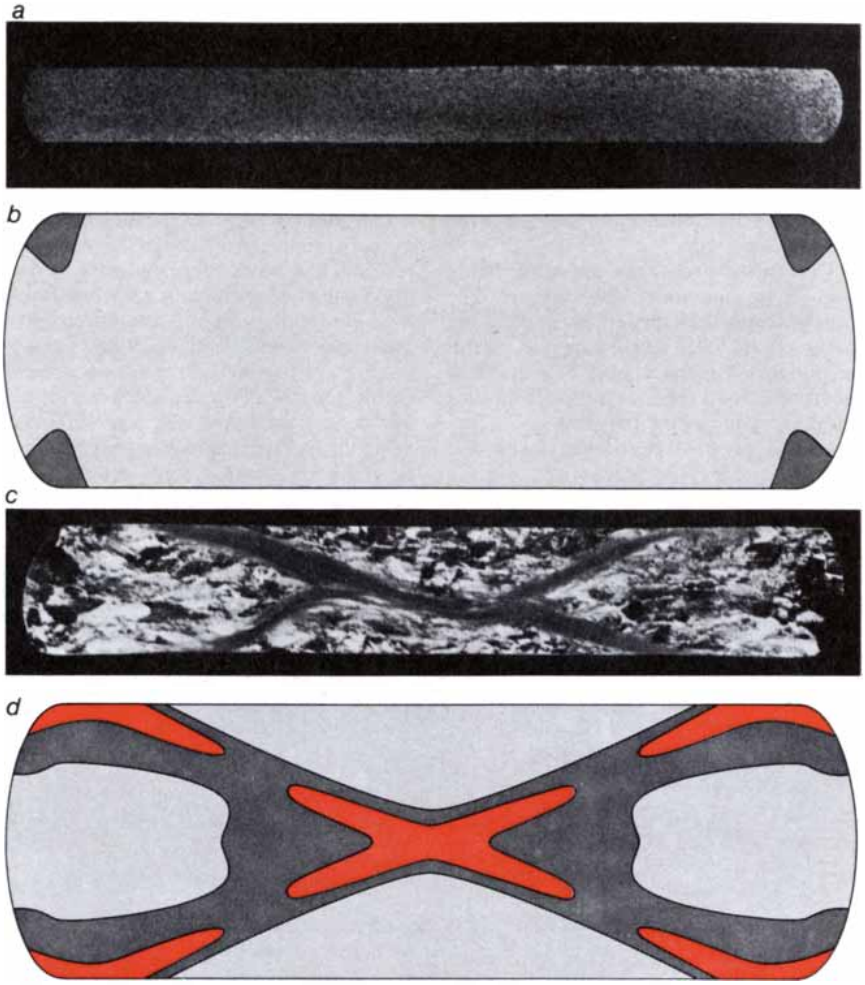


HOT-FORGING DEFECTS include shear bands that can develop when cold dies act against hot metal. Examples appear here (*a, b*) in cross sections of hot-forged bars of a titanium alloy that had been preheated to a temperature at which deformation resistance is highly sensitive to temperature. In a mechanical press fast working leads to small amounts of heat transfer between the dies and the metal and thus to modest shear bands (*a*); in a hydraulic press slow working leads to more heat transfer and hence to severer shear bands and to shear cracks (*b*). In contrast, a relatively uniform deformation can be achieved (*c*) if the workpiece is preheated to a temperature at which resistance to deformation is not highly sensitive to temperature and if the forging is done rapidly, as it is in a mechanical press. The diagram (*d*) depicts a nonuniform structure containing shear bands or shear cracks (*gray*) and dead-metal zones that are cooler (*black*).

problems we have described. Even then care must be taken because the working stress in the hot-working regime is typically quite sensitive to temperature, with large decreases in flow strength as the temperature of the workpiece rises. The effect is particularly important in forging metals such as the titanium alloys employed in jet-engine and airframe parts. When the stress is extremely sensitive to temperature, shear bands (and shear cracks along them) may develop during forging. We and our colleague Albert Hoffmanner at the Battelle Memorial Institute Columbus laboratories and Alan H. Rosenstein of the Air Force Office of Scientific Research were among the first to recognize the source of these defects and to establish ways to avoid them.

The primary source of shear bands in hot-forging is the transfer of heat between the dies and the workpiece. It causes the metal that is in contact with the usually much cooler dies to chill, thereby raising its working stress and the load needed to deform it above the levels required for the other parts of the workpiece. Dead-metal zones and shear bands are formed.

One way to minimize the occurrence of shear bands in conventional hot-forging is to raise the preheat temperature of the workpiece to a level where the stress is not too dependent on temperature. Another is to reduce the transfer of heat by introducing lubricants, which act as thermal insulators between the dies and the workpiece, or by reducing the contact time by forging with faster rams. One can also heat the dies to the same, or almost the same, temperature as the workpiece. This technique, known as isothermal hot-forging, is currently receiving much attention. In the absence of significant amounts of heat transfer, parts can be effectively hot-forged in slow hydraulic presses. Because of the low speed of the ram the working loads are smaller than they are in conventional hot-forging done with mechanical presses or hammers. Moreover, the range of temperatures at which metal can be worked is considerably larger than it is in conventional hot-forging. The technique must be applied carefully with metals that exhibit work softening during hot-working. Deformation in such metals may localize, leading to defective forgings that contain shear bands and shear cracks.



ISOTHERMAL HOT-FORGING is a recently developed technique in which the dies are heated to a temperature approximating the temperature of the workpiece in an effort to minimize defects. The technique yielded a uniform forging in a titanium bar (*a*); a computer simulation of the deformation process (*b*) predicted that only at the corners would the rates of deformation be somewhat higher than elsewhere. This forging technique is not always successful, as is indicated by another titanium bar (*c*), in which shear bands appeared. The reason was flow softening. The computer simulation employed to predict localization of flow (*d*) shows large deformation rates and large gradients in them, as was observed in the forging. In the simulations the range of deformation rates is from low (*gray*) through intermediate (*black*) to high (*color*).

With the demands made on forging today the design of preforms and dies is often too complicated to allow the development of simple and general methods for predicting working loads, the flow of metal and the potential for defects. Until recently the solution of these design problems depended on trial-and-error methods in the hands of experienced people. Such methods are both expensive and time-consuming, and they often call for elaborate experimental setups and complex measurements. For these reasons an increasing amount of research is being done to develop mathematical techniques by which computer programs can simulate forging operations in various conditions. That is one of the main objectives of our colleagues in the Cooperative Center for Metal Processing at the Battelle Columbus laboratories. The work is led by Taylan Altan, who pioneered the application of computer models to forging problems.

In the past the mathematical modeling of forging processes was done by various approximate methods that lacked generality and often failed to give accurate estimates of the required forces and energy. Over the past 15 years a better approach, the finite-element method, has been applied successfully to a wide range of forging problems. In this method a cross section of the material to be deformed is divided into a two-dimensional network of discrete elements and the deformation at selected points in each element is determined by the application of some variational principle. Through the analysis of an aggregate of elements, one at a time, the pattern of deformation in a complex shape can be determined. The accuracy of the predictions increases with the number of elements.

Until recently the application of the finite-element method to practical forging problems was limited by the need for a large computer and an experienced user. Now at Battelle a general-purpose computer program oriented toward users has been developed by Soo-Ik Oh. It can solve forging problems for parts with complex shapes and does not require an excessive amount of computer time. Moreover, its simulation of forging provides detailed mappings of the stress, strain and temperature fields in the workpiece and the dies. We have employed the program to predict shear bands in isothermal hot-forging. Harold Gegel of the Air Force Wright Aeronautical Laboratory applied the program to the prediction of secondary tensile stresses and hole formation in the hot-forging of titanium compressor disks for jet engines.

Computer programs are also being used to design dies. The designs take into account the corrections needed because of the thermal shrinkage of metal in hot-forging, the elastic deflection of the tooling and other factors. Computer models can predict the stresses in the critical areas of dies, thereby forewarning the designer of places where a die might break. The models also enable him to predict where deformation may be uneven or holes may develop; he can then change the design of the preform.

With the advent of interactive graphics (the technique that enables one to call for and to change drawings of designs on the cathode-ray-tube display of a computer terminal) the designer can quickly see the results of his ideas. Quite recently the work of designers and diemakers has been made somewhat easier by the introduction of stand-alone computer-aided design and drafting systems. Such a system includes an independent computer and advanced software for geometrical design and display. When the designer has worked out a plan for dies on the screen, the system produces either a complete set of engineering drawings or a magnetic tape that can be sent to the plant for automatic machining of the dies.

Computer programs that design forging equipment and simulate forging operations give engineers a powerful tool. Increasingly computer programs will eliminate the old trial-and-error methods and make possible better designs of preforms and dies. The users of computer models will be able to predict with good accuracy the flow of metal and the potential for defects in forging, enabling them to change designs to avoid such problems and thereby preventing bad parts from being put into service. The hidden costs of trial-and-error design will be eliminated and productivity in the forging industry will be increased.

HYDRAULIC PRESS at the Alcoa plant in Cleveland is one of the two largest forging presses in the U.S. It has a capacity of 445 meganewtons (50,000 tons). Hydraulic pumps drive the press; constant oil pressure is maintained by means of the accumulator tanks at the left.

At just $4,748, the Toyota Corolla Tercel is a good, honest value. And Toyota's lowest-priced car!

But the real front-page news about this front-wheel-drive is how much you get for what you don't pay.

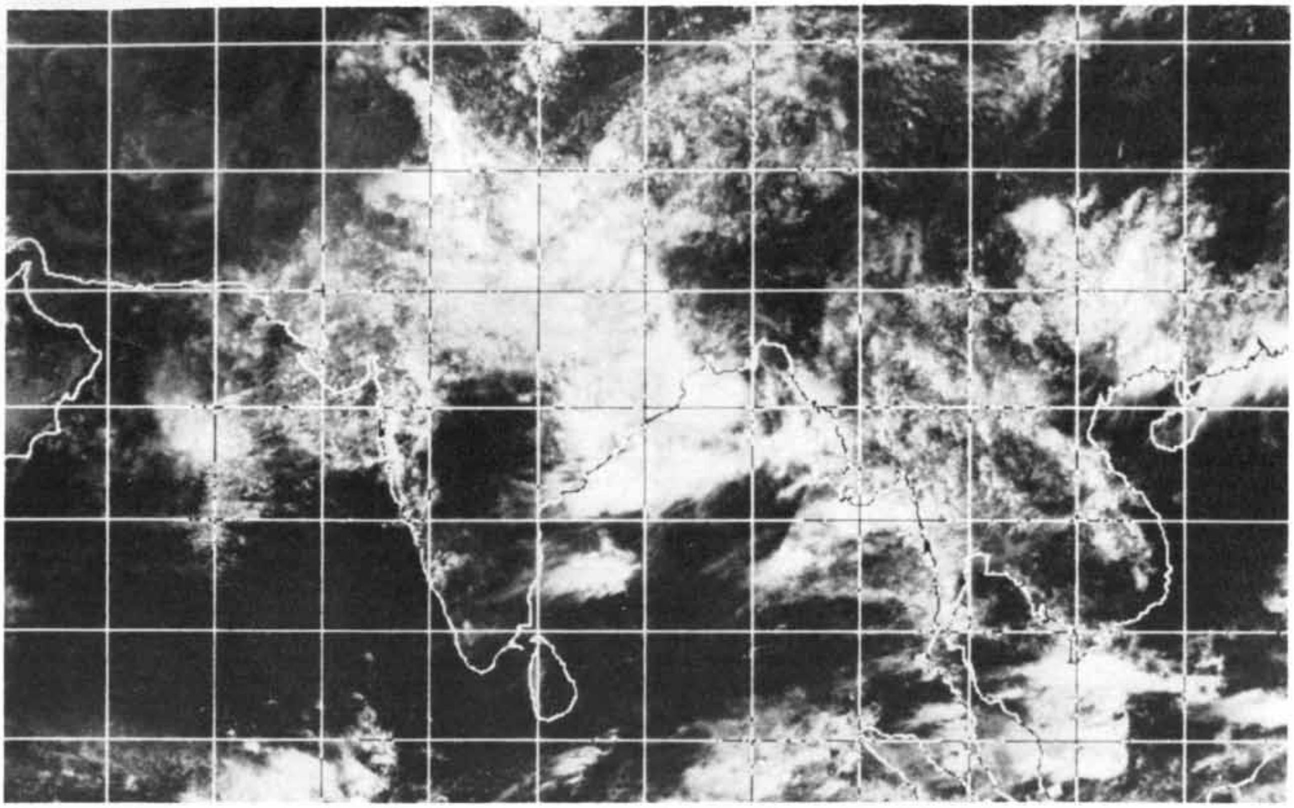
Front-wheel-drive traction, of course. And some very forward-thinking engineering. Like a unique engine mounting that helps make servicing easier. And an astonishing amount of interior head room. Hip room. And up front, leg room that's unsurpassed by any other small front-wheel-drive car!

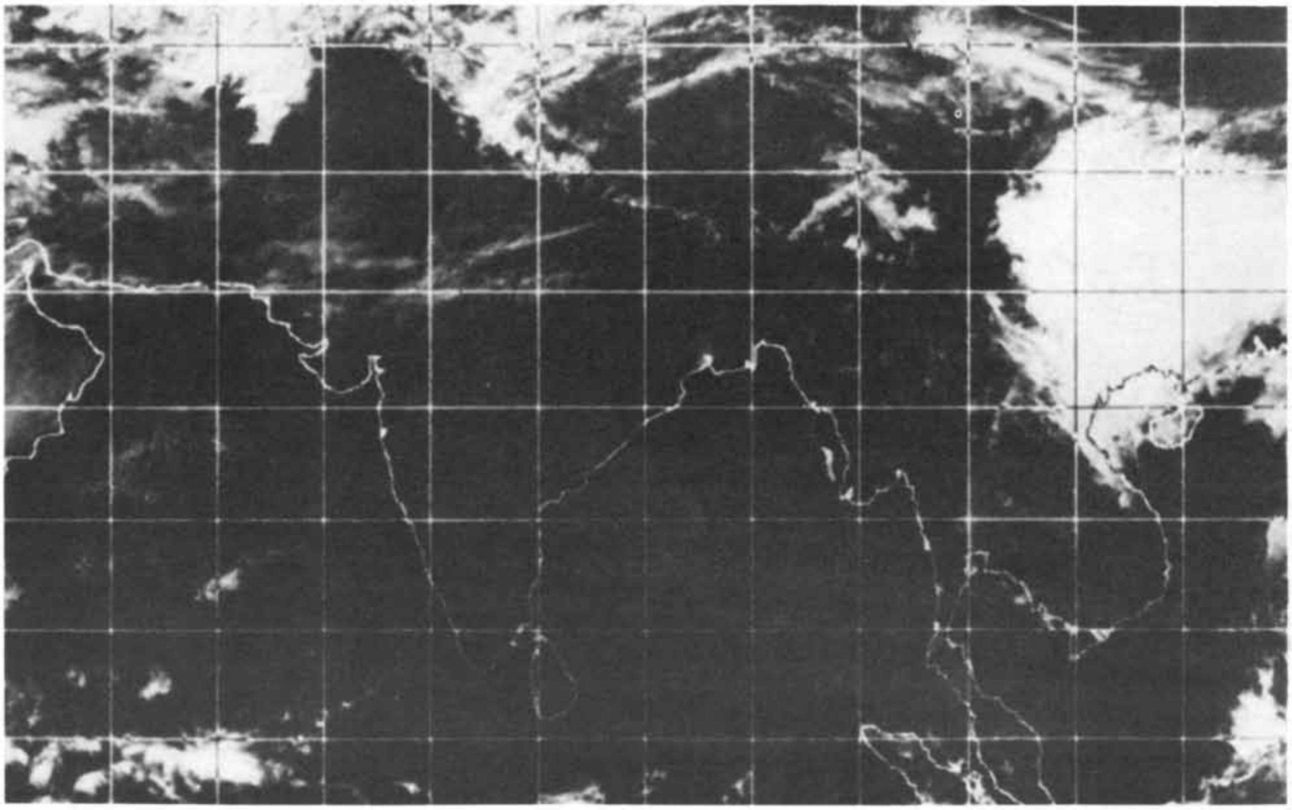
Reclining front bucket seats and power-boosted flo-thru ventilation keep Corolla Tercel passengers out front in comfort.

Finally, the news from the mileage front—Corolla Tercel is rated at 47 EPA Estimated Highway MPG, (36) EPA Estimated MPG. Remember: Compare this estimate to the EPA "Estimated MPG" of other cars with manual transmissions. You may get different mileage depending on how fast you drive, weather conditions and trip length. Actual highway mileage will probably be less than the EPA "Highway Estimate."

The front-wheel-drive Corolla Tercel. For good old-fashioned value per dollar, it's a front runner!

THE FRONT-WHEEL-DRIVE COROLLA TERCEL. AT $4748,* IT'S WAY OUT FRONT.





SUMMER MONSOON CIRCULATION over India and Southeast Asia transfers moist air from the equatorial ocean to the land (*top*). The circulation is driven by differences in atmospheric pressure between the warm air over the land and the cool air over the ocean. The intensity of the circulation is enhanced when water vapor carried by the air condenses and releases energy. In the winter or dry season the circulation is reversed and the cloud cover over the land is minimized (*bottom*). Monsoons are not unique to the Indian Ocean; they develop wherever there is a seasonal wind reversal caused by differential heating of the atmosphere. The two photographs are composite images generated by a computer from data collected by the *NOAA-5* satellite on August 3, 1977, and on December 1, 1977.

Monsoons

By concentrating the solar energy reaching the ocean onto the land these seasonal winds bring water to half of the people of the earth. Computer simulations may soon predict their dry and rainy phases

by Peter J. Webster

If the earth were a simpler planet, capped over most of its Northern Hemisphere with a single continent and covered elsewhere with one vast ocean, the weather patterns near the coast of the continent might not be much different from those that regulate life on three of the continents of the real earth. There would be two major seasons on the coastal plain, a rainy season and a dry one. Within the rainy season periods of drenching squalls would alternate with equal periods of sunny weather every week or two. The inhabitants of the coastal areas and the southern interior of the mythical continent would become accustomed to dependable cycles of seasonal change.

In at least one important way life on the mythical planet would be quite different from life on the real earth: on the mythical planet it would be possible to forecast major weather patterns accurately. Indeed, the mythical planet is a simplified model of the earth that has been developed for the simulation of global weather patterns with the aid of a computer. On the model planet one can predict the onset of the rainy season, the alternation of dormant (dry) and active (rainy) weather within the rainy season and the approximate date of the cessation of the rains at the beginning of the dry season. A farmer on the model planet with access to such information could time his plantings and choose his crops to ensure adequate precipitation and maximize his chance of a successful harvest. For the two billion people of the real earth who depend on seasonal rains for drinking water as well as for agriculture such accurate forecasting would have a profound effect on everyday life.

The terrestrial weather patterns simulated by the model planet are monsoons. The term monsoon has most often been applied to seasonal changes on the shores of the Indian Ocean, and in particular to a wind system in the Arabian Sea that blows from the southwest during one half of the year and from the northeast during the other half. The term is thought to stem from the Arabic word *mausim,* meaning season. As the underlying mechanisms of monsoons have come to be understood, the term has come to signify any annual climatic cycle with seasonal wind reversals that generally cause wet summers and dry winters. The largest and most vigorous monsoons, however, are found in the regions of the earth where they were first named: on the continents of Asia, Australia and Africa and in the adjacent seas and oceans.

Although the definitive characteristic of the monsoon is a seasonal pattern, fluctuations are observed on time scales ranging from days to decades. The short-term variations include not only the active and dormant phases in the rainy season but also individual disturbances in the active phase. During an active phase the weather is unstable, with frequent storms that carry the deluges often associated with the monsoon. During a dormant phase the weather is dry, hot and stable, and notable for an absence of tropical storms. Over much longer periods there are variations in annual precipitation that can lead to years of drought or flood. The superannual cycles are still too poorly understood for forecasting to be practical, although years of flood or drought can be expected about 30 times per century. Recent developments in the theory of moist processes in the atmosphere, however, may soon make predictions of the active and dormant phases feasible.

Practical knowledge and the general predictability of monsoon phenomena played an important social and economic role in many ancient civilizations of the Eastern Hemisphere. Long before the arrival of Europeans, merchants had plied trade routes between ports in Asia and eastern Africa, adapting their commerce to the seasonal rhythms. In 1498 an Arab pilot showed the Portuguese explorer Vasco da Gama the trade route to India from the east coast of Africa, and the monsoon winds became the basis of a lucrative trade and cultural exchange between East and West. The European traders and adventurers returned home with fragmentary information about the southwesterly winds of summer and the northeasterly winds of winter.

Provided with such observations of low-latitude weather, European scholars were able for the first time to consider the circulation of the atmosphere on a global scale. Two of the most important early studies were done in the late 17th and early 18th centuries by Edmund Halley and George Hadley. Halley attributed the monsoon circulation primarily to the differential heating and cooling of the land and the ocean. Differential heating, he reasoned, would cause pressure differences in the atmosphere that would be equalized by winds. Hadley noted that the rotation of the earth would change the direction of such winds, causing winds moving toward the Equator to veer to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. Although more recent work has refined the understanding of the two processes, they are still considered fundamental causes of monsoon phenomena.

There is a third factor, however, that determines many of the distinctive features of the monsoons. The range of temperature and pressure over much of the surface of the earth lies near to the triple point of water. The triple point of a substance is the combination of temperature and pressure at which its solid, liquid and gas phases can coexist. For water the triple point lies at a temperature of .01 degree Celsius and a pressure of 6.104 millibars. Water molecules whose environment is near to the triple point can be freely interconverted among all three states. In contrast, carbon dioxide molecules cannot enter the liquid state at the temperatures and pressures ordinarily encountered at the earth's surface; the pressure must be raised considerably before solid carbon dioxide (dry ice) will melt instead of sublimating into its vapor phase.

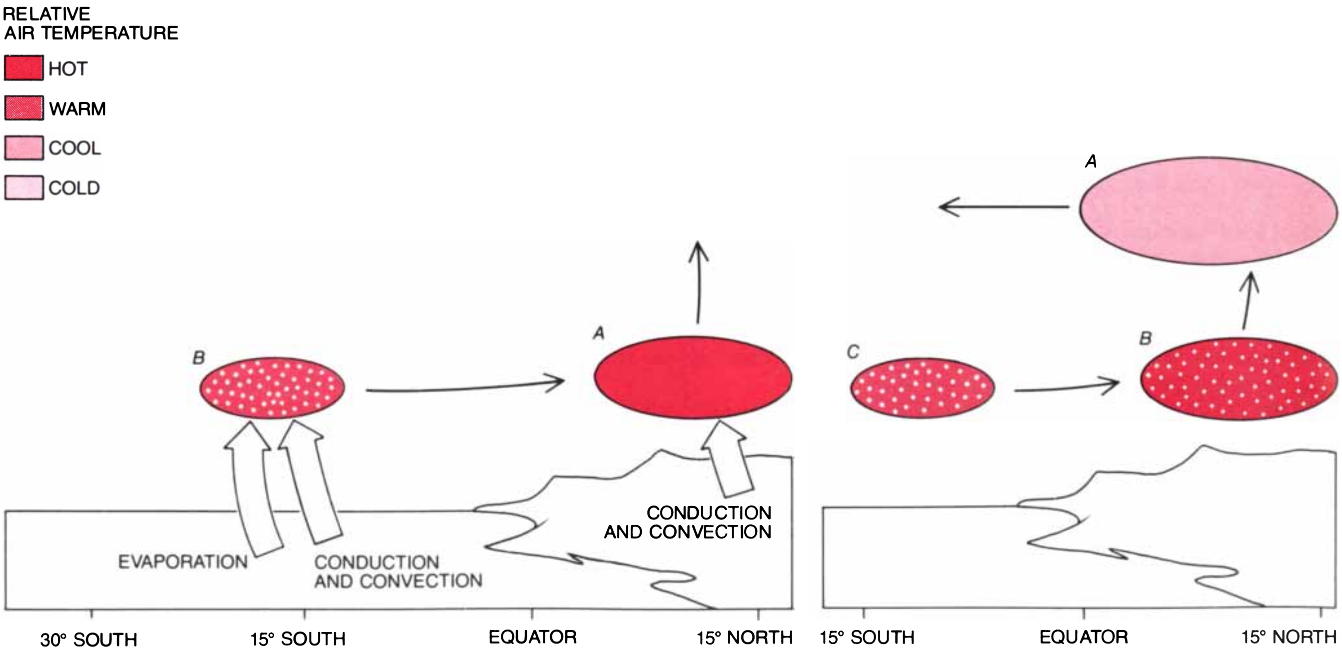
The ability of water to readily evaporate and condense in the atmosphere has a profound effect on monsoon circulation. It is primarily in understanding the effects of moist processes that modern meteorology has made its major theoretical contribution. When water changes from a solid to a liquid, energy must be supplied to break down the crystalline structure of ice so that the molecules can move about more freely in the liquid state. Similarly, energy is required in order to transform the phase from the liquid to the vapor. The energy expended to evaporate water is stored as kinetic energy of the molecules of the water vapor; when the molecules condense again, the energy is released. Thus during a phase change energy in the form of heat is added to or subtracted from a substance without changing its temperature. One can appreciate the effect by noting that although the ice in a glass of ice water will melt in a warm room, the temperature of the water does not change as long as some ice remains.

The importance of moist processes stems from the fact that the water evaporated at any given time from the world's oceans stores about a sixth of the solar energy reaching the surface of the earth. When the water condenses again and falls as rain, the energy stored in the vapor phase of the rainwater is released. In monsoon circulation part of the enormous reservoir of solar energy collected over the oceans can be released over land when the water in moist ocean air condenses over the land. It is the release of this energy that is responsible for the power and the duration of the monsoon rainy season and for the variation within the rainy season between active and dormant phases.

To understand more precisely how moist processes function in monsoon circulation it is necessary to appreciate the workings of the basic driving mechanism of the circulation, which was first discussed by Halley in 1686. The atmosphere is heated differentially because land areas tend to be warmer than the oceans in the summer but cooler in the winter. The land and the oceans respond differently to solar radiation for two reasons. One of the intrinsic properties of water is its high capacity to store heat compared with the capacity of many other substances. The specific heat of a substance is a measure of its heat capacity: the specific heat is the amount of energy that must be supplied to raise the temperature of one gram of the substance by one degree C. The specific heat of water is more than twice that of dry land, although the specific heat of land is considerably increased when the land is wet with rain. Therefore in response to the same amount of solar radiation the temperature of a given mass of dry land will increase more than twice as much as the same mass of the ocean.

The second reason for the greater heat capacity of the ocean is its efficiency in mixing heat energy to lower depths and so distributing heat throughout a large mass of water. Wind stirring at the surface of the ocean creates turbulent eddies that have the effect of conveying warm surface water to lower levels in the summer; the warm water is replaced by cool subsurface water that is heated in turn. In the winter the heat accumulated during the summer is released by the reverse process. As the surface of the water cools because of the decrease in solar radiation, the surface water sinks and is replaced by warmer water that rises from below.

Because of mixing and the high specific heat of water the temperature of the ocean surface varies less than that of the land. The oceans act as an enormous flywheel for storing heat energy, and because of the great inertia of the system the cycle of maximum and minimum surface temperatures lags about two months behind the corresponding cycle of solar heating. In the spring—the beginning of the annual cycle of the monsoon—heat energy reaching the surface of the ocean or of the land is conducted upward into the atmosphere in warm, buoyant and turbulent bubbles of air. The rate of heat transport is proportional to the temperature difference between the ground and the atmosphere. As the bubbles rise they mix with cooler air and transfer their heat to the column of air above the heated surface. This form of heating and heat transfer is called sensible heating because the heated substance must be in contact with the heat source. Sensible heating causes the initial differential heating of the atmo-



STAGES IN THE DEVELOPMENT of a summer monsoon are controlled by the interaction of moist processes with the force that drives higher-density air toward regions of lower density. As solar radiation warms the land and the ocean the overlying air is heated by conduction and expands. Because the land heats faster than the ocean the warmer air over the land rises as buoyant, turbulent bubbles and is replaced by denser ocean air. The latter carries moisture evaporated from the ocean, which stores solar energy in the form of latent heat sphere over the land and the ocean and generates the potential energy that powers the monsoon system.

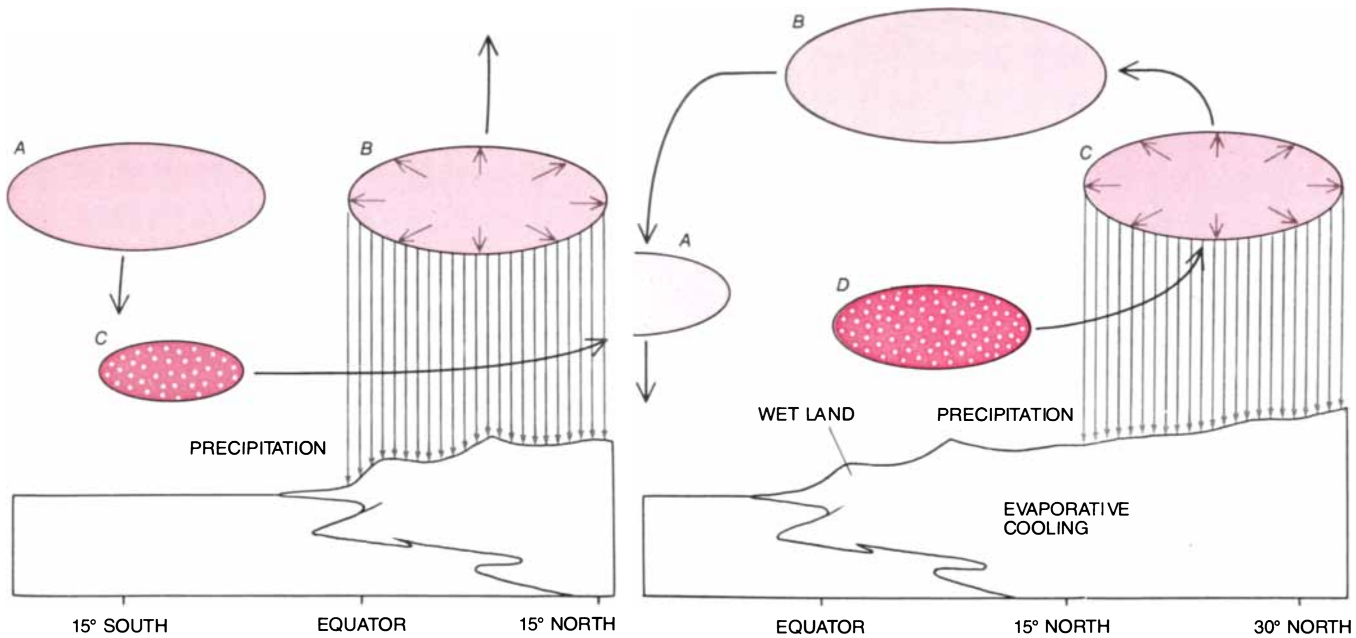
Monsoon winds are driven by the conversion of part of the potential energy of an atmospheric system into kinetic energy. The potential energy of a system under the action of gravity is proportional to the vertical distance between its center of mass and some convenient reference level, such as the surface of the earth. The potential energy can be increased by raising the center of mass of the system. This can be done either by tapping the kinetic energy of the system and thereby depleting its motion or by supplying energy from some external source. On the other hand, if the center of mass is lowered, the potential energy decreases and a corresponding amount of kinetic energy becomes available for fluid motion.

