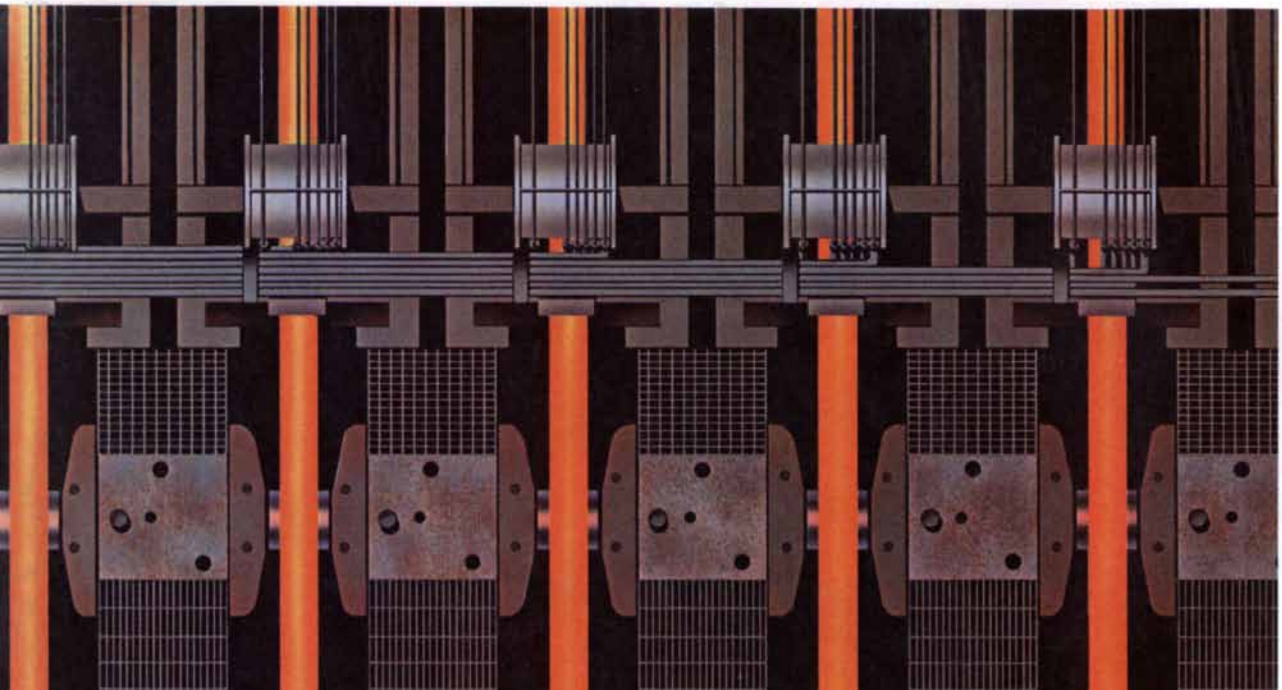
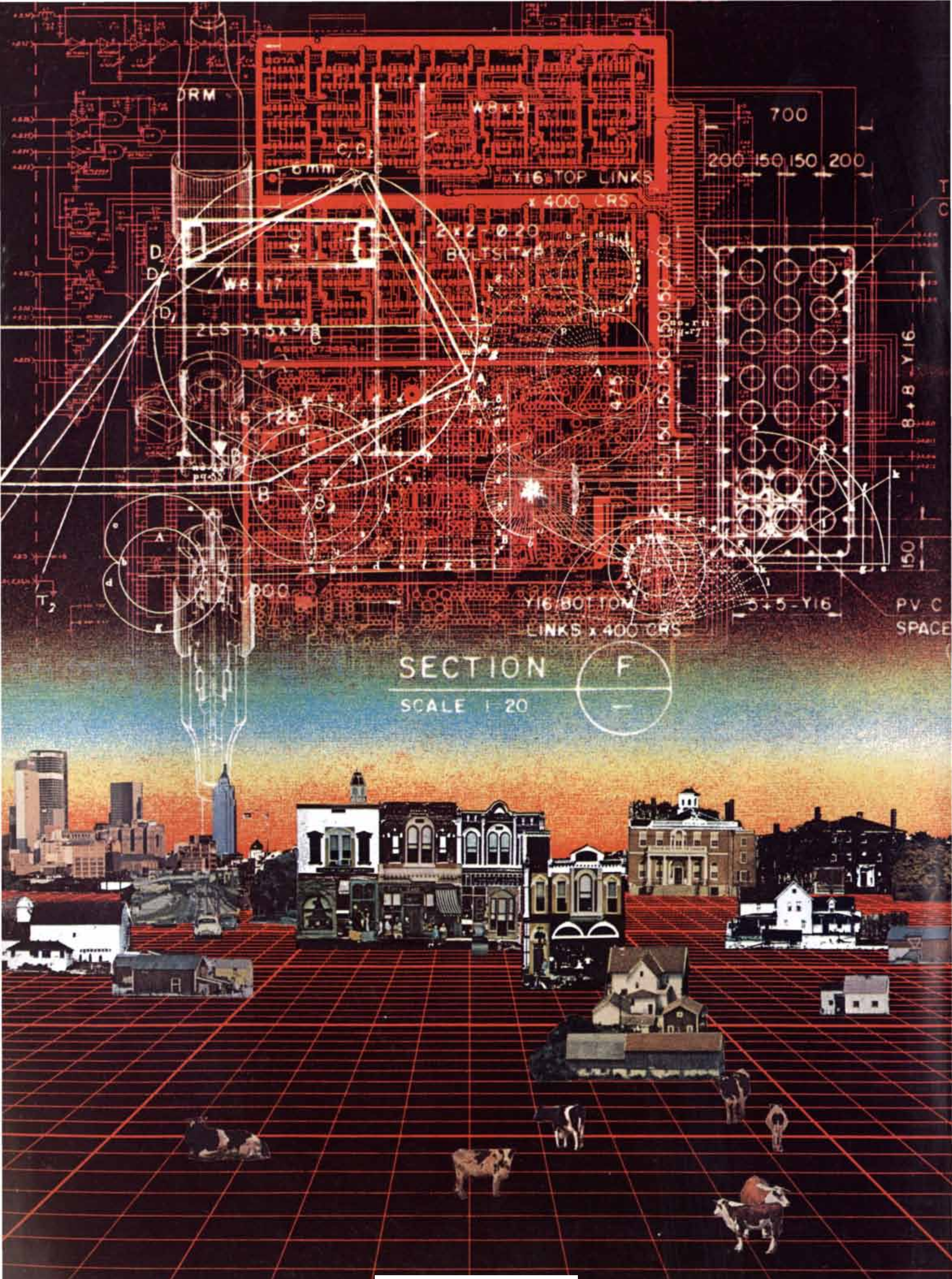
SCIENTIFIC AMERICAN



STEEL MINIMILLS

$2.50

May 1984


SECTION F
SCALE 1:20
700
200 150 150 200
Y16 TOP LINKS
x 400 CRS
Y16/BOTTOM
LINKS x 400 CRS
5+5-Y16
8+8-Y16
PV C
SPACE

Established 1845 SCIENTIFIC AMERICAN May 1984 Volume 250 Number 5

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1984 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: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.25 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036-8733/84. $1.25 + 0.00.

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

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

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name ______________________

New Address

Street ______________________

City ______________________

State and ZIP ______________________

Old Address

Street ______________________

City ______________________

State and ZIP ______________________

THE COVER

The painting on the cover shows a portion of the five-strand continuous-casting operation of a steel minimill, the Raritan River Steel Co. plant in Perth Amboy, N.J. In continuous casting newly made molten steel is poured from a ladle into a tundish, a structure resembling a funnel. From there it descends into a vertical, water-cooled mold and then through a zone where it is cooled by sprayed water. The strands are moved through curved sections assisted by motor-driven withdrawal straighteners. At the top of the painting they are moving horizontally through a cooling area (*downward as one looks at the picture*). At a length of 50.5 feet the strands are cut by torches; the pipes and drumlike objects near the top of the painting are associated with the torches. When the strands have cooled and solidified, they are pushed sideways off the cooling area and stored until they are sent to a rolling mill in the plant for shaping into rods. Minimills are distinguished from conventional steel mills by the fact that their raw material is steel scrap rather than iron ore and that a minimill makes only a limited line of products (see "Steel Minimills," by Jack Robert Miller, page 32). At a time when the large companies commonly referred to collectively as Big Steel are having economic difficulties, minimills are flourishing.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
18–22	Ilil Arbel
32	Barrie Rokeach
34–39	Jerome Kuhl
40–41	Kathleen Crane, Lamont-Doherty Geological Observatory
42	Andrew Tomko
43	William F. Haxby, Lamont-Doherty Geological Observatory (*top*); Andrew Tomko (*bottom*)
44–47	Andrew Tomko
48	Enrico Bonatti, Lamont-Doherty Geological Observatory (*top*); Dee Breger, Lamont-Doherty Geological Observatory (*bottom*)
49–51	Andrew Tomko
53	M. M. Perry and A. B. Gilbert, Agricultural Research Council Poultry Research Centre, Edinburgh (*top*); Hans J. Geuze, State University of Utrecht (*bottom*)
54–56	Bunji Tagawa
57	Frederick R. Maxfield, New York University Medical Center
58	Bunji Tagawa
72	Ilil Arbel
87–97	Edward Bell
109–110	Tom Prentiss
111	Janice Moore, courtesy of *The Journal of Parasitology*
112–114	Ilil Arbel
117–126	Ian Worpole
131	William C. Johnson, University of Pittsburgh
132–134	Ilil Arbel
136	J. M. Adovasio, University of Pittsburgh
138–147	Jerome Kuhl
148–153	Michael Goodman

LETTERS

Sirs:

Creationists appear to be winning some important battles with the more rational sectors of our society ["Texts for Today," "Science and the Citizen," SCIENTIFIC AMERICAN, March]. But will they really win in Texas by keeping evolution out of textbooks? A forbidden topic amplifies manyfold the limited interest and curiosity that are usually the reward of a routine classroom subject. Just as students in a period of the repression of knowledge about sex were led to Havelock Ellis, so students in a period of the repression of knowledge about evolution may well be led to Charles Darwin.

One might predict that an abridged edition of *The Origin of Species* will become a best seller in Texas. The publisher that takes advantage of this potential market might well make the book of a size and thickness to fit conveniently inside the official textbooks for surreptitious reading. If the publisher can find space for a glossary, a listing of the postulates of Darwin's two major theories (the theory of descent with modification and the theory of natural selection) and a brief outline of the modern theory of the mechanism of evolution, biology could thrive in Texas as never before. Creationists, of all people, should know that a forbidden topic is like a forbidden apple.

Creationists are advancing the study of biology in another way. By forcing biology teachers to examine more closely the meaning of the term "theory" creationists have already helped to clarify many topics in biology. An explicit recognition of the structural character of a theory gives a better logical order to facts and ideas. And a better understanding of theory should enable many teachers to recognize that there is even a partial fallacy in the statement you cite: "At no point do 'theories' become 'facts.'" Teachers may remember the cellular theory of sexual reproduction of the past century—with its postulates about gametes, fertilization and zygote—and will recognize that these postulates are now generally accepted as facts. They may also remember the 1978 Nobel-prize acceptance speech of Peter Mitchell, in which he listed the postulates of the chemiosmotic theory and then pointed out that these postulates were now widely accepted as facts. By having given closer attention to "theory" teachers will know there are theories with a very limited range of applicability, theories with a very wide range and an array of theories in between. At the same time they will probably have learned the several meanings of "fact." Creationists are willy-nilly living up to their name by creating a better presentation of the science of biology.

RALPH W. LEWIS

Michigan State University
East Lansing

Sirs:

On reading "The Packing of Spheres," by N. J. A. Sloane [SCIENTIFIC AMERICAN, January], it struck me as strange that a fine magazine would print such an unresearched and inaccurate article. If Sloane had consulted the work of the late R. Buckminster Fuller, he would have found it much easier to solve the dilemma: "What makes it so hard to solve the sphere-packing problem in three dimensions?"

In Fuller's book *Synergetics* (Macmillan, 1975) Sloane would have found the solution he was looking for. Since circles and spheres are in reality polyhedrons, and since nature relies on a 60-degree coordinate system instead of modern man's 90-degree system, the problem becomes less elusive. If Sloane had done a better job of researching his article, he would have realized that Fuller had spent the better part of his life solving problems such as this one in his structural analysis of geodesic domes.

WILLIAM MAKSYM

Babylon, N.Y.

Sirs:

Although R. Buckminster Fuller repeatedly refers in his book *Synergetics* to the face-centered cubic lattice as being the densest packing of spheres in three dimensions, there is no indication he could prove that assertion.

N. J. A. SLOANE

AT&T Bell Laboratories
Murray Hill, N.J.

Sirs:

I have read the article "Labor-intensive Agriculture," by Philip L. Martin [SCIENTIFIC AMERICAN, October, 1983], and the letters in response to the article. I believe many people have missed the point of mechanization, namely that many tasks of farm labor are downright demeaning. Human beings should not have to do them. Anyone who thinks otherwise either has never done farm labor or is a fool. I know what I am talking about; I have worked in agriculture my entire adult life.

GREGG MANSTON

San Luis Obispo, Calif.

AC Plasma Display Manufacturing

IBM

To: Ted
From: Bill
Subject: IBM Technology

I've been reviewing some of our past and present technological achievements, and it occurred to me that the scientific, engineering, and academic communities might like to know more about them. Will you select a topic from the following list?
Thanks.

- Vacuum tube digital multiplier
- IBM 603/604 calculators
- Selective Sequence Electronic Calculator (SSEC)
- Tape drive vacuum column
- Naval Ordnance Research Calculator (NORC)
- Input/output channel
- IBM 608 transistor calculator
- FORTRAN
- RAMAC and disks
- First automated transistor production
- Chain and train printers
- Input/Output Control System (IOCS)
- STRETCH computer
- "Selectric" typewriter
- SABRE airline reservation system
- Removable disk pack
- Virtual machine concept
- Hypertape
- System/360 compatible family
- Operating System/360
- Solid Logic Technology
- System/360 Model 67/Time-Sharing System
- One-transistor memory cell
- Cache memory
- Relational data base
- First all-monolithic main memory
- Thin-film recording head
- Floppy disk
- Tape group code recording
- Systems Network Architecture
- Federal cryptographic standard
- Laser/electrophotographic printer
- First 64K-bit chip mass production
- First E-beam direct-write chip production
- Thermal Conduction Module
- 288K-bit memory chip
- Robotic control language

*Bill—
We have a great story to tell about our manufacturing innovations. The mass production of our large-screen plasma display is an excellent example.
Ted*

Figure 1. *The IBM 3290 Information Panel uses alternating current (AC) plasma technology, making possible high information content and distortion-free images. The screen, which is 10.7 inches by 13.4 inches, can display up to 10,000 characters and can simultaneously display four applications from one or more computers. Its great versatility allows it to mix graphics, images, and text.*

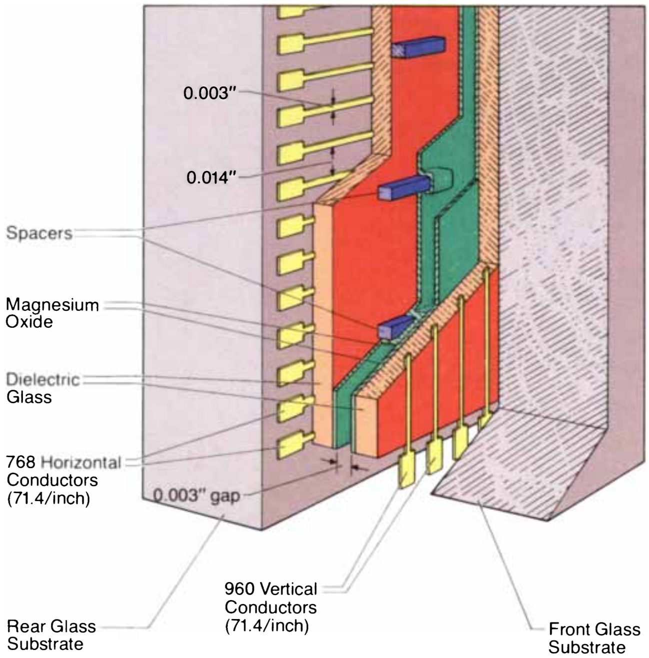


Figure 2. *This cross section of the plasma panel shows the narrow conductor lines on opposing glass substrates. Unique points on the panel can be ionized by applying low voltages to the appropriate horizontal and vertical conductors.*

Visual display terminals have had a profound impact on data processing. The IBM 3270 family of cathode ray tube (CRT) terminals has become widely accepted in the industry as a basic input/output device for mainframe computers.

To display more data and to provide more advanced function without increasing space requirements, IBM has developed a new terminal using alternating current (AC) plasma display technology, invented at the University of Illinois. As a result of IBM's many manufacturing innovations, the IBM 3290 Information Panel, introduced in March 1983, is the industry's first mass-produced, large-screen plasma display terminal for commercial use.

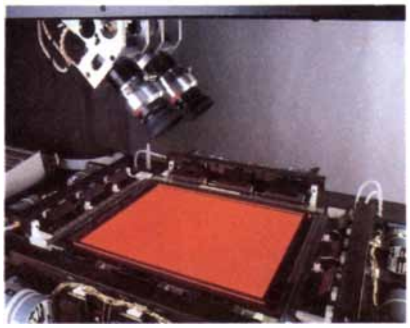
HOW IT WORKS

In the IBM 3290, the plasma panel is a sealed sandwich of two glass plates: the rear plate is embedded with 768 parallel horizontal conductors and the front plate with 960 vertical conductors, thus forming a large grid. The narrow space separating the two plates is filled with inert neon-argon gas, which glows as electrical voltages are selectively applied to any of the over 700,000 intersections on the grid. This locally ionized gas, called a plasma, produces tiny dots of orange light. When combined in matrix patterns, these precisely located dots form images. Because this plasma technology operates

AC Plasma Display Manufacturing

Figure 3. *The proximity printer shown here is one of many tools developed by IBM to mass-produce the AC plasma display panel. This machine automatically prints the hundreds of conductor lines on the glass plate by using highly collimated light to expose the conductor pattern through a mask. This projection printing system produces an excellent image and lowers the number of defects.*

Figure 4. *To assure consistently high quality in mass production, each AC plasma display panel is completely evaluated by this automatic tester developed by IBM. The tester has a camera system that scans and tests the patterns on each panel.*



in memory mode, the images do not have to be refreshed, eliminating any susceptibility to flicker.

MANUFACTURING INNOVATIONS

IBM manufacturing engineers had to find many answers to the challenges of mass-producing large-screen AC plasma panels. For example, special techniques were required to place 2,400 feet of very narrow conductors on each panel. To ensure high yields, engineers improved the method of photoprinting the conductor pattern and devised a way to repair open and shorted lines.

The large area of the new plasma display placed more stringent requirements on both the materials and processes used to fabricate the device. The panel—a composite of glass, metal, and thin-film oxide layers—is made by sequential thermal process steps, with each step conducted at a temperature suitably lower than the prior process step. To reduce material interactions, IBM developed lower-temperature dielectric glass and seal material.

To maintain a uniform chamber gap between the sandwiched glass plates, engineers also developed a new metallic spacer technology. The spacers—about the thickness of a human hair and a quarter inch long—are automatically bonded by a tool that uses a laser to keep placement tolerances within several ten-thousandths of an inch. The metallic spacers are nearly invisible in an operating display and do not interfere with the ionization process.

Many engineers at IBM contributed to the innovations that enabled the mass production of the plasma panels used in the IBM 3290 Information Panel. Their contributions are only part of IBM's continuing commitment to research, development, and manufacturing. IBM®

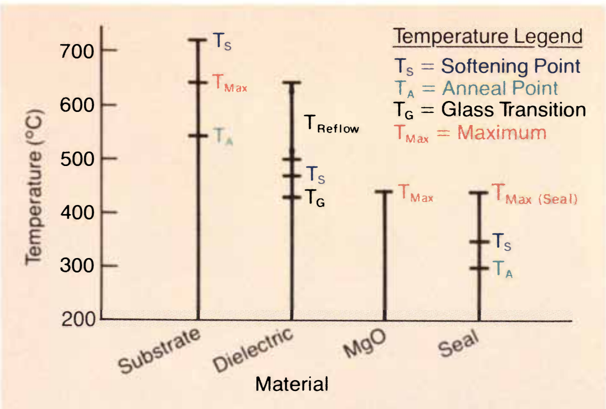


Figure 5. *To manufacture the AC plasma panel, IBM developed a lower-temperature dielectric glass and seal material to fit the thermal hierarchy requirements shown here.*

For free additional information on AC plasma display technology, please write: IBM Corporation, Dept. 31H/978E Neighborhood Road, Kingston, NY 12401


Product of France. Made with fine cognac brandy 80 proof. ©1983 Carillon Importers, Ltd., N.Y.
Grand Marnier Liqueur
Beyond the best is the only place to be.
Grand Marnier.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MAY, 1934: "Missionaries, government officials and scientists whose work takes them into regions where mosquitoes carrying yellow fever may bite them are now being adequately protected against this toll-taking disease. For two years medical research scientists of the Rockefeller Foundation, working in New York at the Rockefeller Institute for Medical Research under the leadership of Dr. Wilbur A. Sawyer, have applied their vaccination technique. Yellow-fever virus made safer by at least 100 passages through white mice is used. When three years ago vaccination against yellow fever was announced to the medical world, it was not known how long the immunity conferred by inoculations with immune blood serum would last. Experience has shown it to be exceedingly efficient. Protection lasts at least two years."