During the summer monsoon differential heating raises the potential energy of the ocean-land system by setting up a difference in pressure between parcels of air over the two regions. Because the air initially over the ocean is cool it remains denser than the air over the land. The force generated by the pressure gradient, which tends to equalize pressure differences, causes the denser, cooler air from the ocean to move toward the land and undercut the warm air over the land. The warm air is therefore forced to rise. The combined rising of the warm air and sinking of the cool air lowers the center of mass of the atmospheric system, so that the air motions release potential energy. At the same time the steady input of solar energy tends to raise the potential energy because of the continued differential heating of the land and the ocean. The winds of the monsoon are the result of the tendency of the atmosphere to minimize the pressure gradient between the land and the ocean. They can therefore be understood from an energy standpoint as the result of the conversion of solar energy into potential energy and then from potential energy into kinetic energy.

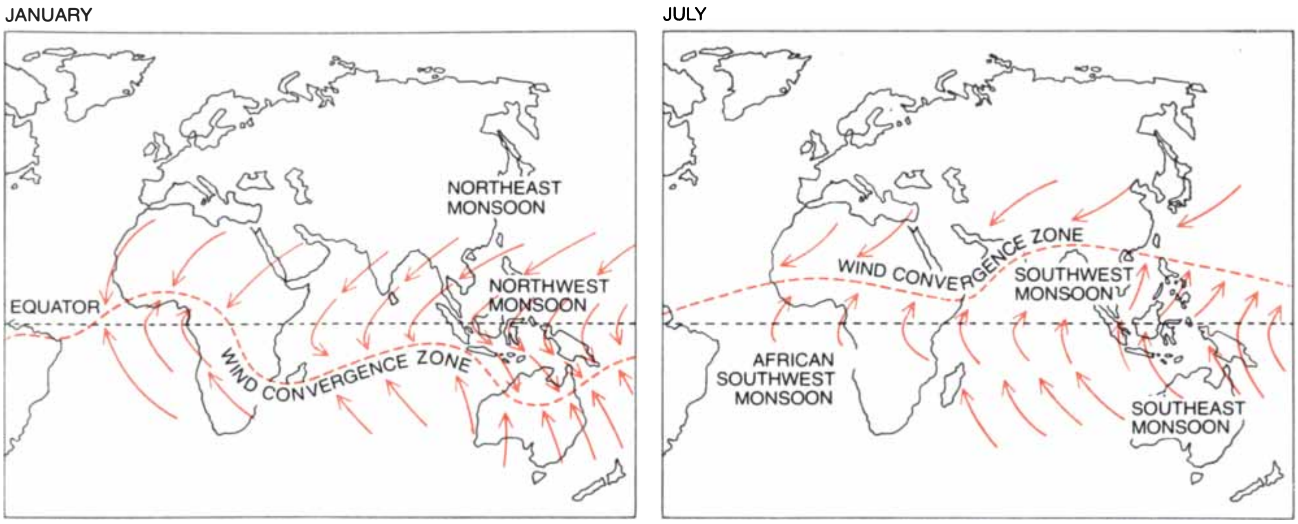
The circulation of the monsoon winds is deflected by the rotation of the earth through the action of the noninertial force known as the Coriolis force. The deflection distinguishes monsoon winds from diurnal sea breezes, which also stem from differential heating. The latter arise and diminish too quickly to be much affected by the Coriolis force. Hadley's description of the Coriolis force is sufficient for winds moving from the poles of the earth to the Equator, but the general effect can be succinctly described for winds moving in any direction. The Coriolis force deflects winds to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. The magnitude of the deflection depends on the latitude of the wind motion: the deflection is maximum at the poles and zero at the Equator. It is directly proportional to the trigonometric sine of the latitude.

The summer monsoon will ideally continue in a steady state until the balance between the potential energy generated by solar radiation and the release of the potential energy by the atmospheric system is upset. In Asia, for example, solar heating diminishes substantially after the autumnal equinox and the temperature of the adjacent oceans starts to fall. Simultaneously regions in the Southern Hemisphere, particularly over the Indonesian archipelago, become the centers of maximum heating. As the temperature difference between the Asian land masses and the surrounding oceans decreases, the potential energy of the system runs down. The monsoon is said to retreat and the winter dry season begins in the Northern Hemisphere.

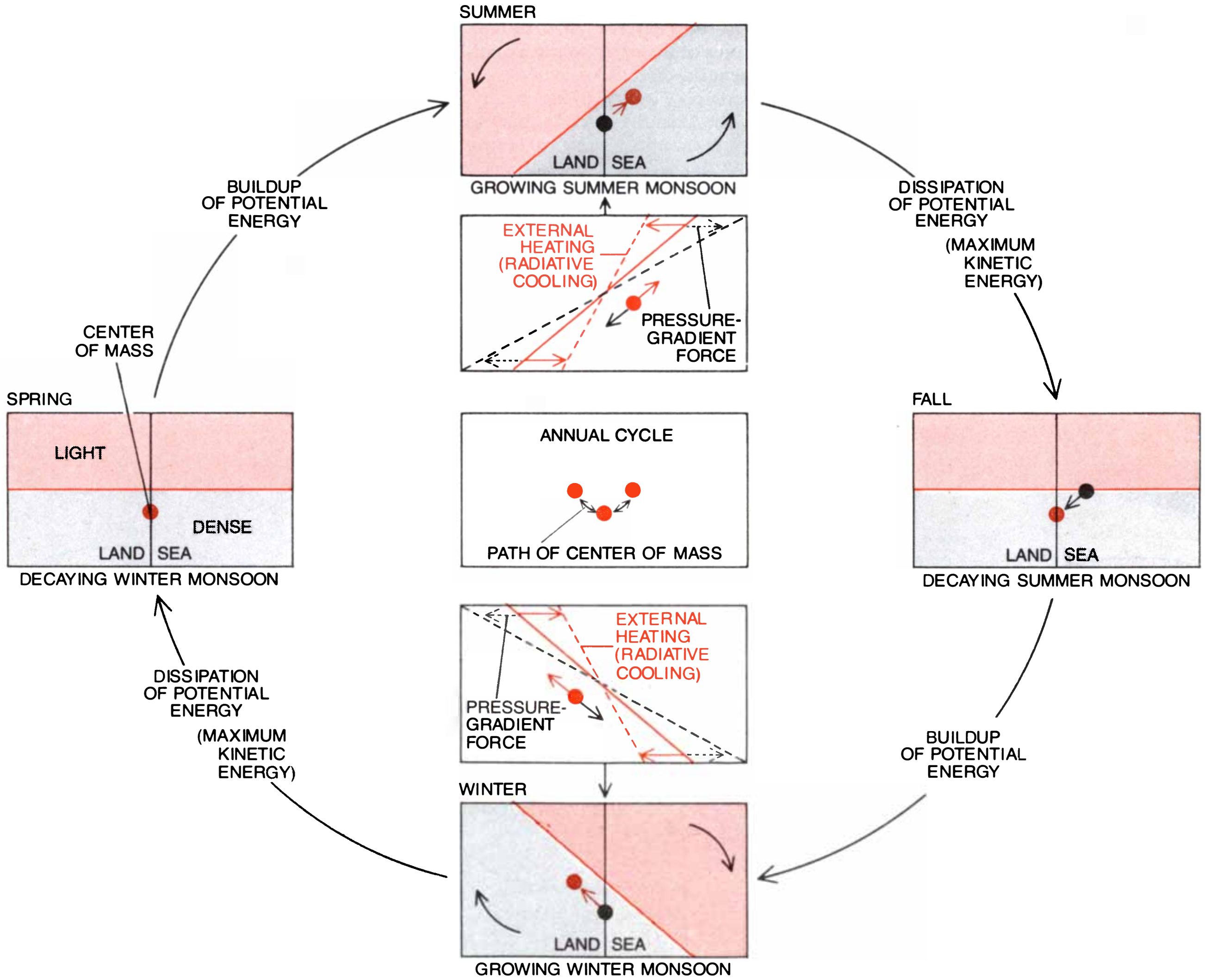
With the onset of winter both the land and the oceans of the Northern Hemisphere lose heat by radiation to space. The radiative losses are attenuated by clouds, but the cloud cover is usually less extensive over the land than it is over the ocean. The resulting increased heat loss from the land and the greater heat capacity of the ocean restore the temperature differences to the system and raise the potential energy once again. Cold, high-pressure air over northern Asia moves toward the Equator in order to restore equilibrium and is deflected to the right by the Coriolis force. The cold air mass, moving from the northeast along the surface, is balanced by warm air from the south that moves northward in the upper troposphere. The rotation of the earth deflects



(*colored stipple*). When the moist air moves inland, it too rises and its water vapor condenses, releasing latent heat. The additional heat causes the air to expand and rise further, reducing the pressure in its wake and so intensifying the monsoon airflow. Rainfall cools the land because evaporation of the water absorbs a portion of the incoming solar radiation. Hence the region of maximum ground heating moves inland and the region of maximum precipitation follows. Air pressure from dense to light is shown by relative sizes of the air parcels.



SEASONAL WIND REVERSAL is characteristic of monsoon circulation throughout the world, but it is most pronounced in the regions around the Indian Ocean. Note that in crossing the Equator the winds change direction, an effect of the rotation of the earth. The zone where the surface winds converge is primarily in the Southern Hemisphere in January but migrates northward with the sun in July.



TEMPERATURE DIFFERENCE between the air over the land and the air over the ocean raises the center of mass of the atmospheric system and so increases its potential energy. The temperature difference is maintained by external heating in the summer and by radiative cooling in the winter. The flow of air caused by the pressure gradient between the two air masses tends to decrease the potential energy of the system and to increase its kinetic energy. Monsoon circulation results from the balance between the two effects. When the monsoon is in equilibrium (summer and winter), the gain of potential energy by external heating or radiative cooling exactly equals the loss caused by the action of the pressure-gradient force. In fall and spring the center of mass falls and the potential energy of the system is dissipated.

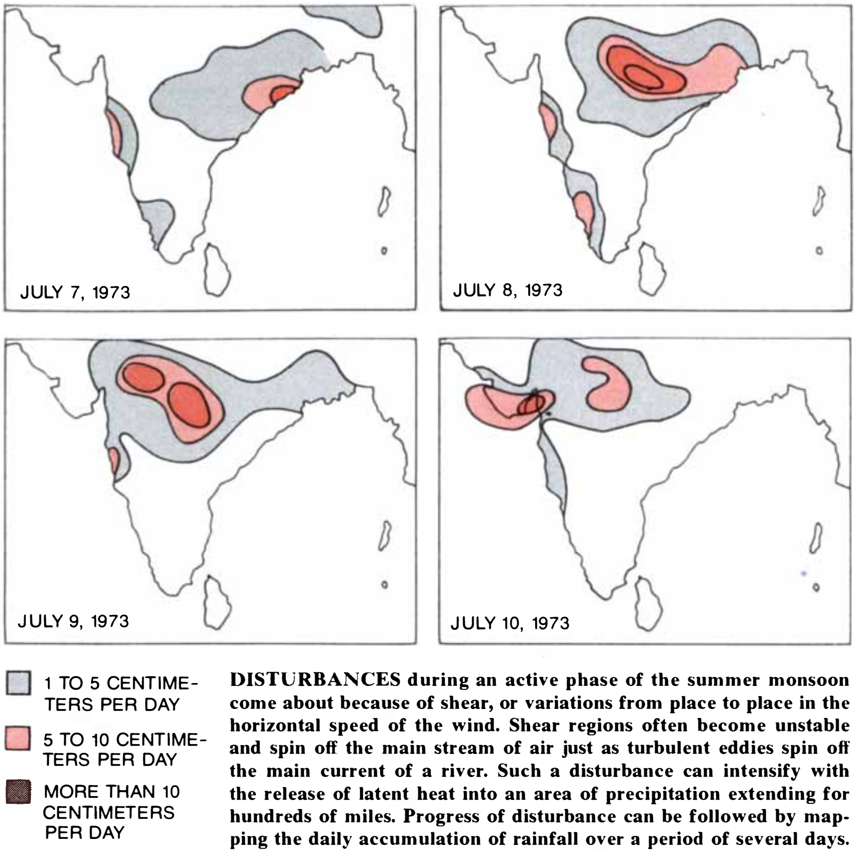
the warm upper air to the east, creating an intense jet stream over Asia and Japan in which speeds commonly reach 100 meters per second, or more than 200 miles per hour. The jet-stream flow often becomes unstable over the central north Pacific and spawns the low-pressure winter storms prevalent in the middle and high latitudes of the Western Hemisphere.

The northeast monsoon of the winter continues in a steady state much like the southwest monsoon of the summer, until solar heating in the spring dissipates the potential energy that powers the winter monsoon. As the temperature of the land again overtakes the surface temperature of the ocean, the potential energy builds up and the cycle begins once more.

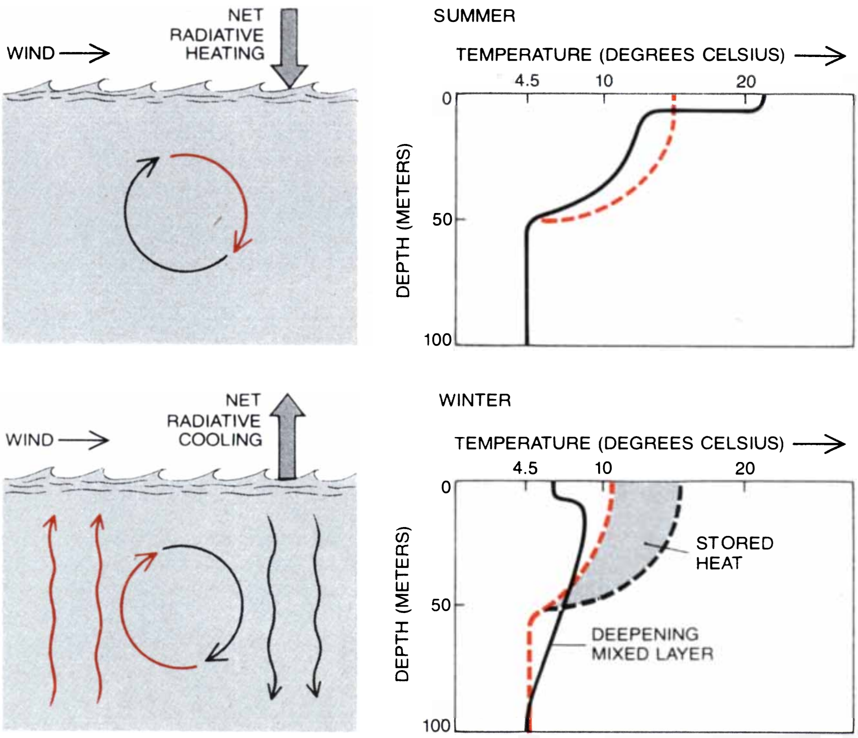
What role do moist processes play in the annual circulation? During the summer monsoon water vapor evaporated from the ocean is borne along with the air moving toward the land. If a parcel of air is displaced vertically in such a way that heat energy neither leaves it nor enters it, the temperature and pressure of the parcel undergo what is called adiabatic change. If the parcel rises, it moves into a region of lower pressure. A pressure gradient then exists between the air parcel and its new environment; in order to equalize the pressures the parcel tends to expand. To expand, however, the parcel must do work on its environment at the expense of the kinetic energy of its molecules. Reducing the kinetic energy of the molecules entails reducing the temperature. If there is no heat exchange across the boundaries of the parcel, the process is called adiabatic cooling. Conversely, in adiabatic heating the temperature of a subsiding air parcel increases as the parcel is compressed by its environment, even though no heat energy is added to (or subtracted from) the parcel.

An air parcel carrying moisture from the sea is warmed by conduction and by upward-moving convective air currents over the land, and it begins to rise to higher altitudes of lower pressure. As the parcel ascends it cools adiabatically and the water vapor condenses into raindrops. In the course of condensation the solar energy that has maintained the water in the vapor phase is released. The released heat energy, called the latent heat, is taken up by the air molecules and thereby causes a nonadiabatic temperature change in the air parcel.

The heat liberated in this way adds considerable buoyancy to the rising column of air over a warm continent. The air parcel rises still higher, further reducing the pressure over the land and bringing on a more vigorous influx of moist air from the ocean. Hence one effect of moist processes in the atmosphere is to increase the strength of



DISTURBANCES during an active phase of the summer monsoon come about because of shear, or variations from place to place in the horizontal speed of the wind. Shear regions often become unstable and spin off the main stream of air just as turbulent eddies spin off the main current of a river. Such a disturbance can intensify with the release of latent heat into an area of precipitation extending for hundreds of miles. Progress of disturbance can be followed by mapping the daily accumulation of rainfall over a period of several days.



SOLAR HEAT reaching the surface of the ocean during the summer is distributed throughout a layer about 50 meters thick as a result of turbulence induced by the wind. In the winter the cycle is reversed and the cool surface water is driven down by the wind and by its own greater density. It is replaced by the warmer, less dense water that accumulated below the surface the preceding summer. Broken lines represent the average distribution of heat energy as a function of depth. Mixing is the chief reason the ocean temperature lags behind the solar cycle.

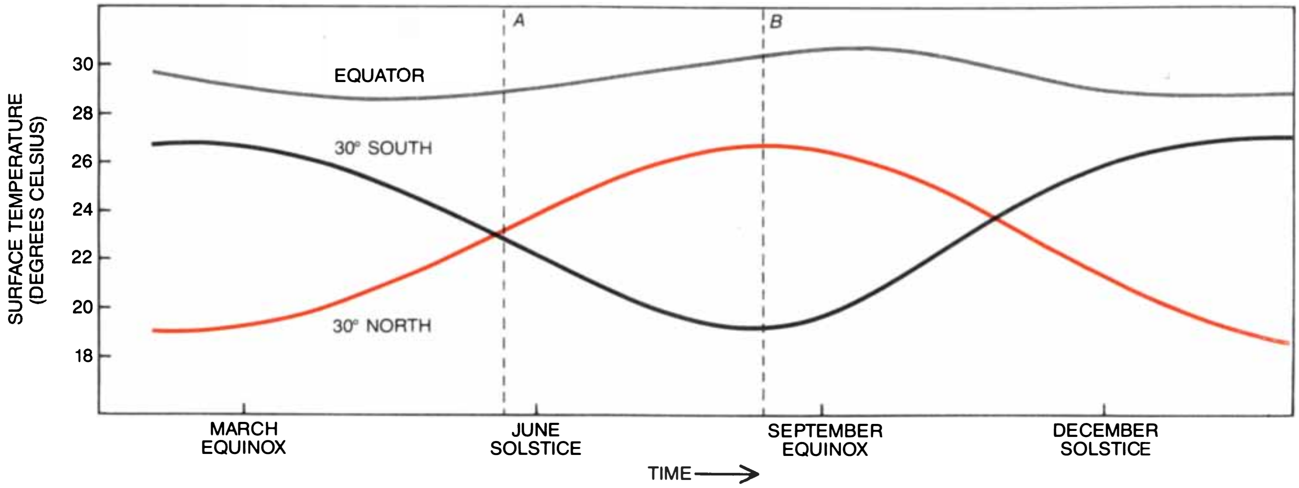
the monsoon circulation. Without moist processes the circulation would still exist, but it would be weaker.

Because of the rising column of air, temperatures in the upper troposphere over southern Asia are much warmer in the subtropics than they are over the Equator. The result is a pressure-gradient force toward the Equator at high levels, the opposite of the force near the surface. The rotation of the earth deflects the upper-level flow westward into an intense east-to-west jet stream in which winds may reach speeds of 50 meters per second, or more than 100 miles per hour. The jet stream extends across the Indian Ocean and Africa, where it crosses the Equator and merges with the winter westerlies of the Southern Hemisphere. The winds finally subside over the subtropical high-pressure belt formed during the winter in the Southern Hemisphere. The general features of the upper-air return flow were predicted by Halley solely on the basis of rational deduction.

The influence of moist processes is most apparent in the effects of such processes on the timing of the events that make up the seasonal monsoon. Moist processes virtually define the time of its onset; they appear to determine its maximum intensity, and they control its retreat. The onset of the period of precipitation comes rather abruptly a few weeks before the summer solstice. The circulation does not reach maximum intensity, however, until from eight to 10 weeks after the solstice.

The reason for the delay is that the average precipitation on a land mass is directly related to the temperature of the ocean upwind from the land. The higher the surface temperature of the ocean is, the higher the temperature of the air over the surface is and the more water vapor the air can carry. When air from the ocean carries more water vapor, more energy is available for release when the water condenses over the land. Hence the intensity of the monsoon circulation increases. Because the oceans in the summer hemisphere reach their maximum temperature some two months after the summer solstice the vapor content of the monsoon winds does not reach its maximum value until middle or late August in the Northern Hemisphere and late February or early March in the Southern Hemisphere.

The retreat of a monsoon is the gradual cessation of precipitation sometime after the autumnal equinox. Not only is the differential heating between the oceans and the continents reduced, but so also is the energy pumped into the system by the transport of water vapor. The cooler air over the ocean holds less water vapor, so that the latent heat released by precipitation over the land gradually decreases.

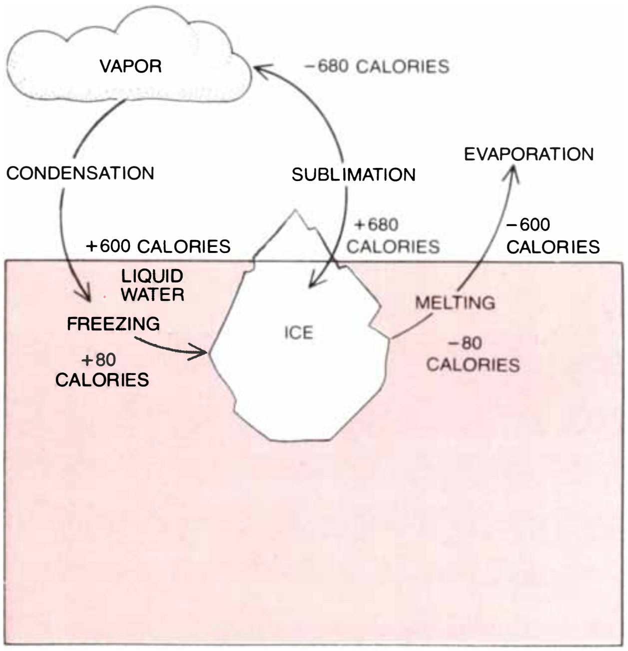


MAXIMUM AND MINIMUM surface temperatures of the ocean do not coincide with the extremes of the solar cycle. Instead they lag behind the sun by from eight to 10 weeks. Because the potential energy of the monsoon system is roughly determined by the difference in temperature between the Northern and the Southern hemispheres, the potential energy reaches a maximum about two months after the solstices and a minimum about two months after the equinoxes. The maximum ocean temperature also determines the maximum rate of evaporation, so that at the time of greatest potential energy moist processes are also most pronounced. Monsoons are most intense then.

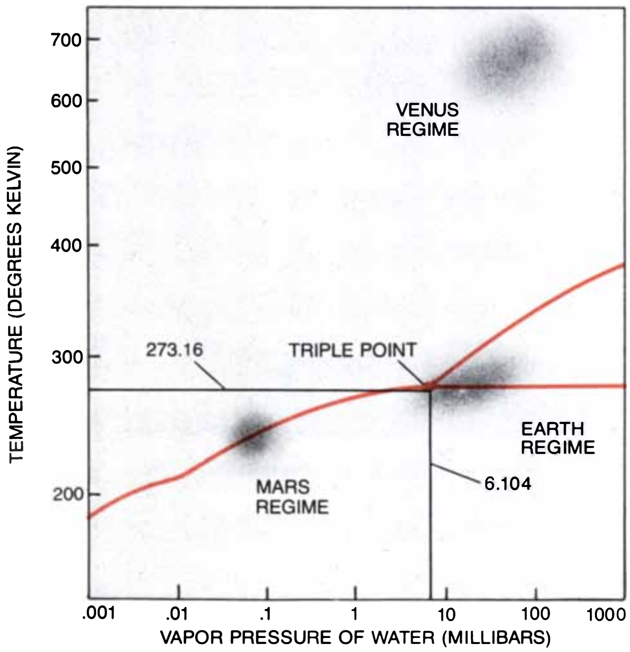
It is now possible to outline the theory of moist processes on which the computer model of monsoon circulation is based. Recall that the model planet is covered by an ocean except for a single continent that symmetrically caps the Northern Hemisphere. The coastline of the continent coincides with the parallel of latitude at 14 degrees north. In other major respects the model planet is identical with the earth: it lies the same distance from the sun, it rotates about the same inclined axis with the same period and its atmosphere is physically and chemically like that of the earth.

In the model it is assumed that the atmosphere responds to a nonadiabatic temperature change in a parcel of air by restoring the temperature of the parcel adiabatically to its original value. The assumption closely matches the behavior of the real tropical atmosphere. For example, when heat is added nonadiabatically to a parcel of air by conduction or as a result of the condensation of water vapor, the air rises to a region of lower pressure and cools adiabatically. On the other hand, if the parcel loses heat through radiation (a nonadiabatic process), the air sinks and contracts adiabatically until its temperature returns to its former value.

To simulate more realistically the



EXTERNAL SOURCE OF HEAT is required in order to change the phase of water from solid to liquid or from liquid to vapor, but because energy is conserved the heat energy is stored as latent heat in the higher-temperature phase. In a phase transition from vapor to liquid, from vapor to solid or from liquid to solid the stored energy is released. The temperature of a substance remains constant during a change of phase even though heat energy is exchanged in the process.

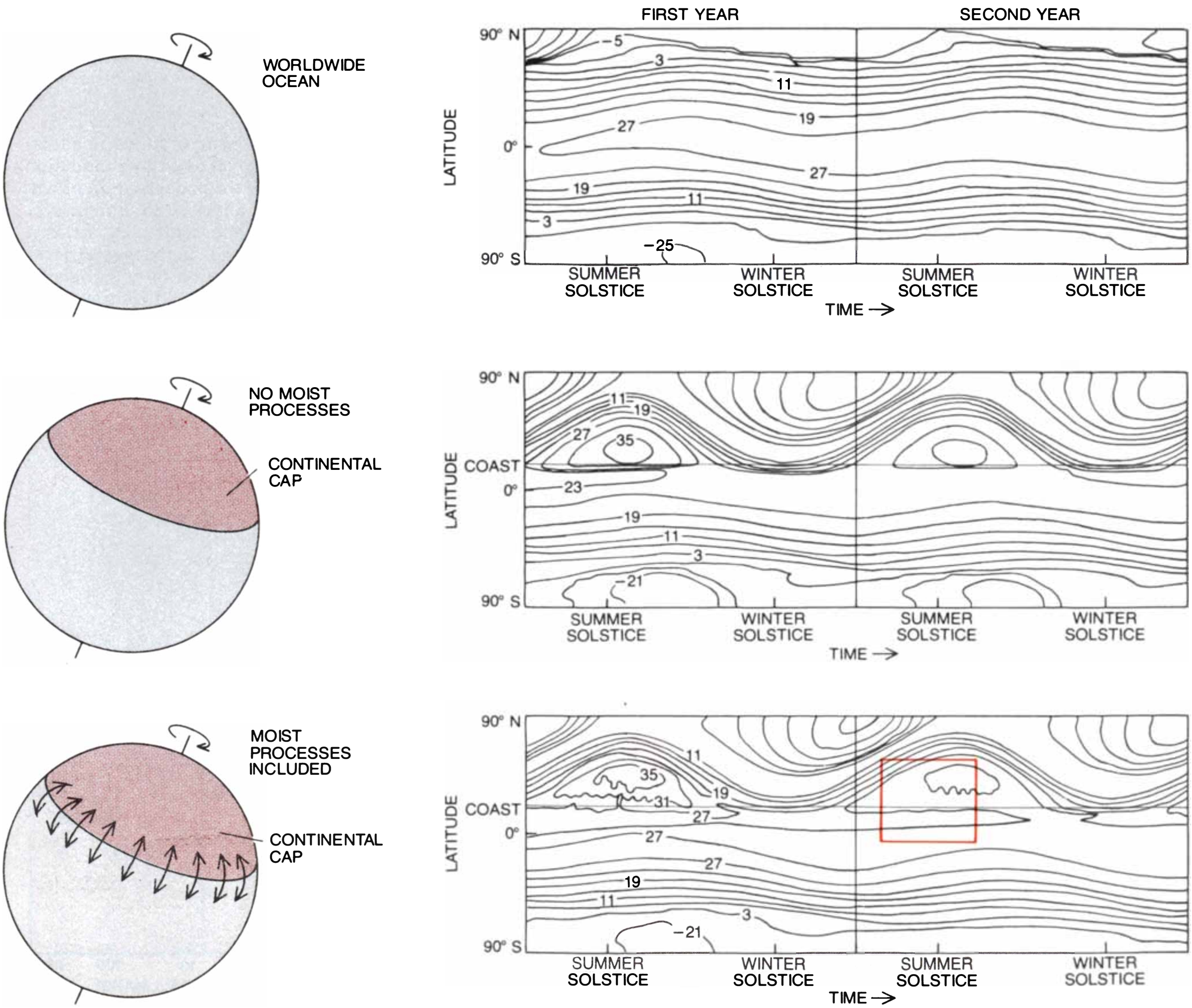


TRIPLE POINT OF WATER is the unique combination of temperature and pressure at which the solid, liquid and vapor phases of water can coexist. Of all the planets only the earth maintains conditions close to the triple point. Water whose environment is near to the triple point can be freely interconverted among the three phases. For this reason solar energy can be collected and stored in a higher-energy phase, then transported and released over a relatively small region.

tendency of the atmosphere to compensate for nonadiabatic heat input one must take into account a complicated nonlinear system. Vertical air motion over a heated continent is a result of interdependent forms of nonadiabatic heating, chiefly sensible heating and latent heating, as well as the adiabatic response. When precipitation begins over land, the ground is moistened and part of the solar radiation previously engaged in heating the surface is diverted into evaporation of the soil moisture. The land cools and sensible heating of the air directly above the wet land is reduced.