"With regard to the prediction of earthquakes, seismology has not yet reached the stage where we can foretell quakes ahead of time, but investigations in this direction being carried out in Japan give hope that the time is not far distant when such prediction will be possible. It has been noticed that in earthquake regions the earth shows evidence of tilt or gradual rising for some years before the quake occurs, much as the inner tube of a tire or the bladder of a football rises gradually through a break in the cover before finally bursting. The tilt of the ground is being carefully observed and measured, and it is hoped that it will finally give the clue to the forecasting of earthquakes."

"On a clear summer evening, as we look northeastward and up into the sky, we may exclaim, 'Why, there are clouds about after all. Look at the dark strip across the Milky Way.' Sure enough, in the constellation Cygnus a narrow band lies athwart the luminous background. It was long supposed that these dark areas represented actual gaps in the star clouds of the galaxy, through which we looked into the inky depths of empty space. The distinction of discovering their real nature belongs to the late Edward Emerson Barnard, who advanced and conclusively proved the thesis that they are produced by actual clouds of absorbing material far out in interstellar space. It is really very remarkable that so much of the matter in space is condensed into luminous stars. Whether the obscuring clouds represent something that never concentrated into stars, or something that once formed part of a star and has been ejected in some way, we do not yet know."

"Recently the Comstock Prize of $2,500 was presented by the National Academy of Sciences to P. W. Bridgman, professor of mathematics and natural philosophy at Harvard University, for his brilliant achievements in advancing our knowledge of the behavior of matter. The National Academy says of his work: 'Most of it falls into three categories: the first, so peculiarly Bridgman's own, the behavior of materials under high pressure; the second, the properties of single crystals at normal pressures, and third, the application of thermodynamics to electrical phenomena.' It was Professor Bridgman who devised and used apparatus for applying pressures up to 600,000 pounds per square inch, the highest pressures ever attained artificially."

SCIENTIFIC AMERICAN

MAY, 1884: "The colossal Statue of Liberty, which is to be placed as a pharos in Upper New York Bay, is at last finished. It is erected to its full height in the court of the *ateliers* of MM. Gayet-Gauthier, whence it will soon be taken in three separate pieces and transported to America. It is the greatest work of this kind that has ever been accomplished. The colossal statues of Memnon are only 62 feet high; the Olympian Zeus, by Phidias, was only 43 feet high; the Colossus of Rhodes, 130 feet; the statue of Nero by Zenodore, about 118 feet; the statue of Arminius, in Westphalia, is about 92 feet high, including the sword, and the statue of St. Carlo Borromeo, on Lake Maggiore, measures 77 feet. Much greater are the dimensions of the Statue of Liberty, by M. Bartholdi. This statue measures 152 feet from the base to the top of the torch carried in the right hand. The total weight of the statue is 100,000 pounds, of which 40,000 is copper and 60,000 pounds iron."

"What should be done with the wires that, like threads in a huge web, hold the New York streets in a mesh-like tangle is a question as perplexing as it is important. The Board of Aldermen says that such of them as pertain to electric lighting plants must be buried after May, 1885; one branch of the legislature says that both electric-light and telephone wires must be underground by June 1, 1885, in cities of 50,000 inhabitants or more. No reasonable person will deny that electric mains carrying powerful and dangerous currents should be placed out of reach. But telegraphic and telephonic wires are dangerous neither to life nor to property, and so far as they are concerned the question would seem to be: Is their unsightliness and the annoyance attendant upon their repair of more importance to the general public than the additional expense of their use that it is but reasonable to look for when the costly subways and tunnels shall have been built?"

"The marvelous advance in the construction of business and residence dwellings in the large capitals of the world is clearly due to the use of the elevator. The substitution of steam for human strength in lifting people from the ground to the upper floors of structures has radically changed not only the appearance of the streets but also our methods of living and doing business. What a change has occurred in our leading business streets! Of course, high buildings have their disadvantages. They imperil life if they are not incombustible. They exclude light and air from neighboring houses, and then they are manifestly out of place on narrow streets. They are, however, so comfortable, convenient and profitable that no law can stand, even if enacted, limiting their height. No legislation will avail against the inevitable."

"M. Olszewski recently stated in the *Comptes rendus* that he has liquefied hydrogen by the aid of liquid nitrogen, his previous use of liquid oxygen being unsatisfactory. The nitrogen was compressed to 60 atmospheres and cooled in a gas tube to -142 degrees centigrade for a considerable time by the aid of ethylene evaporating in a vacuum, and in this way it was liquefied. Hydrogen contained in a glass tube of about 4.5 millimeters internal diameter was plunged into the liquid. While the nitrogen evaporated in the vacuum and the pressure of hydrogen fell from 160 to 40 atmospheres, the hydrogen was observed to condense into a colorless transparent liquid running down the sides of the tube."

"The conditions to be fulfilled by a modern ironclad are complex. She must have thick armor, heavy guns, great speed and be capable of steaming long distances. This appears to be an almost impossible combination of good qualities. The more credit, therefore, is due to the Brazilian officers who have succeeded in turning out a 6,000-ton craft that really appears to embody the impossible. It is to be hoped that peace may reign between the United States and Brazil; if not, the *Riachuelo* alone could destroy every ship of war the United States possesses, fighting them half a dozen at a time."

THE AUTHORS

JACK ROBERT MILLER ("Steel Minimills") is a consulting engineer and economist who has been in the iron and steel industry for more than 40 years. He holds degrees in electrical engineering from the Cooper Union Institute of Engineering and in economics from New York University. In addition to working for several major steel companies in the U.S. he has served as a staff member of the United Nations Centre for Industrial Development, the World Bank, the Inter-American Development Bank and the Battelle Institute. He writes: "After 40 years as a consultant I have worked on nearly 50 overseas projects and at least as many in the U.S. In 1976 I had to retire for medical reasons. After six years of unwilling retirement I went back to work. In 1983 I reestablished my consulting office in St. Petersburg, Fla., but with a difference: we are a group of retirees who do not want to be retired. We have been around the steel industry too long and know too much about its problems and their solutions to leave it so soon."

ENRICO BONATTI and KATHLEEN CRANE ("Oceanic Fracture Zones") are members of the staff of the Lamont-Doherty Geological Observatory of Columbia University who are collaborating on the work described in their article. Bonatti, a native of Italy who is a senior research scientist at the observatory, writes: "I obtained degrees in geological science at the University of Pisa and the Scuola Normale Superiore in Pisa. In 1960 I came to the U.S. on a Fulbright fellowship. I spent one year at Yale University and four years at the Scripps Institution of Oceanography. I was professor of marine geology at the University of Miami for several years before moving to Lamont in 1976." Crane holds a B.A. from Oregon State University and a Ph.D. from the Scripps Institution of Oceanography. She writes: "During my 10 years in oceanography and 25 expeditions at sea I became interested in the vertical dynamics of ocean crust. I joined forces with Enrico Bonatti in the Red Sea, where a newly forming midocean ridge is ramming into the Egyptian coast, thereby pushing up the island of Zabargad. The island is part of an uplifted paleoshear zone disguised as a lovely subtropical desert island: the nirvana of all oceanographers who have spent too much time at sea."

ALICE DAUTRY-VARSAT and HARVEY F. LODISH ("How Receptors Bring Proteins and Particles into Cells") are molecular biologists who worked together on the subject of their article when both of them were at the Massachusetts Institute of Technology. Born in France, Dautry-Varsat received her master's degree in solid-state physics at the University of Paris. After getting it she decided to pursue molecular biology and came to the U.S., where she obtained a second master's degree in molecular biology at the State University of New York at Stony Brook. She returned to France to earn her doctorate at the Pasteur Institute in Paris. Thereafter she worked at the Medical Research Council Laboratory of Molecular Biology in England and at M.I.T. before joining the staff of the Pasteur Institute. Lodish is professor of biology at M.I.T. He was graduated from Kenyon College and went on to get his Ph.D. from Rockefeller University. He joined the faculty of M.I.T. in 1968. This year he will move his laboratory to the Whitehead Institute for Biomedical Research on the M.I.T. campus.

JOHN E. HOPCROFT ("Turing Machines") is professor of computer science at Cornell University. He has three degrees in electrical engineering: a B.S. from Seattle University (1961) and an M.S. (1962) and Ph.D. (1964) from Stanford University. After getting his doctorate he was on the faculty of Princeton University for three years before going to Cornell, where he has remained except for a year spent as visiting associate professor at Stanford.

JANICE MOORE ("Parasites That Change the Behavior of Their Host") is on the faculty of Colorado State University. She writes: "As an undergraduate at Rice University I took a parasitology course for reasons I have forgotten and finished it with a permanent sense of wonder at such diversity and adaptation. I completed a master's degree in insect behavior at the University of Texas at Austin and set out to study parasites. The field of parasite ecology was viewed then as an interdisciplinary one, and I was urged to choose a more mainstream project. Instead I switched schools twice and worked at several editorial and technical jobs. I then met some members of the biology faculty of the University of New Mexico, and with their encouragement I entered a kid-in-a-candy-shop world where I have remained."

ALAN H. GUTH and PAUL J. STEINHARDT ("The Inflationary Universe") are physicists who share an interest in the early history of the universe. Guth went to the Massachusetts Institute of Technology as an undergraduate and a graduate student; his Ph.D. in physics was awarded by M.I.T. in 1972. He writes: "I held postdoctoral positions at Princeton University, Columbia University, Cornell University and the Stanford Linear Accelerator Center (SLAC). During most of that time I worked on rather abstract mathematical problems in elementary-particle theory and knew no more about developments in cosmology than the average layman does. While I was at Cornell, however, Henry Tye, a fellow postdoctoral worker, persuaded me (with great difficulty) to join him in studying the production of magnetic monopoles in the early universe, and that was how my career changed direction. I continued the work in the following academic year at SLAC. Shortly thereafter I returned to M.I.T. as associate professor of physics, the job I now hold." Steinhardt was graduated from the California Institute of Technology with a B.S. in 1974. His M.A. (1975) and Ph.D. (1978) in physics are from Harvard University. From 1979 to 1981 he was a junior fellow in the Society of Fellows at Harvard. In 1981 he moved to the University of Pennsylvania, where he is associate professor of physics.

J. M. ADOVASIO and R. C. CARLISLE ("An Indian Hunters' Camp for 20,000 Years") are members of the department of anthropology at the University of Pittsburgh. Adovasio is professor of anthropology and chairman of the department. He was graduated from the University of Arizona with a B.A. and went on to earn his Ph.D. from the University of Utah in 1970. His scientific interests include the excavation of closed archaeological sites such as caves and rock shelters and the techniques by which baskets were made in ancient times. Outside the laboratory Adovasio is a rider of high-performance motorcycles and a competitive weight lifter. Carlisle is assistant to Adovasio, research assistant instructor of anthropology and editor in the university's Cultural Resource Management Program. He is currently working on his doctorate at Pittsburgh.

ANDREW CLIFF and PETER HAGGETT ("Island Epidemics") are geographers who have worked together since 1968 on the quantitative aspects of geography, in particular the problem of how cultural innovations, economic forms and diseases are transmitted spatially in human populations. Cliff was educated at King's College of the University of London, Northwestern University and the University of Bristol. His Ph.D. in geography was awarded by Bristol in 1969. In 1972 he joined the faculty of the University of Cambridge. Haggett is professor of urban and regional geography at Bristol. He was educated at Cambridge and taught there and at University College London before joining the Bristol faculty in 1966.

COMPUTER RECREATIONS

In the game called Core War hostile programs engage in a battle of bits

by A. K. Dewdney

Two computer programs in their native habitat—the memory chips of a digital computer—stalk each other from address to address. Sometimes they go scouting for the enemy; sometimes they lay down a barrage of numeric bombs; sometimes they copy themselves out of danger or stop to repair damage. This is the game I call Core War. It is unlike almost all other computer games in that people do not play at all! The contending programs are written by people, of course, but once a battle is under way the creator of a program can do nothing but watch helplessly as the product of hours spent in design and implementation either lives or dies on the screen. The outcome depends entirely on which program is hit first in a vulnerable area.

The term Core War originates in an outmoded memory technology. In the 1950's and 1960's the memory system of a computer was built out of thousands of ferromagnetic cores, or rings, strung on a meshwork of fine wires. Each core could retain the value of one bit, or binary digit, the fundamental unit of information. Nowadays memory elements are fabricated on semiconductor chips, but the active part of the memory system, where a program is kept while it is being executed, is still often referred to as core memory, or simply core.

Battle programs in Core War are written in a specialized language I have named Redcode, closely related to the class of programming languages called assembly languages. Most computer programs today are written in a high-level language such as Pascal, Fortran or BASIC; in these languages a single statement can specify an entire sequence of machine instructions. Moreover, the statements are easy for the programmer to read and to understand. For a program to be executed, however, it must first be translated into "machine language," where each instruction is represented by a long string of binary digits. Writing a program in this form is tedious at best.

Assembly languages occupy an intermediate position between high-level languages and machine code. In an assembly-language program each statement generally corresponds to a single instruction and hence to a particular string of binary digits. Rather than writing the binary numbers, however, the programmer represents them by short words or abbreviations called mnemonics (because they are easier to remember than numbers). The translation into machine code is done by a program called an assembler.

Comparatively little programming is done in assembly languages because the resulting programs are longer and harder to understand or modify than their high-level counterparts. There are some tasks, however, for which an assembly language is ideal. When a program must occupy as little space as possible or be made to run as fast as possible, it is generally written in assembly language. Furthermore, some things can be done in an assembly language that are all but impossible in a high-level language. For example, an assembly-language program can be made to modify its own instructions or to move itself to a new position in memory.

Core War was inspired by a story I heard some years ago about a mischievous programmer at a large corporate research laboratory I shall designate *X*. The programmer wrote an assembly-language program called Creeper that would duplicate itself every time it was run. It could also spread from one computer to another in the network of the *X* corporation. The program had no function other than to perpetuate itself. Before long there were so many copies of Creeper that more useful programs and data were being crowded out. The growing infestation was not brought under control until someone thought of fighting fire with fire. A second self-duplicating program called Reaper was written. Its purpose was to destroy copies of Creeper until it could find no more and then to destroy itself. Reaper did its job, and things were soon back to normal at the *X* lab.

In spite of fairly obvious holes in the story, I believed it, perhaps because I wanted to. It took some time to track down the real events that lay behind this item of folklore. (I shall give an account of them below.) For now it is sufficient to note that my desire to believe rested squarely on the intriguing idea of two programs doing battle in the dark and noiseless corridors of core.

Last year I decided that even if the story turned out not to be true, something like it could be made to happen. I set up an initial version of Core War and, assisted by David Jones, a student in my department at the University of Western Ontario, got it working. Since then we have developed the game to a fairly interesting level.

Core War has four main components: a memory array of 8,000 addresses, the assembly language Redcode, an executive program called MARS (an acronym for Memory Array Redcode Simulator) and the set of contending battle programs. Two battle programs are entered into the memory array at randomly chosen positions; neither program knows where the other one is. MARS executes the programs in a simple version of time-sharing, a technique for allocating the resources of a computer among numerous users. The two programs take turns: a single instruction of the first program is executed, then a single instruction of the second, and so on.

What a battle program does during the execution cycles allotted to it is entirely up to the programmer. The aim, of course, is to destroy the other program by ruining its instructions. A defensive strategy is also possible: a program might undertake to repair any damage it has received or to move out of the way when it comes under attack. The battle ends when MARS comes to an instruction in one of the programs that cannot be executed. The program with the faulty instruction—which presumably is a casualty of war—is declared the loser.

Much can be learned about a battle program merely by analyzing its actions mentally or with pencil and paper. To put the program to the test of experience, however, one needs access to a computer and a version of MARS. The programs could be made to operate on a personal computer, and Jones and I have prepared brief guidelines for those who would like to set up a Core War battlefield of their own. (For a copy of the guidelines send your name and address and $2 for postage and handling to Core War, *Scientific American,* 415 Madison Avenue, New York, N.Y., 10017. Delivery may take a few weeks.)

Before describing Redcode and introducing some simple battle programs, I

find the to these by 9 a.m.?

6. Which South American firms import electronic components from the U.S.?

7. How can I find all the orthodontists in Toledo for my direct mail campaign?

8. What are the mechanical properties of shape memory materials?

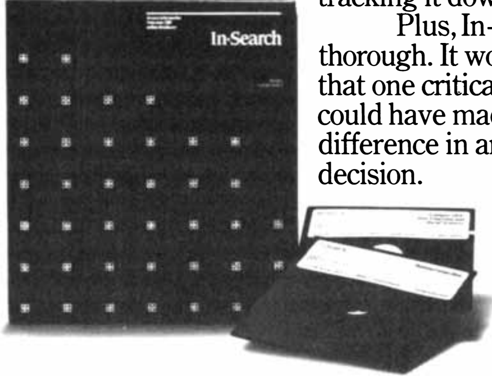
9. Has the cost-effectiveness of teleconferencing changed in the last year?

10. Who are some of the experts in medical sonar scanning techniques?

Not anymore. Just enter the words or subject you'd like to research and In-Search takes it from there, bringing you articles or references containing the information you need. Simple, on-screen graphics guide you through every step.

Fast Answers.

In-Search can find in minutes what would take an individual days or weeks to find using conventional methods. You can now spend more time *using* information because you spend less time tracking it down.



Plus, In-Search is thorough. It won't overlook that one critical fact that could have made the difference in an important decision.

Once you have the information, you can store it on disk or print it out. In-Search is compatible with most word processing programs.

Any Questions?

To learn more about In-Search, call **(408) 986-1200** today. We'll send you an offline demonstration diskette designed for IBM, IBM-compatible, or Texas Instruments personal computers. Use your VISA or MasterCard or charge the purchase of your $5.00 diskette. (California residents please add applicable sales tax.)

In·Search™, Menlo Corporation
4633 Old Ironsides, Suite 400, Santa Clara, CA 95050

☐ Enclosed is $5.00 for my In-Search demonstration diskette.
☐ Please send me a free booklet with more details.

Name ____________________
Address ____________________
City __________ State ______ Zip ______
My occupation is: ____________________

Menlo Corporation™

should say more about the memory array. Although I have noted that it consists of 8,000 addresses, there is nothing magical about this number; a smaller array would work quite well. The memory array differs from most computer memories in its circular configuration: it is a sequence of addresses numbered from 0 to 7999 but it thereupon rejoins itself, so that address 8000 is equivalent to address 0. MARS always reduces an address greater than 7999 by taking the remainder after division by 8000. Thus if a battle program orders a hit at address 9378, MARS interprets the address as 1378.

Redcode is a simplified, special-purpose assembly-style language. It has instructions to move the contents of one address in memory to another address, to alter the contents arithmetically and to transfer control forward or backward within a program. Whereas the output of a real assembler consists of binary codes, the mnemonic form of a Redcode instruction is translated by MARS into a large decimal integer, which is then stored in the memory array; each address in the array can hold one such integer. It is also MARS that interprets the integers as instructions and carries out the indicated operations.

A list of the elementary Redcode instructions is given in the top illustration on page 19. With each instruction the programmer is required to supply at least one argument, or value, and most of the instructions take two arguments. For example, in the instruction JMP −7 the mnemonic JMP (for "jump") is followed by the single argument −7. The instruction tells MARS to transfer control to the memory address seven places before the current one, that is, seven places before the JMP −7 instruction itself. If the instruction happened to be at address 3715, execution of the program would jump back to address 3708.

This method of calculating a position in memory is called relative addressing, and it is the only method employed in Redcode. There is no way for a battle program to know its own absolute position in the memory array.

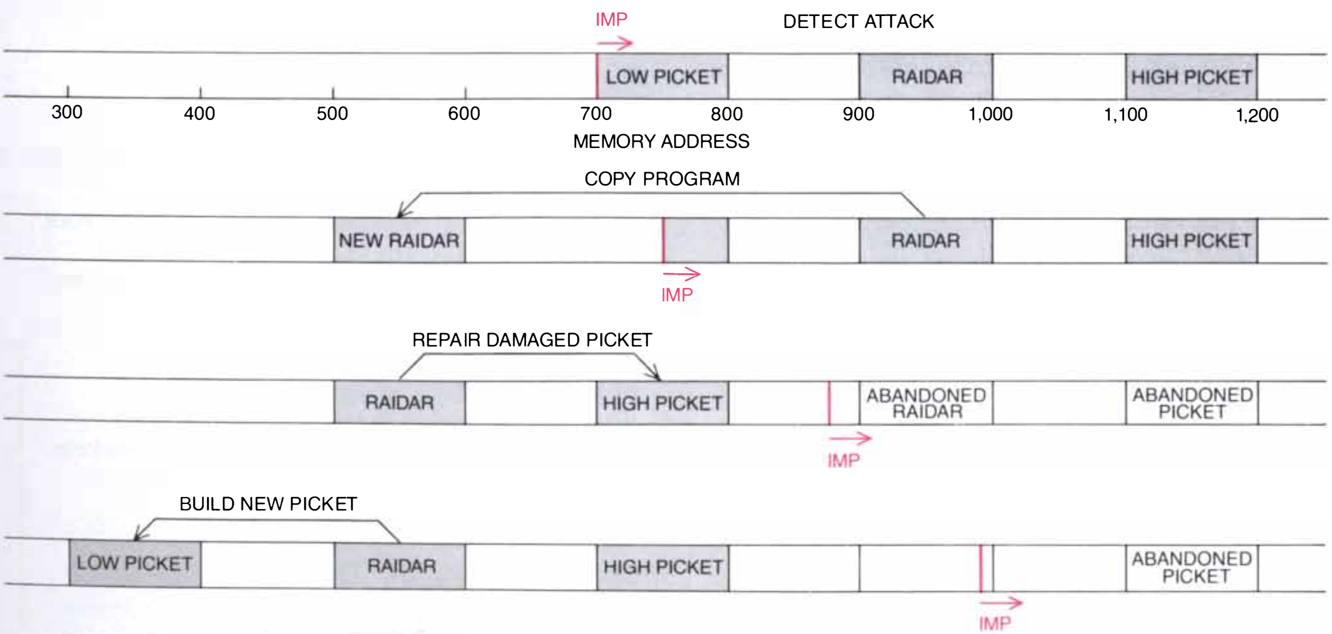
The instruction MOV 3 100 tells MARS to go forward three addresses, copy what it finds there and deliver it 100 addresses beyond the MOV instruction, overwriting whatever was there. The arguments in this instruction are given in "direct" mode, meaning they are to be interpreted as addresses to be acted on directly. Two other modes are allowed. Preceding an argument with an @ sign makes it "indirect." In the instruction MOV @3 100 the integer to be delivered to relative address 100 is not the one found at relative address 3 but rather the one found at the address specified by the contents of relative address 3. (The bottom illustration on page 19 gives more detail on the process of indirect addressing.) A # sign makes an argument "immediate," so that it is treated not as an address but as an integer. The instruction MOV #3 100 causes the integer 3 to be moved to relative address 100.

Most of the other instructions need no further explanation, but the data statement (DAT) requires some comment. It can serve as a work space to hold information a program may need to refer to. Strictly speaking, however, it is not an instruction; indeed, any memory location with a zero in its first decimal position can be regarded as a DAT statement and as such is not executable. If MARS should be asked to execute such an "instruction," it will not be able to and will declare that program the loser.

The decimal integer that encodes a Redcode instruction has several fields, or functional areas [*see middle illustration on page 19*]. The first digit represents the mnemonic itself, and two more digits identify the addressing mode (direct, indirect or immediate). In addition four digits are set aside for each argument. Negative arguments are stored in complement form: −1 would be represented as 7999, since in the circular memory array adding 7999 has the same effect as subtracting 1.

The instructions making up a simple battle program called Dwarf are listed in the illustration on page 20. Dwarf is a very stupid but very dangerous program that works its way through the memory array bombarding every fifth address with a zero. Zero is the integer signifying a nonexecutable data statement, and so a zero dropped into an enemy program can bring it to a halt.

Assume that Dwarf occupies absolute addresses 1 through 4. Address 1 initially contains DAT −1, but execution begins with the next instruction, ADD #5 −1. The effect of the instruction is to add 5 to the contents of the preceding address, namely the DAT −1 statement, thereby transforming it into DAT 4. Next Dwarf executes the instruction at absolute address 3, MOV #0 @−2. Here the integer to be moved is 0, specified as an immediate value. The target address is calculated indirectly in the following way. First MARS counts back two addresses from address 3, arriv-



Raidar, a sophisticated battle program, eludes the simpler Imp in the memory array of Core War

At what point do you become a serious photographer?

It starts with a glimmer. A spark in your mind, hinting that photography can be more than just a quick way to freeze a memory.

As that feeling grows, things like composition, angle, and lighting become an exciting challenge. Time consuming? Yes. Worth it? Definitely. And you notice that while others might hope for interesting pictures to happen, you can *make* them happen

Soon (though you can't exactly remember when), you begin to refer to your pictures as photographs. And nobody's laughing. Because in your photographs, you're capturing not just people and places, but feelings.

It is possibly at this point that you smile at how far you've come. Yet can't help wondering how far you can go

It is at this point that you can fully appreciate the Olympus OM-4.

In an age when most camera innovations are designed to give you less to do, the OM-4 has something that lets you do more. It's called Multi-Spot Metering. And it gives you what photographers have dreamed of since the beginning of photography itself: creative control of light.

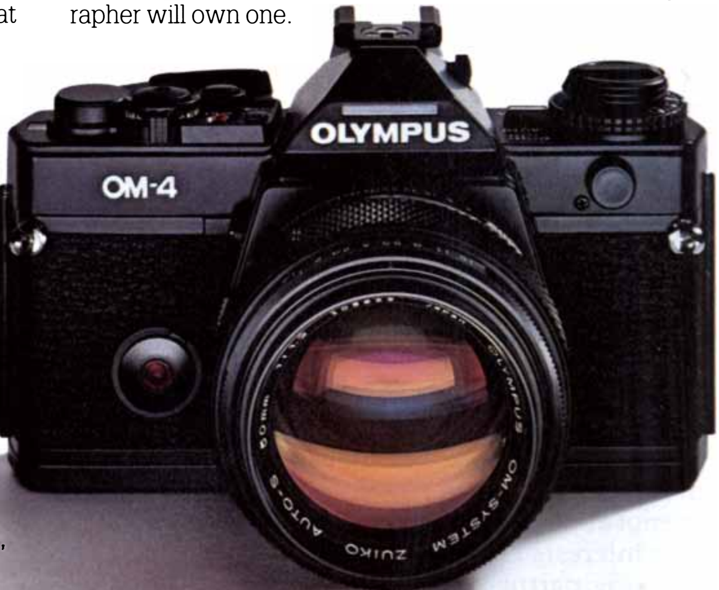
You see, most cameras, as sophisticated as they are, have an averaging meter, which averages all the highlights and shadows in your scene and gives you an exposure somewhere in between. You can't tell an averaging meter to give special attention to the areas you feel are most important. So, very often, you end up with a photograph that's, well, average

And that's where the Olympus OM-4 with Multi-Spot Metering makes a world of difference. Because now you can decide which part of your photograph deserves the best exposure. Just center your subject in the microprism section of the viewfinder, push the spot button to set in the reading, recompose, and shoot for the perfect exposure. In fact, you can take up to 8 individual readings for each shot; the OM-4's computer will balance them perfectly. And you can store a reading in the OM-4's memory for up to an hour when you're planning a whole sequence of shooting.

With the OM-4, you'll capture backlit, spotlit, and high contrast scenes as never before. Without bracketing. Without compensating. Without wasting a shot. The OM-4 even features Highlight Control and Shadow Control for the most dramatic effects imaginable.

You really have to see the OM-4 to believe it. Visit your Olympus dealer, or write for a brochure: Olympus, Box S, Woodbury, NY 11797. In Canada: W. Carsen Co. Ltd., Toronto.

The Olympus OM-4. We think every serious photographer will own one.



OLYMPUS®

WHEN YOU HAVE MORE TO SAY THAN JUST SMILE.

WHAT SHOULD YOU
WORKING PARTNERSHIP

RESULTS WITH ASSURANCE.

Deeds more than words have characterized our working partnerships with customers large and small. We feel the best assurances we can offer prospective customers are the productive experiences and results of our present customers.

"At Mead Johnson, two HP lab automation systems paid for themselves in just six months of operation."

Pat Bryan
Manager, Pharmaceutical Quality Control

"At M/A-COM, HP computer-based instrument systems speed final test 12-fold."

Dr. Alan Carlson
Manager, Automatic Test Equipment Group

Your long-term needs addressed.

Working partnership implies a long-term relationship, in which the initial sale is only the beginning. Our objective is to expand the relationship wherever the "fit" is good. It precludes solving all problems with only HP-available solutions. An exclusive relationship is not prerequisite. We are not timid about working with other organizations to achieve a more effective solution. And we are not afraid to say "no" in the best interests of the customer. Working partnership is a two-way street. It demands accuracy as well as advocacy.

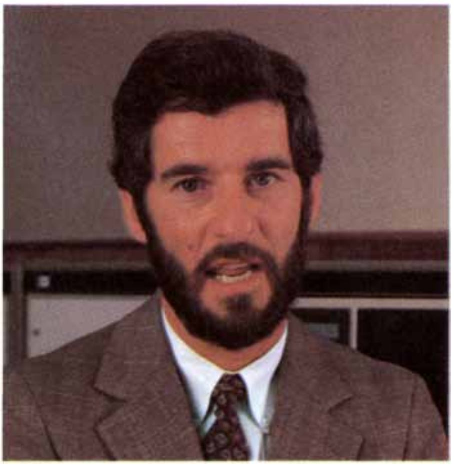
More than computers.

Hewlett-Packard is recognized for the range of its computer solutions in business, industry, science, and medicine throughout the world. Beyond computers, HP offers an enormous variety of instrumentation and computer-based systems. These enable us to work with customers to achieve productivity solutions in such areas as office automation; computer-aided design, manufacturing, and test; and laboratory automation. They also include hospital systems for intensive care and financial management. HP brings this range and versatility to a working partnership with customers.

Quality beyond products.

While quality has been directly related to products, we stretch the term to include all aspects of our relationship with customers. Why? Because the same quality and reliability engineered into HP products is built into our customer support programs, worldwide. From the outset, a local support team of specialists who understand your needs will help you develop a cost-effective strategy to meet your goals. HP

EXPECT FROM A
WITH HEWLETT-PACKARD?

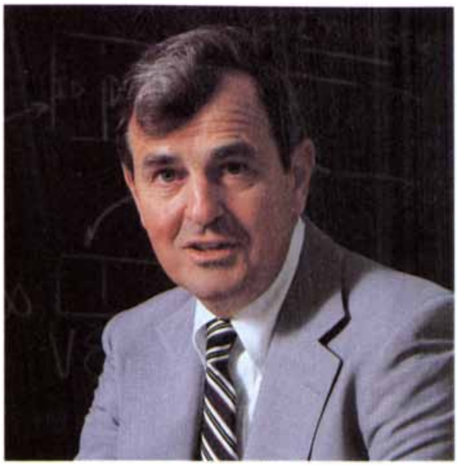


"At City National Bank, four HP computers deliver mainframe capability at one-eighth the cost."

Newton Gore
Sr. Vice President

"At Likoff Cardiovascular Institute, an HP computer processes electrocardiograms in minutes instead of hours."

Dr. Gary Anderson
Heart Station Director
Hahnemann University



"At 3M, an HP computer network helps increase productivity, control inventories, and optimize distribution."

Ewald F. Fischer
Manager, Information Systems

support takes many forms: help in selecting the right equipment, training programs, remote diagnostics that reduce service calls, service contracts that guarantee covered systems will be up and running at least 99 percent of the time—and a great deal more.

R & D that protects your investment.

Your cost of software is likely to exceed your cost of hardware many times over. This is one reason that much of HP's R & D concentrates on technology that will preserve your software investment while increasing the performance and reducing the cost of the hardware. What you buy today from Hewlett-Packard will not be obsolete tomorrow. In a working partnership, results as well as costs are measured over time. The best technology preserves and enhances your investment.

Satisfaction. Based on 45 years of meeting exacting customer needs.

From its beginnings, HP has built its business by satisfying the performance expectations of scientists and engineers. They can't accept doubtful performance, uncertain quality, or occasional support in meeting their demanding goals. It's against this background of unequivocal performance that HP builds working partnerships with its customers. After all, partnership is inherent in our name—and is fundamental to our dealings.

In summary, a working partnership with Hewlett-Packard simply means:

Productivity. Not promises.

HEWLETT PACKARD

0002404

Introducing the *1985*

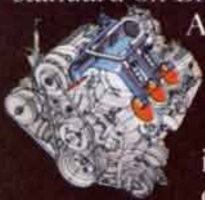
Ninety Eight Regency

Precise road management engineered in a luxury automobile.

High-tech engineering, luxury for six. That's the new Ninety-Eight Regency Brougham. Each wheel has its own independent suspension to give you a smooth, controlled ride. You feel the road, but not the bumps.

Four-Wheel Independent Suspension

A spirited new 3.8 liter multi-port fuel injected engine is standard on Brougham models.



3.8L Multi-Port Fuel Injected Engine

A fuel metering system precisely regulates fuel for optimum performance. An optional electronic instrument panel for precise readouts is available after June, 1984.

Even a voice information system and auto calculator can be added to give you additional assists while driving. An electronic load leveler automatically keeps your Ninety-Eight level, with varying passenger or trunk loads.

Electronic Instrument Panel

This Ninety-Eight also features a new 3-year/36,000-mile, limited new-car warranty. A deductible may apply. See your dealer for details and a test drive.

Electronic Load Leveler

Some Oldsmobiles are equipped with engines produced by other GM divisions, subsidiaries or affiliated companies worldwide. See your dealer for details.

There is a special feel in an *Oldsmobile* ing at address 1. It then examines the data value there, namely 4, and interprets it as an address relative to the current position; in other words, it counts four places forward from address 1 and hence deposits a 0 at address 5.

The final instruction in Dwarf, JMP −2, creates an endless loop. It directs execution back to absolute address 2, which again increments the DAT statement by 5, making its new value DAT 9. In the next execution cycle a 0 is therefore delivered to absolute address 10. Subsequent 0 bombs will fall on addresses 15, 20, 25 and so on. The program itself is immobile but its artillery threatens the entire array. Eventually Dwarf works its way around to addresses 7990, 7995 and then 8000. As far as MARS is concerned, 8000 is equal to 0, and so Dwarf has narrowly avoided committing suicide. Its next missile again lands on address 5.

It is sobering to realize that no stationary battle program that has more than four instructions can avoid taking a hit from Dwarf. The opposing program has only three options: to move about and thereby elude the bombardment, to absorb hits and repair the damage or to get Dwarf first. To succeed through the last strategy the program may have to be lucky: it can have no idea where Dwarf is in the memory array, and on the average it has about 1,600 execution cycles before a hit is received. If the second program is also a Dwarf, each program wins 30 percent of the time; in 40 percent of the contests neither program scores a fatal hit.

Before taking up the other two strategies, I should like to introduce a curious one-line battle program we call Imp. Here it is:

MOV 0 1

Imp is the simplest example of a Redcode program that is able to relocate itself in the memory array. It copies the contents of relative address 0 (namely MOV 0 1) to relative address 1, the next address. As the program is executed it moves through the array at a speed of one address per cycle, leaving behind a trail of MOV 0 1 instructions.

What happens if we pit Imp against Dwarf? The barrage of zeros laid down by Dwarf moves through the memory array faster than Imp moves, but it does not necessarily follow that Dwarf has the advantage. The question is: Will Dwarf hit Imp even if the barrage does catch up?

If Imp reaches Dwarf first, Imp will in all probability plow right through Dwarf's code. When Dwarf's JMP −2 instruction transfers execution back two steps, the instruction found there will be Imp's MOV 0 1. As a result Dwarf will be subverted and become a second Imp

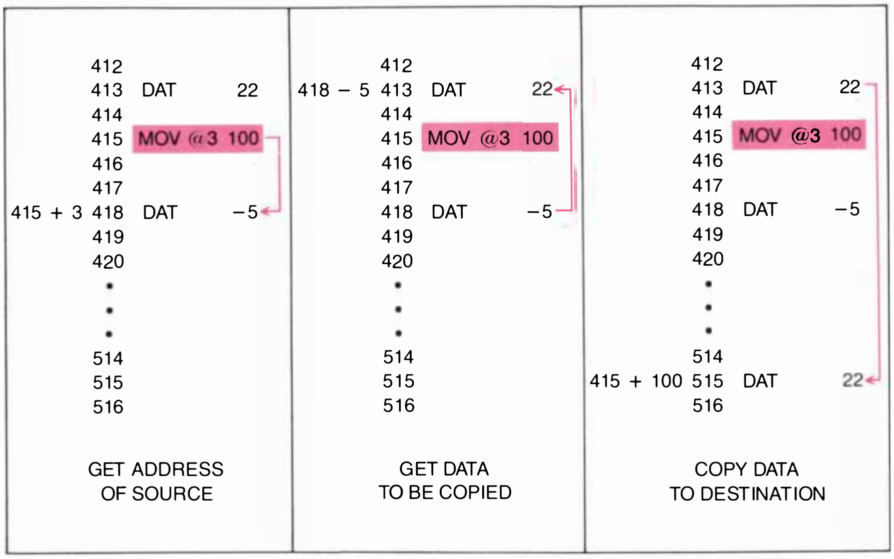
INSTRUCTION	MNEMONIC	CODE	ARGUMENTS		EXPLANATION
Move	MOV	1	A	B	Move contents of address *A* to address *B*.
Add	ADD	2	A	B	Add contents of address *A* to address *B*.
Subtract	SUB	3	A	B	Subtract contents of address *A* from address *B*.
Jump	JMP	4	A		Transfer control to address *A*.
Jump if zero	JMZ	5	A	B	Transfer control to address *A* if contents of address *B* are zero.
Jump if greater	JMG	6	A	B	Transfer control to address *A* if contents of *B* are greater than zero.
Decrement: jump if zero	DJZ	7	A	B	Subtract 1 from contents of address *B* and transfer control to address *A* if contents of address *B* are then zero.
Compare	CMP	8	A	B	Compare contents of addresses *A* and *B*; if they are unequal, skip the next instruction.
Data statement	DAT	0		B	A nonexecutable statement; *B* is the data value.

The instruction set of Redcode, an assembly language for Core War

MNEMONIC	ARGUMENT *A*	ARGUMENT *B*	OPERATION CODE	MODE DIGIT: ARGUMENT *A*	MODE DIGIT: ARGUMENT *B*	ARGUMENT *A*	ARGUMENT *B*
DAT		−1	0	0	0	0000	7999
ADD	#5	−1	2	0	1	0005	7999
MOV	#0	@ −2	1	0	2	0000	7998
JMP	−2		4	1	0	7998	0000

ADDRESSING MODES:			
	IMMEDIATE	#	0
	DIRECT		1
	INDIRECT	@	2

The encoding of Redcode instructions as decimal integers



The three-step mechanism of indirect relative addressing

ADDRESS	CYCLE 1	CYCLE 2	CYCLE 9
0			
1	DAT −1	DAT 4	DAT 14
2	ADD #5 −1	ADD #5 −1	ADD #5 −1
3	MOV #0 @−2	MOV #0 @−2	MOV #0 @−2
4	JMP −2	JMP −2	JMP −2
5		— 0	— 0
6			
7			
8			
9			— 0
10			
11			
12			
13			
14			— 0
15			
16			
17			

Dwarf, a battle program, lays down a barrage of "zero bombs"

endlessly chasing the first one around the array. Under the rules of Core War the battle is a draw. (Note that this is the outcome to be expected "in all probability." Readers are invited to analyze other possibilities and perhaps discover the bizarre result of one of them.)

Both Imp and Dwarf represent a class of programs that can be characterized as small and aggressive but not intelligent. At the next level are programs that are larger and somewhat less aggressive but smart enough to deal with programs in the lower class. The smarter programs have the ability to dodge an attack by copying themselves out of trouble. Each such program includes a segment of code somewhat like the one named Gemini, shown in the lower illustration on page 22. Gemini is not intended to be a complete battle program. Its only function is to make a copy of itself 100 addresses beyond its present position and then transfer execution to the new copy.

The Gemini program has three main parts. Two data statements at the beginning serve as pointers: they indicate the next instruction to be copied and its destination. A loop in the middle of the program does the actual copying, moving each instruction in turn to an address 100 places beyond its current position. On each transit through the loop both pointers are incremented by 1, thereby designating a new source and destination address. A compare instruction (CMP) within the loop tests the value of the first data statement; when it has been incremented nine times, the entire program has been copied, and so an exit from the loop is taken. One final adjustment remains to be made. The destination address is the second statement in the program and it has an initial value of DAT 99; by the time it is copied, however, it has already been incremented once, so that in the new version of the program it reads DAT 100. This transcription error is corrected (by the instruction MOV #99 93) and then execution is transferred to the new copy.

By modifying Gemini it is possible to create an entire class of battle programs. One of these, Juggernaut, copies itself 10 locations ahead instead of 100. Like Imp, it tries to roll through all its opposition. It wins far more often than Imp, however, and leads to fewer draws, because an overwritten program is less likely to be able to execute fragments of Juggernaut's code. Bigfoot, another program employing the Gemini mechanism, makes the interval between copies a large prime number. Bigfoot is hard to catch and has the same devastating effect on enemy code as Juggernaut does.