The reduction of sensible heat does not significantly alter the driving mechanism of the monsoon, because the sensible heat is only about one-tenth the magnitude of the latent heat released in precipitation. The reduction does have the effect of shifting the maximum of the total heat (sensible heat plus latent heat) toward the interior of the land mass. The region where the vertical velocity of the circulation is at its maximum follows the shift of the maximum heating, and so the condensation of moisture in the air moves inland. The monsoon cell leaves in its wake a land surface saturated with water and a relatively dry subsiding air mass. The accumulated ground water slowly evaporates and the temperature of the drying land begins to rise. As sensible heating increases over the land near the coast a low-pressure area develops again and the moist winds from the ocean are deflected from their inland course. A second monsoon cell overwhelms the inland circulation and precipitation again falls near the coast. The cycle is then repeated.

In the model the vertical wind velocity over an area of the continent can show the alternation of active and dormant phases of the monsoon. The ascending regions represent the disturbed, or active, phase. The wake, where subsidence dominates and storms are suppressed, is the dormant phase. If one examines the distribution of temperature and vertical wind velocity for several



COMPUTER SIMULATIONS of monsoon processes on three simplified models of the earth show how surface temperature (*contour lines*) varies with time and with latitude. On a model of the earth covered entirely by an ocean the surface temperature changes quite slowly with time. When the model planet is capped with a single continent and moist processes are not included in the simulation, the ocean temperature lags many weeks behind the temperature on land. The maximum land temperature lies just north of the coast. When moist processes are simulated, the land is cooled just north of the coast and the surface temperature varies cyclically during the summer monsoon. Such terrestrial features as the distance from the sun, the inclination of the earth's axis and the composition of the atmosphere are assumed to be the same in the model as they are on the real earth. The simulation can also show how vertical wind velocity can vary with time and latitude (not shown). In the atmosphere the vertical wind velocity is about a thousandth of the horizontal wind velocity.

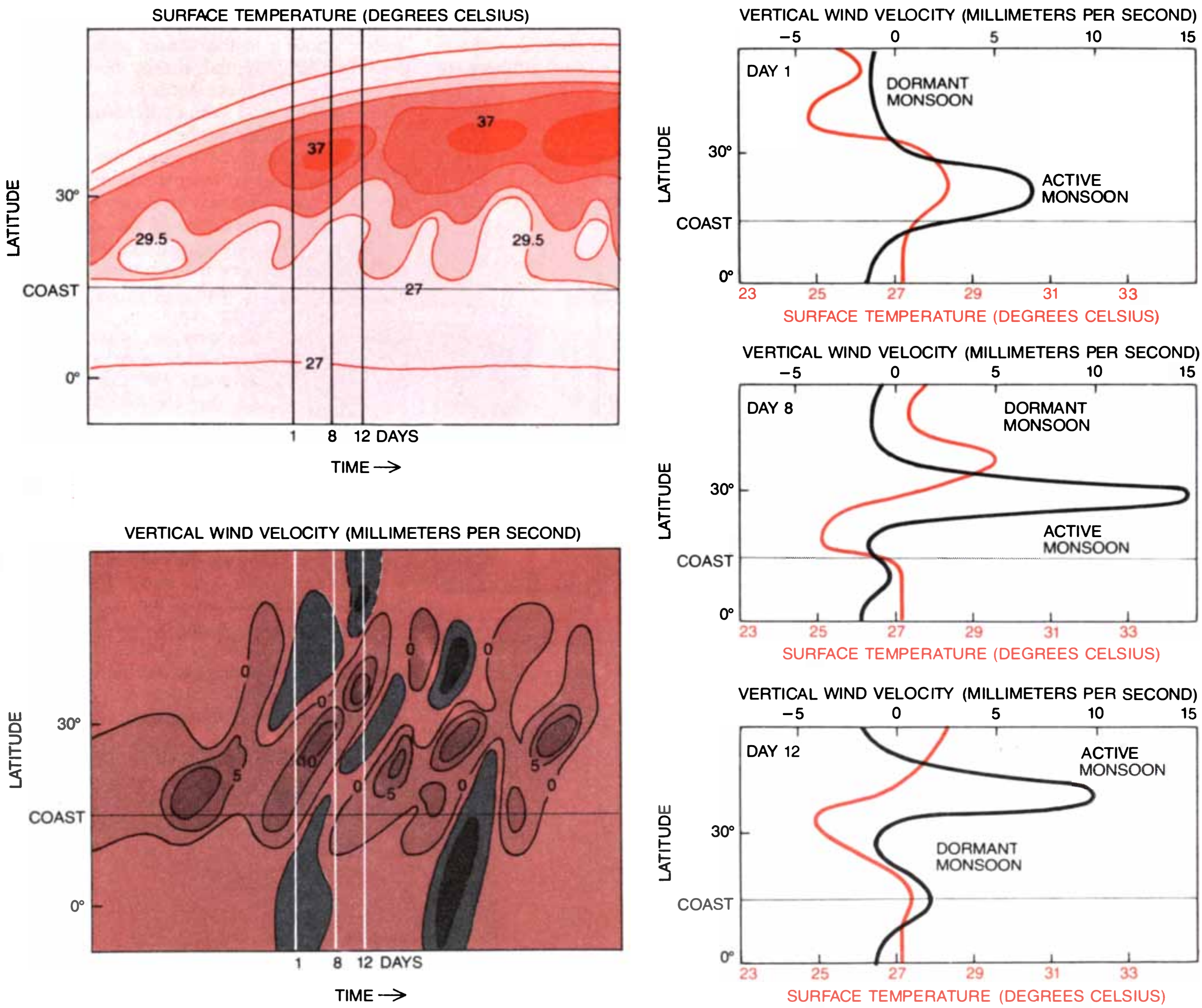
days, one can see how an active phase of the monsoon moves inland from the coast and is followed by a dormant period [*see illustration below*].

Satellite observations provide some justification for the theory of the active and the dormant phases of the monsoon that is incorporated in the mathematical model. The locus of maximum cloudiness over the Indian Ocean can be seen to move gradually northward as a monsoon cell progresses. Associated with the band of clouds are the active monsoon disturbances and intense precipitation. In the cloudless region behind the band and over most of India is a break in the active phase that heralds a period of high temperature but no rain. The cycle is observed to have a period of from 15 to 20 days, in rough agreement with the predictions of the model.

There are three separate forecasting problems that match the three time scales of monsoon phenomena. For the briefest events, such as individual disturbances in the active phase, mathematical analogues of the physical weather system can be devised. When data describing the current weather conditions are supplied, a computer simulation can extrapolate the data forward in time according to the rules of the mathematical analogue. Numerical weather prediction of this kind is an established technique that can attain reasonably good accuracy a few days in advance.

The major drawback of the method is that the initial data must constitute a complete description of the state of the atmosphere in a given region, including the variables that characterize moist processes. The collection of such data is probably beyond the capabilities of the present observational network. Remote sensing of the atmosphere by satellites and floating buoys, working in conjunction with the network of ground stations, may someday provide adequate data for forecasting.

For predicting seasonal or annual trends mathematical analogues are of little use. Here too the data are too sparse, but, more important, the compu-



ACTIVE AND DORMANT PHASES of the summer monsoon arise in the mathematical simulation when moist processes are included in the model of the earth with a continental cap. The distribution of surface temperature with time and latitude is an enlargement of the area outlined in color in the bottom diagram on the opposite page. The distribution of vertical wind velocity is also shown. Rising air currents (*colored region*) are characteristic of the active monsoon, whereas subsiding air (*gray region*) suppresses storms and corresponds to the dormant monsoon. In the diagrams at the right the distribution of temperature and vertical wind velocity is shown in cross section for three days in the cycle. To emphasize the passage of the active and dormant phases the graph of temperature changes shows the difference between the actual temperature and the average temperature for that latitude over a 30-day period. The temperature falls in the wake of the active monsoon on days eight and 12 and rises near the coast on day 12 as the ground dries. The active cycle begins again.

tations are much too cumbersome to allow exact numerical climate forecasting. Nevertheless, the study of long series of data indicates that certain climatic developments are related to certain precursor events. For example, the winter climate of North America seems to be related to anomalies in the temperature distribution at the surface of the North Pacific. Because of the importance of moist processes to monsoons, it may be worthwhile to seek similar correlations, say between extreme monsoon phenomena and abnormal sea-surface temperatures.

It is in the prediction of the events of intermediate time scale, namely the active and the dormant phases of the monsoons, that forecasting would probably have its greatest social and economic impact. Predicting the transition from one phase to another requires a forecast for a period of weeks, too long for exhaustive numerical techniques. If the theories of the alternation of active and dormant phases are correct, however, the difficulty may be overcome. It may be possible to develop a simpler but still effective mathematical analogue that ignores events on a shorter time scale and focuses on the more slowly varying elements of the system.

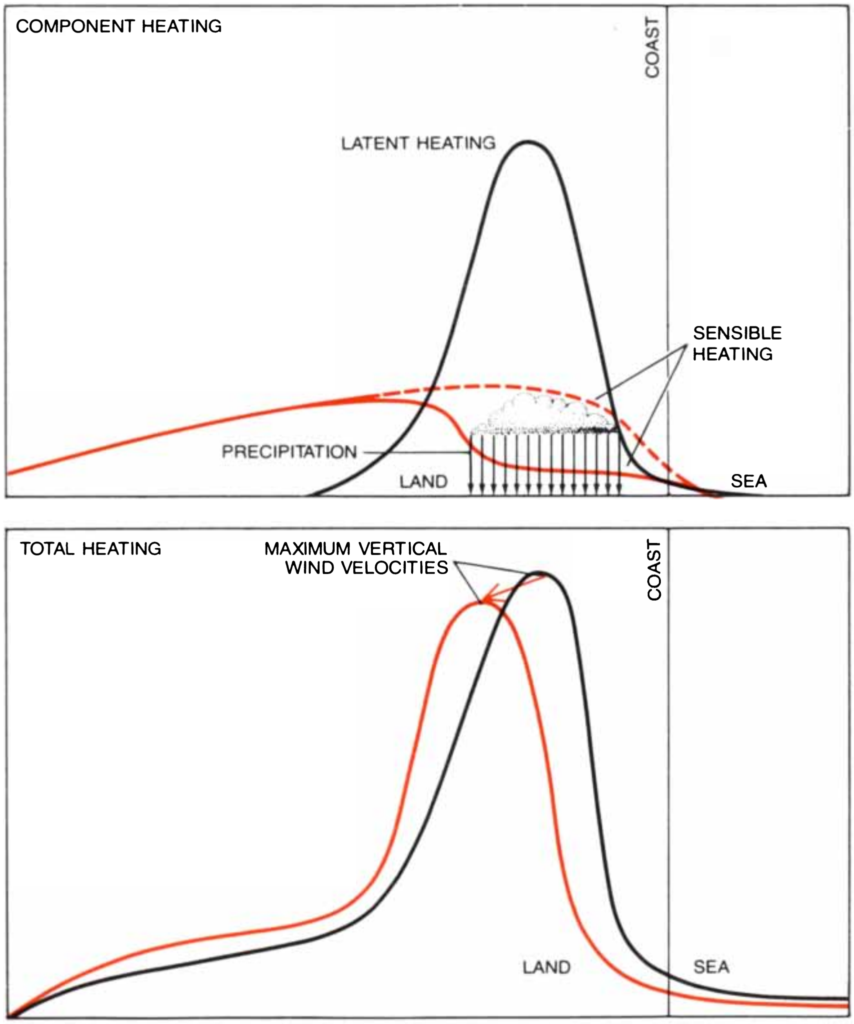
Within the context of the general theory of monsoon circulation there are local and continental variations that must be taken into account. The Himalaya-Tibetan plateau, for example, appears to accelerate the onset of the Asian monsoon and to increase its ultimate intensity. Satellite data indicate that the central and southeastern parts of Tibet remain free of snow throughout most of the year. Hence the plateau must heat rapidly during the spring in the Northern Hemisphere. The precise way in which the plateau then influences the atmosphere is still unclear and has been the subject of a number of theoretical studies.

In contrast, the summer monsoons of Australia and Africa tend to be considerably weaker than their Asian counterparts. Over northern Australia precipitation decreases rapidly inland, so that only a narrow region along the northern coast receives significant monsoon rainfall. Similarly, the central arid lands of the African Sahel receive only spasmodic rainfall during the summer. West Africa and Australia are geographically alike as monsoon regions, and neither one is influenced by a dominant mountainous structure comparable to the Himalayas.

It may seem puzzling that there is no major monsoon system in the Americas. The equatorial region of the Western Hemisphere is dominated by the Amazon basin, whose overlying air mass is subject to sensible and latent heating for much of the year. In North America radiative cooling in the winter generates considerable potential energy between the two hemispheres. But a flow of air between them still cannot develop because it is blocked by the Andes.

The Coriolis force causes the cold North American air to enter the Pacific Ocean as the northeast trade winds. If the mountainous barrier of the Andes did not limit circulation, differential surface heating between the Amazon basin and the Pacific Ocean would cause a flow of air into the Amazon, leaving a low-pressure area over the South Pacific. Such a low-pressure area would cause the northeast trade winds to cross the Equator, and the Coriolis force would turn them toward the southeast, completing the path of interhemispheric airflow. Because of the mountains, however, the cross-equatorial flow is weak in the eastern Pacific and the trade winds continue instead across the Pacific to Indonesia. Furthermore, the eastern Pacific is dominated by cold water. Even if the Andes did not exist, the moisture content of the air would be too low to build up the high energy associated with the Asian monsoons. Air circulating in the Amazon basin is moistened by evaporation from the Atlantic Ocean.

A number of international field experiments have been undertaken to provide detailed data for research on monsoons. The most important of these were the International Indian Ocean Experiment that was conducted between 1959 and 1965 and the Monsoon Experiments (MONEX) of 1978 and 1979. The latter were part of the Global Weather Experiment, and it included separate studies of the summer and the winter monsoon circulation. In each experiment satellites, ships and research aircraft were employed in order to obtain a three-dimensional picture of an evolving monsoon. One of the most important applications of the data will be in testing theories of the monsoon's active and dormant phases.



TOTAL HEATING of a column of air over the land is the sum of three contributions: radiative heating (not shown), sensible heating (from the sun-warmed land) and latent heating (from the condensation of water vapor). Precipitation tends to reduce the sensible heating of the column, causing the maximum total heating to shift inland. The region where the ascending air has its maximum velocity also shifts inland and a slow migration of the active monsoon ensues.

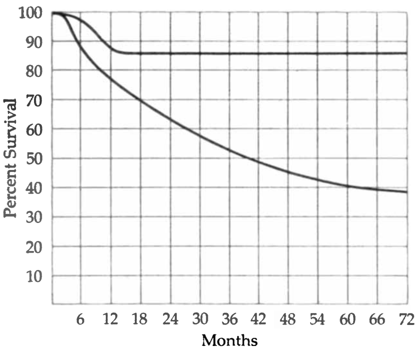
Medicine is always changing, now there's a text that changes right along with it.

SCIENTIFIC AMERICAN *Medicine.*

The answers that you get from SCIENTIFIC AMERICAN *Medicine* are current, because each month new chapters replace the old. You will know where to look for the answers, because this advanced text follows the familiar structure of medical knowledge, changing as medicine changes.

Comprehensive. Leading authorities from Harvard and Stanford cover the full range of medicine for you the seasoned clinician.

Complimentary CME. At no extra cost, we offer a CME program through which you can earn 32 top credits* per year.



Top curve represents survival in chronic hepatitis patients treated with steroids.

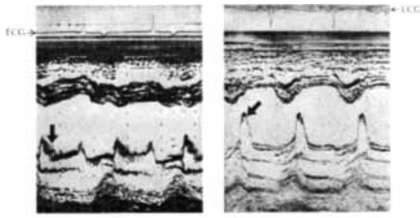
CT scan reveals large meningioma and mild hydrocephalus.

Trial Offer. We invite you to try SCIENTIFIC AMERICAN *Medicine* for two months with our compliments. Send us the coupon and you will receive the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $185 for the first 12 months (renewal is $160); otherwise, return the volumes; there will be no charge.

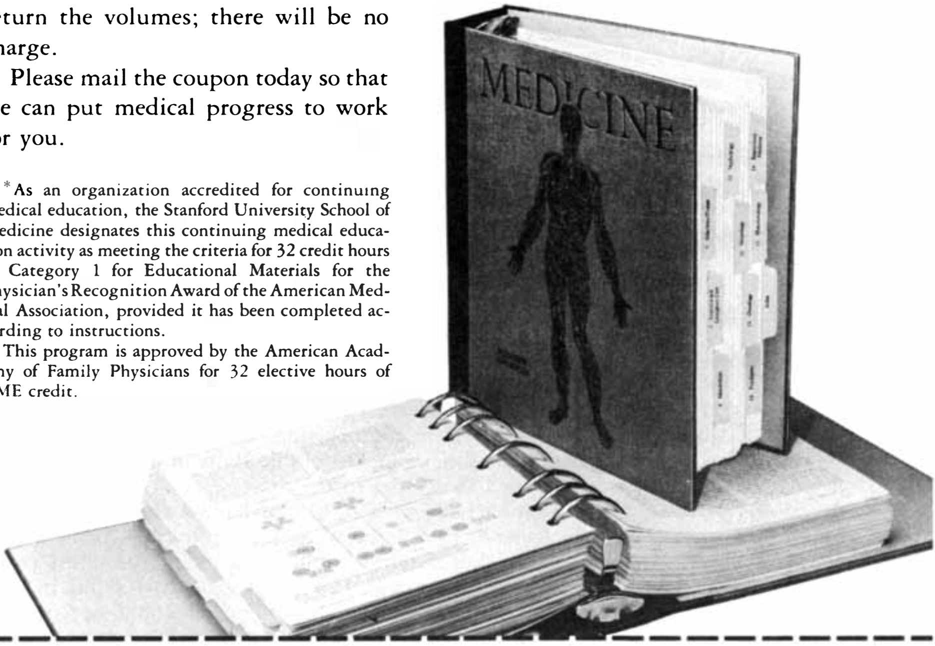
Please mail the coupon today so that we can put medical progress to work for you.

*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

This program is approved by the American Academy of Family Physicians for 32 elective hours of CME credit.



Echocardiograms from patients with aortic regurgitation.



SCIENTIFIC AMERICAN **MEDICINE**

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine.* On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $185 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

- ☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
- ☐ I shall also enroll in the CME Program
- ☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $185*
- ☐ Please bill me

*Please add sales tax for California, Illinois, Massachusetts, Michigan, Ohio, and New York

Name ______________________ MD Specialty ______________

Address ______________________

City ______________ State ______________ Zip ______________

Signature ______________________

Subscribers outside the U.S. and possessions will be charged extra for shipping and handling; updates will be sent by surface routes unless airmail delivery is requested. Please allow 6-8 weeks for delivery.

A1

A Mesolithic Camp in Ireland

Although Ireland seems to have been unpopulated in the Ice Age, recent excavations show that hut-dwelling hunters and gatherers came on the scene in early postglacial times, some 9,000 years ago

by P. C. Woodman

Men of the Old Stone Age left abundant evidence of their presence in Britain, but these ice-age hunters apparently never got to Ireland. The island was one of the last parts of Europe to be occupied by man. Its first inhabitants were hunters and gatherers of postglacial times whose period, sandwiched between the end of the Paleolithic and the first appearance of Neolithic farmers, is known in the Old World as the Mesolithic.

Until recently the principal evidence of these first immigrants' presence in Ireland has consisted of flint tools, many of them discovered on the surface of former beaches that are now raised well above sea level. These artifacts have been tumbled and eroded by wave action. As a result they are not easy to analyze. They have been found mainly along Ireland's northeast coast, where flint is naturally abundant. In the 1930's Hallam L. Movius, Jr., of Harvard University excavated a number of Mesolithic sites in this area and concluded that the people who had made the tools had established themselves in Ireland by about 6000 B.C. Apart from the flints, however, there was little evidence of the immigrants' material culture.

The accepted image of these earliest Irishmen, colored by the impoverished quality of their material remains, was that of a people who had barely eked out an existence along the seacoast and had never successfully penetrated the island's interior. They appeared to fill a niche at the edge of western Europe similar to that of the aboriginal Tasmanians in the Pacific, the Patagonians in subpolar South America and the Hottentot "strandloopers" of South Africa. Because their resources seemed so limited it was assumed that they had crossed over to Ireland by the shortest sea route: from Argyll in Scotland to the northeast coast. It was also assumed that they were confined to that corner of Ireland because of the ready availability of flint in the area.

This perspective of the Irish Mesolithic has now been revolutionized, largely because of the findings at a single site on the north coast: Mount Sandel. The excavations there have provided new information about Mesolithic subsistence, have greatly expanded knowledge of the hunter-gatherers' material equipment and have shown that Ireland was populated at least 1,000 years earlier than had been thought, or about 9,000 years ago. These are not inconsiderable results from a site that at first seemed to be short on promise.

Mount Sandel (the site is named after a nearby earthwork, Mount Sandel Fort, thrown up in the Iron Age) lies on the estuary of the River Bann near the town of Coleraine. The fort is on the top of a 100-foot bluff that overlooks the Bann, a river that runs from Lough Neagh northward to the sea. Here the Bann, although it is several miles from the coast, is a tidal stream. For at least a century amateur collectors have turned up Mesolithic flints in the fields beside the fort.

The Mount Sandel area has been of interest to archaeologists because the tools found there differ from those considered typical of the Mesolithic in Ireland. The tools that could be collected by the thousands along the northeast coast are mostly rather large flint blades. Those found at Mount Sandel, however, included smallish flint axes and the tiny shaped pieces of flint known as microliths. Artifacts such as these, although they are typical of the Mesolithic tools found in many parts of northwestern Europe, were almost unknown in Ireland.

A few years ago, when it became known that a housing development was to be built in the Mount Sandel area, I was offered the opportunity to excavate in advance of the construction work on behalf of the Historic Monuments Branch of the Government of Northern Ireland. No one expected that much evidence of the Mesolithic presence there would remain because the fields had been farmed intensively for generations. Nevertheless, it was hoped that whatever our findings might be they would clarify the relation between the Mount Sandel microliths and the more abundant large-blade tools of the Irish Mesolithic found elsewhere.

As it turned out, so much had survived at the site that the excavation took 10 times as long as had been expected. Our 40 weeks of work were spread over five seasons. Eventually the digging covered an area, with the sampling excavations included, that totaled 1,000 square meters. The work called for the help of many specialists, enlisted both locally and from abroad. Among them were Louise van Wijngaarden-Bakker of the University of Amsterdam, who analyzed the faunal remains at the university's Van Giffen Institute. J. P. Paals, also of the University of Amsterdam, and Michael Monk of University College in Cork have analyzed botanical materials other than pollens; the pollen analysis has been undertaken by Alan Hamilton of the New University of Ulster and Rob Scaife of the Institute of Archaeology of the University of London. Rick Batterbee of University College London has studied the geological past of the Bann estuary and Fred Hammond of the University of Cambridge and Jim Cruickshank of the Queen's University of Belfast have analyzed the soils at the site. Ronald Doggart, also of the Queen's University of Belfast, has studied the magnetic susceptibility of hearth areas, and John Dumont of the Baden-Powell Institute of the University of Oxford has examined the flint tools for evidence of microwear.

Our excavation soon exposed a number of Mesolithic pits, hearths, areas of charcoal-stained soil and postholes. The fact that many of these features were at the edge of a small hollow had preserved them in spite of generations of plowing. At first the pits and postholes appeared to be an incomprehensible muddle. It became possible, however, to show that many of the pits had been dug long after the postholes. When these later pits were excluded from the plan, we were able to discern four central hearth areas in the hollow, each area associated with the postholes of a sheltering hut. The huts had not all been built at the same time; they were occupied in succession.

The general configuration of the postholes, many of them about 20 centimeters deep and set in the ground at an angle, together with the chance event that the perimeter of one hut coincided with the rim of the hollow, indicated that the huts were about six meters across. The angle of the holes further suggested that the huts were dome-shaped. They were probably built by setting up a roughly circular array of sturdy saplings that were then bent inward so that their tops could be lashed together. This method of construction would have got around the need for internal roof supports. Lighter branches could then have been woven into the robust framework and the resulting structure could have been weatherproofed with some kind of covering.

The builders cleared the ground inside these spacious huts down to the level of the subsoil. Thus it is possible that the covering on the half of the hut that faced into the worst of the winter weather—the north side—consisted of the removed sod. Alternatively the sod could have been used to build a bank around the entire circumference of the hut. In either case the size of the postholes suggests that the huts' south side was somewhat more lightly constructed. This may indicate that the more sheltered southern half was weatherproofed with skins or with bark held in place by stone ballast.

It is worth noting in this connection that clear traces of substantial shelters such as these are scanty at other Mesolithic campsites in northwestern Europe. Usually the archaeological evidence indicates the presence of nothing more than light tents. This suggested to us that perhaps the hut builders at Mount Sandel remained in residence for a good part of the year.

Next to the hearth that belonged to what was possibly the last hut to be built in the hollow our digging revealed the remnants of what seems to have been a small work area. Around it were found quantities of the waste flint that was produced in the process of making tools, mixed with worn microliths that had evidently been discarded once replacements were available. A number of small flint awls also turned up beside the hearth; Dumont's microwear examinations suggest that these piercing tools had seen much use.

A scattering of flint axes, scraping tools and blades were found around the periphery of this same hut. The blades, which still carry traces of the iron pigment red ocher, are one of our more enigmatic discoveries. They may have been used in some form of hide preparation that involved the pigment or perhaps simply in the manufacture of some kind of red paint.

Scattered around the hollow were various types of pits. One type, generally about 60 centimeters deep and one meter in diameter, had in most instances been dug at some distance from the huts. Around the edges of one pit we noted several small holes, as if some kind of lining had once been pinned to the pit wall. Unfortunately nothing was found in any of the pits that gave an obvious clue to their function.

Speculating on the purpose of the pits, we were able to exclude the possibility that they had been produced naturally by the wrenching of the root system that accompanies the toppling of a tree. Both their shape and the fill they contained rule out any such origin. They do not appear to have been borrow pits, dug to provide earth to bank up around the

POSTHOLES OF VARIOUS SIZES pock the floor of the Mount Sandel site. The large cavities are pits; many of them had been filled with trash taken from the many hearths at the site. The excavation required more than 40 weeks' work spread over a five-year period.

All you need when you need to straighten out your trip. In English.

© American Express Company 1980

You need someone who not only understands English, but also someone who understands your travel problems. And that's exactly who you'll find at the more than 1,000 Travel Service Offices of American Express Company, its subsidiaries and Representatives around the world. They know how to untangle a snarled itinerary. To get you emergency funds with the Card. To replace a lost Card, fast. The only card with all the resources and people of American Express to help you at home and abroad. The American Express® Card. Don't leave home without it.™

AMERICAN EXPRESS®

hut walls. Nor is there any evidence that they are the result of flint quarrying. The regularity of their shape and the fact that many of them were recut on almost the identical site rules out either possibility.

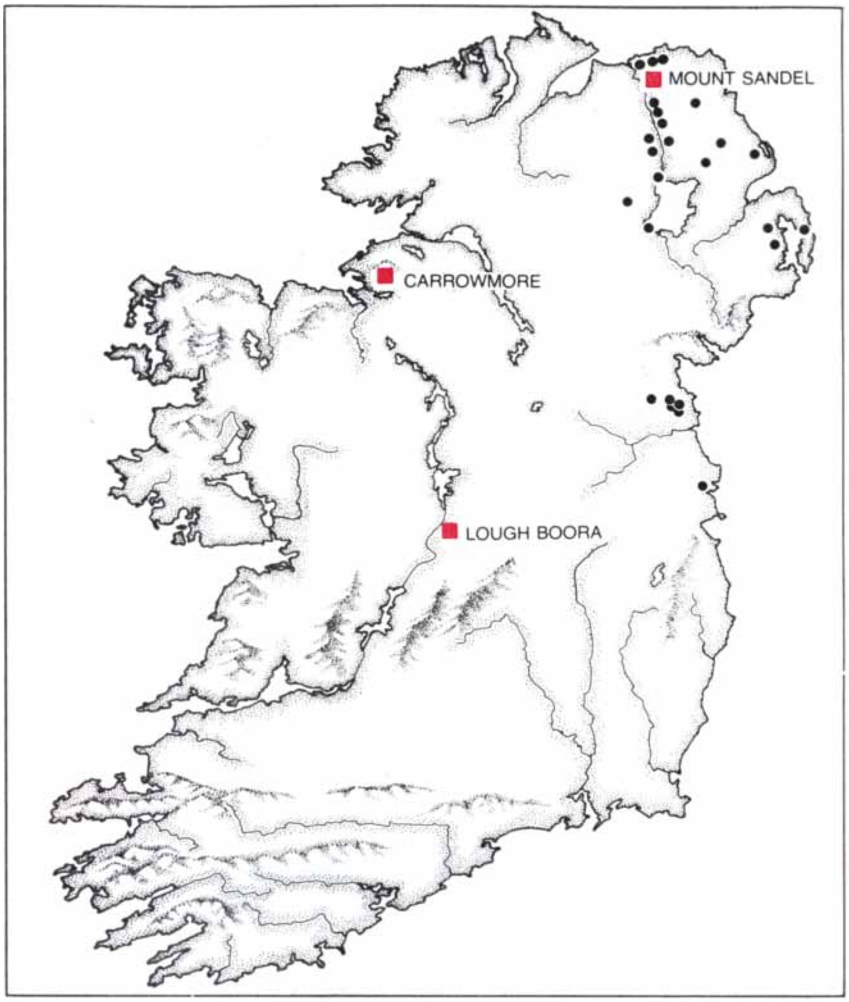
The hearths, some of them bowl-shaped, inside the huts incorporated burnt animal bones, the carbonized shells of hazelnuts and flint tools that showed evidence of burning. Analyses of the hearth debris show high concentrations of trace elements associated with food preparation, such as copper and phosphorus. These findings confirmed our assumption that the hut fires had served for cooking as well as for warmth.