Neither Bigfoot nor Juggernaut is very intelligent. So far we have written only two battle programs that qualify for the second level of sophistication. They are too long to reproduce here. One of them, which we call Raidar, maintains two "pickets" surrounding the program itself [*see illustration on page 18*]. Each picket consists of 100 consecutive addresses filled with 1's and is separated from the program by a buffer zone of 100 empty addresses. Raidar divides its time between systematically attacking distant areas of the memory array and checking its picket addresses. If one of the pickets is found to be altered, Raidar interprets the change as evidence of an attack by Dwarf, Imp or some other unintelligent program. Raidar then copies itself to the other side of the damaged picket, restores it, constructs a new picket on its unprotected side and resumes normal operation.

In addition to copying itself a battle program can be given the ability to repair itself. Jones has written a self-repairing program that can survive some attacks, although not all of them. Called Scanner, it maintains two copies of itself but ordinarily executes only one of them. The copy that is currently running periodically scans the other copy to see if any of its instructions have been altered by an attack. Changes are detected by comparing the two copies, always assuming that the executing copy is correct. If any bad instructions are found, they are replaced and control is transferred to the other copy, which then begins to scan the first one.

So far Scanner remains a purely defensive program. It is able to survive attacks by Dwarf, Imp, Juggernaut and similar slow-moving aggressors—at least if the attack comes from the right direction. Jones is currently working on a self-repair program that keeps three copies of itself.

I am curious to see whether readers can design other kinds of self-repairing programs. For example, one might think about maintaining two or more copies of a program even though only one copy is ever executed. The program might include a repair section that would refer to an extra copy when restoring damaged instructions. The repair section could even repair itself, but it might still be vulnerable to damage at some positions. One measure of vulnerability assumes that a single instruction has been hit; on the average, how many such instructions, if they are hit, ultimately cause the program to die? By this measure, what is the least vulnerable self-repairing program that can be written?

Only if reasonably robust programs can be developed will Core War reach the level of an exciting game, where the emphasis is shifted from defense to offense. Battle programs will then have to seek out and identify enemy code and mount an intensive attack wherever it is found.

I may have given the impression that Redcode and the entire MARS system are fixed. They are not. In spare moments we have been experimenting with new ideas and are certainly open to suggestions. Indeed, we have been experimenting so much with new programs and new features that some battles remain to be fought in our own system.

One idea we have been playing with is to include an extra instruction that would make self-repair or self-protection a little easier. The instruction PCT *A* would protect the instruction at address *A* from alteration until it is next executed. How much could the vulnerability of a program be reduced by exploiting an instruction of this kind?

In the guidelines offered above we de-

7978	MOV	0	1	
7979	MOV	0	1	
7980	—	0		
7981	MOV	0	1	
7982	MOV	0	1	
7983	MOV	0	1	
7984	MOV	0	1	
7985	—	0		
7986	MOV	0	1	
7987	MOV	0	1	
7988	MOV	0	1	
7989	MOV	0	1	
7990	—	0		
7991	MOV	0	1	
7992	MOV	0	1	
7993	MOV	0	1	
7994	MOV	0	1	} IMP
7995				
7996				
7997				
7998				
7999				
0				
1	DAT		7994	DWARF
2	ADD	#5	−1	
3	MOV	#0	@−2	
4	JMP	−2		
5	—	0		
6				
7				
8				
9				
10	—	0		
11				

Imp v. Dwarf: Who wins?

scribe not only the rules of Core War but also how to set up a memory array and write a MARS system in various high-level languages. We also suggest how to display the results of a Core War battle. For now the following rules define the game with enough precision to enable pencil-and-paper players to begin designing battle programs:

1. The two battle programs are loaded into the memory array at random positions but initially are no closer than 1,000 addresses.

2. MARS alternates in executing one instruction from each program until it reaches an instruction that cannot be executed. The program with the erroneous instruction loses.

3. Programs can be attacked with any available weapon. A "bomb" can be a 0 or any other integer, including a valid Redcode instruction.

4. A time limit is put on each contest, determined by the speed of the computer. If the limit is reached and both programs are still running, the contest is a draw.

The story of Creeper and Reaper seems to be based on a compounding of two actual programs. One program was a computer game called Darwin, invented by M. Douglas McIlroy of AT&T Bell Laboratories. The other was called Worm and was written by John F. Shoch of the Xerox Palo Alto Research Center. Both programs are some years old, allowing ample time for rumors to blossom. (Darwin was described in *Software: Practice and Experience,* Volume 2, pages 93–96, 1972. A vague description of what appears to be the same game is also given in the 1978 edition of *Computer Lib.*)

In Darwin each player submits a number of assembly-language programs called organisms, which inhabit core memory along with the organisms of other players. The organisms created by one player (and thus belonging to the same "species") attempt to kill those of other species and occupy their space. The winner of the game is the player whose organisms are most abundant when time is called. McIlroy invented an unkillable organism, although it won only "a few games." It was immortal but apparently not very aggressive.

Worm was an experimental program designed to make the fullest use possible of minicomputers linked in a network at Xerox. Worm was loaded into quiescent machines by a supervisory program. Its purpose was to take control of the machine and, in coordination with Worms inhabiting other quiescent machines, run large applications programs in the resulting multiprocessor system. Worm was designed so that anyone who wanted to use one of the occupied machines could readily reclaim it without interfering with the larger job.

One can see elements of both Darwin and Worm in the story of Creeper and Reaper. In Core War, Reaper has become reality.

	DAT		0	/pointer to source address
	DAT		99	/pointer to destination address
LOOP	MOV	@−2	@−1	/copy source to destination
LOOP	CMP	−3	#9	/if all 10 lines have been copied . . .
LOOP	JMP	4		/. . . then leave the loop;
LOOP	ADD	#1	−5	/otherwise, increment the source address . . .
LOOP	ADD	#1	−5	/. . . and the destination address . . .
LOOP	JMP	−5		/. . . and return to the loop
	MOV	#99	93	/restore the starting destination address
	JMP	93		/jump to the new copy

Gemini, a program that copies itself to a new position in the memory array

BOOKS

Calculating prodigies, the causes of color, a tragedy of a black biologist, vegetables

by Philip Morrison

THE GREAT MENTAL CALCULATORS: THE PSYCHOLOGY, METHODS, AND LIVES OF CALCULATING PRODIGIES, PAST AND PRESENT, by Steven B. Smith. Columbia University Press ($25). This art is old. The first calculating prodigy of whom we know was John Wallis, linguist, cryptanalyst and mathematician, whose treatise on limits led the young Newton to his earliest great success. In 1670, not for the first time, Wallis' journal recounts how he satisfied the curiosity of a learned visitor: "I did that night propose to myself in the dark without help to my memory a number in 53 places... of which I extracted the square root in 27 places... when I did from memory dictate them" about a month later. The recorded number does contain an error, but the root is correct, rounded off to the nearest integer, and allows reconstruction.

The art lives. A photograph and an interview record the engaging personality of Hans Eberstark, who also contributes a personal comment to this remarkable study. Eberstark is a linguist too, a simultaneous translator by trade, working either way between German, English, French, Spanish, Italian, Portuguese and Dutch, with a dozen other tongues at some level, not to omit his "favorite language," Surinamese, on which he wrote his thesis years ago. He was good at mental arithmetic at school, but he entered the lists of the prodigies only at age 22, when he read of the wonderful feats of the Indian calculator Shakuntala Devi and offered to match her work for the newspapers. His performance is certified by the author (himself a scholar of linguistics): "He quickly and accurately mentally multiplied the following numbers... 77,567 times 43,559."

The numbers were proposed in writing; Eberstark finds that easier than working without the problem in view. It takes Eberstark the better part of a minute to multiply numbers with five or six digits; the faster Wim Klein (he too writes a brief foreword) can multiply two nine-digit numbers correctly in 48 seconds. These are typical results. Mrs. Devi, however, is reported in print as having correctly multiplied within 28 seconds two 13-digit numbers chosen at random, her feat having been performed in June, 1980, in London. That speed is so far beyond any other ever reported that skeptical Smith finds it "unbelievable."

This lively and definitive volume rewards the reader in two distinct ways. Not only is the substance timely, strange and fascinating, but also the scholarship is a model of insight and care. Twenty careers are presented in some detail, from Jedediah Buxton, much studied by *Gentleman's Magazine* in the 1750's, to the only American now appearing as a professional calculator: magician and graduate student Arthur Benjamin. The printed record requires intelligent sifting. The unfamiliarity of the data, the absence of expertise among witnesses, the problems of editing and proofreading numbers in the usual prose texts, not to mention the usual environment of entertainment and even mystery—all lead to more than a little misstatement and hyperbole. The saving grace of these data is of course the necessary sharp consistency of arithmetic results. The success of the book owes much to the way Smith has dealt with the reports of two or three centuries.

The first quarter of the book seeks psychological generalization, the second the calculational methods used or surmised; the rest is organized by the individual calculators. Motivation is the key. If every normal child does learn the native language, the greatest of feats, at least the child is strongly moved to do so and is surrounded by people talking about matters of importance and interest. Few of those people are doing mental calculations. Again and again the prodigies tell us that numbers are their friends. That implies willing immersion in numbers; the experience must resemble speech among peers every waking hour more than it does spending an hour a day in class to pick up a second language.

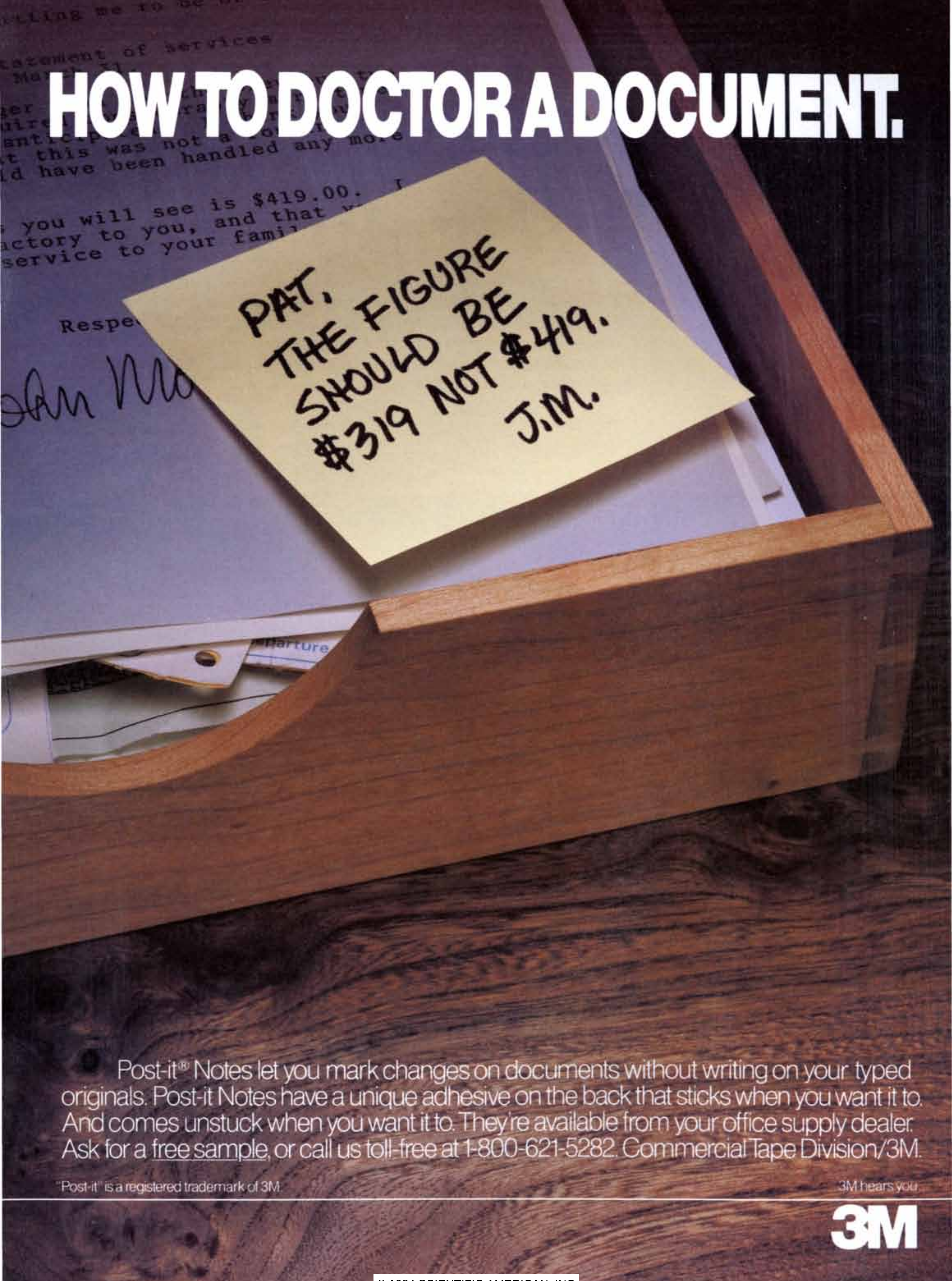
Since Alfred Binet's time we have known there are two types of calculating prodigies, those who mentally see numbers and those who hear them. More than one has performed prodigiously before learning to read and write numbers. There are those who see every number as it is written in their own hand and those who hear the numbers "in my own voice." Eberstark is a visual calculator, but his specialty of memorizing numbers—he knows 11,944 digits of pi (not a record)—is something he does by an auditory mnemonic method. "I convert figures into words... in a new language." In his polyglot code the first 10 decimals of pi are pronounced "ayfisktapses."

It is not wise to omit the intrinsic structure of arithmetic, which provides many paths to one goal. The paper algorithms we all learned at school tend to minimize writing; they do not well serve men and women who can remember long number strings with ease and pleasure. Multiplication is usually carried out by the prodigy from left to right: 348 times 461 is done as 300×400, adding 300×60, then adding 300×1, 40×400, 40×60 and so on until the full sum is formed. "True, the method here exhibited requires a much larger number of figures than the common Rule, but... pen, ink and paper cost Zerah very little," wrote Vermonter Zerah Colburn in 1833 about his own prodigious childhood.

It is not hard to show that the calculators neither need nor use giant multiplication tables already memorized; even 100×100 tables are unlikely. Some have been clever at factoring; others use a few well-memorized logarithms and derive the rest from those. There is a new rivalry in extracting integer roots of perfect powers. Undoubtedly the widespread use of computers has stimulated this craft, by providing easy access to lots of good material both for building algorithms and for setting problems. The 13th root of a 100-digit number is now a standard time trial for *The Guinness Book of World Records.* Klein can do that in less than two minutes. He gives his method, using logarithms, interpolations and partial steps of successive approximation. The feat is hard work, but there is little doubt that it plays somewhat on the credulity of most nonarithmetic observers. Fifth roots have been favorites for a century, since the units digit of the root agrees with the units digit of the power.

A brief account of other calculators ends the text, with a list of eight about whom nothing substantial is to be found. Portraits of some two dozen great mental calculators are another offering of this comprehensive account. Among the greats of mathematics Euler, Ampère and Gauss were all calculating prodigies, it is held, although without the best evidence. Euler is said to have calculated the first six powers of all numbers below 20 in his head one night to help sleep come, and to have recited them days later.

Jedediah Buxton, an illiterate farm la-


HOW TO DOCTOR A DOCUMENT.
PAT, THE FIGURE SHOULD BE $319 NOT $419. J.M.
Post-it® Notes let you mark changes on documents without writing on your typed originals. Post-it Notes have a unique adhesive on the back that sticks when you want it to. And comes unstuck when you want it to. They're available from your office supply dealer. Ask for a free sample, or call us toll-free at 1-800-621-5282. Commercial Tape Division/3M.
"Post-it" is a registered trademark of 3M.
3M hears you.
3M

Today's Chevrolet

Welcome aboard Starship Camaro.

Camaro Berlinetta Climb in. Buckle up. Adjust the retractable, touch-sensitive instrument pods so your hands never have to leave the wheel to control vital functions. Turn the key and watch the system monitor perform seven preflight tests as the engine sparks to life. Blip the throttle and watch the vacuum-fluorescent tach dance to the rhythm of your right foot.

To orchestrate your voyage, dial up a symphony on the pivoting, pod-style Delco AM/FM stereo with vacuum-fluorescent display and optional graphic equalizer. (Radio may be deleted for credit.)

Now, put it in gear, give it some gas and watch the digital speedometer numbers multiply. Your journey has begun.

Let's get it together...buckle up.

borer in Derbyshire in the reign of George II, could cipher wonderfully as well. Probably somewhat retarded mentally, he worked daily with spade in hand but took time off to reckon sums for the bemused Georgian gentlemen who invariably bought him beer and ale. (He kept a mental account of these treats, cited here as it was recorded, with the name of the noble "D. of Kingston" at the top, a generous patron who had over the years stood Buxton no fewer than 2,130 free pints.) Buxton spent two and a half months squaring a certain number derived from one of the outlandish problems he was set in sport. The 78-digit answer is reported; it is correct in all digits but one, which he missed by one unit.

Here again it is the consistency of number that compels us to credit these reports, because lacking modern computing aids no gentleman then had the leisure and curiosity to simulate or even test Buxton's prowess. Buxton walked once to London to see the king, who was on holiday, but at least he was then presented before the Royal Society. He lived to the age of 70, still on the land. One hopes that some of today's number crunchers, armed with their swift, capacious memories of silicon, will raise a mug or two in memory of old Jedediah, who managed so much with so little, and who boasted that he never drew more than one breath to a pint.

THE PHYSICS AND CHEMISTRY OF COLOR: THE FIFTEEN CAUSES OF COLOR, by Kurt Nassau. John Wiley & Sons, Inc. ($43.95). "Every gaudy color / Is a bit of truth," wrote the poet Nathalia Crane. Fascinated by color, this physical chemist at AT&T Bell Laboratories has compiled under that epigraph a readable and diversified volume to winkle out truth deep below any merely adequate recipe. Grass is green, to be sure, because of the properties of one abundant organic pigment, but from Nassau's book we learn more: just how chlorophyll absorbs, how the clear sky shines so blue, how the gem ruby gleams so red and how pure gold glitters so yellowly. The unity of color is in fact an electronic unity; the notion that we cannot see electrons is belied by the plain fact that all we ever do see is the light sent to our eyes one way or another by electrons.

The work opens with an introduction to color and its perception; there follow a dozen chapters each treating the colors formed by one or another physical process, from atomic excitations in thin gases to the remarkable interference effects in opals, beetle wings and liquid crystals. Two closing chapters stroll through the colorful world, sampling art and gemstones, lightsticks and lasers. The exposition is conceptual, often graphical, no calculational monograph for the expert. Each chapter seeks to summarize for a serious general reader all the diverse physics needed in adequate detail, not neglecting essential quantity. The focus lies within quantum mechanics and is perforce a little abstract, but the talk is good physical talk of charge binding and its energy levels, not the more formal discourse on wave functions and matrix elements.

A section of illustrations in color exhibits the phenomena gaudily enough. The very first picture shows an array of six gemstones, each one much the same striking blue. Together they display no fewer than six distinct causes of color as the author distinguishes them. The first is a blue beryl, called by the name of the Brazilian mine that in 1914 yielded wonderful deep-blue stones that disappointingly faded once they were brought into the light of day. That color arises as a result of a radiation-induced hole in the crystal lattice where an electron can be caught in a shallow energy trap. Next to this Maxixe beryl lies a blue spinel; its color is the result of the electric fields within the regular array of crystal atoms, splitting into a group of levels the perturbed state of an impurity atom.

Third is another spinel whose blue is more artfully produced; a sandwich of clear stone has received a filling of organic dye. The dyestuff is blue because of the potential for shifting electron position within the pattern of the molecular bonds. Large dye molecules still overtax the computers if sure prediction of color is asked of the wave equation; nonetheless, "the problem of the origins of color in organic molecules can now be considered to be solved." Next stands a cobalt compound, where that familiar blue arises again from a crystal-field splitting, not in any impurity but in an ion intrinsic to the crystal itself.

Finally, a transparent cut sapphire and an opaque oval of lapis lazuli show what is called charge-transfer color. The sapphire is blue because an electron can pass from an iron ion to a titanium ion, whenever the two occur as neighboring impurities in the aluminum oxide lattice. This is a kind of photochemical reaction, but one that rapidly reverses. It took 60 years to work out why neither iron nor titanium alone would produce blue sapphires out of the colorless crystal corundum, whereas a few parts in 10,000 of the two together yielded the precious deep blue. The lapis, which is not unlike the inorganic pigment ultramarine, gains its color by charge transfer among three close-lying sulfur atoms bound in the complex crystal, bearing a single negative charge. The case is not unlike that of a dyestuff.

The fateful Hope diamond could not leave the Smithsonian to pose next to its blue counterparts. If it had been on hand, it would have demonstrated a seventh origin of blue gems. A rare impurity atom of trivalent boron dopes the diamond, the colorless prototype of a wide-gap semiconductor. The missing electron at the boron site is a hole, able to accept an electron from the normally filled band. Then almost any color of visible light can be absorbed to promote some electron to fill the hole, although blue is the least absorbed. The result is the prized blue of the rare Hope, whose trace of boron has also turned it into a better conductor of electricity than a more commonplace water-white diamond.