Similar rubbish was found in several of the pits outside the huts. Doggart's magnetic studies indicated that it had been intensely burned. There was no evidence, however, that the pits had ever served as hearths. The only logical conclusion is that the Mesolithic hut dwellers cleared the ash from their hearths from time to time, carried it outside and perhaps occasionally dumped it into the nearest empty pit.

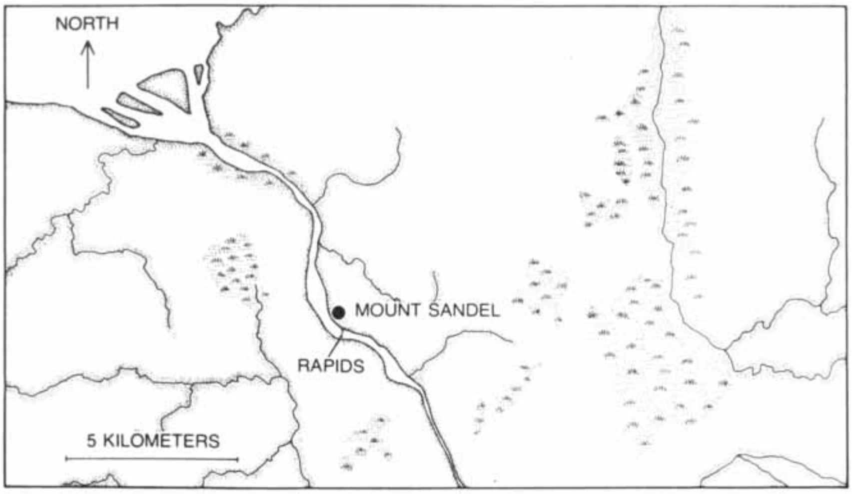
Evidence of toolmaking outside the hut area was unearthed to the west of the hollow. There was an extended scatter of waste flint and of the cores from which the finished tools were struck, but there were few finished tools. Apparently this was a place where the cores were roughed out, usually from a piece of a naturally occurring flint nodule; small blades were then struck from the prepared cores. The richest concentrations of finished tools came from the layers in the pits that also held the hearth rubbish such as animal bones and nutshells.

The commonest flint tools at Mount Sandel are microliths. They fall into four broad classes: needle points, rods, obliquely trimmed blades and, most numerous, small elongated blades, many with one edge retouched. We named these last tools scalene triangles even though few of them were completely triangular. Many of the triangles had probably formed the cutting edge or the tip of composite implements such as knives, arrows, spears and harpoons. Evidently after such wood or bone implements had outlived their usefulness they were thrown into a hut fire, stone triangles and all. For example, it was common to find a little group of similar triangles lying together in one layer of a pit, suggesting that they had all once been part of a single composite implement. A count of the individual microliths found in such groups suggests that composite projectile heads were commonly fitted with six to 10 such flint barbs or cutting edges.

It is less clear what some of the other



IRISH MESOLITHIC REMAINS, once chiefly represented by surface finds of wave-rolled flint implements on raised beaches, have now been found at numerous sites (*black dots*) in the east and northeast and at one central site, Lough Boora, near the River Shannon (*colored rectangle*). The two other colored rectangles designate Mount Sandel on the River Bann and Carrowmore on the west coast. Mount Sandel and Lough Boora indicate that the Irish Mesolithic is 1,000 years older than was previously believed, and Carrowmore may prove to be even older.

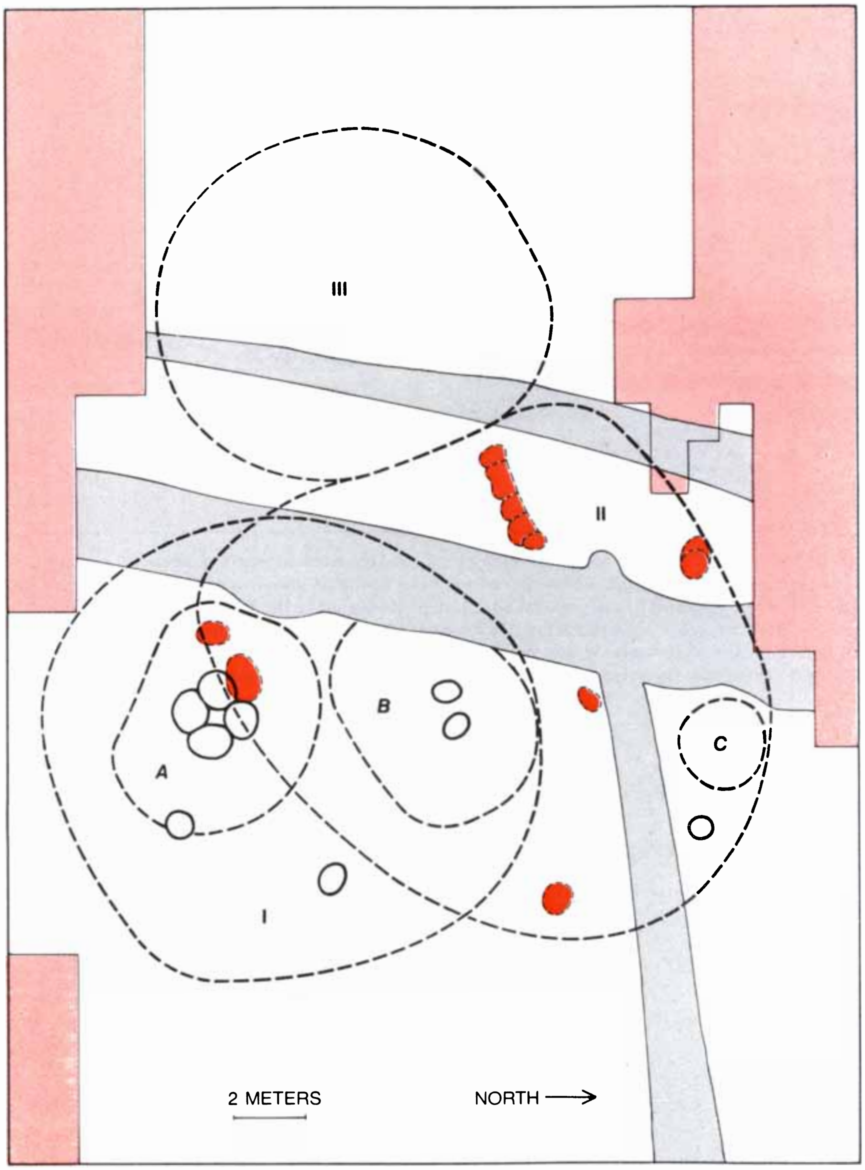


LOWER REACHES of the Bann valley in the vicinity of Mount Sandel are bounded by steep bluffs, but the land above the bluffs is relatively flat and occasionally boggy. During the Mesolithic the bogs were either small lakes or dry land. The river at Mount Sandel was a series of narrow rapids where residents of the campsite probably caught salmon and eels in season.

microliths were for. Those we call needle points might have been the tip of arrows or perhaps the tip of projectiles that had a cutting edge made up of triangles. The little rods are enigmatic. Dumont's microwear studies, although they are not yet complete, reveal that the rods are heavily worn. Most of the rods were found in pits inside the huts; perhaps the rods served in some way for food processing. The same pits have yielded large quantities of fishbones, and one group of rods came from a pit layer that also held the seeds of apples and water lilies.

Larger flint tools such as axes and picks were also unearthed in considerable numbers. Among the axes are many that were fashioned by shaping an entire core; implements of this kind are common at Mesolithic sites in northwestern Europe. A type of splay-edged chisel or adze was made by removing a large flake, possibly from a preshaped core, in such a way that one of the original edges of the flake formed a broad and strong cutting edge.

The flint picks, which were also fashioned out of a core, may have been the implements that dug the many pits at Mount Sandel. Most of the picks we found were broken. Two axes of the polished type, a fairly standard Irish Mesolithic implement, were also unearthed, as were some sturdy flaked flint awls. Other standard tools, however, such as burins for working in wood and bone and scrapers for the preparation of hides, were scarce.



MAJOR FEATURES of the Mount Sandel site appear in this plan. The boundaries shown as broken lines mark the site's three active areas. Area III, to the west, was a flintworking zone where the blades that would be made into microliths were struck from cores of flint. Areas I and II overlap; both have yielded evidence of hut building associated with hearths. The northernmost hearth was outside rather than inside the hut. Most of the largest pits (*color*) were found in Area II. The details of hut area *A*, with its five hearths, are illustrated on page 126.

Generally the acid soils of northeastern Ireland make short work of buried animal bones. Some of the pits at Mount Sandel, however, contained quantities of burnt and fragmented bone. From many of these fragments van Wijngaarden-Bakker was able to reconstruct identifiable bones. Hence we know that of the mammals killed by the Mesolithic hunters of Mount Sandel by far the most numerous were wild pigs. Evidently hares were also caught, but not often. One of the mammal bones may be that of a dog.

Fishbones and bird bones were numerous, particularly in the pits inside the huts. The large majority of the fishbones are those of salmonid species; most of them are probably those of the Atlantic salmon, *Salmo salar*. Other fish species that were caught in lesser numbers included sea bass, eel and an occasional flounder. Among the birds eaten at Mount Sandel were duck, pigeon, dove, grouse, goshawk and capercaillie. The presence of the bones of this last bird, a large species of grouse common to northern Europe, tells us something about the Mesolithic landscape of Ireland. The bird's natural habitat is pine forest, and when pine forest died out in Ireland some 4,000 years ago, the capercaillie disappeared. Nine thousand years ago, however, much of Ireland would still have been covered with Scotch pine.

Such an example of environmental change demonstrates the need to keep it in mind that the Mesolithic hunters of Mount Sandel lived in a postglacial world. The landscape, the animals and even the sea level of today bear little or no resemblance to those of early postglacial Ireland. For example, it is unlikely that 9,000 years ago this part of the Bann valley had a tidal stream. Although the land, only recently relieved of its burden of glacial ice, was lower than it is today, the level of the sea was lower still. Batterbee's preliminary geomorphological analyses suggest that the Bann estuary was some distance farther downstream and that opposite Mount Sandel the river was a series of rapids. The rainfall then was less than it is today, so that in the summer months the Bann would have been a much narrower and more accessible river.

What can be said about the postglacial conditions 9,000 years ago, both in Ireland in general and at Mount Sandel in particular? As I have already indicated, the trees would have been predomi-

WHAT'S BETTER THAN SPEED READING?

SPEED LEARNING

(SPEED PLUS COMPREHENSION)

Speed Learning is replacing speed reading. It's easy to learn...lasts a lifetime...applies to everything you read...and is the only accredited course with the option of college or continuing education credits.

Do you have too much to read and too little time to read it? Do you mentally pronounce each word as you read? Do you frequently have to go back and reread words or whole paragraphs you just finished reading? Do you have trouble concentrating? Do you quickly forget most of what you read?

If you answer "yes" to any of these questions — then here at last is the practical help you've been waiting for. Whether you read for business or pleasure, school or college, you will build exceptional skills from this major breakthrough in effective reading, created by Dr. Russell Stauffer at the University of Delaware.

Not just "speed reading" — but speed reading-thinking-understanding-remembering-and-learning

The new *Speed Learning Program* shows you step-by-proven-step how to increase your reading skill and speed, so you understand more, remember more and use more of everything you read. The typical remark made by the 75,000 slow readers who completed the *Speed Learning Program* was: "Why didn't someone teach me this a long time ago?" They were no longer held back by the lack of skills and poor reading habits. They could read almost as fast as they could think.

What makes Speed Learning so successful?

The new *Speed Learning Program* does not offer you a rehash of the usual eye-exercises, timing devices, costly gadgets you've probably heard about in connection with speed reading courses or even tried and found ineffective.

In just a few spare minutes a day of easy reading and exciting listening, you discover an entirely new way to read and think — a radical departure from anything you have ever seen or heard about. Research shows that reading is 95% *thinking* and only 5% eye movement. Yet most of today's speed reading programs spend their time teaching you rapid eye movement (5% of the problem) and ignore the most important part (95%) *thinking*. In brief, *Speed Learning* gives you what speed reading *can't*.

Imagine the new freedom you'll have when you learn how to dash through all types of reading material *at least* twice as fast as you do now, and with greater comprehension. Think of being able to get on top of the avalanche of newspapers, magazines and correspondence you have to read . . . finishing a stimulating book and retaining facts and details more clearly and with greater accuracy than ever before.

Listen-and-learn at your own pace

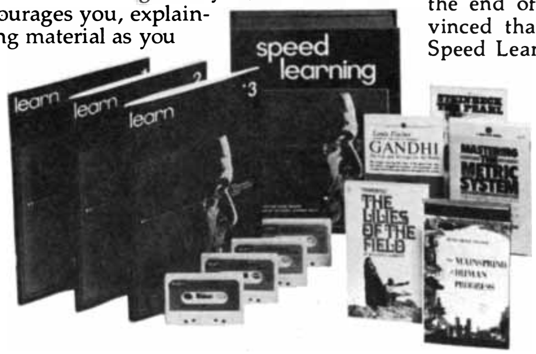
This is a practical, easy-to-learn program that will work for you — no matter how slow a reader you think you are now. The *Speed Learning Program* is scientifically planned to get you started quickly . . . to help you in spare minutes a day. It brings you a "teacher-on-cassettes" who guides you, instructs, encourages you, explaining material as you read. Interesting items taken from *Time* Magazine, *Business Week, Wall Street Journal, Family Circle, N.Y. Times* and many others, make the program stimulating, easy and fun . . . and so much more effective.

Executives, students, professional people, men and women in all walks of life from 15 to 70 have benefited from this program. Speed Learning is a fully accredited course . . . costing only 1/5 the price of less effective speed reading classroom courses. Now you can examine the same, easy, practical and proven methods at home . . . in spare time . . . without risking a penny.

Examine Speed Learning FREE for 15 days

You will be thrilled at how quickly this program will begin to develop new thinking and reading skills. After listening to just one cassette and reading the preface you will quickly see how you can achieve increases in both the speed at which you read and in the amount you understand and remember.

You must be delighted with what you see or you pay nothing. Examine this remarkable program for 15 days. If, at the end of that time you are not convinced that you would like to master Speed Learning, simply return the program and owe nothing. See the coupon for low price and convenient credit terms.



Note: Many companies and government agencies have tuition assistance plans for employees providing full or partial payment for college credit programs.

In most cases, the entire cost of your Speed Learning Program is Tax Deductible.

COLLEGE CREDITS

You may obtain *2 full semester hour credits* for course completion, wherever you reside. Credits offered through Whittier College (California). Details included in your program.

CONTINUING EDUCATION UNITS

National Management Association, the world's largest association of professional managers, awards 3.0 CEU's for course completion. CEU's can be applied toward the certificate in Management Studies.

PROFESSIONAL SOCIETIES

Speed Learning is offered internationally to members of professional associations such as: American Chemical Society, Foundation for Accounting Education, Institute of Electrical and Electronics Engineers and dozens more. Consult your Education Director for information.

BUSINESS, INDUSTRY, GOVERNMENT

Many companies and government agencies offer Speed Learning as a wholly-paid or tuition reimbursement program. Consult your Training or Personnel Director for details.

31SAI

learn INCORPORATED 113 Gaither Drive, Mt. Laurel, NJ 08054

☐ **YES!** Please send the Speed Learning program @$119.95 plus $4 postage and handling
NJ residents add 5% sales tax

Check method of payment below:

☐ Check or money order enclosed (payable to Learn Incorporated)
☐ Charge my credit card under the regular payment terms
☐ Visa ☐ Master Card Interbank No.________ ☐ American Express
Card No.________ Exp. Date________

I understand that if after 15 days I am not delighted in every way, that I may return the materials in their original condition for a full refund. No questions asked.

Name________
Address________
City________ State________ Zip________
Signature________

If you don't already own a cassette player, you may order this Deluxe Cassette Recorder for only $49.95. (Includes handling and delivery.)
Check here to order ☐

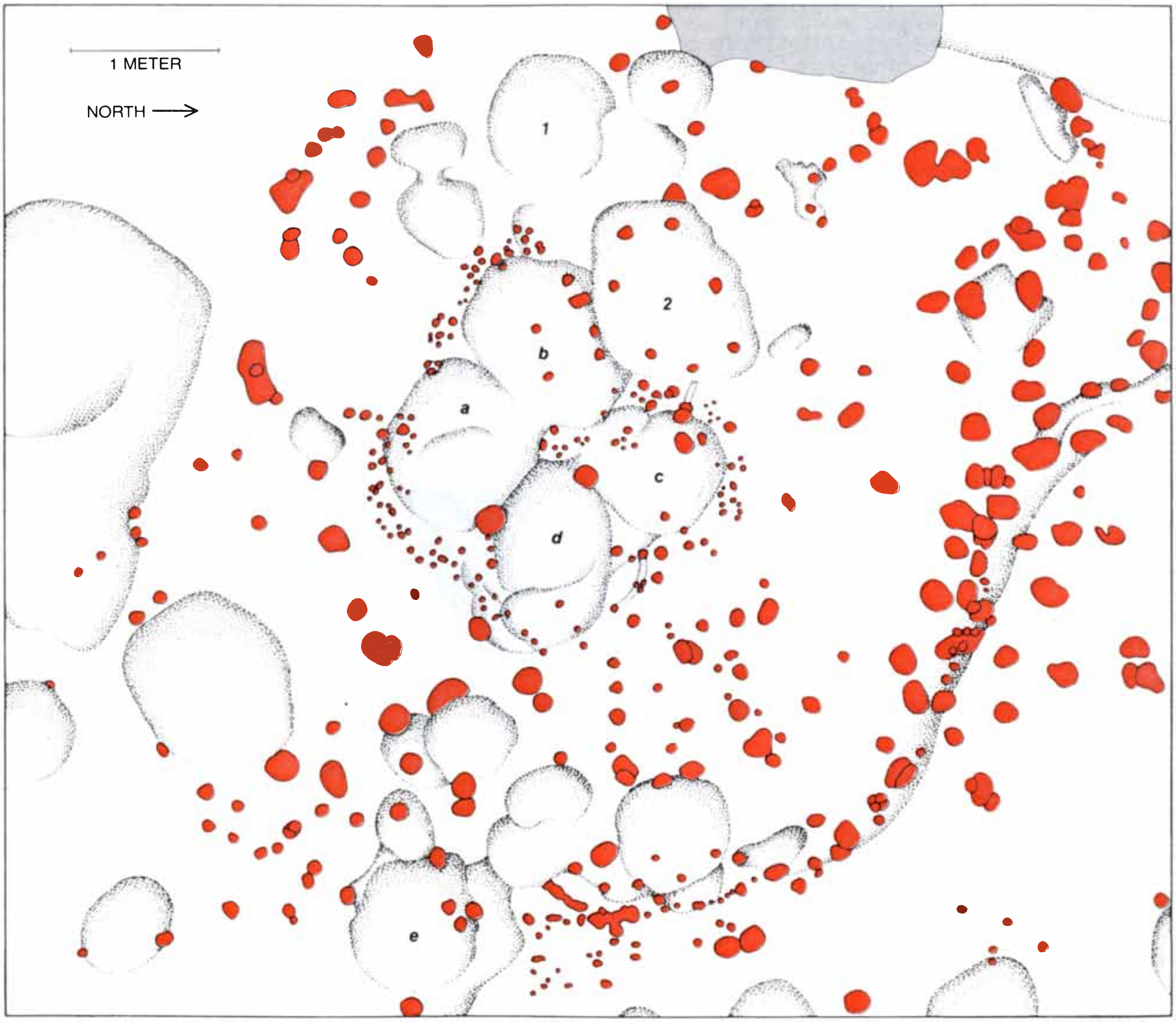
Outside USA add $10 per item—Airmail extra

nantly Scotch pine rather than those of the dense deciduous forest that covered much of Ireland in early historical times. Birch would also have been present, and willows would have grown in the lower-lying damp areas. Hazel, already established in the south, would have just reached the Mount Sandel area. Climatological studies indicate that the weather was drier than it is today, so that many of the present-day boggy areas in the Bann valley would have been small lakes or would have held no water at all.

In glacial times both Ireland and Britain were connected by land bridges to the mainland of Europe. Ireland may have been an island, however, from the beginning of postglacial times and was certainly an island not long afterward. This insularity had only a slight effect on vegetation, but its influence on the animal population was dramatic. Such large mammals as the elk, the roe deer and wild cattle failed to reach Ireland, as did such smaller mammals as the otter and the beaver. So did such freshwater fishes as the pike and possibly the tench and the bream.

The island's mammals, however, did include the red deer, the brown bear and, as I have noted, the pig, the hare and possibly the dog. Several migratory fishes such as the salmon and the eel were in the rivers for seasonal runs in the summer and fall, and a range of marine foods would have been available along the coast. These resources, combined with the availability of seasonal plant foods, should have kept the Mesolithic hunter-gatherers of Ireland reasonably well fed from late spring to fall. Winter, however, would have been lean.

At what times of year were the Mount Sandel huts occupied? Could the site have served as a year-round settlement? Our excavations have uncovered much evidence with seasonal connotations. For example, the salmon bones are evidence of summer occupation. Today the main salmon run up the Bann is in midsummer: June, July and August. In Mesolithic times there may also have been an earlier run, say in April or May. Although the lower water temperatures of the early postglacial period make it unlikely that the salmon run would con-



MAIN HUT AREA includes five hearths; four of them (*a–d*) are in the central part of the area and one (*e*) is at its eastern edge. The hut postholes are shown in color. The presence of the five hearths suggests that the site was occupied by a succession of hut builders, each group possibly occupying its structure for more than a single season. Two pits (*1, 2*) were dug within the hut area. The huts were about six meters wide; the edge of one hut is outlined by postholes along the northeast edge of the hollow where the structures stood. Carbon-14 analysis indicates that the huts were built in about 7000 B.C. or soon thereafter. Mesolithic campsites as well preserved as this one are rare.

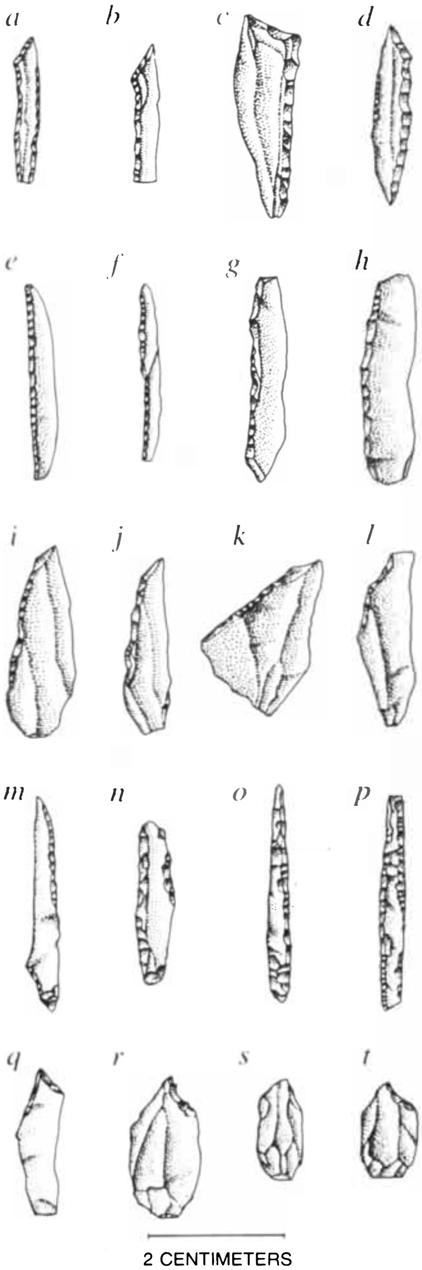
tinue beyond the fall, eels do run downstream in the fall; September, October and November are the best months of the eel run. Hence the eel bones at the site are evidence of fall residence at Mount Sandel. So is the presence of hazelnuts, which are ready for picking by mid-fall, and of water-lily seeds, which are best collected in September.

As for the evidence of winter residence, most of the pig bones are from young animals that were probably killed in late winter. Although the site lacks the large and diversified rubbish deposits that make possible firm judgments on seasonal occupation at many Mesolithic sites in Europe, what has been unearthed from the pits and the hearths shows an apparent continuity of occupation at Mount Sandel from perhaps as early as the spring, although more likely from the summer, on through to the fall, and also an occupation in late winter. That is a substantial part of the year, and it is conceivable the occupation was in fact nearly year-round. For example, hazelnut shells are found with the pig bones, a fact that implies storage of the nuts for winter eating. Were some of the larger pits used for storage?

Why did the Mesolithic occupants of Mount Sandel choose it as a place to build? Two prime factors in selecting the location must have been its proximity to the Bann rapids and its access both to the estuarine stretch of the river farther downstream and to the open sea. The residents probably fished mainly at the rapids, but the presence at the site of wave-rolled flint nodules, of elongated and beveled pebbles of a kind often associated with the preparation of shellfish and perhaps also of the few bones of the sea bass (*Labrax*) is evidence that they also exploited marine food resources.

Hazelnut shells have been found in such large numbers at Mount Sandel that it is tempting to consider them a major food resource. As their presence among the pig bones suggests, they may have been stored in the fall to help bridge the hungry winter months. A considerable effort would have been required to collect the quantity of hazelnuts necessary for this purpose and bring them back to the encampment. Studies of food-collecting societies still extant show that this kind of activity, often women's work, usually proceeds in the immediate vicinity of an encampment. It is possible that stands of hazel scrub occupied the slopes leading down from the bluff to the river at Mount Sandel. If, as seems probable, hazel shrubs were not particularly abundant elsewhere in the area, their presence near the encampment would have been another prime factor in the Mesolithic group's decision about where to camp.

One major problem of interpretation remains. We know that in the fall wild pigs tend to congregate in large numbers where ripe nuts are available. The animal bones at Mount Sandel, however, give no evidence that pigs were much hunted at that time of year. Was the Mount Sandel camp abandoned for part

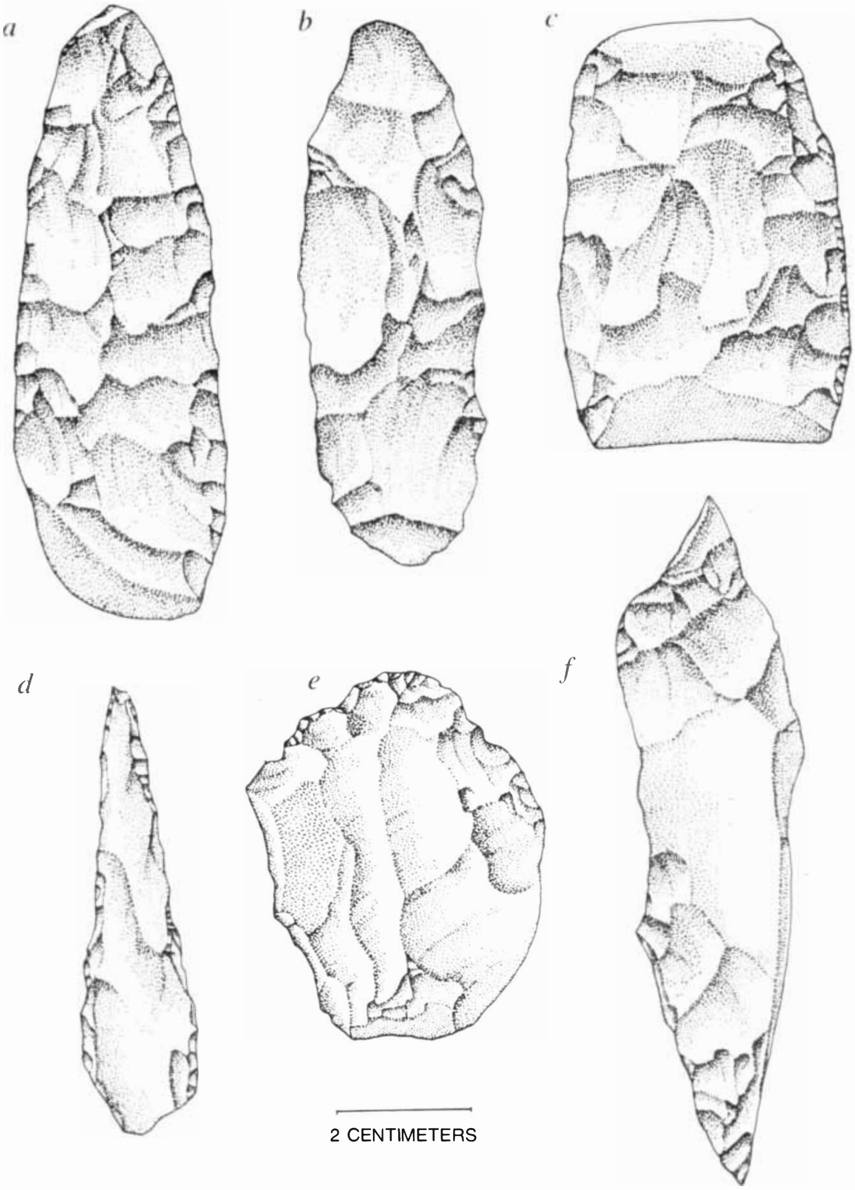


TWENTY MICROLITHS, the commonest form of flint artifact at Mount Sandel, are seen here at their actual size. They fall into four categories. Scalene triangles (*a–d*) were arrow tips or the barbs and cutting edges of composite implements. Rods (*e–h*) may have been used in food preparation. Obliquely trimmed blades (*i–l*) could have been small knives. Needle points (*m–p*) may have been the tips of arrows or other projectiles. The microburins (*q–t*) are not tools but waste products.

of the fall? Or was the site occupied continuously from summer through winter even though the surviving faunal remains provide neither a continuous nor an entirely representative record of the fall occupation? In any event it is clear that even when quantities of animal bones are found, they may not answer all questions.

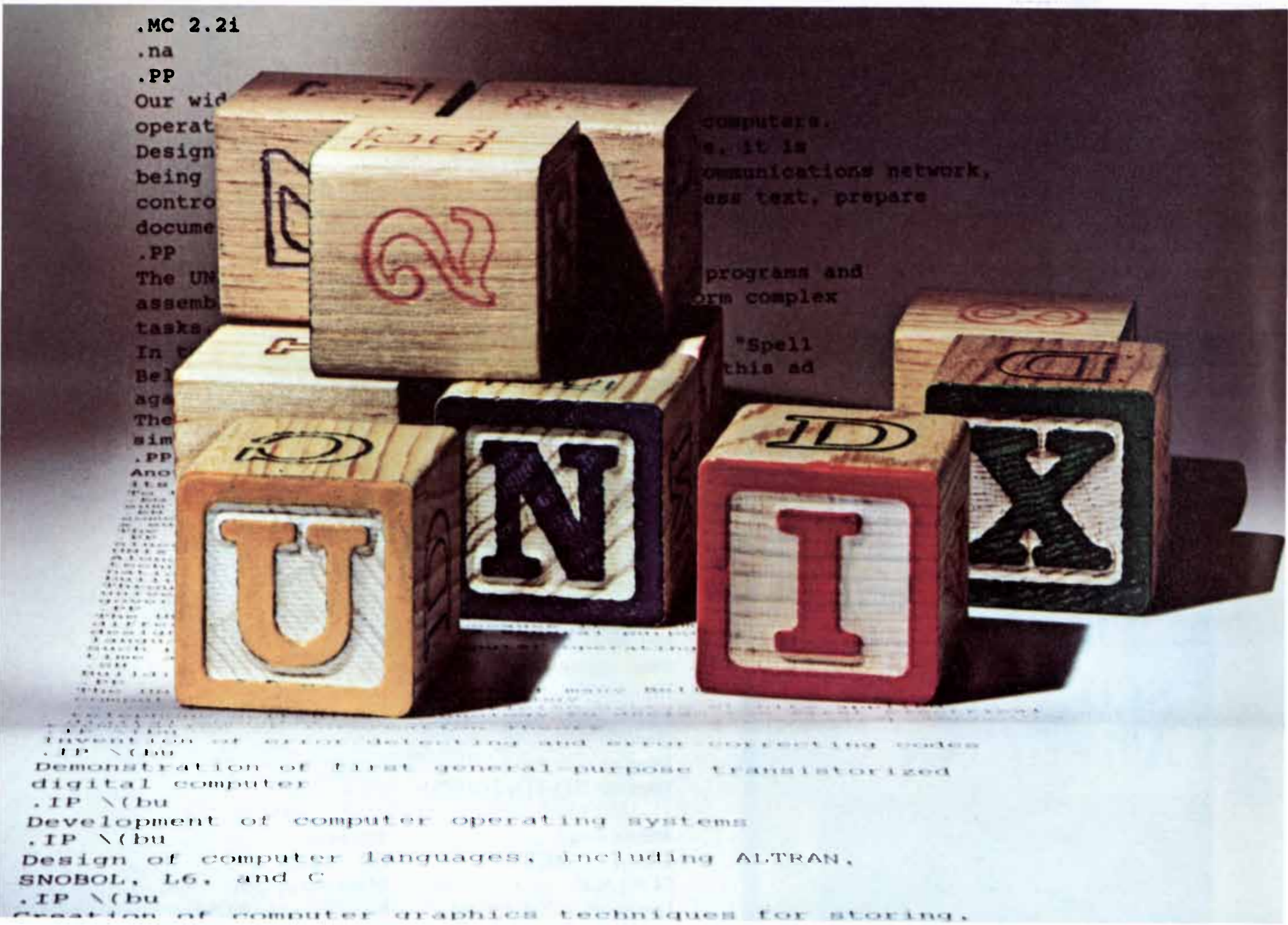
The inhabitants' appetite for hazelnuts proved to be a boon for our work. The charred nutshells were the principal source of the charcoal samples we relied on for dating the Mount Sandel site by carbon-14 analysis. The samples yielded a range of dates that fell largely between 7000 and 6500 B.C., with the majority lying between 7000 and 6800 B.C. It is these measurements that indicate Mount Sandel was occupied about a millennium earlier than any other human habitation known in Ireland until quite recently.

The Mount Sandel encampment appears never to have included more than one hut at a time, although each structure was certainly occupied for much of the year and possibly for consecutive seasons. How many inhabitants did the encampment have? Most rule-of-thumb estimates of population based on hut size are made on ethnographic parallels in climates more tropical than the climate of postglacial Ireland. For Mount Sandel an analogy with the winter camps of the Eskimos seems more appropriate. Such an analogy would suggest a hut population of at most eight to 12 people. The inhabitants of the site



SIX LARGER FLINT TOOLS are axes or choppers (*a, b*), made from cores by striking off flakes; a chisel or adze (*c*), made from a heavy flake; an awl (*d*); a scraper (*e*), possibly for preparing hides, and a pick (*f*), perhaps the kind of tool used to dig pits at Mount Sandel. Flake chisels, like the needle points on the preceding page, are tools peculiar to Mesolithic Ireland.

An easier way to use computers...from a pioneer in computing technology.



The computer printout in the picture above contains the instructions and input needed for the *UNIX* system to format and produce the text on the opposite page.

Our widely accepted *UNIX*™ operating system is simplifying the use of computers. Designed to handle a variety of applications, it is being used to manage and maintain the telecommunications network, control experiments, develop software, process text, prepare documents, and teach computer science.

The *UNIX* system allows users to take small programs and assemble them like building blocks to perform complex tasks. In text processing, for example, the command "Spell Bell Labs Ad" tells a computer to proofread this ad against a dictionary filed in its memory. The program that performs the task was created by simply combining several smaller *UNIX* programs.

Another useful feature of our *UNIX* operating system is its ease in typesetting mathematical expressions. To typeset an equation like

$$\sum_{i=1}^{\infty} \frac{1}{x_i} = \pi$$

someone need only type: "Sum from i = 1 to infinity 1 over x sub i = pi." The computer does everything else.

Since 1969 the Bell System has installed more than 1100 *UNIX* systems. Along with other Bell Labs innovations in computing technology, these systems are enhancing the way the nation's telecommunications network is planned, designed, built, and operated. Through licensing agreements with Western Electric, universities have installed over 800 *UNIX* systems, and government and industrial facilities are using over 400.

The *UNIX* operating system can be used with computers of different manufacturers because it is small, cleanly designed, and written in a general-purpose programming language. Such portability in a computer operating system saves time and money.

Building on the past

The *UNIX* system is just one of many Bell Labs advances in computing science and technology over the years. Among our innovations:

- Application of telephone switching concepts and technology to early computers
- First demonstration of remote computer terminal and data link
- Conception of electronic analog computer
- First design of AND and OR gates for diode circuitry
- Formulation of Information Theory
- Invention of error-detecting and error-correcting codes
- Demonstration of first general-purpose transistorized digital computer
- Development of computer operating systems
- Design of computer languages, including ALTRAN, SNOBOL, L6, and C
- Creation of computer graphics techniques for storing, manipulating, and presenting information
- Development of Fast Fourier Transform
- Design of central processors for switching systems having virtually no downtime

Looking ahead

Computing technology is having a major impact on the telecommunications business. It's increasing the Bell System's ability to provide new services and handle existing ones more efficiently. Today, for example, the nationwide telecommunications network links thousands of software-controlled electronic systems, making it by far the world's largest distributed processing network. We and our partners—Western Electric and the telephone companies of the Bell System—are putting technology to work so that the network will continue to evolve and expand its information-handling capabilities. The object, of course, is to meet the fast-growing and changing needs of Bell System customers.

For information about employment, write:

Bell Laboratories, Room 3C-303, 600 Mountain Avenue, Murray Hill, N.J. 07974

Please include a resume.

An equal opportunity employer.

Bell Laboratories

From Science: Service

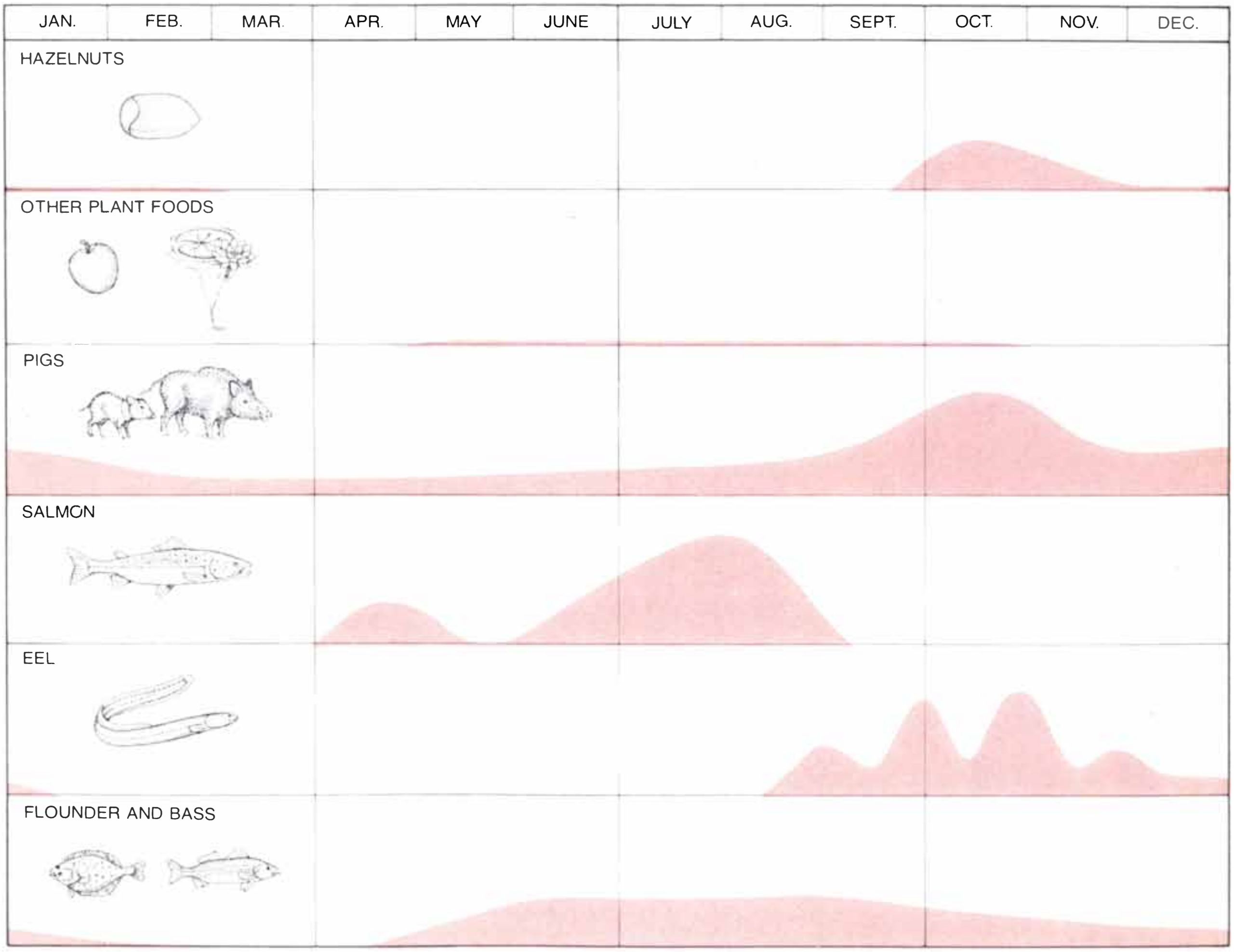
therefore cannot be seen as a totally isolated group. Merely to survive they would necessarily have had to be part of a larger community.

There is evidence to support the existence of some such larger community. For example, some of the flint tools at Mount Sandel, particularly the needle-point microliths and the broad flake adzes, are unknown outside Ireland. Because it takes time for a local stone industry to develop one can hardly view the inhabitants of Mount Sandel as members of an initial colonizing population. Other supporting evidence includes the fact that a number of the microliths unearthed at the site were made from black chert rather than from flint. Black chert is not present in northeastern Ireland but is found farther south, in areas where rocks of the Carboniferous period occur. The lack of any waste chert at Mount Sandel suggests either that these artifacts reached the encampment through exchange networks or that some of the Mount Sandel people visited areas where black chert was available. Whichever was the case the implication is the same: man was already established in Ireland by the time the first hut was built at Mount Sandel.

Since the completion of the Mount Sandel excavations in 1977 Michael Ryan of the National Museum of Ireland has uncovered additional evidence of Mesolithic hunters' early presence in Ireland. At Lough Boora in County Offaly, in what was then the island's supposedly inhospitable interior, Ryan has found a series of small Mesolithic settlements on the shore of a lake now filled with peat. His site has yielded the same range of microliths and other stone artifacts as Mount Sandel and nearly the same range of carbon-14 dates.

Even more recently in County Sligo in the west of Ireland the Carrowmore project, directed by Göran Burenhult of the University of Lund in Sweden, has uncovered tentative evidence of what may prove to be an even earlier Mesolithic presence, perhaps dating back to 7500 B.C. Both of these recent finds, taken together with the discovery of microliths along the east coast of Ireland, have thoroughly overturned the old notion that man first entered Ireland in the extreme northeast after a short sea crossing from Argyll.

Not only the latest evidence but also what the excavations at Mount Sandel have revealed about the economic strategies of Mesolithic hunters and gatherers suggest a picture of Ireland's first settlers totally different from that of hungry strandloopers who were unable to penetrate the island interior and did little more than scavenge a living along shoreline and riverbank. A millennium or more has been added to Ireland's prehistory and, not least, meaningful assemblages of artifacts can now take the place of the wave-rolled flints that for so long represented almost the totality of early Irish material culture.



SEASONAL SUPPLIES OF FOOD in the vicinity of Mount Sandel included hazelnuts in the fall, apples and water lilies in late summer and other food plants starting in late spring. The majority of the pig bones at the site are those of young animals, suggesting that pigs were hunted in late winter. Sea bass and flounder would have been fairly abundant from spring through fall. The salmon run upstream and the eel migration downstream would also have provided food from spring through fall. Storing nuts might have allowed year-round occupancy.

Who will be first to electronically forecast the future?

It could be you and Hughes Support Systems.

In fact, a big part of our business is to simulate tomorrow's possibilities today — and in a variety of ways.

Our logistics support, our computerized training systems, our electronic simulators are all firsts that in a specialized way forecast, plan or prepare for almost every kind of crisis.

So that everyone — from pilots in hazardous skies, to troops at the front, to the technicians who support them — is ready for the future when it arrives.

With Hughes Support Systems, you could be part of state-of-the-art electronics here at home, or part of a field support team that's swept to almost any site on earth. You could be involved in an almost unlimited array of support programs and systems and devices that affect defense avionics, missiles, and electro-optics — and the people who depend on them throughout the free world.

An it's all part of Hughes, with 1,500 projects and a backlog of over $6 billion.

Who will be first with the electronics and support of tomorrow? Who will be first to teach the technology of the future?

It could be you and Hughes.

At Hughes Support Systems, we'll introduce you to people, ideas and jobs that could change your world. And maybe ours.

It could be you and Hughes **Support Systems**

Call or send resume to:

Hughes Support Systems
B130/MS4, Dept. 224
P.O. Box 90515
Los Angeles, CA 90009
(213) 670-1515, ext. 5444

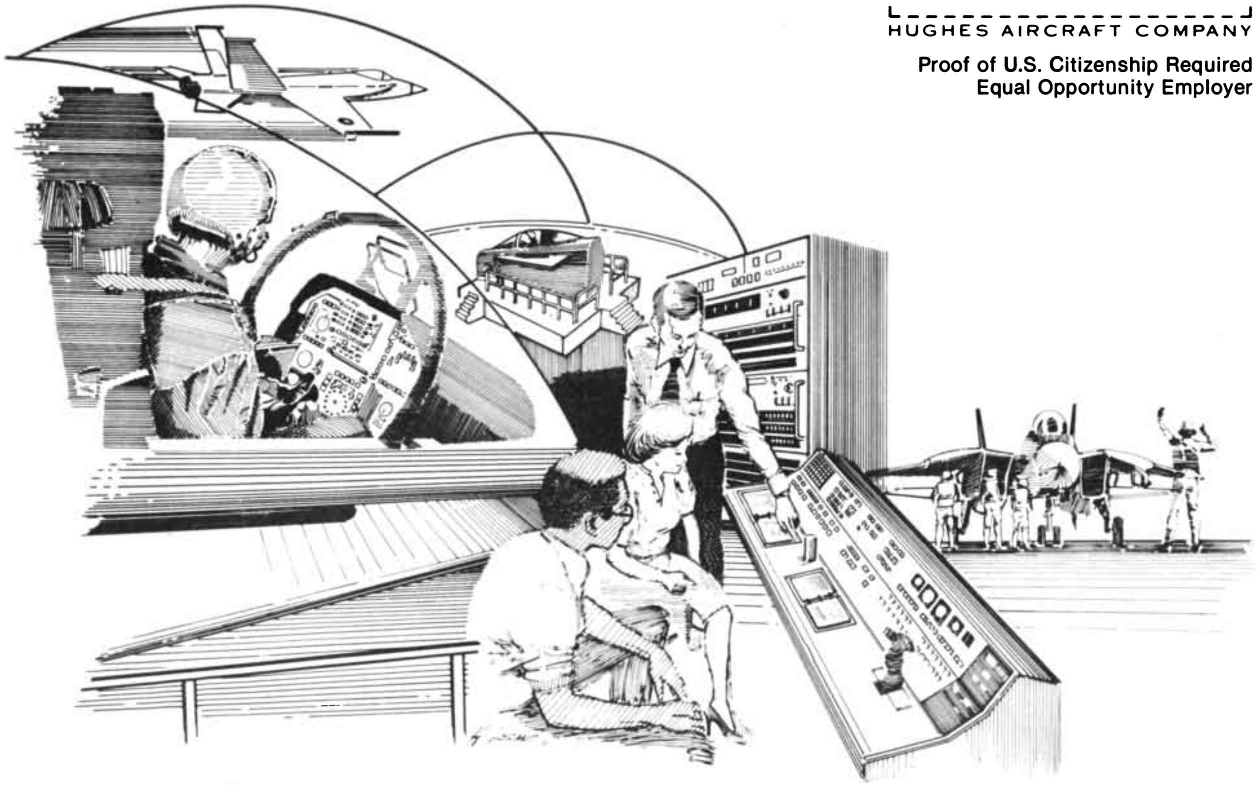
Current openings:

Field Engineering
Product Support
Logistics Support Analysis
Maintainability
Electro-Optical Testing
Real-Time Programming
Analog/Digital Circuit Design
Automatic Test Systems Design
Systems Engineering
Project Engineering
Calibration
Engineering Writing
Training Engineering

HUGHES

HUGHES AIRCRAFT COMPANY

Proof of U.S. Citizenship Required
Equal Opportunity Employer



Sadi Carnot

He is known for his analysis of an ideal heat engine, but his chief interest was the practical application of steam power. Furthermore, the cycle he described is not an ideal one for most modern engines

by S. S. Wilson

Nicolas Léonard Sadi Carnot was a remarkable member of an extraordinary family. His father, Lazare Nicolas Marguerite Carnot, was a distinguished general who served as minister of war under Napoleon and wrote treatises on military strategy, mechanics, geometry and the foundations of the calculus. Sadi Carnot's younger brother, Lazare Hippolyte, was a journalist and a radical politician. Sadi's nephew, Marie François Sadi Carnot, was president of the French Republic from 1887 until 1894, when he was assassinated by an Italian anarchist. Sadi Carnot himself has entered history as the natural philosopher (or what would now be called a physicist) who in 1824 founded the science of thermodynamics with the seminal work *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance* (*Reflections on the Motive Power of Heat and on the Machines Adapted to Develop This Power*). In particular Carnot is thought to have defined an ideal cycle of operation for a heat engine, that is, a cycle with the highest possible efficiency. The thermal efficiency is the ratio of the work output to the heat input; it is expressed in terms of the absolute temperature of the source of heat and the absolute temperature of the component of the engine that rejects the heat.

This view of the substance of Carnot's work is misleading. Carnot's main interests did not lie in the realm that is now proper to the physicist. His analysis of heat engines was intended not as a theoretical treatise but as a popular work to promote the improvement and the wider use of steam engines and other heat engines in France. The cycle Carnot described is not an ideal one for the operation of actual heat engines. Moreover, he did not define the thermal efficiency of the cycle because he did not completely understand the equivalence of heat and work as forms of energy and because the concept of absolute temperature was not known then.

Carnot's work has been misunderstood chiefly because few have read it. Although *Réflexions* was published by Bachelier, the leading scientific publisher in France, and was favorably reviewed, it soon went out of print. It was not translated into English until 1890. Today most engineers and scientists are familiar with Carnot's work not through his own writings but through the mid-19th-century commentaries of Benoît Paul Émile Clapeyron, Rudolf Julius Emanuel Clausius and William Thomson (who later became Lord Kelvin). It was Thomson who expressed the thermal efficiency of Carnot's hypothetical cycle in terms of absolute temperature. The commentaries put too much emphasis on the role of the Carnot cycle as an ideal for heat engines, to the neglect of the rest of Carnot's work, which is full of sound practical advice. I suspect that if the advice had been widely read and followed by Carnot's contemporaries, it would have accelerated the Industrial Revolution in France.

Carnot was born in 1796, near the end of the period in European history that has been designated the Age of Reason. Two years earlier the École Polytechnique had been founded to train engineers for the army. In its first 35 years a succession of brilliant workers in mathematics and the physical sciences were trained there and taught there. Among them were Joseph Louis Lagrange, Jean Baptiste Joseph Fourier, Claude Louis Berthollet, André Marie Ampère, Pierre Louis Dulong, Augustin Louis Cauchy, Charles Bernard Desormes, Nicolas Clément-Desormes, Gaspard Gustave de Coriolis, Siméon Denis Poisson, Joseph Louis Gay-Lussac, Augustin Jean Fresnel, Jean Baptiste Biot, Jean Louis Marie Poiseuille and Clapeyron. Carnot was also trained there.

The work of these investigators contributed to the belief that all aspects of the physical world could be understood through reason if sufficient effort was made to apply the scientific method of theoretical analysis supported by experiment. The scientific method succeeded admirably in mechanics, electricity, optics and the analysis of materials. It was also the foundation of the science of engineering, which displaced the empirical approach to technology that had served well until the early 19th century. In 18th-century Britain in particular skilled millwrights and instrument makers had brought about the Industrial Revolution not by theoretical analysis but by empirical rules of thumb, which resulted in improvements in agriculture, in the harnessing of the power of wind, water and steam, in the construction of canals, railways and steamships, in the smelting of iron with coke and in the mechanization of textile production and of other industries. The Industrial Revolution took hold in Britain long before it came to France, the rest of Europe and the U.S.

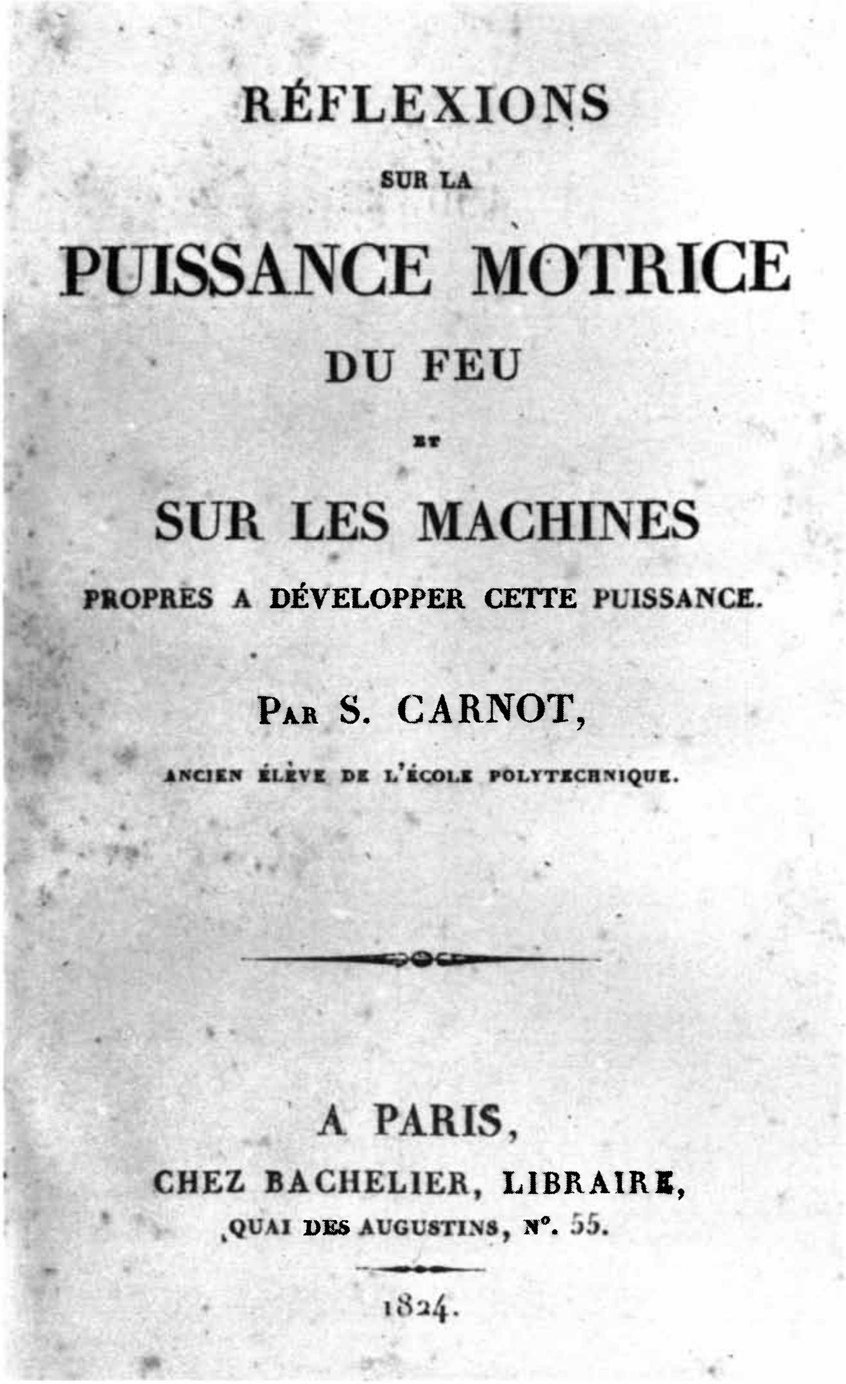
The delay in the industrialization of France stemmed in part from political instability, including the vacillating power of the monarchy and the aristocracy, the rise and fall of Napoleon and the dissolution of the French empire after Napoleon's defeat at Waterloo. Technological factors also contributed to the delay, however. The coal industry was much smaller in France than it was in Britain. The first passenger railways in France were not built until the 1840's, and even then the work was overseen by a British contractor, Thomas Brassey.

Sadi Carnot was born in the Palais du Petit Luxembourg in Paris, where Lazare Carnot lived as a member of the five-man Directoire that governed France between the end of the revolution in 1795 and the rise of Napoleon in 1799. Lazare withdrew from public life in 1807 to teach Sadi and Hippolyte mathematics, physics, languages and music. He enrolled Sadi in the École Polytechnique in 1812, when he was 16, the minimum age for admission.

The following year the armies allied against the French invaded France and Sadi wrote to Napoleon asking that the

PORTRAIT OF SADI CARNOT was painted by Louis Léopold Boilly in 1813, when Carnot was 17. He is wearing the uniform of the École Polytechnique in Paris, a school founded in 1794 to train engineers for the army; he had a brief army career before he took up his work on heat engines. While Carnot was still a student he petitioned Napoleon to allow the students of the École Polytechnique to participate in the defense of the empire, which was under attack by an alliance of several European armies. Napoleon granted the request, and when Paris was besieged in March, 1814, Carnot and his fellow students fought bravely but unsuccessfully on the outskirts of the city.

students be allowed to join in the defense of the empire: "Sir: The country needs all its defenders. The pupils of the École Polytechnique, faithful to their motto, ask to be permitted to hasten to the frontiers to share the glory of the brave men who are consecrating themselves to the safety of France. The battalion, proud of having contributed to the defeat of the enemy, will return to school to cultivate the sciences and prepare for new services." Napoleon granted the request, and in March, 1814, the students fought bravely but in vain to keep the alliance of armies from entering Paris.


RÉFLEXIONS

SUR LA

PUISSANCE MOTRICE

DU FEU

ET

SUR LES MACHINES

PROPRES A DÉVELOPPER CETTE PUISSANCE.

PAR S. CARNOT,

ANCIEN ÉLÈVE DE L'ÉCOLE POLYTECHNIQUE.

A PARIS,

CHEZ BACHELIER, LIBRAIRE,

QUAI DES AUGUSTINS, N°. 55.

1824.