Color is the stamp of symmetry in many electron energy levels. The blue-to-pink color change in cobalt salts familiar as an old-fashioned humidity indicator is thus geometric in origin. In the dry blue salt there is a tetrahedral arrangement of ions around the cobalt; the hydrated salt has instead an octahedral field, and the new splitting yields a pink. In a similar way the red ruby of corundum is close electronic kin to the green emerald of beryl. In both jewels the impurity ion is the same, trivalent chromium, replacing aluminum in the two colorless lattices. Even the ionic environment is similar; the grand distinction in color arises from the slight reduction of the ambient electric field in the case of the emerald. The level diagrams and absorption curves are striking. Still more wonderful, nature has provided a true intermediate between the opposites of ruby red and emerald green; it is the alexandrite, a gem that holds a chromium-ion impurity within a colorless beryllium aluminate lattice. There the splitting field is intermediate; the resulting color is blue green as viewed by blue-rich daylight but (as a photograph demonstrates strikingly) deep red under red-rich candlelight.

This review will not list all 15 of the color causes discussed in the book. Note only that biological coloration joins molecular properties with intricate geometric form at the level of spatial microstructure; the dual nature of arrays of color such as those of bird feathers is made manifest by such simple treatments as wetting and bleaching. The gem opal shares that extra level of structure, along with a few other peacocklike phenomena in the mineral kingdom. The book overall is richer in example than any review ought to be. A satisfying amount of specific detail accompanies most of the many examples; the exposition is thinnest, perhaps, in the auxiliary accounts of color perception itself, and around the foundations of quantum theory. The entire book is a bountiful sheaf of topics, tests and examples for students and hobbyists.

BLACK APOLLO OF SCIENCE: THE LIFE OF ERNEST EVERETT JUST, by Kenneth R. Manning. Oxford University

Press ($29.95). Like Porgy a generation later, young Ernest Just took ship from the Charleston docks to New York to enter the wider and more hopeful world. The 16-year-old son of a free black couple of middle-class status, himself a new graduate of the state normal school, Ernest made his way to New England as a scholarship student by his brilliance and energy. Acclaimed as classicist, journalist and poet, remarkably handsome and grave of bearing at 23, he stood first in the Dartmouth graduating class of 1907, the only magna cum laude of his year, a biology major with a demonstrated capacity for research. He had won everything save a chance "to enter the white professional world."

He had set his mind on science, although medicine was plainly the easier road to a career. His first job was as an instructor at Howard University, in English and then in zoology. He used his Dartmouth connection to win a chance to work summers as research assistant in the laboratory of Frank Rattray Lillie, professor at the University of Chicago and second director of the Marine Biological Laboratory at Woods Hole, Mass. Their close relationship endured for a lifetime, subtly shifting as Just repeatedly sought independence of mind and job.

Just earned his Ph.D. at Chicago in experimental embryology, with a thesis on the mechanics of fertilization, the process with which most of his 80-odd scientific papers were seized. He studied the early development of the eggs of marine invertebrates—sea worms, starfish and sea urchins—at Woods Hole until he left there for good in 1930. For a decade he appeared in research as a gifted young collaborator of Lillie's. By the end of World War I he had built a reputation of his own, particularly for his authoritative and meticulous knowledge of and control over the handling of those eggs and sperm, materials central at Woods Hole. He knew the effects of temperature, he understood the essential nature of careful experimental controls, he pioneered in handling the living cells without damage to them. His expertise was so much in demand that it cost him heavily in time from his own research. Howard University was no place for research; without financial support, he was heavily burdened with teaching duties. He might aspire to a deanship, even the presidency, at Howard, but hardly to science. His haven of investigation, his "Mecca, Woods Hole," would have to nourish him from early spring to fall every year.

This was the Woods Hole of A. H. Sturtevant, T. H. Morgan and Calvin Bridges, great names of chromosome genetics in America. It was the Woods Hole of K. S. Cole and Selig Hecht, the American pioneers of biochemical and biophysical neurology, and of the cytologist E. B. Wilson. It was also the Woods Hole of the quixotic and impulsive Jacques Loeb, a weighty figure in behavior, physiology and public debate. Loeb befriended Just effectively, saw him demonstrate error in some of Loeb's rasher conclusions and then turned against him in a virulent letter of 1923. That letter "squelched" any chance of a research position at the Rockefeller Institute for Medical Research, a position that might have been "symbolic for the whole black race." Loeb had written that he had tried "to help and encourage" Just, but that he had come to conclude "the man is limited in intelligence, ignorant, incompetent, and conceited."

The book is written so quietly, with such unfailing sympathy, yet without sentiment or partisanship, that it reads very much like one of the novels from America between the wars, like some newfound Willa Cather. That tone sounds the clearest when Just first visits Europe, a black American abroad. It was pre-Depression 1928; after years of indecision, the Julius Rosenwald Fund had agreed to support Just's department at Howard, for books, equipment and research, over the next five years. No other black in science had ever had such practical recognition. First of all, Just went off to spend half a year at the famous Zoological Station on the Bay of Naples. Everyone was kind; his research went like clockwork. The women scientists—one of them recalls his visit with the title phrase "black Apollo"—were taken by Just's charm and good looks.

Just returned an advocate for life in Europe, and he even tried to raise American funds for the Zoological Station. In the first half of 1930 he went again to Europe, this time as guest professor at the Kaiser Wilhelm Institute for Biology. We know how golden a time it was for him, starved for scientific community and respect, because there is a witness to the seminar of June, 1930, when Lillie's 60th birthday was celebrated at Woods Hole. Five old Lillie students spoke. The last of them was Just, his most prized student and perhaps his best friend. Just closed his technical talk, and added: "'I have received more in the way of fraternity and assistance in my one year at the Kaiser Wilhelm Institute than in all my other years at Woods Hole put together.' A peculiar silence fell over the room." He never came to Woods Hole again.

The Depression and the death of Julius Rosenwald ended Just's Howard grant in 1933. He had fallen in love at first sight with an elegant young German philosophy student in the summer of 1931. After that he made seven trips to Europe; his wife and children in Washington faded from his life. The echoes in America were ominous, and pained friends and foes. Saddest of all, by 1936 Just wrote a letter from the blue Bay of

Naples: "Il Capo del Governo d'Italia, S. E. Signore Mussolini.... As an American negro who for more than 25 years has contributed to the progress of biological science without having attained a place... which such service deserves, I desire earnestly the opportunity to continue my labors in Italy and thereby cooperate [as] your energetic leadership accelerates Italy to a magnificent destiny." The Man of Destiny never replied.

By 1938 Ernest and his Hedwig were living together in France, all but penniless. Just's big book was about to appear in the U.S., but history put forward urgent demands. France fell, and foreigners had to leave. There was nowhere to go in 1940 but America, with Hedwig, now his wife by virtue of a Riga divorce and soon to be a mother. Ill and unhappy, Just went back to teaching at Howard. At 58 he died painfully of cancer of the pancreas before war came to the U.S.

Just belongs to an era of biology when bricks were perforce made with little straw. Embryonic development is far from our understanding still, but only the broadest strokes could be made out before the electron microscope and biochemical cycles and protein sequences and the double helix. Ernest Just indeed saw the key place of the cell surface in the economy of the cell; his 1939 book bore the title *The Biology of the Cell Surface.*

Black Apollo of Science, however, is not a book on the inner logic of biology. It is a reminder of how much the men and women who construct the new ideas and artifacts of science are part of a world they never made. From this story, true and mythical at once, readers can gain a new insight into the heavy costs still being paid day by bitter day for that most peculiar of the institutions of America's wealth, two centuries of African slavery and its unending reverberations.

WORLD VEGETABLES: PRINCIPLES, PRODUCTION AND NUTRITIVE VALUES, by Mas Yamaguchi. Avi Publishing Company, Inc. Westport, Conn. ($35). This compendious book has no entry for ketchup, a vegetable by bureaucratic quibble alone, but just what is included is by no means easily defined. Ordinary American usage offers some guide: oranges, grapes and apples are out, tomatoes are in. Starchy roots, tubers and fruits, such as the potato and the breadfruit, belong. Most vegetables are food crops that yield succulent roots, bulbs, tops or fruits; the dry seeds of the cereals are not included, although of course those grasses are the major crops of our agricultural species. About 15 percent of our harvested area is given over to vegetables, including pulses, roots and tubers but excluding oilseeds.

Three-fourths of this text by a specialist at the famous school of agriculture of the University of California at Davis is a muster of some 160 or 180 crops, mainly grouped by botanical family, with information on origin, botany, culture, harvest, storage and nutritive value, together with some economic data deriving mainly from the *FAO Production Yearbook*. Most of the plants are shown in black-and-white photographs. The importance of the crop worldwide generally governs the length of the treatment; many a starchy underground vegetable and many a species of legume, say cassava and black gram, are given reasonable space even though they may not be found at the produce counters in your neighborhood supermarket.

The fuzziness of what we call vegetables is made evident by the sensible inclusion of sweet corn, a subspecies of the great American cereal staple, maize, one eaten not as flour or meal but as succulent immature kernels. The U.S. harvests most of the sweet corn grown worldwide, about three million tons per year, the yield of 1,000 square miles of good cropland, more than that devoted to peas, carrots or onions. Of course, that is still only 1 percent of the broad acres of field corn grown for grain. In the U.S., however, field corn is consumed mostly at a second stage, after being transformed into pork and beef.

The flavor of the onion family arises when the enzyme alliinase in the bulbs and stems breaks down various sulfur compounds to yield volatile organic sulfides. The onion lachrymator, thiopropanyl sulfoxide, is set free by the same enzyme, but the garlics lack the propyl radicals that the onions hold in abundance. Color also is a biochemical indicator: the yellow- and orange-fleshed carrot cultivars are high in the valuable carotenes, precursors to vitamin *A*. Carotenes are lacking in the white-fleshed varieties. The startling reds that brighten the carrots sold in northern India, for example, are pigments of no nutritive significance.

Spinach is an ancient annual crop of cool seasons. It bolts—goes to seed—in long days, particularly if the weather warms. Its tall relative, orache, perhaps of eastern Indian origin, is known in the U.S. as mountain spinach, because it is hardy and slower to set seed in the western highlands, where the common species bolts readily. It has varieties with yellowish and even blood red leaves.

A long table of vegetable legumes compares the properties of about 30 important species, alfalfa to fenugreek. Few crops are older than the lentil, whose shape gave its name early in the Renaissance to the familiar glass optical component. Black gram, a proteinaceous bean with a black seed coat, is a staple in southern India, where it is used as enrichment for rice in a variety of fermented dishes. The winged bean is a twining perennial, probably originating in New Guinea, much cultivated now from Sri Lanka to the Philippines. It bears well in hot regions where the invaluable soybean will not grow. The entire plant is edible: pods, mature seeds, shoots, flowers, leaves and tuberous roots. Enthusiasts see in it a world crop of the future, although current varieties mature too slowly after flowering (whose date is fixed by 12-hour days) to do well in temperate climates.

The volume closes with mushrooms. The classical species cultivated in the U.S. is the white agaric. It is cultured in the dark on horse manure or on various composts of hay and straw, bean meal and inorganics. In China the straw mushroom has long been reared on a compost of rice straw. The black mushroom called shiitake is an important commercial domesticate in Japan, grown on moist hardwood logs. It has lately become a U.S. crop as well.

The yam, so important as a starchy staple in tropical climates around the world from the Gulf of Guinea to Guyana and New Guinea, is the topic of a chapter. A similar treatment is given the remarkable çrucifers: mustards, cabbages and turnips. All are kin but show a bewildering variety of habit. Among the cool-climate coles "at the seedling stage, the different crops are hard to distinguish." You might reap on maturity a leafy cabbage head or a curly kale, a curded cauliflower or a floweret of broccoli, the budded stems of Brussels sprouts or the beetlike kohlrabi.

A sketch of the history and importance of vegetable crops, of their nutritive value and toxicant and medical importance, of the environment for growth and the means of light and temperature control occupies the first fourth of the text. Giving a few sentences for each technique, one chapter lists mulches and active protection schemes against frost, along with the structural hot caps, plastic tunnels, cold frames and lath houses of the advanced grower. This well-tested textbook, meant as a survey for agricultural students, is not unlike a good vegetable soup, a nourishing start, full of diverse tasty morsels. It can be recommended for the general reader and for first reference; the bibliographies point topic by topic to the intensive literature of this important discipline, both in books and in periodicals.

TYPICAL MINIMILL is operated by the Florida Steel Corporation in Tampa. The plant's raw material is steel scrap bought locally. The scrap is melted in electric furnaces, which are in the second building from the left, and cast into billets in a continuous-casting machine in the same building. The billets are rolled into reinforcing rods in the rolling mill, which is the long building at the left. Water for cooling the continuous caster comes from the rectangular reservoir at the top center and is recycled there. A "baghouse," perpendicular to the building containing the electric furnaces, serves for filtering smoke and dust. Finished rods, along with billets being shipped for rolling elsewhere, are visible as a blue material in some of the railroad cars. The plant's annual capacity is approximately 254,000 metric tons.

Established 1845 SCIENTIFIC AMERICAN May 1984 Volume 250 Number 5

Steel Minimills

At a time in the U.S. when some large steel plants that start with iron ore are being shut down, smaller plants that start with scrap steel and market a limited line of products locally are doing well

by Jack Robert Miller

Since 1977 the 10 American companies that make up what is often called Big Steel have closed, idled, transferred or sold some 20 plants or parts of plants, shrinking their production capacity by about 10 million metric tons, or 8 percent, from 123 million metric tons per year. Another consequence was a reduction in the number of workers in the steel industry from 453,000 in 1979 to 247,000 at the end of 1982, a decline of 45 percent in three years. In 1982 the steel industry as a whole lost $3.2 billion, far more than it had in any previous year. The focus in Big Steel is on "restructuring," which has been described by David M. Roderick, the chairman of the U.S. Steel Corporation, as a "state of accelerating self-liquidation." More specifically, restructuring includes facilities closed, projects deferred or canceled, production consolidated and products added or dropped.

In contrast, the segment of the industry represented by what are called minimills has expanded and prospered since about 1960. In that year there were 10 or 12 minimills sharing about 2 percent of the American steel market; at the beginning of this year, 50 minimills accounted for between 15 and 18 percent of the market. Their total capacity (reckoned in terms of plants that can produce up to 800,000 metric tons per year) approximates 14 million metric tons per year. The focus among the operators of minimills is on the expansion of plants and markets and on enlarging the range of products.

The Big Steel companies tend to blame much of their trouble on a surge of imports of low-cost steel from other countries. Some observers believe Big Steel has contributed to its own problems by sluggishness in adopting new technology and by complacency in marketing. In a recent interview with *Financial Times* the president of a minimill company put the argument as follows: "The big companies kept their prices high and gave up markets when sales were booming. They are now stuck with old plants and facilities that were obsolete years ago. The only thing left for them is massive restructuring."

Although the advice is pertinent, Big Steel may rightly discount it for its source. Notwithstanding the fact that minimills have become more important and more visible in the steel industry since the late 1970's, they operate at the margins of the industry. They make low-end-of-the-market, low-value-added-by-manufacture light structural rods, bars and shapes for sale to the local construction market, competing in it with imported steel. Their raw material is mainly local steel scrap, which they melt in electric furnaces. Therefore except for the attention they pay to the adaptation of new technology to their operations they represent something less than a complete model for a national steel industry.

A steel industry capable of beginning with iron ore in the ground and ending with steel in a large variety of shapes and sizes requires a much greater investment per ton of capacity than a minimill does. Data on the capital costs were assembled by the Office of Technology Assessment of the U.S. Congress in 1980. In 1978 dollars the estimates range from $154 to $320 per metric ton of annual capacity for a minimill and from $956 to $1,500 for an integrated plant (so named because it is capable of conducting a series of operations beginning with iron ore and ending with steel in a number of usable shapes).

The first step for almost all integrated operations is mining the ore. The second is preparing the raw materials. Ore of lower grade than was commonly available before World War II must be beneficiated, that is, its content of iron must be increased to between 65 and 72 percent by the removal of waste materials such as sand, clay and metallic impurities. The concentrated ores are ground and mixed, and the mixture is formed and hardened into pellets from three-eighths to five-eighths of an inch in diameter. The concentrated fines may also be sintered, that is, fused into porous lumps that are then broken up into one- and two-inch pieces. Pellets ship well and sinter does not. A pelletizing plant is therefore often built near the mine, and the pellets are transported by rail or ship to the steel plant. A sintering plant, however, is usually a standard part of an integrated mill's iron department.

At the integrated plant the preparation of raw materials also includes the crushing of limestone and the charring of coal (in the absence of oxygen to prevent burning) to make coke. The limestone and the coke are sized: broken into lumps of convenient size.

The third step is ironmaking. The prepared ore, limestone and coke are mixed and charged into the top of a blast furnace while a hot blast of air is blown in at the bottom. The heated air burns the coke, which generates carbon monoxide that rises upward in the furnace shaft to meet the downward-moving ore. The hot gas melts and reduces the ore. The result is liquid pig iron.

In the fourth step the pig iron is transferred to and refined in a steelmaking furnace, usually a basic-oxygen furnace or an electric-arc furnace. In the refining phase unwanted elements are removed from the pig iron and precise adjustments are made in the amounts of wanted elements, particularly carbon, which is brought down to between 1.7 and .3 percent. From the furnace the molten "raw" steel is poured into a ladle and from the ladle into ingot or slab-ingot molds made of cast iron, where it solidifies into ingots or slab-shaped blocks of steel respectively weighing as much as 60 and 100 tons.

Except when the ingots or slabs are sent to a yard for surface treatment or for storage, they are taken (as soon as possible after they are stripped from the molds) to the rolling mill. There ingots are put into soaking pits and slabs into reheating furnaces, where they are heated to a uniform temperature. They are then passed into a breakdown mill (known as a roughing mill and also a bloomer or slabber) in which the ingots receive a preliminary shaping before they are put through a series of intermediate and finishing mill stands.

In the 1970's a few integrated mills installed continuous-casting machines that receive the molten steel direct from the ladle (eliminating the ingot and the soaking-pit or reheating-furnace stage) and deliver slabs, blooms or billets to the rolling mills. Continuous billet-casting has been standard practice in the minimills for about 15 years.

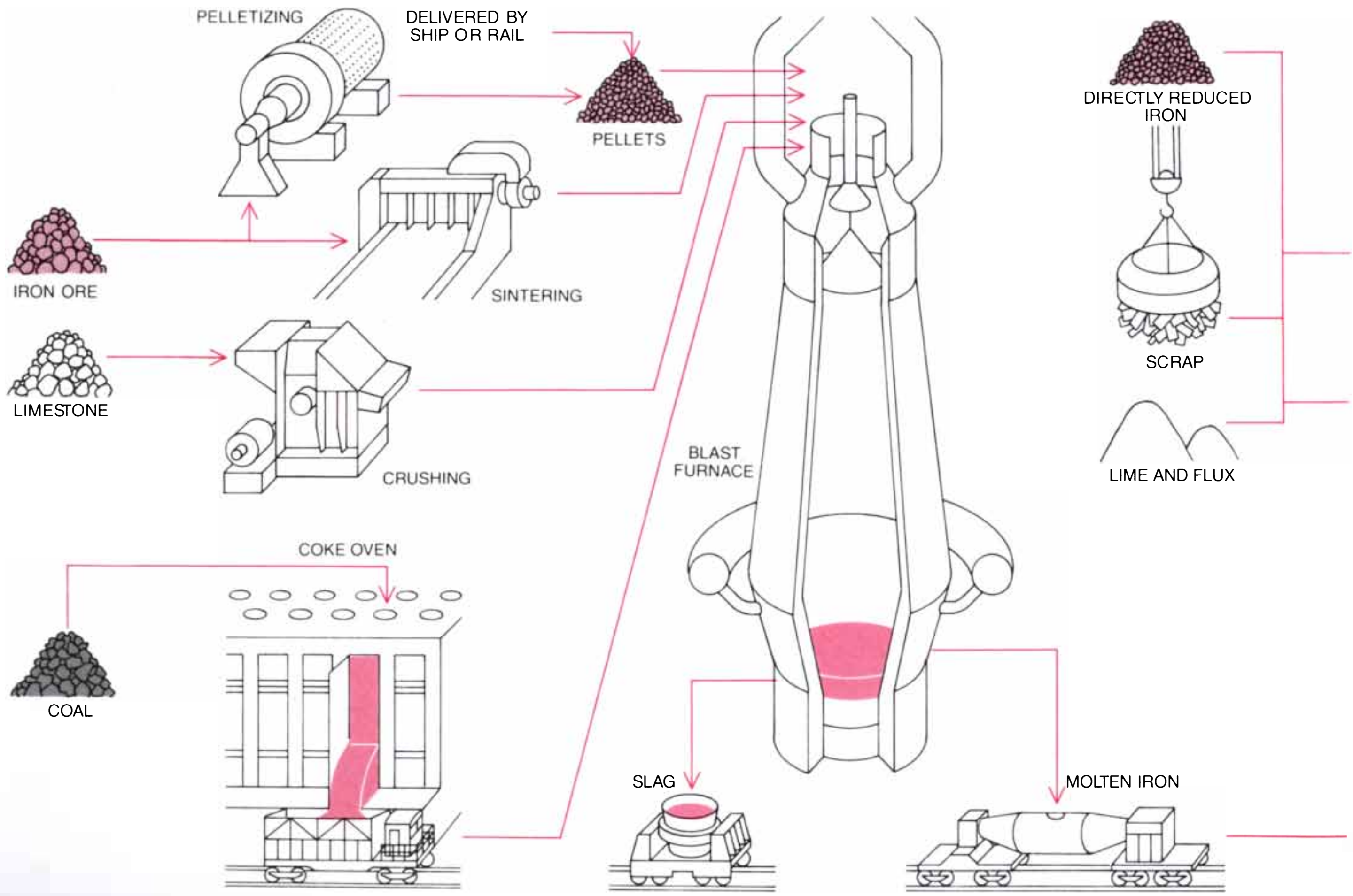
In the rolling mills slabs are rolled into plates for ships, for large industrial pressure and storage vessels and for large-diameter welded pipe; into wide sheet for automobile manufacture, and into narrower strip for smaller manufactured products and for tinplate and galvanized sheet. Larger blooms and billets are rolled into large-cross-section structural shapes; smaller billets are rolled into light structural shapes and concrete-reinforcing bar, into rods for drawing into wire and into tube rounds for making seamless tube.

The minimill gets its name not so much because it is small as because its operations encompass only part of the massive integrated steelmaking enterprise. To achieve the necessary economies of scale in reducing production costs integrated mills have been built to produce from three to 10 million metric ingot-tons per year. In the U.S. most minimills have a capacity of some 250,000 metric tons per year; a few of them turn out 600,000 tons or more. On the other hand, there is a 50,000-ton plant in Hawaii, and units with half that capacity are found in several countries outside the U.S.

The economics of the minimill have been described by Joseph C. Wyman, a steel-industry analyst with Shearson/American Express, Inc., who wrote (in 1981) that typically the minimill "has no investment in raw materials and transport facilities" and does not have to create an intermediate product such as pig iron. "It relies for its raw material on investment by others: scrap dealers and utility companies. It is also 'mini' in its product mix; it typically produces only a very limited range of shapes and sizes within a very narrow group of products."

Unencumbered by the large capital investment required for an integrated plant, the minimill is freer to put capital into new technology. By doing so the minimills have been able to recapture markets that Big Steel has abandoned to imports.

In a minimill the first step is to melt steel scrap in an electric furnace. The



CONVENTIONAL STEEL MILL is termed an integrated plant because it begins with iron ore and ends with finished or semifinished steel products. Mixtures of lump ore and pellets (and perhaps sinter), sized limestone and coke are charged together into a blast furnace to make pig iron. In the refining step the objective is to remove carbon from the iron to make steel, which has a carbon content of 1.7 percent or less. Pig iron, scrap, lime and fluxes are put into a steelmaking furnace (the two types depicted are an electric-arc furnace and a basic-

molten steel goes to a continuous-casting machine, from which it emerges as a continuous billet or bloom. That is cut off at convenient lengths, which go to a rolling mill to be formed into the shapes making up the mill's limited line of products.

The furnace, continuous-casting machine and rolling mill might all be found in a large integrated plant; I describe them here as they function in a minimill. A modern electric furnace is a substantial installation. Its main part is a round vessel, from three to eight meters in diameter, lined with a refractory material. Three graphite electrodes enter through holes in a roof. The electrodes are connected to a transformer rated at from 130 to 700 kilovolt-amperes per metric ton of furnace charge. Under standards proposed by the International Iron and Steel Institute in 1982 a unit of 700 kilovolt-amperes is classified as an ultrahigh-power furnace. Heat for the system is provided by arcs struck between the electrodes and the pieces of metal scrap that constitute the charge of the vessel.

The roof and the electrode assembly can be moved off and away from the furnace. The open furnace can then be loaded with steel scrap delivered from a large steel charging bucket carried by an overhead crane. Pig iron and, more recently, directly reduced iron may be added to the metallic charge to dilute the effect of impurities in the scrap such as copper, lead, chromium and zinc.

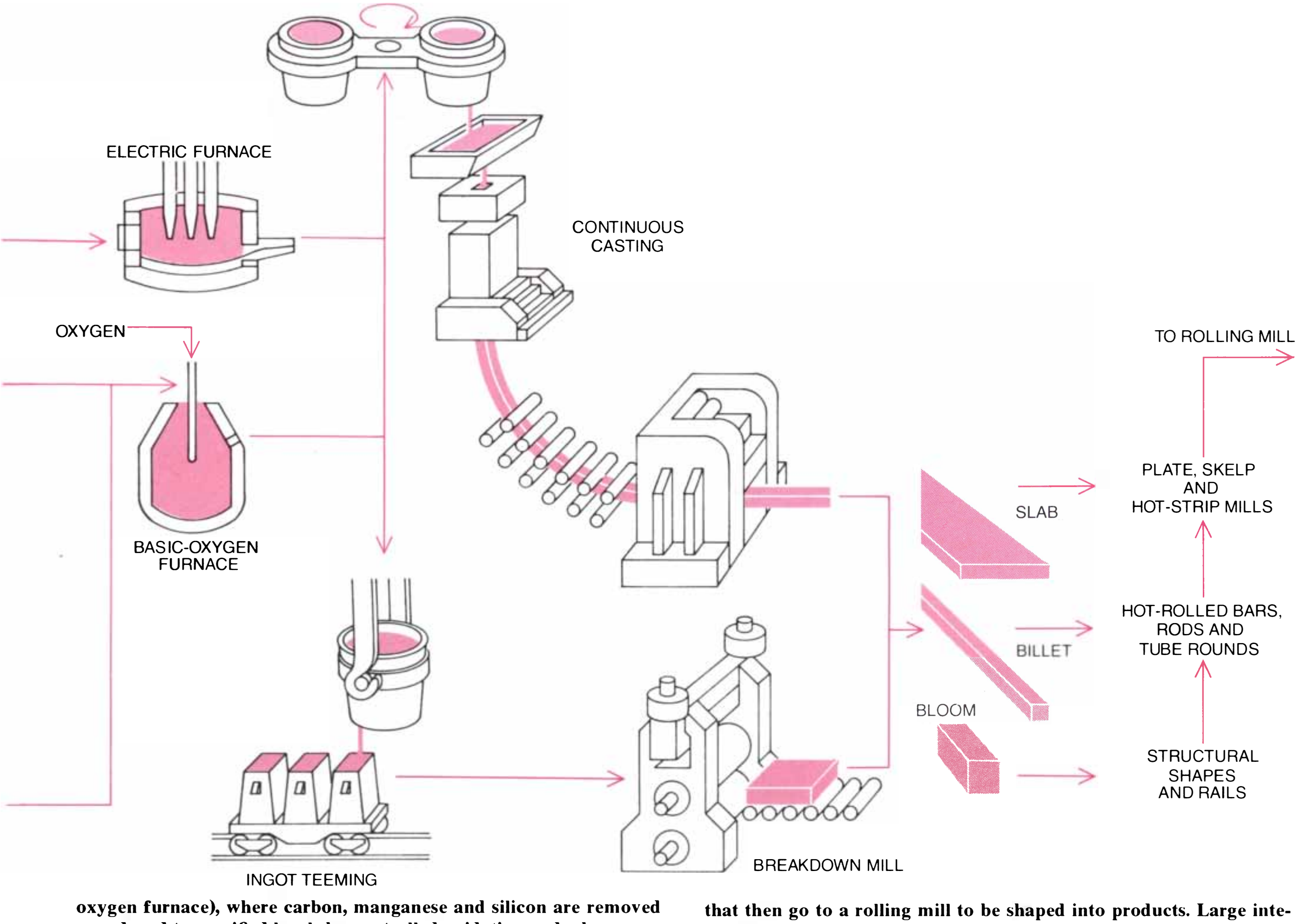
Oxygen may be employed to speed up the melting of the charge or to remove carbon from the molten bath of metal; if it is, it is usually injected into the furnace with a water-cooled lance inserted through a port. A modern furnace can refine a heat, or charge, of from 80 to 100 tons in a tap-to-tap time (the period between corresponding points of one cycle and the next) of from 100 to 180 minutes, which is equal to a rate of from 45 to 65 metric tons per hour. The consumption of electric energy by such a furnace is about 550 kilowatt-hours per metric ton.

At the end of a heat the roof and the electrode assembly are moved aside and the furnace is tilted to pour its charge into a ladle held and positioned by an overhead crane. The steel is now ready to be cast.

In the first minimills the casts were small ingots of billet size, which were sent directly to the reheating furnace of the rolling mill. After reheating they were pushed into the first rolling stand of the mill. A major modification introduced by minimills in 1962 was the replacement of billet-size ingot practice by continuous-casting machines.

In such a machine molten steel from a ladle is teemed into an elevated tundish, a structure that functions rather like a funnel. The tundish has an opening in the bottom through which the liquid steel flows into a water-cooled copper mold. The metal and the mold move downward together for about 25 millimeters as the skin of the strand of steel freezes. Then the mold is moved rapidly upward, breaking loose from the solidified skin of steel and returning to its starting position to grasp the strand of metal again.

The steel billet continues to move downward, passing through several sets of water-sprayed guides that gradually curve to a horizontal conveyor at floor



oxygen furnace), where carbon, manganese and silicon are removed or reduced to specified levels by controlled oxidation and where contaminants such as phosphorus and sulfur are also removed. The steel is cast into ingots, which are broken down into slabs, billets or blooms that then go to a rolling mill to be shaped into products. Large integrated steel mills are increasingly installing continuous slab-casting machines to replace the ingot-mold procedure. An integrated plant has more extensive operations and makes more steel than a minimill.

level. The continuous strand, now fully solidified, moves through a leveler that acts on it with flattening rolls. In some plants the strand is next cut into suitable lengths. A recently developed practice is to feed it directly and continuously into the reheating furnace of the rolling mill, and an even newer practice is to bypass the furnace.

The products made in the rolling section of a minimill are simple and limited in variety. They include wire rods, reinforcing rods (for concrete) and bar products in a variety of forms. Among the bar forms are flats, rounds and squares, the terms referring to cross-sectional shape. A minimill may also make other things: light I beams and T beams, angles (with a right-angle cross section) and channels (showing a shallow *U* in cross section). When the longest part of a shape seen in cross section is 75 millimeters or less, the product is a light section; when it is longer, the product is a heavy section. Bars made of carbon steel and rolled hot are known as merchant bars. When the melt is alloyed by the addition of small amounts of manganese, chromium, nickel or other elements singly or in combination, the product is an alloy steel. One such product made increasingly by minimills is described as SBQ grade (standing for special bar quality).

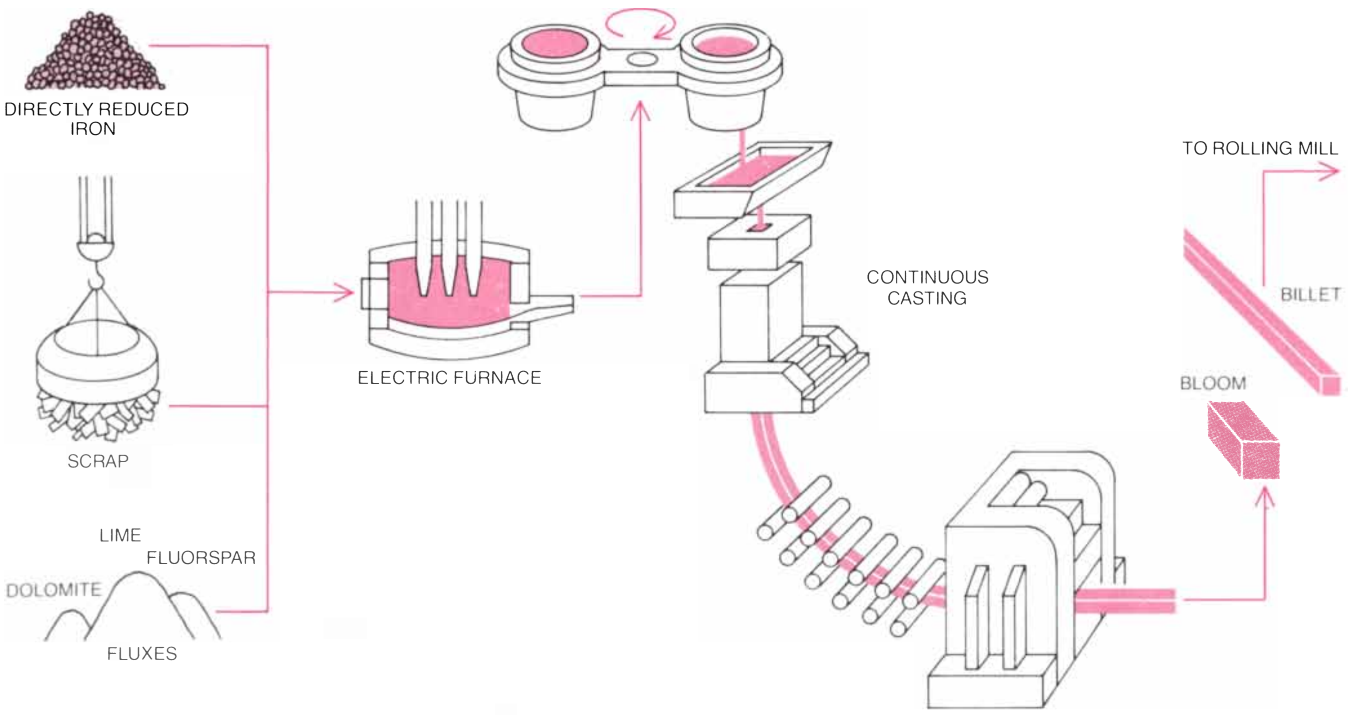
Bar and rod mills have been improved steadily over the past 10 years, largely because of their growing role in minimills. Dimensional controls have become more precise; the surface quality of the products has been improved, and such variables as the temperature of the steel and the pressure of the mill rolls have been made more responsive to control by computer. Moreover, mill speeds have been greatly increased. Finishing rates close to 100 meters per second (nearly 20,000 feet per minute) are now attainable in rolling rods 5.5 millimeters (.217 inch) in diameter. That is more than double the production pace of the rod mills of a decade ago.

A closer look at two minimill operations will show how minimills in general have paid attention to new technology, efficient operation, plant location and marketing. One is the Raritan River Steel Co. plant that began operations in March, 1980, on a tidewater site in Perth Amboy, N.J., some 20 miles from the Empire State Building in New York and 75 miles from Constitution Hall in Philadelphia. A truck moving at 55 miles per hour on the interstate highway system can connect the plant in less than four hours with customers in Boston, Worcester, Providence, Albany, Binghamton, Scranton, Harrisburg and much of the Baltimore-Washington area. Within that region are 55 million people, constituting nearly 25 percent of the U.S. population and nearly 25 percent of the market for steel. The plant makes rods only.

The plant's facilities are examples of recent technology applied to minimills. Its single electric furnace, six meters in diameter, can produce 135 metric tons of liquid steel from scrap. Heats from the furnace are teemed into a five-strand continuous caster designed to deliver billets 130 millimeters square in 16-meter lengths. The metal is stirred electromagnetically to yield carbon steel of high grade and uniform grain structure. A novel feature of the caster makes it possible to replace a section of mold in one strand without halting the operation of the other strands. The rolling mill may put the billets through as many as 25 passes in 25 roll stands to reduce and elongate them into rods from 5.5 to 12.7 millimeters in diameter.

The nominal capacity of the plant is about 550,000 metric tons of wire rod per year. The plant was built in 1978–79 at a cost of $110 million. The plant employs nonunion labor in an area where labor organizations have been powerful for many years. Considerable attention is given to the training of new workers. When the plant opened, an optimum productivity target of 1.36 man-hours per metric ton was established; by 1983 a rate of 1.63 man-hours per ton was being achieved routinely.

The second example consists of five minimills operated by the Florida Steel Corporation. One of them was the first steel plant ever built in Florida: a small plant erected in Tampa in 1958 to provide some 3,000 metric tons per month of reinforcing bars for the company's own steel-fabricating business. The bars

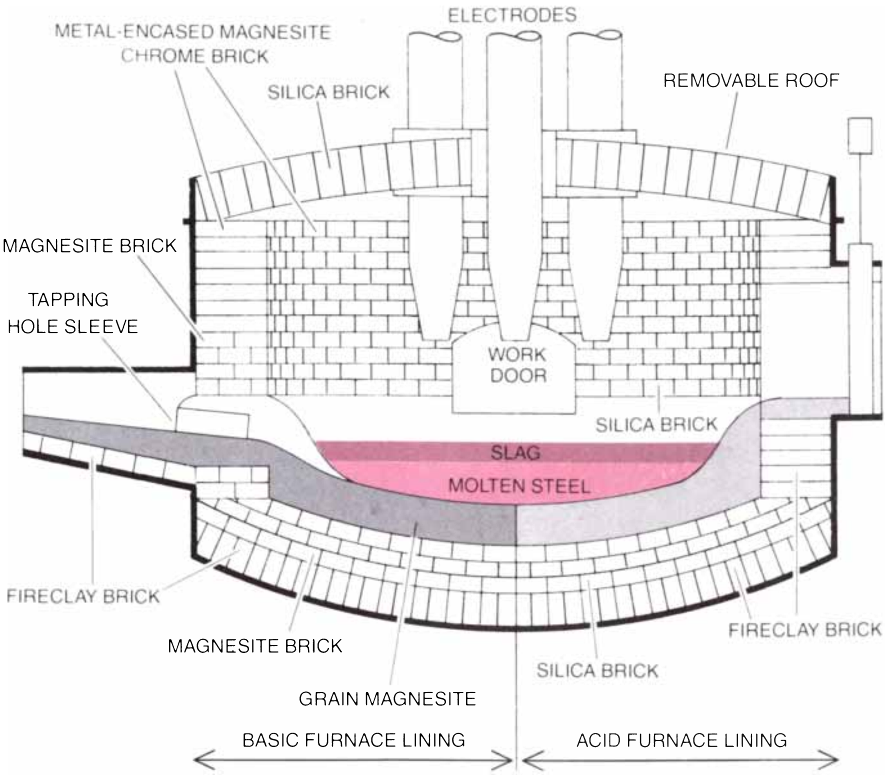


MINIMILL OPERATION dispenses with the ironmaking steps of the conventional steel plant and begins with steel scrap, flux and (occasionally) directly reduced iron. The scrap is melted in an electric furnace, poured into ladles and transferred to a continuous-casting machine. From the casting machine the steel emerges as a continuous billet or bloom that is cut off at suitable lengths and sent into the rolling mill, where usually only a few products are made. They include rods and bar and structural shapes for light construction. Over the past 10 years a few minimills have expanded their line of products to include such items as pipe, plate, narrow strip and narrow sheets.

soon found a large local market, and within a few years the company had become a steel producer first and a steel fabricator second. For 25 years Florida Steel prospered, expanding the Tampa plant eightfold and building four similar plants: at Charlotte, N.C., in 1961; at Indiantown, Fla., in 1970; near Jacksonville in 1975, and at Jackson, Tenn., in 1981. Costs at the Indiantown plant were higher than had been expected, and in 1982 it was closed. The bar-mill capacity at the four operating plants is 250,000 metric tons per year at Tampa and Charlotte, 320,000 at Jacksonville and 350,000 at Jackson.

The company had different reasons for expanding its capacity, and they illustrate three concepts of minimill operation. The mills at Tampa and Charlotte were built to take advantage of locally available markets and local sources of labor and scrap; this is the classic minimill concept. The purpose of the plants at Indiantown and Jacksonville was to meet a growing demand and to establish markets in nearby areas ahead of potential competitors; this is the neighborhood mill. It is still a minimill, but it seeks to establish itself in an area where the demand for steel can be expected to grow. The plant at Jackson was built to serve existing and expected demand for reinforcing bars and merchant bars with a mill that concentrates on specific products to meet a given market; this is the market-mill concept. Again it is still a minimill, but it is narrowly oriented toward one or two particular items.

The plant at Jackson incorporates an advanced type of continuous casting. It is called sequence casting or continuous-continuous casting. The electric furnace and the continuous-casting facilities are designed so that the time needed for making enough hot metal to fill a ladle is a little less than the time needed to pass the steel through the caster. Hence a series of ladles are teemed into the tundish in sequence and the casting procedure does not have to be stopped when a ladle is empty.



ELECTRIC FURNACE is the central part of a minimill. It can be more than eight meters in diameter and can hold more than 150 metric tons of molten metal. Three graphite electrodes enter through separate holes in the roof. When power is turned on, intense heat is generated in the furnace. Electric arcs are struck from one electrode to pieces of the steel scrap that constitute the metallic charge of the furnace and then are struck from the charge to the next electrode. The charge is melted, forming a metallic bath covered with a layer of slag. The refractory lining differs for basic and acidic practice. In the more usual basic practice the chemical reactions (involving a "limy" slag, magnesite refractories and the hot metal) remove carbon, silicon, manganese, phosphorus and sulfur from the bath and yield molten steel. The roof-and-electrode assembly is moved aside and the furnace is tilted to pour the steel into ladles for casting.