CARNOT'S MAJOR WORK, published in 1824, promoted the construction of steam engines and other heat engines in France, whose industrial development was lagging behind England's. Here the title page is shown; the title is translated *Reflections on the Motive Power of Heat and on the Machines Adapted to Develop This Power.* In *Réflexions* Carnot defined a hypothetical cycle of operation for a heat engine that would generate the most motive power of any cycle. The nature of the heat-engine cycle he described has been largely misunderstood.

Napoleon abdicated in April but returned to power a year later; in the subsequent period, known as the Hundred Days, Lazare Carnot served as minister of the interior. After Napoleon abdicated for the second time following his defeat at Waterloo, Lazare was exiled. Sadi, who had been graduated and commissioned in the army as an engineer and officer, was relegated to routine garrison duty. In 1820 he retired from the army on half pay and returned to Paris. There he studied physics and political economy at the Sorbonne, the Collège de France, the École des Mines and the Conservatoire des Arts et Métiers. He also visited factories and workshops to observe how the principles of economics worked in practice.

Not much is known of Carnot's character. He was described as being reserved and as hating publicity, and he is said to have remarked, "Say little about what you know, nothing about what you do not know." Most of the information about his life comes from a sketch Hippolyte wrote in 1878, some 46 years after Sadi's death. Hippolyte noted that Sadi was named for a medieval Persian poet and moralist, Sa'dī of Shīrāz, whose writings Lazare enjoyed. Sadi was of delicate constitution, according to Hippolyte, but built up his strength through varied exercises. He was endowed with an extreme sensibility as well as great energy. If he thought he was combating injustice, nothing could restrain him.

It is not clear how Carnot became interested in steam engines. Robert Fox of the University of Lancaster, who is the leading historian of Carnot's work, thinks Carnot found his inspiration not in the physics of the time but in the engineering tradition. Carnot had long discussions with the chemists and industrialists Charles Bernard Desormes and Nicolas Clément-Desormes, who delivered a joint talk in 1819 on the theory of the expansive use of steam, a common technology that James Watt had patented in England. The expansive use of steam was much discussed in France, particularly because the Woolf compound engine in which it was employed had been introduced there only a few years earlier. Carnot's attempt in *Réflexions* to deduce the maximum power of a steam engine was indeed timely.

The purpose of *Réflexions* was to bring to public notice the potential of the steam engine for improving the standard of living in France. The work began with a detailed account of the importance of the steam engine, notably in England. Carnot wrote:

"The study of these engines is of the greatest interest, their importance is enormous, their use is continually increasing and they seem destined to produce a great revolution in the civilized world.

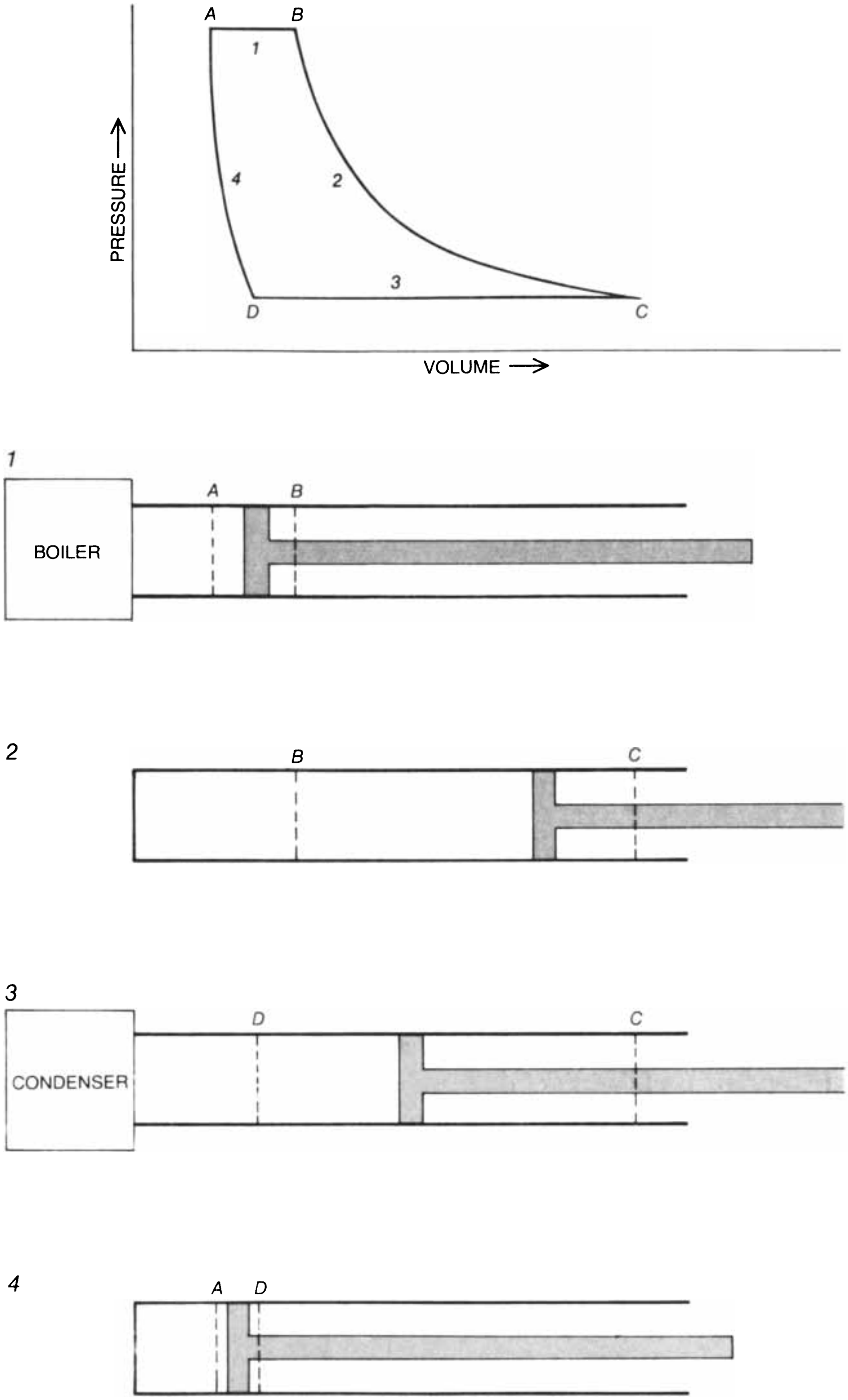
"Already the steam engine works our mines, propels our ships, excavates our ports and our rivers, forges iron, fashions wood, grinds grains, spins and weaves our cloth, transports the heaviest burdens, etc. It appears that it must someday serve as a universal motor and be substituted for animal power, waterfalls and air currents. Over the first of these motors it has the advantage of economy, over the other two the inestimable advantage that it can be used at all times and places without interruption.

"If, someday, the steam engine shall be so perfected that it can be set up and supplied with fuel at small cost, it will combine all desirable qualities, and will afford to the industrial arts a range the extent of which can scarcely be predicted. It is not merely that a powerful and convenient motor that can be procured and carried anywhere is substituted for the motors already in use, but that it causes rapid extension in the arts in which it is applied and can even create entirely new arts.

"The most signal service that the steam engine has rendered to England is undoubtedly the revival of the working of the coal mines, which had declined, and threatened to cease entirely, in consequence of the continually increasing difficulty of drainage and of raising the coal. We should rank second the benefit to iron manufacture, both by the abundant supply of coal substituted for wood just when the latter had begun to grow scarce, and by the powerful machines of all kinds, the use of which the introduction of the steam engine has permitted or facilitated.

"Iron and heat are, as we know, the supporters, the bases, of the mechanic arts. It is doubtful if there be in England a single industrial establishment of which the existence does not depend on the use of these agents, and which does not freely employ them. To take away from England her steam engines would be to take away at the same time her coal and iron. It would be to dry up all her sources of wealth, to ruin all on which her prosperity depends, in short, to annihilate that colossal power. The destruction of her navy, which she considers her strongest defense, would perhaps be less fatal.

"The safe and rapid navigation by steamships may be regarded as an entirely new art due to the steam engine. Already this art has permitted the establishment of prompt and regular communications across the arms of the sea, and on the great rivers of the old and new continents. It has made it possible to traverse savage regions where before we could scarcely penetrate. It has enabled us to carry the fruits of civilization over



CARNOT CYCLE has four stages, each of which is represented by a line in the graph of pressure v. volume and by a diagram of a piston in a cylinder. The letters on the graph correspond to the lettered positions of the piston. In the first stage the cylinder is brought in contact with a heat source, such as a boiler. As the steam in the cylinder expands it draws enough heat from the boiler to maintain a constant pressure and temperature. In the second stage the boiler is removed. As the steam continues to expand the pressure and temperature fall. In the third stage the cylinder is brought in contact with a low-temperature body, such as a condenser. As the steam is condensed it gives up enough heat to the condenser to remain at a constant pressure and temperature. In the fourth stage the condenser is removed. As the steam continues to be compressed its pressure and temperature rise. When the temperature reaches its initial value, the cycle is complete and the engine is in its original condition. No cycle generates more motive power than the Carnot cycle. Carnot did not demonstrate, however, that the cycle has the maximum possible thermal efficiency, that is, the maximum ratio of work output to heat input. The first law of thermodynamics, which expresses the equivalence of work and heat as forms of energy, had not been established then. Evaluating the thermal efficiency of a heat-engine cycle also requires the idea of absolute temperature, which had not yet been developed.

portions of the globe where they would else have been wanting for years. Steam navigation brings nearer together the most distant nations. It tends to unite the nations of the earth as inhabitants of one country. In fact, to lessen the time, the fatigues, the uncertainties and the dangers of travel—is not this the same as greatly to shorten distances?"

Carnot went on to explain the empirical and the historical development of the steam engine. He was aware of the contribution made by each English engineer. Curiously, he ignored the work of his countryman Denis Papin, who invented the steam autoclave, or pressure cooker, and proposed a steam pump. The account continues:

"The discovery of the steam engine owed its birth, like most human inventions, to rude attempts that have been attributed to different persons, while the real author is not certainly known. It is, however, less in the first attempts that the principal discovery consists than in the successive improvements that have brought steam engines to the conditions in which we find them today. There is almost as great a distance between the first apparatus in which the expansive force of steam was displayed and the existing machine as between the first raft that man ever made and the modern vessel.

"If the honor of a discovery belongs to the nation in which it has acquired its growth and all its developments, this honor cannot be here refused to England. Savery, Newcomen, Smeaton, the famous Watt, Woolf, Trevithick and some other English engineers are the veritable creators of the steam engine. It has acquired at their hands all its successive degrees of improvement. Finally, it is natural that an invention should have its birth and especially be developed, be perfected, in that place where its want is most strongly felt."

Carnot's last point was confirmed by subsequent events in the development of the steam engine and other heat engines. While Carnot was writing *Réflexions* engineers in Cornwall, of whom Richard Trevithick was the best-known, were introducing remarkable improvements in the engines that were widely employed there for pumping water out of mines. Thomas Newcomen, Watt and other investigators had this application in mind when they designed steam-driven pumping engines.

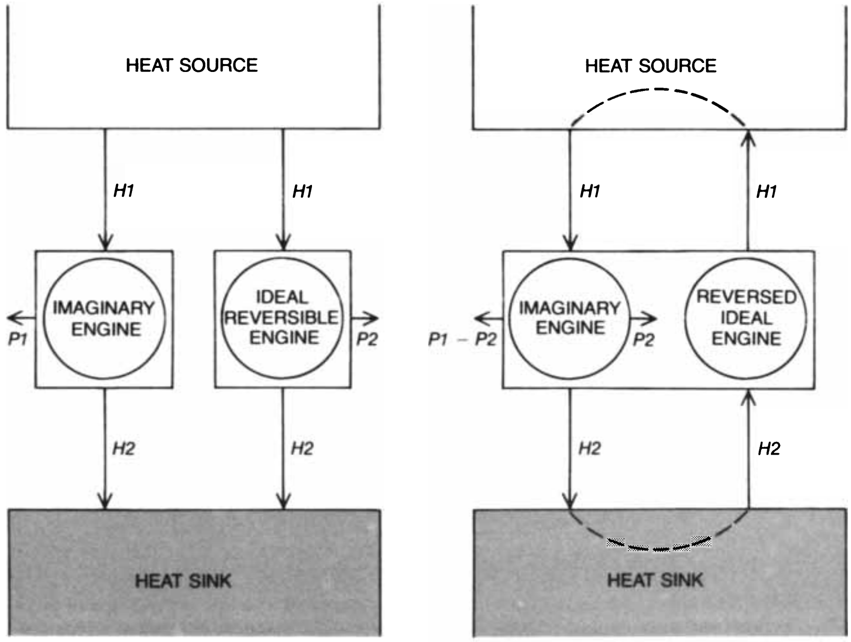
The improvements that led to a particularly successful engine, the Cornish engine, stemmed from several factors, including the need to save coal, which had to be transported to Cornwall, the expiration in 1800 of a patent that Watt and Matthew Boulton had had on a fuel-saving method of condensing steam and the publication beginning in 1810 of *Engine Reporter* by Joel Lean and his relatives. *Engine Reporter* included statistics on how much water the best Cornish engines pumped per bushel of coal. (A bushel of coal is equal to 94 pounds.) The amounts, in millions of foot-pounds of water pumped, rose from 15.7 in 1811 to 20.5 in 1814 to 28 in 1820. The best engine reached 100 million foot-pounds by 1835, and one engine averaged 107 for a month in 1842.

In *Réflexions* Carnot turned next to a general analysis of the power of steam engines and other heat engines. He set out to determine whether there was a limit on the motive power of heat that would prevent heat engines from being indefinitely improved. He also wondered whether some working substance such as alcohol or air might be preferable to steam. Other investigators had examined these questions, but only with respect to a particular design. Carnot set out to find answers that would be applicable to all imaginable heat engines.

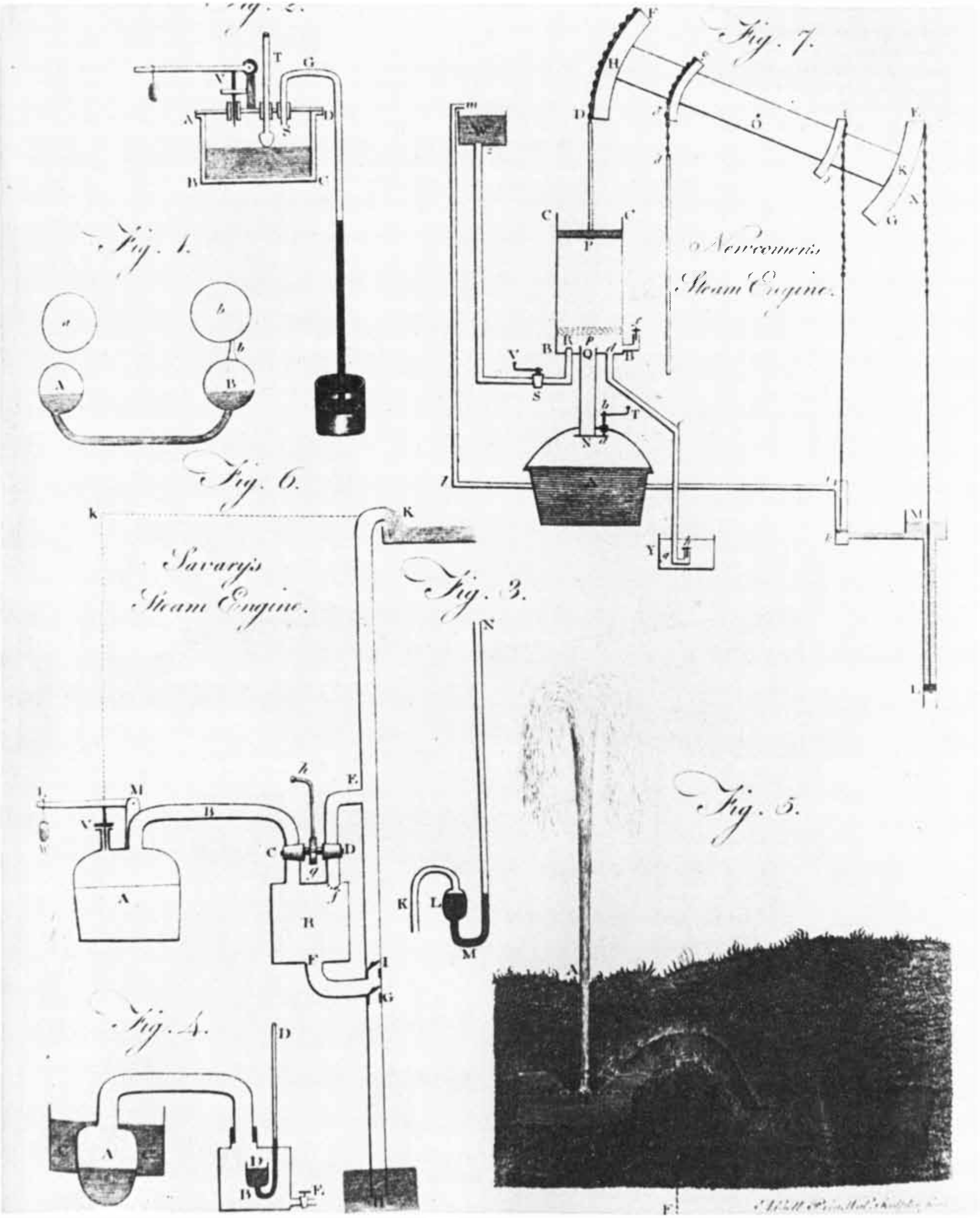
Carnot was able to address these questions even though the laws of thermodynamics were not known then. The first law, which states that energy is conserved in a thermodynamic system, was not established until the work of James Prescott Joule in the 1840's. Fundamental to the first law is the equivalence of heat and work as forms of energy. Such an equivalence ran counter to the caloric theory of heat, which in spite of some opposition was the prevailing theory when Carnot wrote *Réflexions*.

According to the caloric theory, heat is an imponderable, massless fluid that is always conserved. It is testimony to Carnot's genius that he could draw correct and far-reaching conclusions about heat engines without assuming that heat can be converted into mechanical work and mechanical work into heat. Without knowing the first law of thermodynamics he could not have defined the thermal efficiency of an engine, as he is commonly thought to have done. He defined only the maximum "duty" of an engine: its useful output (how much water is lifted multiplied by how high it is lifted) for a given quantity of coal.

Carnot assumed that the motive power of a heat engine results from the transfer of caloric from a hot body (a heat reservoir, or boiler) to a cold body (a heat sink, or condenser). Lazare had



THERMODYNAMIC REVERSIBILITY was the basis of Carnot's proof that no cycle could have a higher output than his cycle. In *Réflexions* Carnot considered heat to be an indestructible fluid known as caloric. Ordinarily caloric flows from a hot body to a cooler one; if the cycle could be reversed, caloric would be transported from a low-temperature body to a high-temperature one, and the cycle would consume as much motive power as it generates when it is run forward. The proof begins by supposing there is a cycle that could yield more motive power. At the left is a diagram of the imaginary cycle and of the ideal reversible cycle operating between a heat source and a heat sink. *P1* is the motive power of the imaginary cycle and *P2* is that of the reversible cycle; *H1* and *H2* represent the flow of heat. At the right is a diagram of the imaginary cycle driving the ideal cycle in reverse. Because *P1* is greater than *P2* there is power left over to do external work and the system is a perpetual-motion machine. The possibility of perpetual motion had long been ruled out (by Carnot's father for one) and so Carnot concluded that the supposition of a cycle more powerful than the ideal reversible cycle was incorrect.

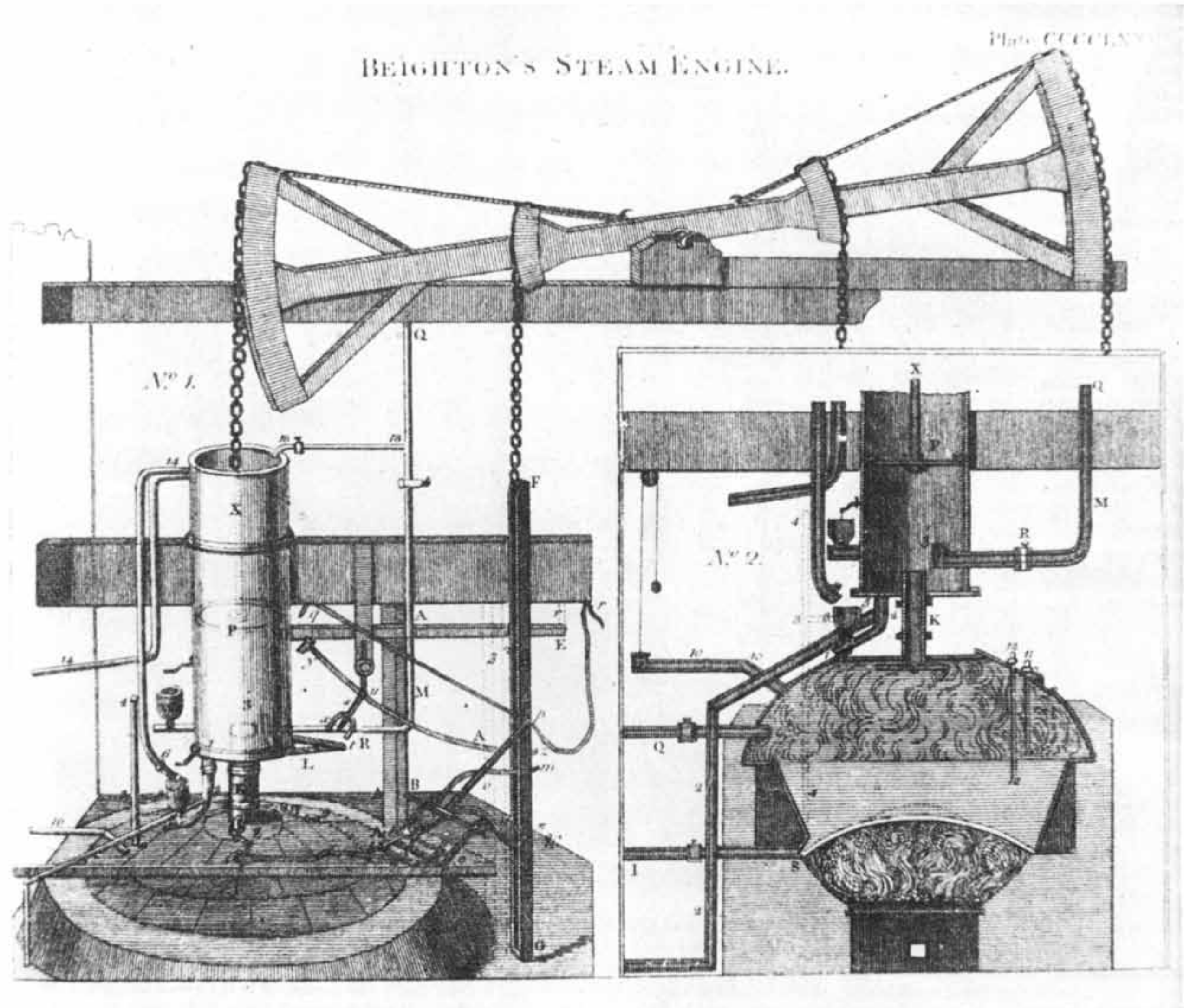


STEAM-DRIVEN PUMPING ENGINES designed by Thomas Savery and Thomas Newcomen were diagrammed in 1797 in the third edition of the *Encyclopaedia Britannica.* In 1698 Savery had received a patent for a "new Invention for Raiseing of Water and occasioning Motion to all Sorts of Mill Work by the Impellent Force of Fire, which will be of great use and Advantage for Drayning Mines, Serveing Towns with Water, and for the Working of all Sorts of Mills where they have not the Benefitt of Water nor constant Windes." The diagram shows an improved version of the Savery engine, which pumps water by means of suction created by the condensation of steam in a closed vessel. In 1712 Newcomen developed the first piston-operated steam engine. The steam was generated in the vessel under the cylinder and entered the cylinder through valve *N.* In order to make the steam condense, cold water was sprayed into the cylinder through valve *S.* In *Réflexions* Carnot cites the *Encyclopaedia Britannica* as a source of information on various kinds of steam-driven engines.

published a treatise in which he analyzed the efficiency of water-driven machines, and in *Réflexions* Sadi established an analogy between the heat engine and the waterwheel. The work done by a heat engine is the result of the passage of caloric through a difference in temperature, as the work done by a waterwheel is the result of the passage of water through a difference in gravitational potential. In the waterwheel the motive power depends on the amount of water and on the distance the water falls. In the heat engine the motive power depends on the amount of caloric and on the magnitude of the temperature difference through which the caloric moves.

Carnot described an ideal cycle for a heat engine consisting of a cylinder, a piston, a working substance, a heat source and a heat sink. The working substance, which can be any material (a solid, a liquid or a gas) that can be made to expand and contract by successive applications of heat and cold, is placed in the cylinder, into which the piston is snugly fitted. The cycle has four stages. First, the working substance is brought in contact with the heat source, so that as the substance expands it withdraws enough heat from the reservoir to maintain a constant temperature. Second, the working substance is isolated from the heat reservoir, with the result that as it continues to expand its temperature falls. Third, the working substance is brought in contact with the heat sink, so that as the substance is compressed it gives up enough heat to maintain a constant temperature. Fourth, the working substance is isolated from the heat sink, and so its temperature rises as it continues to be compressed. When the temperature returns to its initial value, the cycle is complete and the working substance and the engine have been restored to their original condition. The cycle can be repeated indefinitely to produce continuous power. In modern terminology the cycle consists of an isothermal expansion (one carried out at constant temperature), an adiabatic expansion (in which heat is converted into work), an isothermal compression and an adiabatic compression.

In introducing the ideal cycle Carnot wrote: "Imagine two bodies *A* and *B*, kept each at a constant temperature, that of *A* being higher than that of *B*. These two bodies, to which we can give or from which we can remove heat without causing their temperature to vary, exercise the functions of two unlimited reservoirs of caloric. We shall call the first the furnace and the second the refrigerator."



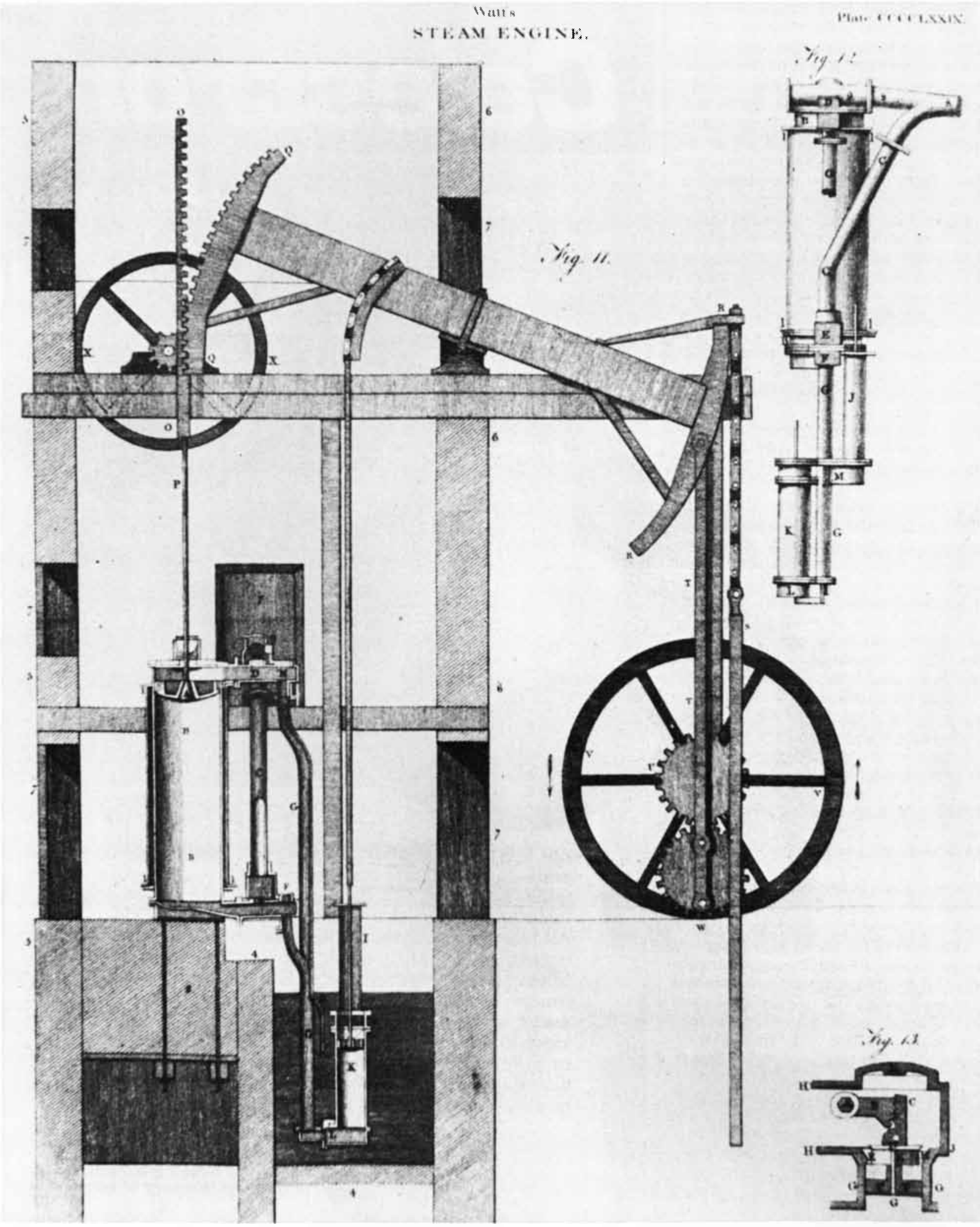
ATMOSPHERIC ENGINE investigated by Henry Beighton is also shown in an illustration from the 1797 *Britannica*. Beighton made the first quantitative analysis of the motive power of a steam engine. He calculated the depths from which engines of various cylinder sizes could draw water out of a mine. In 1721 Beighton's calculations were published in *The Ladies' Diary*, of which he was the editor.

In an actual heat engine the heat source is almost always a finite one, so that as heat is expended the temperature of the source falls. As a result the majority of practical heat engines have a heat-reception process that is not isothermal. This is true of all internal-combustion engines and gas turbines. Even in the steam engine the development has been away from the almost completely isothermal heat-reception process of the early engines. Instead modern designers favor the high-pressure steam cycle, in which the isothermal part of the heat reception is small, and the supercritical steam cycle, which has no isothermal heat reception. The idea of an infinite heat sink is more realistic than that of an infinite heat source: there is usually an unlimited supply of air or water for cooling. Therefore a cycle based on isothermal heat rejection is feasible, and indeed it is achieved by most steam cycles that condense the steam at a constant temperature.

With his cycle Carnot introduced a fundamental thermodynamic concept, that of reversibility. Each repetition of the cycle implies the flow of caloric from a high-temperature body to a low-temperature one, with the attendant production of work. Since caloric was thought to be an indestructible fluid, Carnot reasoned that each stage of the cycle could be reversed by doing work on the engine, so that the caloric is transported from the low-temperature body to the high-temperature one. In the analogous waterwheel the process could be reversed by doing work on the water, so that it is lifted through the gravitational potential difference back to its source at the top of the wheel. Since each stage of Carnot's cycle could be reversed, the entire cycle is reversible. When it is run in reverse, it consumes as much motive power as it generates when it is run forward.

The hub of the theoretical discussion in *Réflexions* is Carnot's demonstration that for a given temperature difference and a given amount of caloric no cycle can generate more motive power than his ideal, reversible cycle. Imagine a cycle that could generate more power. This cycle could be harnessed to drive the ideal cycle in reverse. To be restored to its original state the ideal cycle would need to consume only part of the motive power produced by the imaginary cycle. The rest of the motive power would be available to do external work. To put it another way, if one could connect the ideal reversible cycle to a cycle that was more efficient, one would have a perpetual-motion machine.

The possibility of a perpetual-motion machine had long been rejected by most investigators, including Lazare Carnot, who wrote on the subject. Thus Sadi



WATT ENGINE, developed in 1765 by James Watt, was in essence a Newcomen engine in which the steam was condensed not in the cylinder but in a separate vessel. The introduction of a separate condenser was of the first importance because heat was no longer wasted in the alternate heating and cooling of the cylinder; the cylinder and the piston remained at the temperature of the steam. The separate condenser resulted in a fuel saving of almost 75 percent. In the early 1780's Watt made the first engine that worked on both the upstroke and the downstroke of the piston. He also made one of the first engines that converted reciprocating motion into rotary motion. In the engine in the illustration the conversion was accomplished by a gear system in which a planet gear (*W*) at the end of a rod (*T*) orbits a sun gear (*U*). The sun gear, whose motion is smoothed by a flywheel (*V*), provides rotary motion. The illustration is from the 1797 *Britannica.*

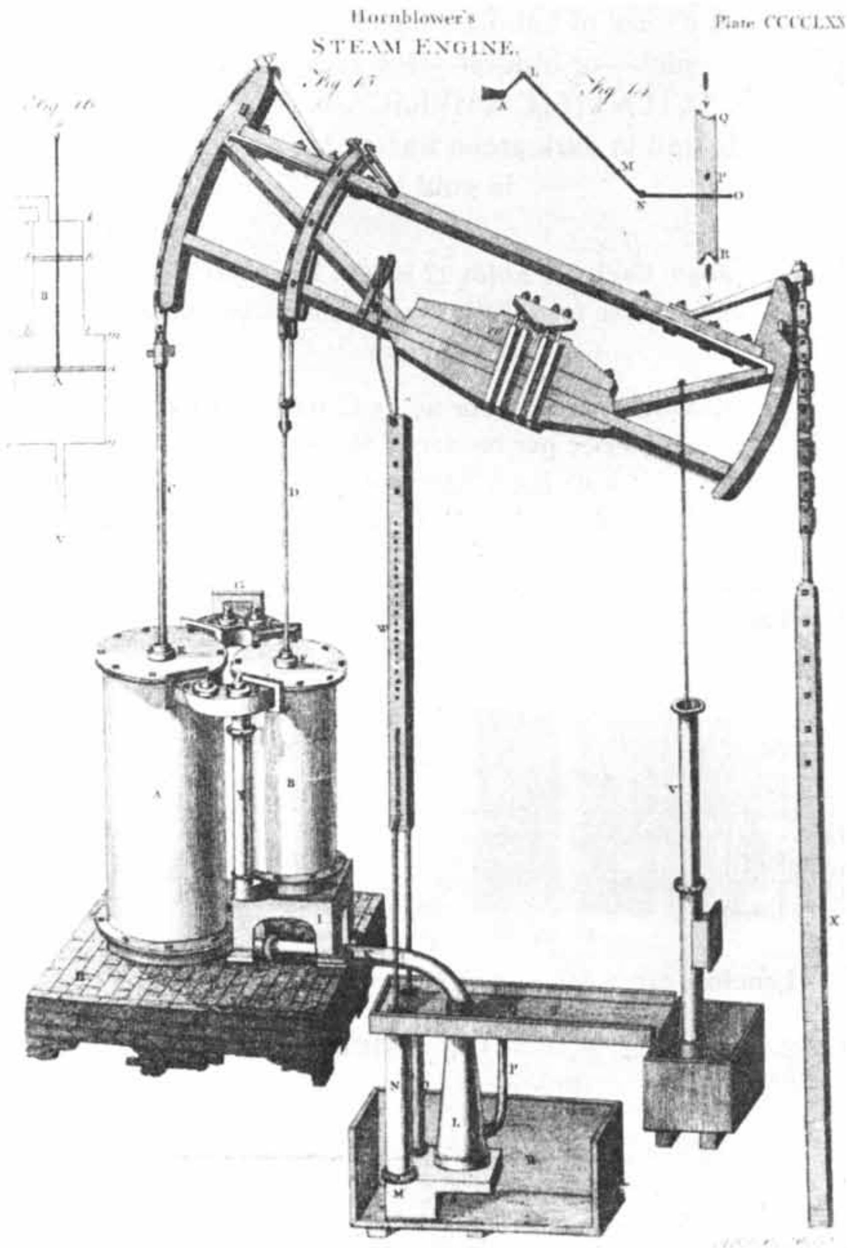
Carnot reasoned that the initial assumption of a cycle that could generate more motive power than the ideal, reversible cycle was incorrect. With this demonstration Carnot in essence expressed the second law of thermodynamics, which was not stated explicitly until the work of Clausius and Thomson in the 1850's. According to this law, no process is possible whose sole result is the absorption of heat from a reservoir and the conversion of this heat into work. The law indicates that a spontaneous thermodynamic process can proceed in one direction only, from a high-temperature source to a low-temperature sink. Energy from an external source is needed to reverse the process.

Carnot demonstrated that the power of the ideal cycle does not depend on the particular working substance, although the nature of the substance is important from a practical point of view in that its properties affect the design of the engine and the temperature at which it can operate. He concluded that the best working substance is the one that expands the most for a given change in temperature. By this standard gases are better prospects than either solids or liquids.

Carnot recognized the potential advantage of air as the working substance. With air there is the possibility of internal combustion, which eliminates the need for the boiler or some other form of heat exchanger. Furthermore, there is no need for a condenser because the air can be discharged into the atmosphere. Carnot knew (and approved of the fact) that such a discharge process was already being exploited, not in air engines but in the high-pressure steam engines pioneered by Trevithick. The absence of a condenser made possible the development of a compact, portable engine and hence of the steam locomotive. Carnot realized the mechanism that compresses the air in an air engine would have to be more elaborate and would consume more power than the corresponding mechanism in a steam engine. Water is virtually incompressible and occupies a small volume compared with the volume of the equivalent mass of steam, and therefore little power is needed to drive the pump that feeds water into the boiler.

Carnot's ideas on air as a working substance foreshadowed the development of the internal-combustion engine. "The use of atmospheric air for the development of the motive power of heat," he wrote, "presents in practice very great but perhaps not insurmountable difficulties. If we should succeed in overcoming them, it would doubtless offer a notable advantage over vapor of water." Since the 1790's combustible gases and powders had been tried as fuels, but it was not until 1859 that an internal-combustion engine, made by Jean Joseph Étienne Lenoir, ran smoothly, and it was another 17 years before the first commercially successful internal-combustion engine was developed by Nikolaus August Otto. Carnot's predictions, both of the difficulties and of the potential advantages of internal combustion, were borne out.

Carnot also foresaw the possibility of igniting the fuel of an internal-combustion engine by compression. In 1891 Rudolf Diesel conceived of an engine working on a Carnot cycle with compression ignition. He hoped to achieve isothermal combustion by controlling the rate at which the fuel was injected. Nevertheless, the first engine he built in 1898 worked on a different cycle, as have all subsequent diesel engines, in which there is no isothermal process at all. The history of the diesel engine confirms the idea that the Carnot cycle is not the basis of an actual engine but rather is a hypothetical cycle with which one can study thermodynamic concepts such as the maximum motive power of an engine.



TWO-STAGE-EXPANSION ENGINE, developed in 1781 by Jonathan Hornblower, consisted of two cylinders. After the steam expanded in one cylinder it was exhausted into the other cylinder so that it could expand further. In principle the two-stage expansion would generate more motive power than the one-stage expansion. In practice the two cylinders were an unnecessary complication because the initial steam pressure was too low. Engines were later developed that effectively employed a compound expansion. Illustration is from the 1797 *Britannica*.

Carnot also outlined the case for a compound cycle. "We might conceive even the possibility of making the same heat act successively upon air and vapor of water. It would only be necessary that the air have, after its use, an elevated temperature, and instead of throwing it out immediately into the atmosphere, to make it envelop a steam boiler, as if it issued directly from a furnace." Such a compound cycle was exploited in the 1920's in the Still engine, a diesel engine that incorporated an exhaust boiler and a reciprocating steam engine. The Still engine was tested in ships and locomotives but failed to catch on. More recently compound cycles have been introduced that combine a gas turbine with a turbine run by steam or another vapor. Such compound cycles are gradually becoming established, some 150 years after Carnot proposed them.

Carnot considered vapors other than steam, but he dismissed them because of practical difficulties and because they offered no theoretical thermodynamic advantage. Indeed, few engines have successfully employed vapors other than steam and none has done so on a large scale. Currently, however, there are efforts to drive a heat engine with a vapor that has a high molecular weight; in a small turbine such a working substance offers practical advantages over steam.

Although Carnot has been regarded as the founder of the science of thermodynamics, he was first and foremost an engineer, and he was acutely aware of the compromises that have to be made in the design and construction of an engine. His *Réflexions* ends on this note:

"The economy of the combustible is only one of the conditions to be fulfilled in heat engines. In many cases it is only secondary. It should often give precedence to safety, to strength, to the durability of the engine, to the small space it must occupy, to small cost of installation, etc. To know how to appreciate in each case, at their true value, the considerations of convenience and economy that may present themselves; to know how to discern the more important of those that are only secondary; to balance them properly against one another in order to attain the best results by the simplest means—such should be the leading characteristics of the man called to direct, to coordinate the labors of his fellow men, to make them cooperate toward a useful end, whatsoever it may be."

Little is known about what Carnot did after he wrote *Réflexions*. In 1828 he worked for a short time as a military engineer in Lyons and Auxonne, and then he returned to Paris to resume his study of engines and the theory of heat. A history of the École Polytechnique published at the time describes Carnot as a builder of steam engines, but there is no other evidence that he actually constructed engines.

Carnot approved of the July Revolution of 1830, which forced Charles X to abdicate, because it seemed to promise a more liberal regime. Indeed, it was suggested that Carnot would be asked to join the new government as a member of the Chambre des Pairs. Nothing came of the suggestion, however, because Carnot quickly grew disenchanted with the government, which restored the monarchy, and he objected to the hereditary nature of the position that had been proposed for him.

Soon thereafter Carnot began to analyze the relation between the temperature and the pressure of a gas, but in June, 1832, he contracted scarlet fever and then "brain fever." On August 24, at age 36, he died in a cholera epidemic. As was commonly done with victims of cholera, he was buried with virtually all his papers and other belongings.

A few fragments of Carnot's notebooks have survived. In 1878 Hippolyte turned them over to the French Academy of Sciences along with the manuscript of *Réflexions.* One fragment, apparently written between 1824 and 1826, indicates that Sadi had come to reject the caloric theory of heat and to understand the equivalence of heat and work. "Heat is simply motive power, or rather motion that has changed form," he wrote. "It is a movement among the particles of bodies. Wherever there is destruction of motive power there is, at the same time, production of heat in quantity exactly proportional to the quantity of motive power destroyed. Reciprocally, wherever there is destruction of heat there is production of motive power." He even outlined experiments that would demonstrate the equivalence of heat and work as forms of energy, but he never carried them out. Some of the experiments were similar to the ones Joule carried out some 15 years later in establishing the first law of thermodynamics.

It is interesting, if idle, to speculate on what Carnot might have achieved if he had lived longer. Would his ideas on heat engines have been put into practice? Or would he have died an unheeded prophet? It is not too late to pay tribute to his foresight and to his analytical power, which had a firm grasp on practical realities. The need to keep human concerns in the forefront of scientific work and the recognition of the empirical nature of technological progress are lessons the modern investigator should keep in mind.

THE AMATEUR SCIENTIST

Interference patterns made by motes on dusty mirrors

by Jearl Walker

Two beautiful examples of optical interference can be observed if you look into a mirror covered with dust or mist as you hold a small light in front of your face. A burning match works well. What you will see are colored rings centered around the reflected image of the light. Another set of interference rings and colors overlaps the first but may not be centered at the same place. This additional pattern is more difficult to see because your eye and the source of the light must be nearly on the same line perpendicular to the mirror. With careful adjustment of the position of the source and your eye the second set of rings can be made to coincide with the first. Under some circumstances the colors in these patterns are stunning.

Both patterns require a mirror with the reflecting surface on the back, which is the case with ordinary glass mirrors. The pattern that is always centered on the reflected image of the light source is usually called the Fraunhofer pattern. It has been thoroughly studied, not only because it affords an excellent demonstration of optical diffraction but also because the interference is responsible for the colored ring occasionally seen close to the moon or the sun.

The other interference pattern that can be seen in a dusty or misty mirror was studied intensely in the 19th century but has been virtually forgotten since then. I shall begin my story with this less familiar interference. Then I shall return to the Fraunhofer pattern and the lunar or solar ring.

Isaac Newton was apparently the first to study colored interference patterns in a dusty mirror. He made a shaft of sunlight fall on a pinhole in a screen in front of a hemispherical concave mirror. Light passing through the pinhole was reflected from the mirror and returned to the pinhole because it was positioned at the center of curvature of the mirror. The interference pattern appeared on the screen around the pinhole.

The irony of much of Newton's work in optics was that he discounted the possibility that light consisted of waves. Holding fast to the notion that light consisted of particles, he failed to explain any of the demonstrations of interference he investigated, including the one with the dusty mirror. The proper understanding did not emerge until Thomas Young convincingly demonstrated the wave nature of light. He explained briefly the colors that could be seen in Newton's dusty mirror. Later work on the pattern was done with flat mirrors by William Whewell in England and by Lambert Adolphe Jacques Quételet in Belgium; as a result the pattern is now known as the Whewell-Quételet interference pattern.

Almost any small particles—room dust, chalk dust, lycopodium powder—on a mirror can generate the Whewell-Quételet pattern. The moisture that condenses when you breathe on the mirror will do. You can also create a thin layer of dried milk globules. Mix one part of milk with three parts of water and coat the front surface of a mirror. Tilt the mirror so that the excess liquid runs off. Allow the remainder to dry. The milk globules left on the mirror serve in the same way as the dust particles. The dusty material does not have to be directly on the mirror. Sir John Herschel demonstrated the interference pattern by tossing wig powder into the air in front of a mirror.

The interference pattern depends on the scattering of light by the particles on the mirror or just in front of it. For the sake of simplicity I shall assume that the observer and the light source are on the same perpendicular line from the mirror and that the light from the source consists of only one wavelength. Later I shall discuss interference arising from white light and explain how the colors are dispersed.

Suppose two rays travel to the mirror from the source as part of one wave front. They are coherent (the light waves of the two rays are in phase) and nearly parallel. Although the light source is incoherent (unless it is a laser), it does emit coherent waves in short bursts over short distances.

One ray scatters from a particle toward the back of the mirror, is reflected and travels to the observer. The other ray enters the mirror without scattering from the particle, is reflected, then scatters from the same particle and travels to the observer. Although the rays were initially in phase, the phase relation now depends on the lengths of the paths they have taken. If the paths are exactly the same length, the rays are once more in phase and they interfere constructively. The observer sees relatively bright light from the position of the particle on the mirror. When the observer and the light source are on the same perpendicular line from the mirror, this situation results with a particle directly in front of the observer, that is, at the foot of the perpendicular line.

A particle slightly off the line sends the observer rays with a different phase relation. If the difference in path lengths results in a phase difference of half a wavelength, the rays interfere destructively. The observer sees a dark spot. Moreover, because of a circle of dust particles around the perpendicular line where the geometry is correct for destructive interference the observer also sees a dark ring around the perpendicular line. It is the centermost of the dark rings in the interference pattern.

Particles a bit farther from the center of the pattern scatter pairs of rays that end up a full wavelength out of phase when they reach the observer. The result is constructive interference and a bright ring. More bright and dark rings are seen somewhat farther from the line. They are caused by progressively larger differences in the path lengths of pairs of light rays.

Interference patterns are normally labeled according to the phase difference between the participating rays. The bright place at which the phase difference is zero is labeled $n = 0$ and is called the zero-order fringe of constructive interference. In the geometry of observer and light source I have described, this fringe is at the foot of the perpendicular line. The next bright fringe, where the phase difference is a full wavelength, is labeled $n = 1$ and is called the first-order fringe of constructive interference. It is the central bright ring in the interference pattern.

In nearly all standard examples of interference the numerical labels for the bright fringes increase with distance from the center of the pattern. The interference pattern from the dusty mirror is an exception, at least in principle. If the bright rings still farther from the center of the pattern could be seen, their order would begin to decrease. The reason is that when two rays return to the observer from a point relatively far from the perpendicular line, they have traveled almost equal distances. The phase difference between them is therefore smaller than the difference between the rays of another pair scattering from a particle somewhat closer to the line. Unfortunately the rings where the numerical sequence reverses cannot be observed in an ordinary mirror.

The width of the interference rings also varies with distance from the center of the pattern, narrowing at first and then (too far from the center to be observed normally) widening. The width of a ring for any particular order of interference depends on several factors. The ring is larger for a longer wavelength, for a larger index of refraction of the glass in the mirror and for a thinner mirror. The width also depends on the relation between the distance of the light source from the mirror and the distance of the observer from the mirror.

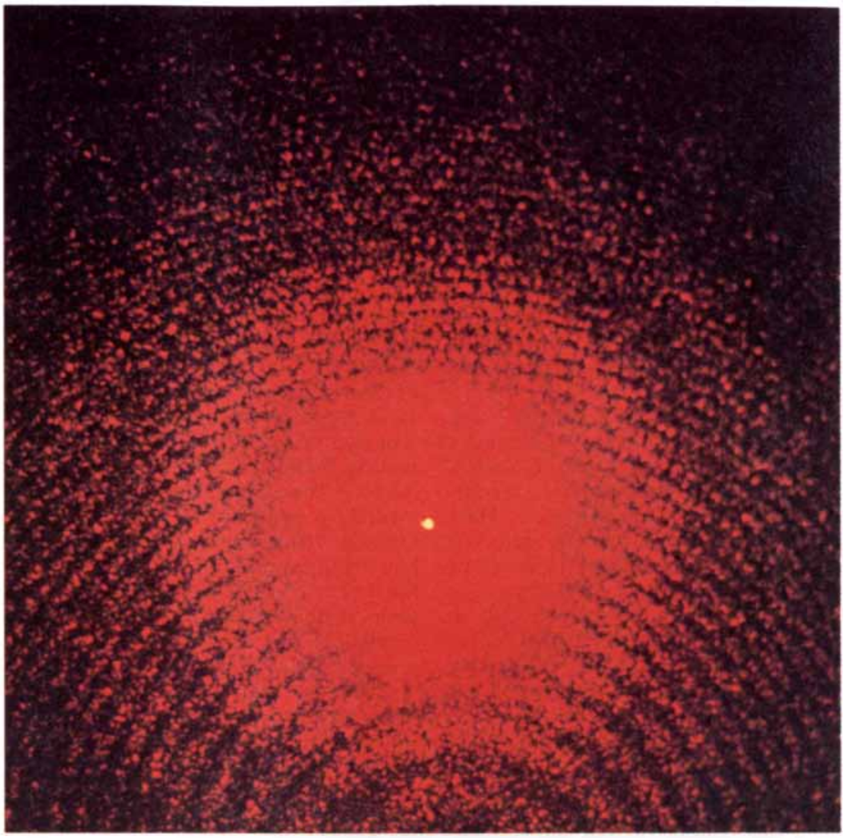
Why do a pair of rays participating in the creation of an interference pattern scatter from the same particle? A ray from one particle could interfere with a ray from another, but the interference would not contribute to a composite pattern. The scattering of light from a particle causes a shift in the phase of the light that depends on the size and shape of the particle. If two rays scatter from the same particle, the phase shift is the same. The phase difference then depends only on the geometric relations of the light source, the observer and the particle.

The type of scattering from a particle on the mirror depends on the size of the particle. Particles larger than about 100 micrometers scatter light through either reflection or diffraction. Smaller particles, such as those of lycopodium powder, scatter light through a more complex process that involves surface waves. The Mie theory of scattering provides the basis for a detailed understanding of this scattering. If the particles are quite close to one another on the mirror, the interference between rays that are diffracting through the space between particles becomes important. I shall not consider such complications here.

In white light the rings are colored. Each interference order has a sequence of colors from blue to red, ranging outward from the center of the pattern. The separation of colors arises because the phase difference between a pair of rays participating in the interference depends on the wavelength.

Suppose a particle slightly off the center of the pattern causes constructive interference of blue light in the innermost bright ring ($n = 1$). Two rays contributing to this ring travel along a path difference equivalent to one full wavelength of blue light. For the same phase difference in red light (a longer wavelength) a pair of red rays must scatter from a particle a bit farther from the center, so that their path difference is equivalent to a full wavelength of red light. Hence the ring is blue on the inside and red on the outside. Intermediate colors lie between the blue and the red.

I have described the general interfer-



A Fraunhofer pattern **(top)** *and a Whewell-Quételet pattern* **(bottom)** *in laser light*

A Fraunhofer pattern from lycopodium powder in white light ence pattern that is seen when the observer and the light source are on the same perpendicular line. If the observer moves, the center of the pattern moves away from the directly reflected image of the light source and is either bright or dark, depending on the displacement. As the observer moves, new rings seem to emerge from the center. The narrower rings of higher order can be seen near the reflected image of the source. Eventually the center of the pattern is so displaced that it disappears. If the observer moves far enough to the side, the entire pattern becomes obscure and disappears.

The structure of the higher-order fringes has been studied by employing a thin layer of mica as a mirror. Such a layer, split from a thicker sheet, is coated on one side so that it reflects. Then the interference pattern can be seen even if both the light source and the observer are at highly oblique angles.

The interference pattern can also be seen with mirrors that are coated for reflection from the front surface. Lightly dust an uncoated sheet of glass and put it in front of the mirror with the dusted side facing the reflecting surface. Insert thin strips of mica or paper between the mirror and the glass as separators. (Without separation a pair of rays normally responsible for the interference would travel the same distance regardless of the location of the dust particle on the glass, and there would be no interference pattern.)

To investigate the interference pattern I held a small flame between one eye and a mirror I misted by breathing on it. The reflective surface was on the back of the mirror. With the room lights off I could see both the Fraunhofer pattern and the Whewell-Quételet pattern. As I moved the flame closer to my eye the rings of the Whewell-Quételet pattern became less curved. With the flame close to my eye they were almost straight lines. When a friend held the flame behind me (still positioned so that my head did not block all the light), the rings again curved.

I then sprinkled a light coating of lycopodium powder over the mirror, gently shaking the mirror to achieve as even a distribution as I could. (If the layer is too thick, no pattern will be seen.) The mirror was illuminated with an incandescent lamp or a sodium lamp placed 10 meters away so that it approximated a point source. (I could have inserted a pinhole between a closer source and the mirror.) The Fraunhofer rings were immediately evident, but the Whewell-Quételet fringes were harder to find. I had to put my head almost in the path of the beam of light before I could see them. A helium-neon laser gave similar results.

To more easily observe the Whewell-Quételet pattern I placed an uncoated sheet of glass in the path of the light falling on the mirror, angling it to cast a reflection of the mirror perpendicular to the beam. When I looked into the glass, I could see both interference patterns on the mirror. By carefully adjusting the angles of the mirror and the uncoated sheet I got the same view of the mirror that I would have if I had been in line with the light source.

Such an arrangement was first employed by Eugen Lommel in Germany 100 years ago to project an image of the interference pattern onto a screen by means of a convex lens. The arrangement will also serve in making a photograph of the interference pattern. The camera lens is adjusted for a large focal length because in principle the rays reflected from the mirror converge to form a real image of the Whewell-Quételet pattern only at a considerable distance from the mirror. With the lens in place the image is brought into focus on the film.

If you use a laser as the light source, be extremely circumspect about the light reflected from the mirror and from the sheet of glass inserted into the beam. Both reflections can be bright enough to hurt your eye.

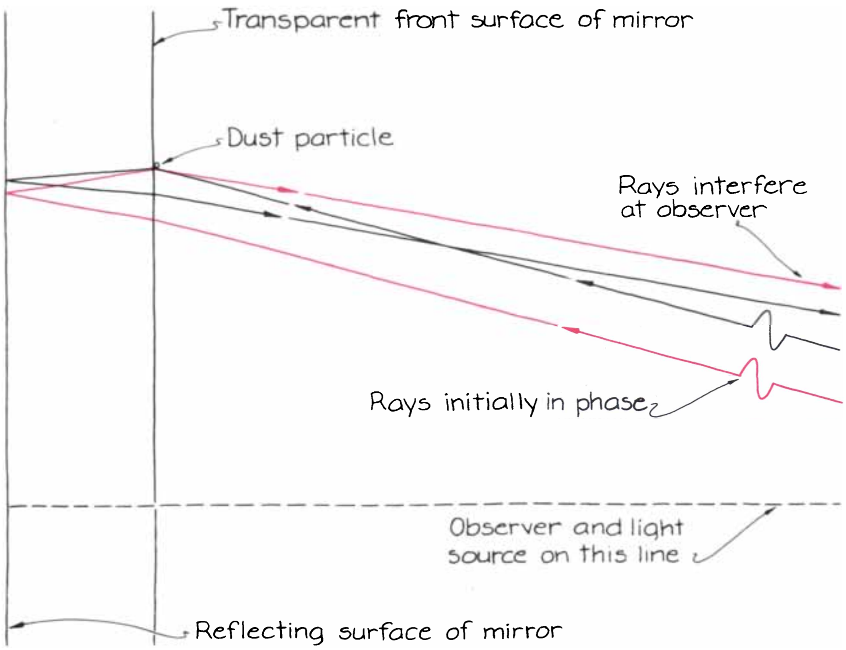
The other major principle of interference—the Fraunhofer pattern—is produced by the diffraction of light by the dust particles on the surface of a mirror coated for reflection from the back surface. To explain what is seen I shall first describe a simpler demonstration of diffraction by small particles. If you look at a small light source through a misty window, the image of the source is surrounded by concentric interference rings. They are Fraunhofer rings resulting from the diffraction of light rays by the water droplets on the window.

When a beam of light passes a small particle such as a dust particle or a water droplet, the light diffracts around the sides of the particle. The diffracted rays create an interference pattern when they are intercepted by a screen after passing the particle. If the particle presents to the light an approximately circular cross section, the pattern consists of concentric bright rings surrounding a relatively bright central spot and interspersed with relatively dark rings.

The bright rings are created when rays of light diffracting around the particle interfere constructively. The dark rings are caused by destructive interference. The kind of interference depends on the geometry of the rays. The difference in their path lengths causes phase differences between them.

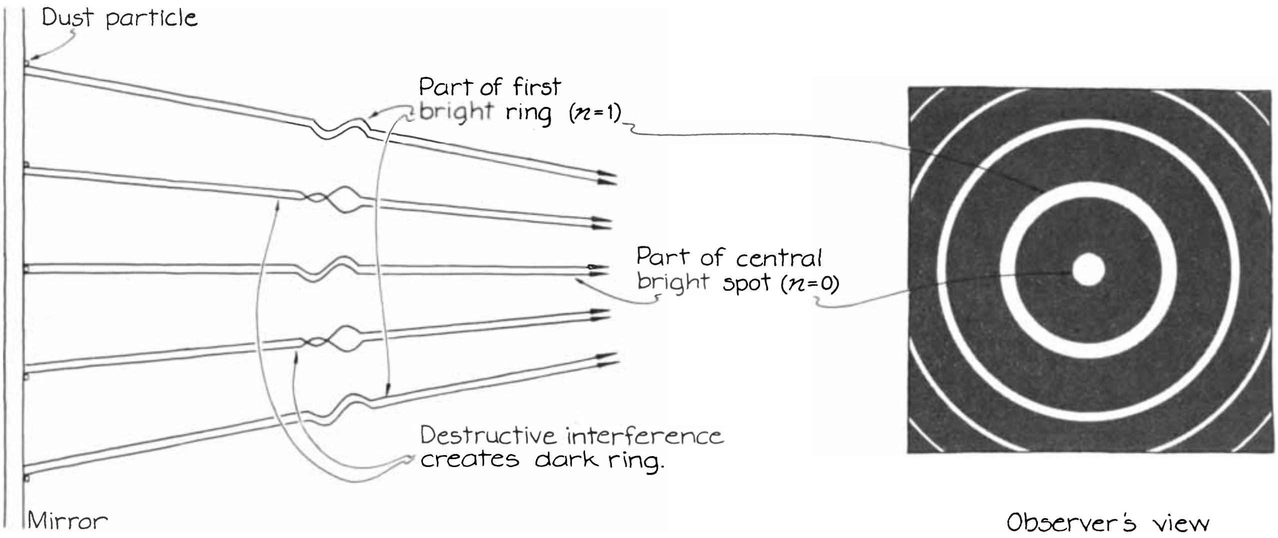
At the center of the pattern all the rays arrive in phase because they all travel the same distance. Rays arriving slightly off the center have path differences resulting in a phase difference of half a wavelength. These rays interfere destructively. Since the same condition develops in a circle around the bright center of the pattern, a dark band appears around the center. The next ring is bright because the rays interfere constructively.