Since minimills depend on a supply of steel scrap, one must ask if the supply can be sustained in the face of the growing role of minimills in the steel industry. James W. Brown and Richard L. Reddy of the Electrode Systems Division of the Union Carbide Corporation looked into the matter a year ago and reported that the domestic supply of scrap appears to be "adequate to support domestic needs and future export levels." Hans Mueller of Middle Tennessee State University believes "the ubiquitous availability of scrap in this country" accounts for the establishment of electric-furnace minimills in the western and southern states, which "were shunned by the large steel companies." Tabulations by the Institute of Scrap Iron and Steel showed a scrap inventory in the U.S. of 620 million metric tons at the beginning of 1982. A backlog of that size, together with the annual additions to the inventory, should be sufficient to meet predictable demand for steel scrap for several decades.

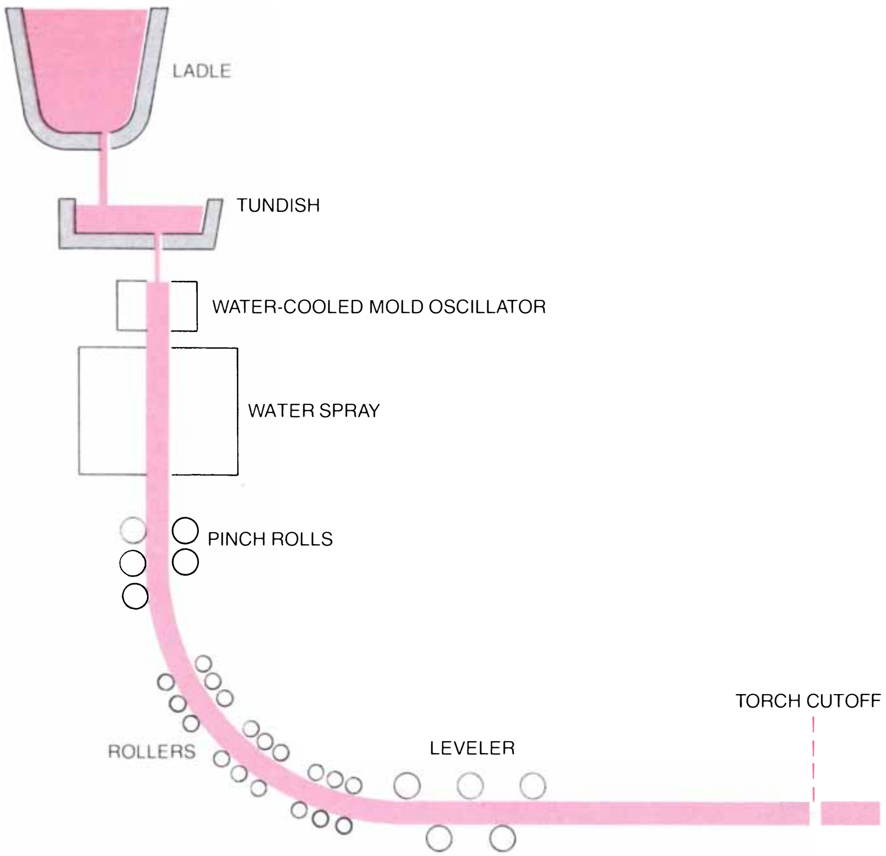
Scrap is a low-cost raw material compared with iron ore and pig iron. A study by Union Carbide last year comparing the cost of scrap per metric ton of raw steel in minimills with the cost of hot metal from blast furnaces in an integrated mill indicates a difference of approximately $100 in favor of the minimill.

The minimill also has advantages over the integrated mill in costs for labor and energy. The minimills are largely staffed by nonunion labor. Management practices have emphasized good relations with workers by means of profit-sharing plans, training courses, promotion programs and consultations with workers on decisions. Although the figures vary, it can be said as an approximation that the total employment cost per metric ton of steel product shipped ranges from $75 to $100 at minimills and from $195 to $295 at integrated plants.

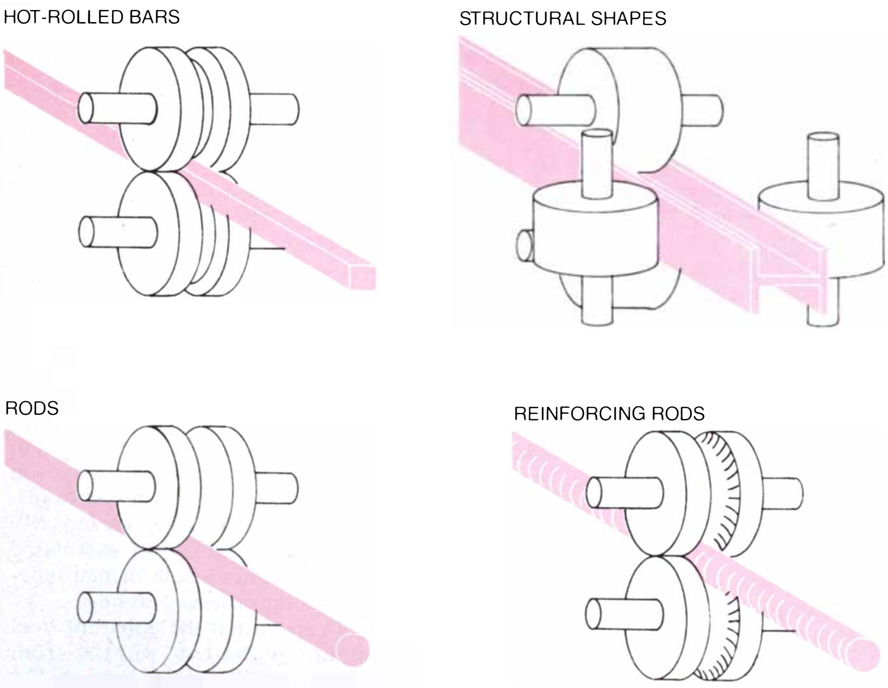
As for energy, Robert W. Bouman of the Bethlehem Steel Corporation has analyzed the requirements in British thermal units. He found that 5.32 million B.t.u.'s are consumed in making a net ton of steel in an all-scrap electric-arc furnace (the minimill situation) and 16.08 million at a blast-furnace, basic-oxygen operation in an integrated plant. "From an energy point of view," he wrote, "it is desirable to maximize the use of scrap to produce steel."

I turn now to some personal predictions of the course of the steel industry over the next seven years. They are based on my own long association with the industry and on recent discussions with a number of people in it. The forecast is mostly limited to the years from 1984 through 1990 because changes in the industry are coming too fast to justify estimates for a longer period. My projections are not precisely calculated forecasts but rather are informed judgments based on observed trends.

First, I expect that the apparent steel consumption in the U.S. will rise from 123 million metric tons in 1985 to 127 million in 1990. (Apparent steel consumption is production plus net international trade.) The projection is based on



CONTINUOUS CASTING is the standard method of casting in minimills. Molten steel delivered to the continuous-casting operation in a ladle by an overhead crane is poured into the tundish, which acts as a funnel, feeding the liquid into a water-cooled copper mold. The metal and the mold move downward for a short distance, allowing the skin of the steel to solidify. Then the oscillator breaks the mold loose and returns it to the starting position. The strand of steel continues downward through water-sprayed guides. Near the bottom the strand is rolled into a horizontal position. Sections are cut off and sent to the reheating furnace of the rolling mill. A newer practice is to send the strand directly from the casting machine to the rolling mill.



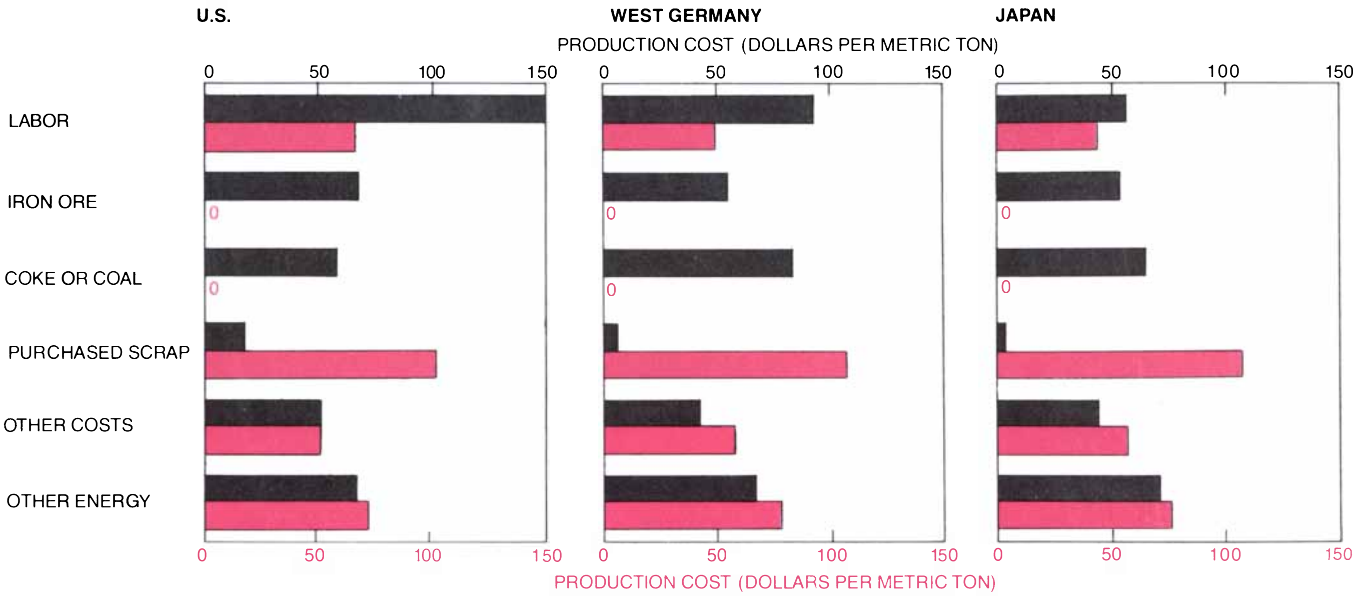
PRODUCTS OF MINIMILLS include bars, rods, smooth and deformed reinforcing rods and light structural shapes. The bar forms include flats, rounds and squares (named for their cross-sectional shape). In some plants the product is rolled into bar form and then drawn into wire.

an estimated annual growth rate of .62 percent, which would be quite modest. I also estimate that the level of imports, which in recent years has been from 19 to 22 percent of the domestic market, will decline to about 16.5 percent by 1985 and will remain there through 1990. Part of the decline in imports will result from increased activity by minimills. Having gained a large share of the market for their present line of products (90 percent of light shapes, 73 percent of concrete-reinforcing bars, 55 percent of wire rods and 30 percent of hot-rolled bars in 1981), minimills are now extending the line to include virtually every kind of steel product: pipe and tube, plate, tinplate, sheet and strip. As before, the minimills will make their market penetration at the expense of imports and of the integrated producers.

Steel made by the electric-furnace method accounted for 28 percent of U.S. steel production in 1980; I envision that share as rising to 32 percent in 1985 and 36 percent in 1990. Scrap for electric-furnace operations is likely to be available in more than adequate amounts at prices ranging from $118 per gross ton in 1985 to $126 in 1990. These average prices will not be exceeded because of the competitive restraints imposed by the availability of directly reduced iron as an alternative charge for an electric furnace. In this process iron ore is reduced to the solid-metallic state by heating it without melting it. The result is a highly reduced iron that can serve as the raw material for steel-production furnaces.

Minimills are expected to continue growing, although because of increased competition from a restructured Big Steel and from other modernized miniplants, it is doubtful they will match their rate of growth over the past decade. The minimill capacity of more than 14 million metric tons in 1983 will increase steadily through 1990. I expect that minimills may account for from 20 to 24 percent of the U.S. market, that is, they may have a total production of some 22 million metric tons in 1985 and 26.5 million in 1990. It also seems likely that many minimills will broaden their range of products and the quality of the steels they make.

Several technological advances can be expected, chiefly in minimills. In the melt shops new techniques will bring tap-to-tap times for 60- and 80-ton heats down to less than an hour. Recent innovations already seen in some electric-furnace shops are rapidly being introduced at most minimill operations. One of these "secondary practices" is ladle refining. In this procedure scrap is rapidly melted in the electric furnace at the maximum input of power. Then, as quickly as possible, the liquid metal is poured into a ladle, where it is refined.



PRODUCTION COSTS per metric ton for integrated steel mills and minimills in 1981 are compared for the production of wire rods. Under each heading such as labor and iron ore the black bar represents the costs for an integrated mill and the colored bar represents the costs for a typical minimill. The minimill makes no outlays for iron or coke. Costs such as depreciation, interest and taxes are not shown.

The procedure allows the electric furnace to be reserved for what it does best: melting at high power.

Foamy-slag practice is another new application, although it was developed in a miniplant electric-furnace shop in Mexico some 25 years ago. Coal or coke is added to the furnace at the start of meltdown; the resulting reaction causes the formation of a layer (from one foot to four feet thick) of bubbling lightweight slag on top of the bath. A foaming slag envelops and shields the electrodes, facilitates heat transfer from the electrodes to the metal, reduces the tendency for hot spots to form in the furnace and minimizes damage to the refractory material.

One of the most effective developments of the past decade has been the use of water-cooled panels in the wall and roof construction of the electric furnace. This step can eliminate as much as 75 percent of conventional wall and roof refractories, resulting in impressive lowering of production costs.

These and other novel techniques are aimed at increasing productivity, lowering production costs and improving the quality of the product. Many of them are now employed in recently built minimills. I expect these new techniques to be standard practice at most minimills within the next three to five years.

Equally important technical improvements can be foreseen in continuous casting. Interest is being shown in horizontal continuous casters, which would reduce the hazard of spills when large ladles full of molten steel are moved by an overhead crane 40 feet or more above the floor level. It would also eliminate the need for the expensive high structures that house vertical continuous-casting machines. Nevertheless, I doubt that horizontal casters will be ready to challenge vertical ones for many years.

In the rolling mill the speed of many existing mills can be increased by more than 50 percent with simple modifications. More significant, some minimills will be able to make thin slabs, from 25 to 40 millimeters thick and from 600 to 1,000 millimeters wide, by vertical continuous casting in the next two to five years and to roll such slabs directly (that is, to pass them directly to the finishing stands of a rolling mill) by 1990.

At least two large U.S. minimills already do direct rolling regularly. As two billet strands (sometimes just a single strand) emerge from the continuous caster they are simultaneously transferred directly to the rolling mill and rolled in it without the usual reheating. (The temperature over the cross section of a billet is made uniform by moving the billet through induction heaters between two of the initial mill stands.) In 1985 this practice is likely to be followed at half a dozen or more minimills.

Late in the 1990's one should see the industry's major technical achievement of the decade: the direct casting of strip steel from three to five millimeters thick, which constitutes the raw material for the many factories that turn steel into a great variety of finished products. One may also see at some U.S. minimills in the 1990's the introduction of plasma torches and plasma melting furnaces. A plasma is an ionized gas consisting of approximately equal numbers of positive ions and electrons. It is formed by passing a compressed gaseous fuel through an electric arc between two water-cooled electrodes. The fuel breaks down and ionizes with the generation of large amounts of heat. Plasma techniques are being worked on in several countries, including the U.S. Plasma torches for auxiliary heating are now in service in Sweden and Japan. In East Germany two plasma furnaces have been producing high-quality alloy steels competitively for the past three years.

Ladle refining, continuous-continuous casting, direct rolling of billets and direct casting and rolling of thin slabs and strip set the stage for continuous steelmaking of a full line of steel items in uninterrupted production operations where raw materials enter at one end and finished steel products emerge at the other. In the minimill linkages between the electric furnace and the continuous caster and between the casting unit and the rolling mill have already been demonstrated. I expect to see continuous steelmaking in one or more U.S. minimills before the end of the century.

Finally, although I expect the minimills' share of U.S. steel production to rise from 20 percent in 1985 to 24 percent in 1990 and then to continue upward, I do not expect the minimill sector to replace the integrated steel plants. It is likely that steel production by minimills will reach a peak between 32 and 40 percent, probably near 34 percent, sometime before 1995. The reason for such a limit is that steel products, including the steel scrap that sustains minimill production in the first instance, come from iron ore. Without the integrated-steel-plant sector, whose blast furnaces start the chain from ore to steel to steel products to scrap, the operation of minimills would be forced to halt in less than 10 years.

Oceanic Fracture Zones

They complicate the pattern discerned in the theory of plate tectonics by dissecting the edges of the plates that compose the ocean floor. Some of them are the width of an ocean basin

by Enrico Bonatti and Kathleen Crane

The realization that the ruggedest topography on the surface of the earth is on the floor of the oceans is little more than two decades old. First the midocean-ridge system was discovered. It is a submarine chain of mountains that traverses the major ocean basins for a length of some 60,000 kilometers, thus constituting by far the longest mountain range on the earth. Then a compilation of topographic data for the major ocean basins revealed a complication in their geometry: the axis of the midocean ridge is offset laterally in many places by distances ranging from a few kilometers to several hundred. The offsets are particularly common along the Mid-Atlantic Ridge: if one traces the crest of the ridge, one encounters offsets at intervals of 50 to 100 kilometers. Most of them are short: one must travel less than 30 kilometers to find where the ridge continues. Some are notably longer. Characteristically each of the longer offsets consists of a deep trough joining the tips of two segments of the ridge. The trough is bounded by elevations running more or less parallel to the trough. They are called transverse or transform ridges. Remarkably, the trough and its flanking ridges can often be traced well beyond the ridge-axis segments they join. The trough and the transverse ridges thus form extensive ocean-floor disruptions now known as an oceanic fracture zone.

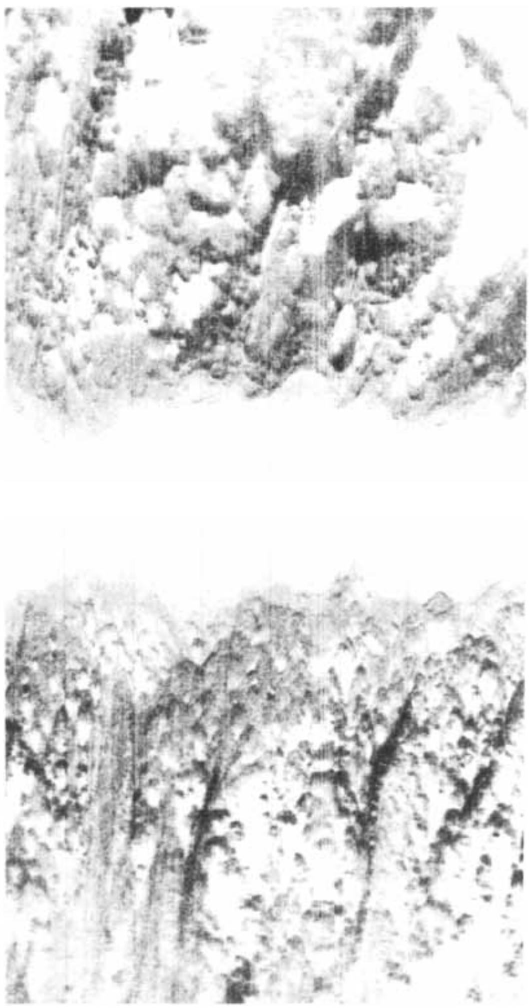
Consider the equatorial Atlantic. Here a set of closely spaced fracture zones dissects the Mid-Atlantic Ridge. The largest of them, the Romanche Fracture Zone, offsets the axis of the ridge by almost 1,000 kilometers. The deepest parts of the floor of the Romanche trench are more than seven kilometers below sea level; the highest parts of the ridges flanking the trench are less than one kilometer below sea level. Hence the vertical relief is more than six kilometers. The Grand Canyon is scarcely a fourth that deep. The Romanche Fracture Zone is flanked by several similar zones that are almost equally impressive, producing a sequence of troughs and transverse ridges several hundred kilometers wide from north to south. The resulting terrain, the Equatorial Megashear Zone, is hardly matched in size and ruggedness anywhere else on the planet.

What geologic processes created the fracture zones of the equatorial Atlantic and similar fracture zones discovered elsewhere in the earth's ocean basins? The answer requires that the events producing an ocean basin be examined. The fracture zones then emerge as places of massive geologic activity. The zones arise as part of the process by which an ocean basin opens. Then at later times they are involved in massive readjustments.

Perhaps the first clue to the nature of fracture zones was the finding that the crest of the midocean ridge is a locus of intense seismic and volcanic activity and of a great flow of heat from the interior of the earth. This finding and others were explained by the theory of plate tectonics. One of the tenets of the theory is that new oceanic crust is created along the crest of the midocean ridges and is added to the great plates that form the floor of the ocean basins. The plates carry crust away from the crests at speeds ranging from one centimeter per year to 20 centimeters per year. The crests are thus the axes of sea-floor spreading. In some cases, including that of the Atlantic, the motion of the oceanic plates pushes continents apart, providing a mechanism for continental drift.

The geologic significance of the lateral offsets in a crest, however, was unclear until 1965, when J. Tuzo Wilson of the University of Toronto introduced the concept of the transform fault into the hypothesis of sea-floor spreading. Within that framework one oceanic plate carries crust away from one side of the crest of a midocean ridge and another plate carries crust away from the other side. Therefore along an offset of two ridge-axis segments, blocks of crust would be sliding past each other as they were carried in opposite directions. The offset is a transform fault: a zone of what geologists call strike-slip motion.