You see a similar pattern when you look through a dusty or misty sheet of glass, provided the light source is small enough or far enough away to occupy a small angle in your field of view. You see not the full diffraction pattern from each particle on the glass but a composite pattern encompassing parts of the pattern cast by each particle. The particles on the axis between you and the light source send the bright centers of their patterns; the particles slightly off

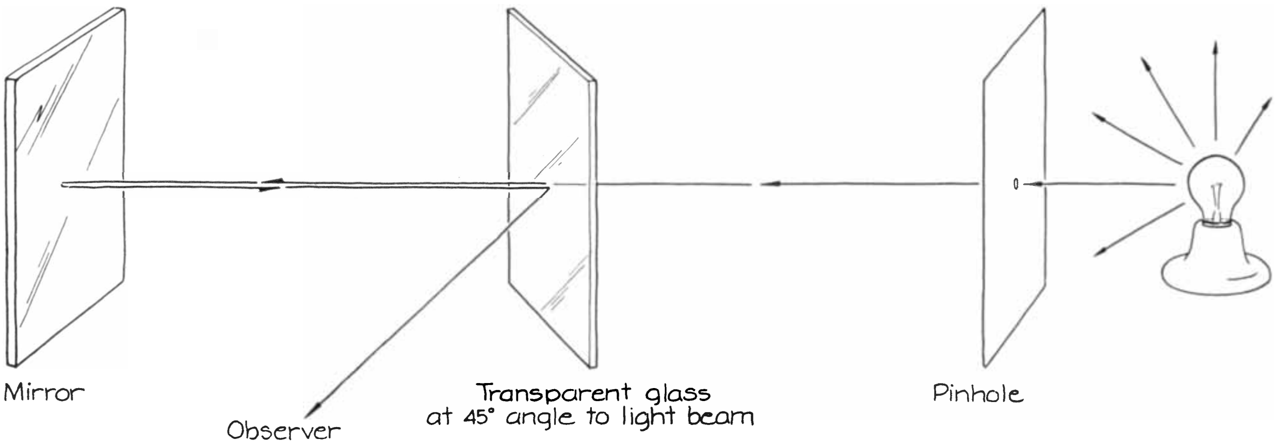


How light is scattered from a dust particle


Mischievous as children protecting a secret, Le-ah and Tiu scampered behind the falls and bade us follow.
Our climb from the village had been long and hot, and we were glad to cool our weary limbs. Suddenly, these bright spring blossoms burst from the chill shower, bearing an oaken scuttle brimming with bottles of both light and dark San Miguel Beer.
Such hospitality we'd never imagined. We savored the cold, clean taste of the San Miguel and set aside all discussion of the repairs our reef-bound schooner would require on the morrow.
Thus did the afternoon drift by: I and the crew and our carefree island guides, sipping San Miguel and splashing in the summer garb of Eden.*
*Inspired by Herman Melville's Typee. And the rich, rewarding taste of San Miguel.
San Miguel
Classic beer of the Pacific.
Imported by San Miguel International (USA).



The mode of formation of a Whewell-Quételet pattern



A setup for viewing the patterns

the axis send part of their innermost dark ring. The result from all the particles is again a pattern of alternating bright and dark rings. The particles on the glass not only diffract the light into the pattern of Fraunhofer rings but also disperse the white light into its component colors so that the rings are colored.

The Fraunhofer ring around the moon or the sun is called a corona. (The term should not be confused with the solar corona seen during a total eclipse of the sun.) The particles responsible for the ring are water droplets or ice crystals in the thin clouds between the observer and the moon or the sun.

I can see a similar corona when condensation mists my eyeglasses or car windows. Distant streetlights and the headlights of cars are surrounded by colored rings. Sometimes I see Fraunhofer rings when distant headlights illuminate the rearview mirror in my car because my mirror is normally somewhat dusty! (Another kind of corona that is seen surrounding distant light sources has nothing to do with the diffraction I am discussing. This faint "entoptic corona" is created in the observer's eye.) I once saw a remarkable example of the diffraction pattern when I was about a meter away from a misty mirror that was illuminated by bright light from a window some five meters behind me. Superimposed on each of my eyes in the mirror was a Fraunhofer pattern.

The corona caused by diffraction is seen in a dusty mirror because the pattern diffracted forward by the dust particles is reflected by the back of the mirror. Regardless of the observer's angle of view, the pattern of rings and colors always surrounds the reflected image of the light source. This consistency is the primary clue for distinguishing the Fraunhofer rings from the Whewell-Quételet rings, which surround the reflected image only if the observer and the light source are on the same perpendicular line from the mirror.

The clarity of the Fraunhofer rings is strongly correlated with the size of the particles. The smaller the particle, the larger the angular spread of the diffraction pattern. A large diversity of particle sizes gives rise to an indistinct Fraunhofer pattern because the resulting diffractions overlap.

I get distinct Fraunhofer rings when I illuminate an uncoated sheet of glass or a mirror that is lightly covered with lycopodium powder. The particles in this powder are fairly uniform in size and therefore generate diffraction patterns of about the same size. I get less distinct (sometimes barely visible) Fraunhofer rings when I illuminate a sheet of glass or a mirror covered with condensation from my breath or from a hot shower. The droplets involved in the interference are apparently more diverse in size. They may be more closely spaced than dust particles usually are, so that additional interference effects may obscure the simple diffraction pattern I have described.

The photograph of the interference patterns in the top illustration on page

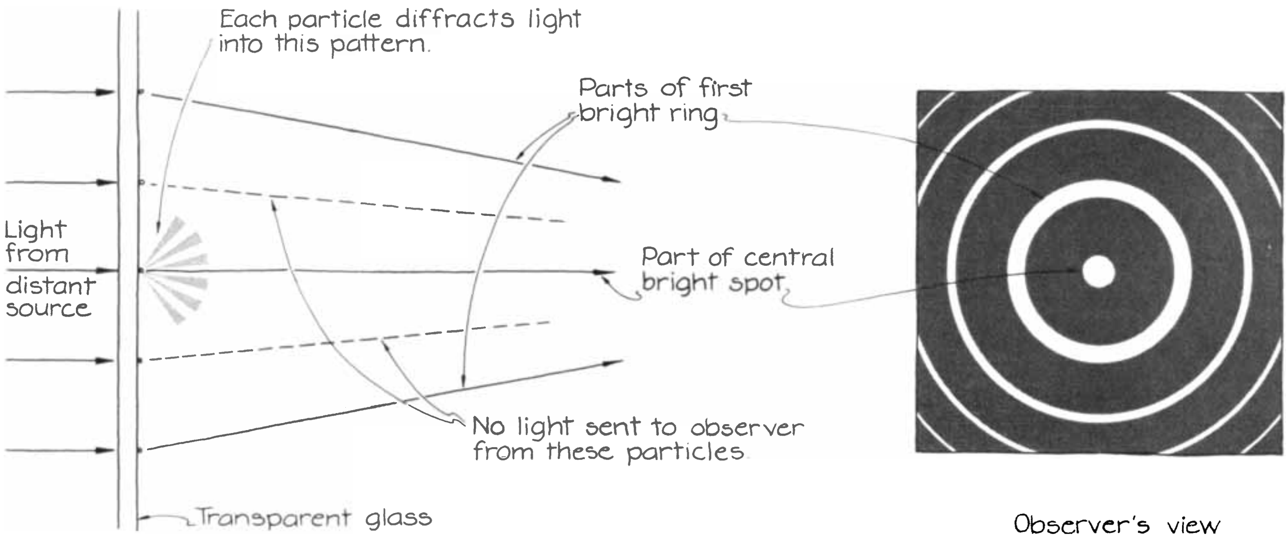
147 was made with light from a helium-neon laser. The laser beam was reflected by a mirror dusted with lycopodium powder. The primary reflection traveled to a screen behind the laser. Surrounding the brightest area of the display is the Fraunhofer pattern. Toward the bottom of the photograph, almost in the shadow of the laser, is the center of the Whewell-Quételet rings.

Is the pattern I have called the Fraunhofer pattern truly reflected by the rear surface of a mirror as I have described? To find out I scraped off part of the reflective coating on the back of an inexpensive mirror. After lightly dusting the front surface with lycopodium powder I illuminated the mirror. The Fraunhofer rings appeared as strongly as usual everywhere except over the nonreflecting area. There they were comparatively weak. A faint amount of pattern remained because the back surface of the nonreflecting area still reflected about 2 percent of the light.

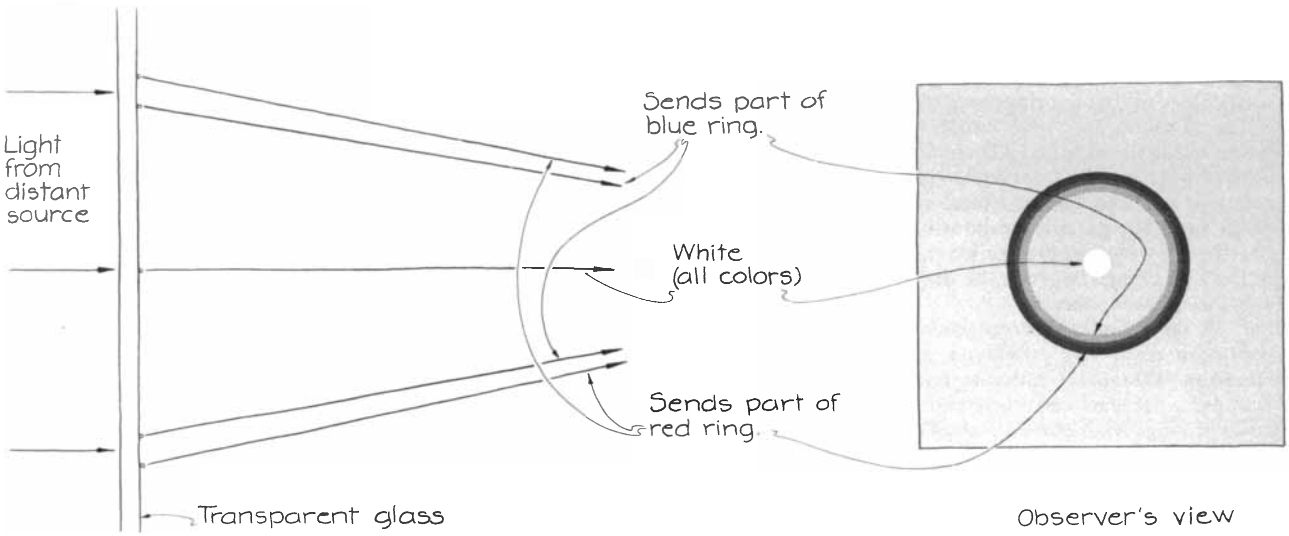
To eliminate virtually all reflection from the scraped area I sprayed the back of the mirror with a flat black paint. After the paint had dried I again dusted the front surface of the mirror. No interference pattern was visible from the scraped area.

I wondered why the source of light for the two interference patterns seen in a dusty mirror had to be either small or distant from the mirror. When I held a large diffuse source of light, such as a light bulb, near a dusty mirror, I could see no pattern. The explanation lies in two features of the light: the angular spread of the rays and the coherence of the light falling on the dust particles. The pair of light rays giving rise to an interference pattern must be coherent and approximately parallel at a particle. If they are incoherent, their phase relation shifts from instant to instant. The amount of interference between them when they reach the observer also shifts, and he does not see a consistent interference. If the rays are not parallel, the interference pattern is obscured by the overlapping of many patterns sent out at slightly different angles from a particle.

An astute observer will note that the Fraunhofer patterns created by water droplets differ in two respects from those created by lycopodium powder. The center spot is surrounded by a dark area in the first instance and by a bright area in the second one. Moreover, the patterns differ in form when the observer's view is not perpendicular to the surface of the mirror or of the uncoated sheet of glass. A careful observer will also note that the patterns have an intriguing granular structure, particularly near the brightest region. These and other puzzles must wait for another time.



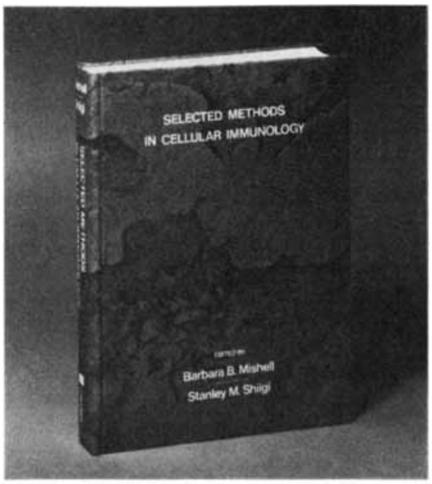
The development of a diffraction pattern from dust on glass



The separation of color in a diffraction pattern

Selected Methods in Cellular Immunology

Edited by
Barbara B. Mishell, *University of California, Berkeley*
Stanley M. Shiigi, *Oregon Regional Primate Research Center*



"*Selected Methods in Cellular Immunology* by B. Mishell and S. Shiigi is almost all that one could ask from a methods book. . . .This is a book that every laboratory doing or planning to do *in vitro* cellular immunology will want to have, and don't bother finding a spot for it on your shelf; this book will be on the bench being used." —*Cell*

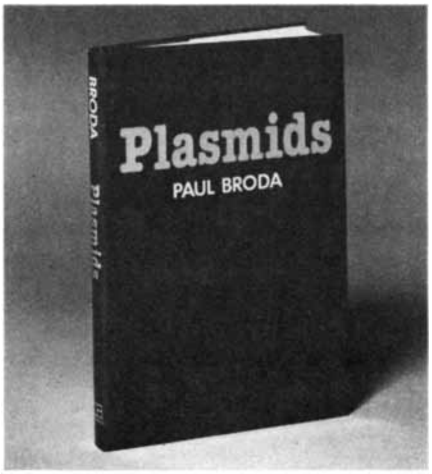
"An excellent book. . . . No laboratory with an interest in cellular immunology should be without it." —*Bioscience*

"An added bonus is that the book is written with such style that it is quite enjoyable to read for general knowledge, as well as for specific procedures." —*American Society for Microbiology News*

1980, 486 pages, 36 illustrations

Plasmids

Paul Broda, *U.M.I.S.T., Manchester, England*



"An up-to-date, comprehensive and comprehensible guide to the subject was badly needed. . . . Here it is—Paul Broda has provided just the thing, and it is not only newcomers to the subject who will welcome his *Plasmids*." —*Nature*

"This book satisfies a real need for current information on plasmids for university lecturers, as a reference text in libraries and for use in advanced graduate student courses on microbial genetics. The author's writing style makes easy reading and his experimental familiarity with plasmids is clearly displayed." —*Search*

1979, 197 pages, 26 illustrations

Please send me the following books. I enclose payment or credit card information with order. Publisher pays postage and handling.

Copies

_____ *Selected Methods in Cellular Immunology*
Hardbound: 1106 $31.95

_____ *Plasmids*
Hardbound: 1111 $19.95

Guarantee: Examine these books for fourteen days. If for any reason you are not satisfied, you may return them for a full and prompt refund.

W. H. Freeman and Company
660 Market Street, San Francisco, California 94104

Name ______________________

Address ______________________

City/State ______________ Zip ________

I enclose payment with order.
(California residents add appropriate sales tax.)

Charge my ☐ Visa/BankAmericard ☐ Master Charge.

Account Number ______________________

Expiration Date ____________

Signature ______________________
All credit card orders must be signed. SA 61

W. H. Freeman and Company is the book-publishing affiliate of *SCIENTIFIC AMERICAN*.

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

THE PARABOLA. Harold R. Jacobs in *Mathematics, a Human Endeavor: A Textbook for Those Who Think They Don't Like the Subject.* W. H. Freeman and Company, 1970.

GALILEO'S DISCOVERY OF THE PARABOLIC TRAJECTORY. Stillman Drake and James MacLachlan in *Scientific American,* Vol. 232, No. 3, pages 102–110; March, 1975.

DO SIMILAR FIGURES ALWAYS HAVE THE SAME SHAPE? Paul G. Kumpel, Jr., in *Mathematics Teacher,* Vol. 68, No. 8, pages 626–628; December, 1975.

PRECISION-GUIDED WEAPONS

INACCURACY OF DEPARTMENT OF DEFENSE WEAPONS ACQUISITION COST ESTIMATES: NINTH REPORT. U.S. House of Representatives, Committee on Government Operations. Government Printing Office, 1979.

NEW TECHNOLOGY AND MILITARY POWER: GENERAL PURPOSE MILITARY FORCES FOR THE 1980S AND BEYOND. Seymour J. Deitchman. Westview Press, 1979.

THE FY1981 DEPARTMENT OF DEFENSE PROGRAM FOR RESEARCH, DEVELOPMENT, AND ACQUISITION. Under Secretary of Defense for Research and Engineering William J. Perry. Government Printing Office, 1980.

JANE'S WEAPON SYSTEMS 1980–81. Edited by Ronald T. Pretty. Franklin Watts Inc., 1980.

THE IMPACT OF NEW MILITARY TECHNOLOGY. Edited by Jonathan Alford. Gower Publishing, 1981.

THE NEWEST STARS IN ORION

DETECTION OF CO BAND EMISSION IN THE BECKLIN-NEUGEBAUER OBJECT. N. Z. Scoville, D. N. B. Hall, S. G. Kleinmann and S. T. Ridgway in *Astrophysical Journal Letters,* Vol. 232, pages L121–L124; September 1, 1979.

INTERSTELLAR MOLECULES. Edited by B. H. Andrew. D. Reidel Publishing Company, 1980.

INFRARED ASTRONOMY. Edited by C. G. Wynn-Williams and D. P. Cruikshank. D. Reidel Publishing Company, 1981.

OUTFLOW OF MATTER IN THE KL NEBULA: THE ROLE OF IRc2. D. Downes, R. Genzel, E. E. Becklin and C. G. Wynn-Williams in *The Astrophysical Journal,* Vol. 244, No. 3, Part 1, pages 869–883; March 15, 1981.

DUTCH ELM DISEASE

ATTRACTANT PHEROMONE OF THE EUROPEAN ELM BARK BEETLE (*SCOLYTUS MULTISTRIATUS*): ISOLATION, IDENTIFICATION, SYNTHESIS AND UTILIZATION STUDIES. G. N. Lanier, R. M. Silverstein and J. W. Peacock in *Perspectives in Forest Entomology,* edited by John F. Anderson and Harry K. Kaya. Academic Press, 1976.

SOME PHYTOTOXIC GLYCOPEPTIDES FROM *CERATOCYSTIS ULMI,* THE DUTCH ELM DISEASE PATHOGEN. Gary Strobel, Neal Van Alfen, Kenneth D. Hapner, Michael McNeil and Peter Albersheim in *Biochimica et Biophysica Acta,* Vol. 538, No. 1, pages 60–75; January 3, 1978.

PHEROMONE-BAITED TRAPS AND TRAP TREES IN THE INTEGRATED MANAGEMENT OF BARK BEETLES IN URBAN AREAS. G. N. Lanier in *Management of Insect Pests with Semiochemicals: Concepts and Practices,* edited by E. E. Mitchell. Plenum Press, in press.

THE RIBOSOME

RIBOSOMES. Masayasu Nomura in *Scientific American,* Vol. 221, No. 4, pages 28–35; October, 1969.

BIOLOGICAL STUDIES. James A. Lake in *Optical Transforms,* edited by H. Lipson. Academic Press, 1972.

NEUTRON-SCATTERING STUDIES OF THE RIBOSOME. Donald M. Engelman and Peter B. Moore in *Scientific American,* Vol. 235, No. 4, pages 44–54; October, 1976.

RIBOSOME STRUCTURE AND tRNA BINDING SITES. James A. Lake in *Transfer RNA: Structure, Properties and Recognition,* edited by Paul R. Schimmel, Dieter Söll and John N. Abelson. Cold Spring Harbor Laboratory, 1979.

THE FORGING OF METALS

THE NATURE OF METALS. A. H. Cottrell in *Scientific American,* Vol. 217, No. 3, pages 90–100; September, 1967.

FORGING MATERIALS AND PRACTICES. A. M. Sabroff, F. W. Boulger and H. J. Henning. Reinhold Book Corporation, 1968.

FORGING EQUIPMENT, MATERIALS, AND PRACTICES. Edited by Taylan Altan. Sponsored by the Air Force Materials Laboratory. The Metals and Ceramics Information Center (Handbook No. MCIC-HB-03), Battelle Memorial Institute, Columbus Laboratories, Metalworking Division, 1973.

THE DEFORMATION OF METALS AT HIGH TEMPERATURES. Hugh J. McQueen and W. J. McGregor Tegart in *Scientific American,* Vol. 232, No. 4, pages 116–125; April, 1975.

MONSOONS

MONSOON METEOROLOGY. Special issue of *Pure and Applied Geophysics,* edited by T. N. Krishnamurthi, Vol. 115, Nos. 5/6, pages 1087–1529; 1977.

THE SCIENCE AND WONDERS OF THE ATMOSPHERE. Stanley D. Gedzelman. John Wiley & Sons, Inc., 1980.

MONSOON DYNAMICS. Edited by Sir James Lighthill and R. Pearce. Cambridge University Press, 1981.

A MESOLITHIC CAMP IN IRELAND

THE MESOLITHIC IN IRELAND: HUNTER-GATHERERS IN AN INSULAR ENVIRONMENT. Peter C. Woodman. British Archaeological Reports, Oxford, 1978.

THE CHRONOLOGY AND ECONOMY OF THE IRISH MESOLITHIC: SOME WORKING HYPOTHESES. Peter C. Woodman in *The Early Postglacial Settlement of Northern Europe,* edited by Paul Mellars. University of Pittsburgh Press, 1979.

EARLY MAN IN BRITAIN AND IRELAND. Alex Morrison. St. Martin's Press, 1980.

EXCAVATIONS AT MOUNT SANDEL. Peter C. Woodman. Northern Ireland Archaeological Monographs No. 2, Her Majesty's Stationary Office, Belfast, in press.

SADI CARNOT

A SHORT HISTORY OF THE STEAM ENGINE. Henry W. Dickinson. Cambridge University Press, 1963.

FROM WATT TO CLAUSIUS: THE RISE OF THERMODYNAMICS IN THE EARLY INDUSTRIAL AGE. D. S. L. Cardwell. Cornell University Press, 1971.

REFLECTIONS ON THE MOTIVE POWER OF FIRE. Sadi Carnot. Peter Smith, 1977.

RÉFLEXIONS SUR LA PUISSANCE MOTRICE DU FEU. Sadi Carnot. Librairie Philosophique J. Vrin, 1978.

THE AMATEUR SCIENTIST

ON THE COLOURS OF THICK PLATES. Sir George Gabriel Stokes in *Mathematical and Physical Papers: Volume 3.* Johnson Reprint Corporation, 1966.

INTERFERENCE IN SCATTERED LIGHT. A. J. de Witte in *American Journal of Physics,* Vol. 35, No. 4, pages 301–313; April, 1967.

THE CORONA. R. A. R. Tricker in *Introduction to Meteorological Optics.* American Elsevier Publishing Company, Inc., 1970.

THE CORONA. Robert Greenler in *Rainbows, Halos, and Glories.* Cambridge University Press, 1980.

Small computer friendliness vs. big computer power.

HP's new computer systems end the compromise.

Whether it's sophisticated experimentation or data reduction and analysis, Hewlett-Packard's newest technical computer systems put high-powered scientific performance into easy-to-configure, easy-to-use workstations. So you can solve your applications problems just the way *you* want to. And a lot sooner than you might have thought possible.

The HP 9826 for complex experiment control.

Designed specifically for I/O-intensive instrument and control applications, the 9826 uses the powerful new MC 68000 chip for ultra-fast data transfer and number crunching. It also has a variety of I/O ports and built-in drivers, so you're free to concentrate on testing, not on system configuration. Which means you can get your solutions up and running in days, instead of weeks or months.

But the 9826 makes it even easier still. Because you can develop your programs in any of three languages: HP-enhanced BASIC, HPL, or PASCAL. And with a CRT display and advanced graphics built right into the system, you can see and interpret your data almost as soon as you get it.

The HP 9826. System prices start at $8950.

The HP 9845 for high-performance computation.

With a broad range of advanced capabilities (including new options offering three times the computational speed of the basic system), the HP 9845 can be configured to meet your precise needs—from finite element analysis to simulations and modeling. And almost anything in between.

You can also add hard discs, printers, plotters and other peripherals to further tailor the system to your application. And with the 9845's advanced datacomm capabilities, you can even share technical information with other computers in a distributed processing environment.

The HP 9845. System prices start at under $20,000.

All of our computers are backed by HP's worldwide service organization. For more information, call your local HP sales office listed in the White Pages. Or write to Hewlett-Packard, Attn: Pat Welch, Dept. 56109, 3404 E. Harmony Rd., Ft. Collins, CO 80525.

Prices U.S.A. list: subject to change without notice. 40107

When performance must be measured by results.

HEWLETT PACKARD

Making sure everything

Lockheed knows how.

Across the U.S., there is a rising, insistent cry for more quality in products. At Lockheed, advances in quality are a way of life. When a fleet ballistic missile is launched, as shown above, there can be no compromise in quality. And that applies to aircraft, spacecraft, and a host of other Lockheed products large and small. The examples below cover some of the advances in quality that Lockheed has made recently in keeping pace with scientific progress.

Giving flaws no place to hide.

The cost and time-consumption of inspecting products like rocket motors and gas turbine engines can be staggering.

Gas turbine engine under Lockheed X-ray inspection.

With rocket motors, it's a search for cracks, voids, flawed bonding, or other critical defects. With gas turbine engines, the vital subject is gas-seal clearances, because these affect fuel efficiency. The engine testing is done at the development stages and the gas seals redesigned if necessary.

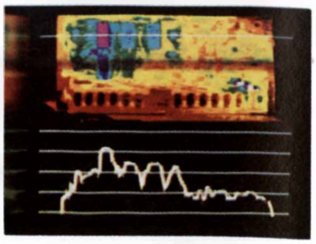
Until recently, such testing combined X-ray with photographic techniques. With rocket motors, however, inspecting by this photographic imagery method required several hundred separate pictures, each needing incremental rotations of the product, exposure times, individual film processing, and painstaking review of every picture.

So Lockheed has developed a real-time X-ray imaging system that permits uniform, 100% inspection (rather than sampling) at far faster speed and lower cost than conventional X-ray methods. Users of the Lockheed system report savings of up to 50% in time and 66% in X-ray materials costs.

Spotting trouble before it becomes a problem.

You're looking at potential trouble in the photograph below. Actually, it's an infrared photograph of a printed circuit board. The magenta coloring at top left identifies an integrated circuit on the board that may be too hot.

Until recently, it was almost impossible to verify the thermal integrity of the hundreds or even thousands of PC boards that go into a modern aircraft. Now Lockheed has developed a unique method of evaluating PC boards completely. This new approach involves thermal imaging. It is a non-contact, non-destructive way of testing electronic equipment.



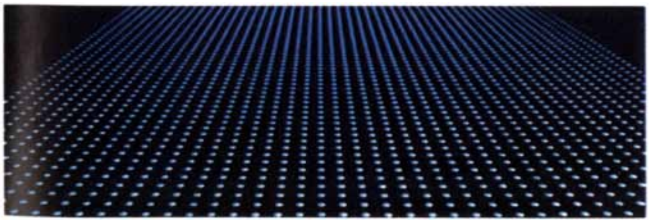
The actual test of a PC board using infrared photography can be made in a few minutes.

works the first time.

Once a "hot" spot is detected, early steps can be taken to correct it.

Lockheed's thermal survey also has a long-term benefit. The infrared photograph provides a permanent temperature record that may be compared with future design changes in the PC board. The result: far higher reliability over the life of these assemblies.

Found: a way to test 200,000 holes.



A big airlifter or jetliner may have as many as 200,000 holes in its airframe. Mainly rivet and fastener holes, they all must be checked. And rechecked after the aircraft has been in service. It's been a costly, time-consuming process.

Now Lockheed has developed two unique systems that cut costs and time enormously while increasing reliability of testing. One, a Capacitance Hole Probe tests newly drilled holes. A probe with 48 sensors is inserted into the hole. The sensors check the hole on six different layers, and a computer makes a written record of the test—all in nine seconds. Formerly, it took several minutes to test a hole with less reliability.

Lockheed has not neglected old holes either. In maintenance inspections, a computer-based Eddy Current Scanner tests an old hole for cracks in two minutes. The recent manual method took 10 minutes with less reliability. The scanner also uses a sensor probe and makes a computer record of the findings.

50,000 fuzes a month.

Aircraft and spacecraft roll down assembly lines in relatively small numbers. But Lockheed also is involved in mass production at a unique facility in Denville, New Jersey. There Lockheed produces 50,000 fuzes a month for U.S. Army shells. It already has manufactured more than 1,000,000 fuzes.

Designed by Lockheed under U.S. Army direction, the facility could produce well over 100,000 fuzes monthly. It is highly mechanized, and its extensive testing program is automated. Computerized stations test 100% of all electrical components as they arrive at the facility. Then, all subassemblies and final assemblies are tested. Lockheed believes this may be the most rigorous testing program of this type in the country. Result: highly reliable fuzes, produced in great quantity with enhanced productivity.

Lockheed

"~~Scotch~~ Jameson Irish and water."

If you like fine Scotch, you'll love light, imported Jameson Irish.

Try a glass of Jameson Irish the way you would your favorite Scotch. With water. Soda. On the rocks.

You'll notice how much it tastes like fine Scotch—only lighter and more delicate.

The dedicated Scotch drinker will instantly appreciate this flavor difference.

Though it may take a little time getting used to saying, "Jameson Irish and water, please."

Jameson. World's largest-selling Irish Whiskey.

80 PROOF • CALVERT DIST. CO., N.Y.C.