In Wilson's model the strike-slip motion occurs at the length of the transform and not beyond it, along the line representing the projection of the trans-



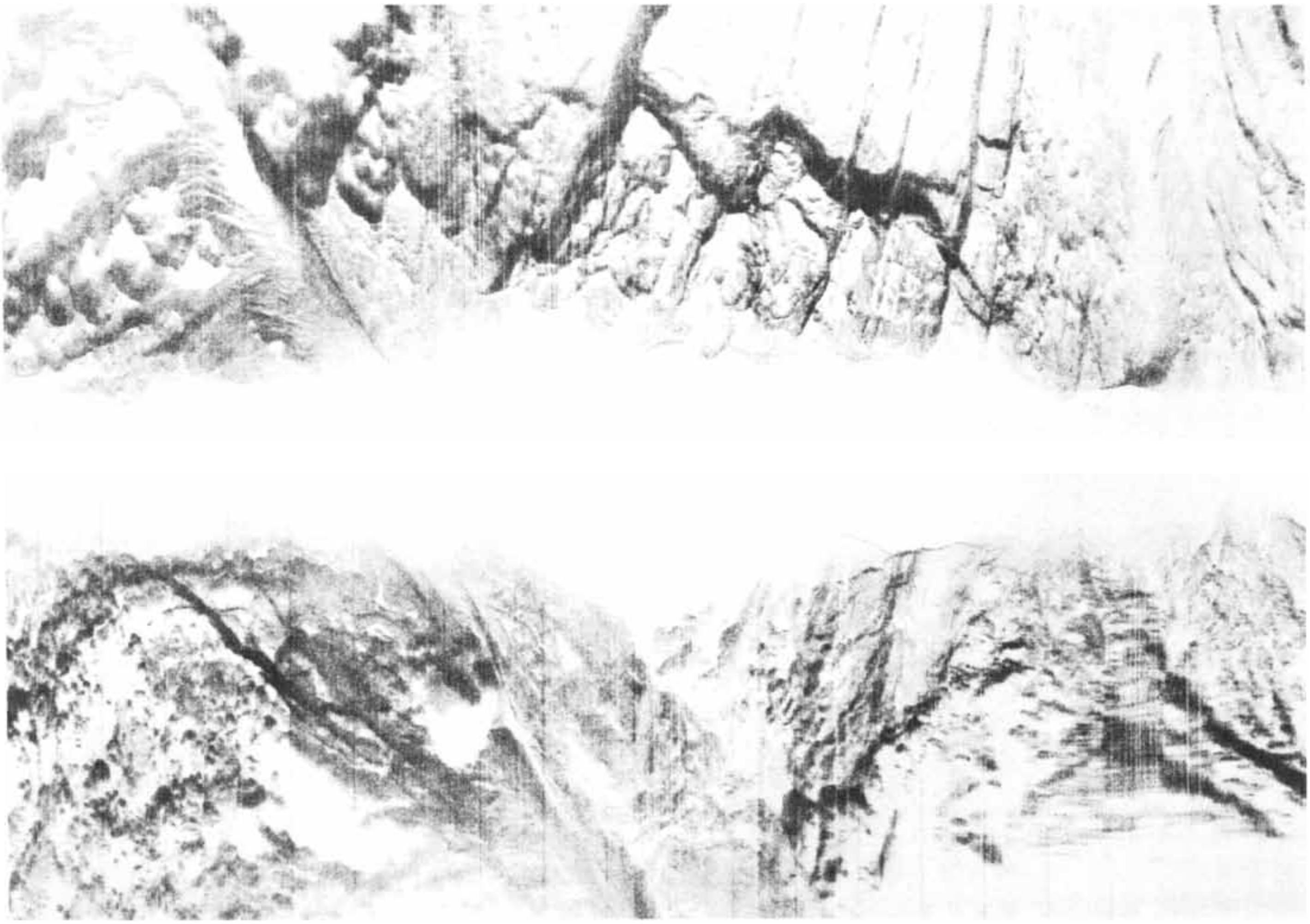
CLIPPERTON FRACTURE ZONE disrupts the floor of the Pacific and produces an offset of 80 kilometers in the submarine ridge known as the East Pacific Rise; the offset is established by the part of the zone called the

form fault across an ocean basin. That prediction is borne out by the observation that earthquake epicenters are common within the transform but uncommon along its projection. In addition the prediction is supported by an analysis of earthquake seismic waves done by Lynn R. Sykes of the Lamont-Doherty Geological Observatory of Columbia University. The waves reveal the direction in which crustal blocks abruptly slip, generating earthquakes in the transform zone. Presumably the direction of slippage is also the direction of the slower tectonic motion. Sykes's analysis indicates that the motion is indeed strike-slip and that it follows the directions Wilson predicted.

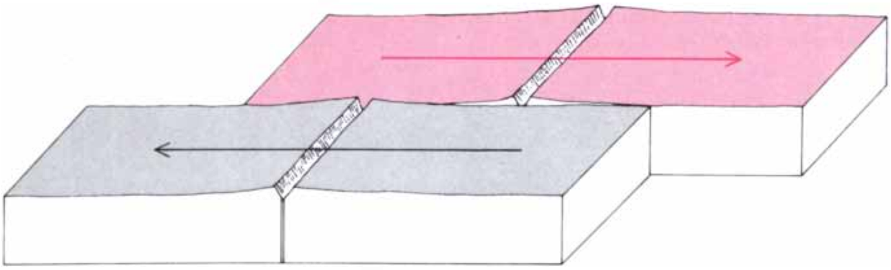
Clearly Wilson's model of oceanic transform faults is essentially correct. Nevertheless, recent research suggests it is an oversimplification. For one thing, the trough and the transverse ridges that are the signature of large-offset transform faults can be traced outside the offset zone, along the projection of the offset. Indeed, the signature of some equatorial-Atlantic transforms can be traced from one side of the Atlantic to the other, that is, from the coast of Africa to the coast of South America. Near the margin of each continent it is buried under sediment, but seismic surveys reveal it is there. The concept that transform faults are places where crustal blocks slide past each other more or less passively is yielding, therefore, to the concept that they are places of complex geologic activity of prime importance in determining the structure and evolution of ocean basins.

What determines the presence of a deep trough along a transform fault? Several factors are probably at work. Fundamentally, new oceanic crust forms along the axis of a midocean ridge because an upwelling of hot basaltic magma, produced in the mantle of the earth some 30 to 60 kilometers below the sea floor, cools and solidifies at the axis of the ridge as it reaches the ocean floor. The level attained by the magma corresponds to the depth of the axis (that is, its distance below sea level) and is governed by the quantity of the magma, by its temperature and viscosity and by the drag due to the interaction of the magma with the relatively cold walls of the conduit where the upwelling takes place. The crust moves away from the axis, and as it moves it continues slowly to cool and contract. As a result the depth of the ocean floor increases both with the floor's age and with its distance from the axis at which it formed. The relation is expressed by a law that is gratifyingly simple and has been well verified in the three major oceans, the Atlantic, the Pacific and the Indian Ocean: the depth of the ocean increases with the square root of the age of the crust.

At a transform fault this simple law is broken: the trough is in essence an anomalous oceanic deep. At a transform fault, however, a hot young ridge-axis segment meets older, colder litho-



Clipperton Transform Fault, which traverses the image diagonally from top to bottom. (It can be seen directly above as a prominent band.) The image itself was made by Sea MARC, a sonar apparatus deployed by one of the authors (Crane), working at the Lamont-Doherty Geological Observatory of Columbia University. Towed between 200 and 500 meters above the sea floor, the Sea MARC apparatus directs a beam of acoustic waves to the left and right of its path. Dark points in a Sea MARC image represent strong reflections of the beam, for example by undersea cliff faces. The image shows a swath of sea floor approximately five kilometers wide by 10 kilometers long.



FIRST HYPOTHESIS about oceanic fracture zones was that they are much like a continental transcurrent fault: they are lines along which two blocks of crust slide past each other. According to this hypothesis, segments of a midocean ridge are carried along with the blocks.

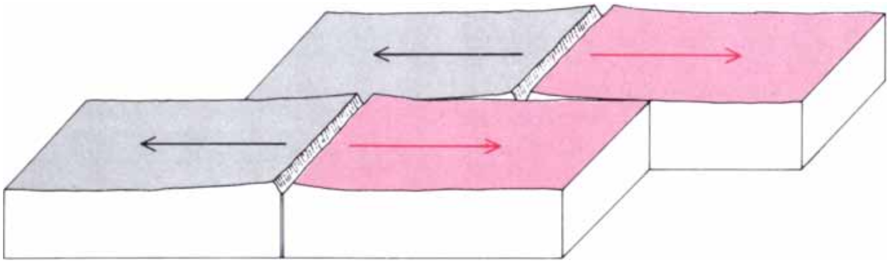
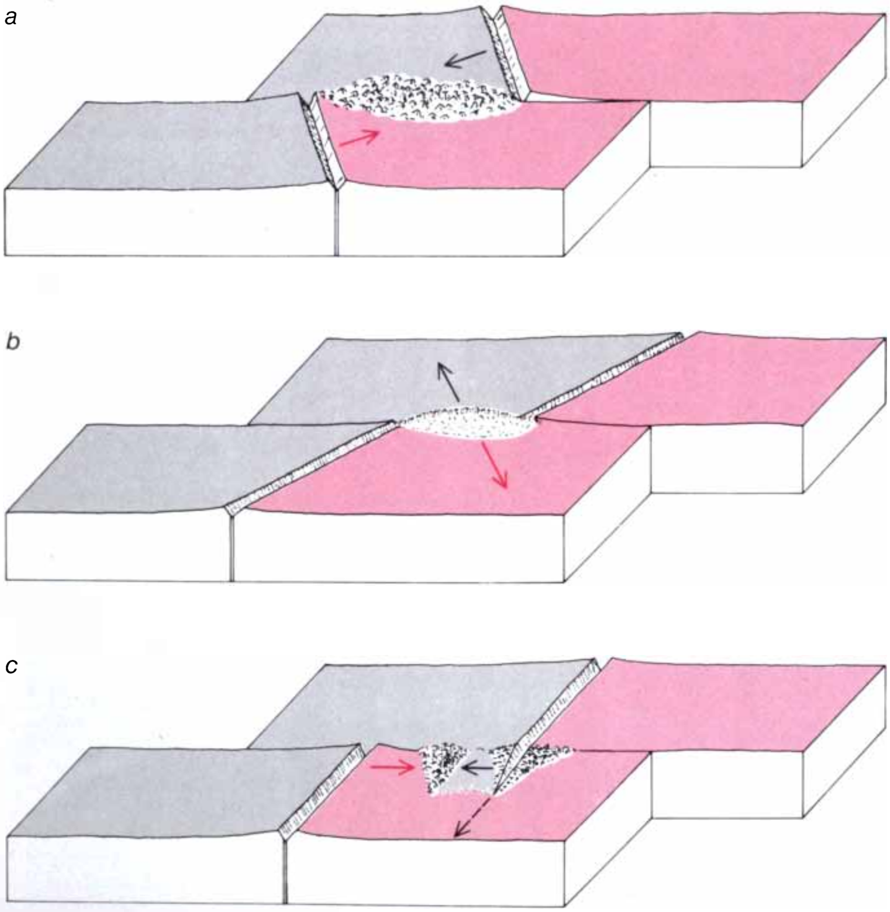


PLATE-TECTONIC THEORY proposed that new oceanic crust forms along each segment of a midocean ridge. Thus the ridge is stationary, and a plate diverges from each side of it. Two plates slide past each other only along the offset between the two ridge-axis segments.



STRESSES IN A FRACTURE ZONE may arise when plates change their direction of spreading. If the direction changes so that plates converge at the fracture zone (*a*), the crust may be compressed, causing the sea floor to rise. If the direction changes so that plates diverge from the zone (*b*), the crust may be stretched and weakened, forming a "pull-apart basin." The propagation of the tip of one of the ridge-axis segments (*c*) may also cause plates to be compressed.

sphere. (The lithosphere includes the crust and the relatively rigid upper part of the mantle. Together they make up the thickness of a tectonic plate.) Thus the hot material upwelling along the axis near a transform fault must interact with a cold surface it does not encounter elsewhere. It must then "freeze" at a deeper level, producing a topographic low at the intersection of the axis and the transform. The spreading of the plates extends the topographic low into a trough that can traverse the entire width of an ocean basin. The trough may even widen during the initial stage of sea-floor spreading, because the cooling of a plate presumably causes it to contract not only vertically but also horizontally.

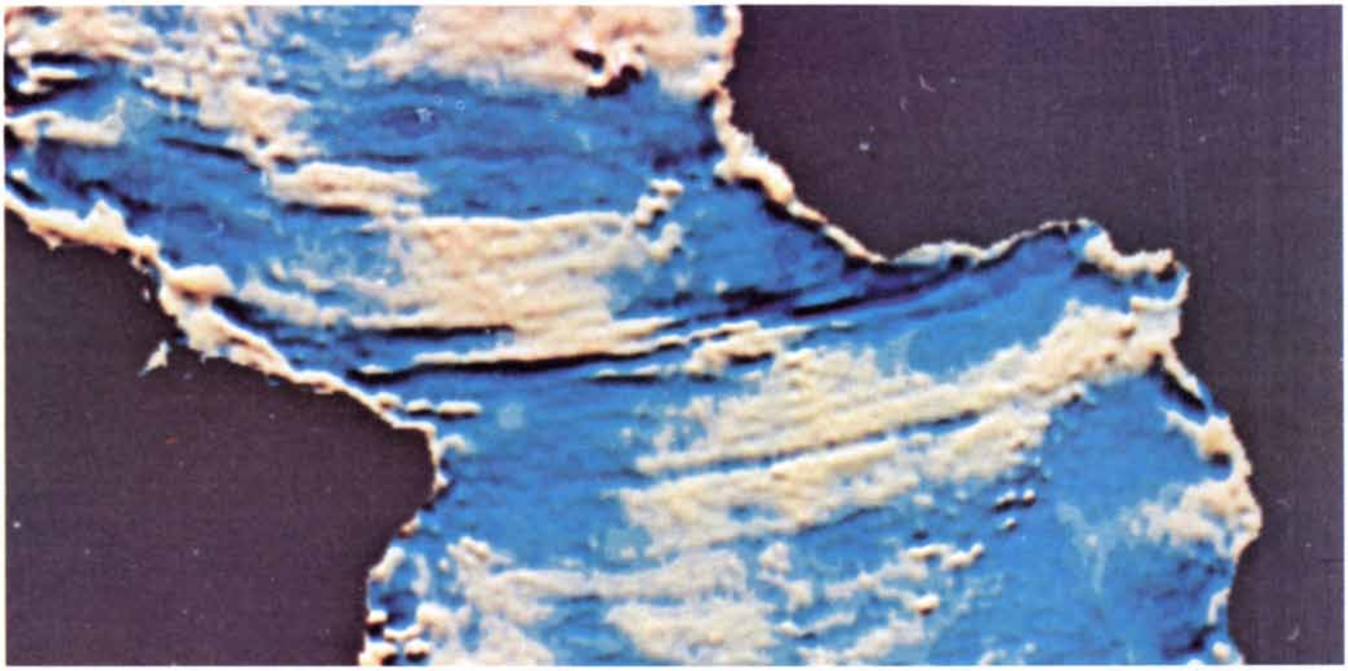
The transverse ridges flanking the trough constitute a further infringement on the law linking the depth of the ocean to the age of the oceanic crust. They are anomalous highs. How, then, do they originate?

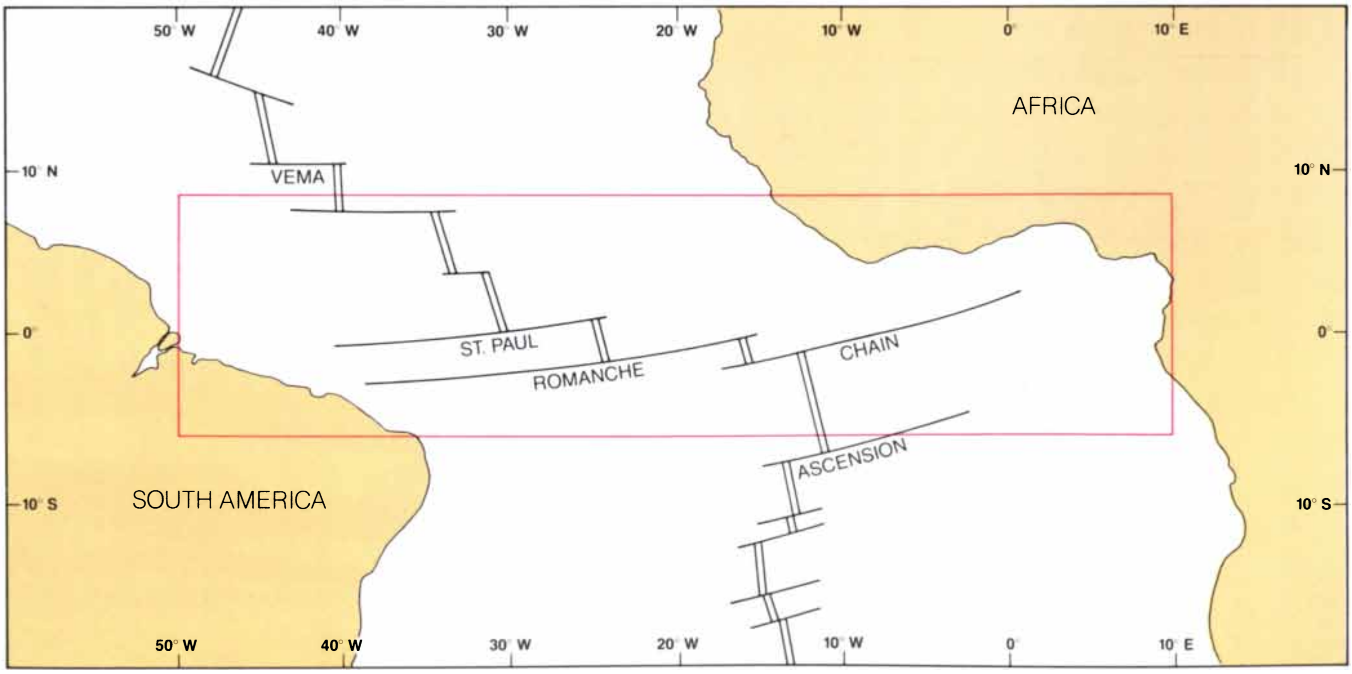
The rocks forming the walls of transform troughs and the slopes of transverse ridges have been sampled extensively in the past decade. Most of the samples have been recovered by scratching the sea floor with a dredge maneuvered from a ship; some were recovered by the submersible research vessel *Alvin* in the course of a study of the Oceanographer Transform Fault (done by us along with Paul J. Fox of the University of Rhode Island and workers from the State University of New York at Albany and from the Woods Hole Oceanographic Institution). The Oceanographer transform is at 35 degrees north latitude, near the Azores in the North Atlantic.

Among the rocks that were sampled two types turned out to be abundant. The first was peridotites. Mineralogically peridotites are rocks consisting chiefly of olivine, pyroxenes and spinels. Olivine and pyroxenes in turn are silicates of elements such as magnesium and iron; spinels are oxides of magnesium, aluminum and chromium. Thus the magnesium content of a peridotite (expressed as a net content of magnesium oxide) is high: it can reach 40 percent. The magnesium content of a basalt (the characteristic rock of the oceanic crust) is no more than a fourth as great. Physically peridotite is a dark greenish rock with a high density, typically 3.2 grams per cubic centimeter. That density is consonant with the hypothesis that peridotites are the chief constituent of the upper mantle. The density of the upper mantle has in fact been inferred from the velocity at which seismic waves travel through it. The result suggests that the density of the upper mantle under the oceans is from 3.1 grams per c.c. to 3.3.

The second abundant rock was gabbros. Mineralogically and chemically they are much like basalt: they consist chiefly of silicates of calcium and aluminum, which form minerals such as feldspar. Physically too they resemble basalt: they are grayish and have a density of about 2.8 grams per c.c. The principal difference between a gabbro and a basalt is that the crystals in gabbro are larger. It is therefore inferred that gabbros and basalt both arise from the partial melting of peridotites in the mantle. The parts of the melt that reach the sea floor cool rapidly when they erupt. The result is basalt. The other parts cool slowly at greater depths. The result in that case is gabbro. Specifically gabbro solidifies under the axis of the midocean ridges, in magma chambers in the lower part of the oceanic crust.

The crucial point is that both types of rock abundant at transform faults—peridotite and gabbro—are normally deep-seated. Peridotite is in the upper mantle, gabbro is in the lower crust. The presence of these rocks suggests, therefore, that transverse ridges are slivers of deep crust and upper mantle that have somehow been uplifted. In this regard the study of sections of transverse ridges made up entirely of mantle-derived peridotite has been particularly revealing. The minerals in peridotite are stable only at the great temperatures and pressures of the mantle. Elsewhere they tend to be changed into other minerals. Moreover, some of the minerals in peri-

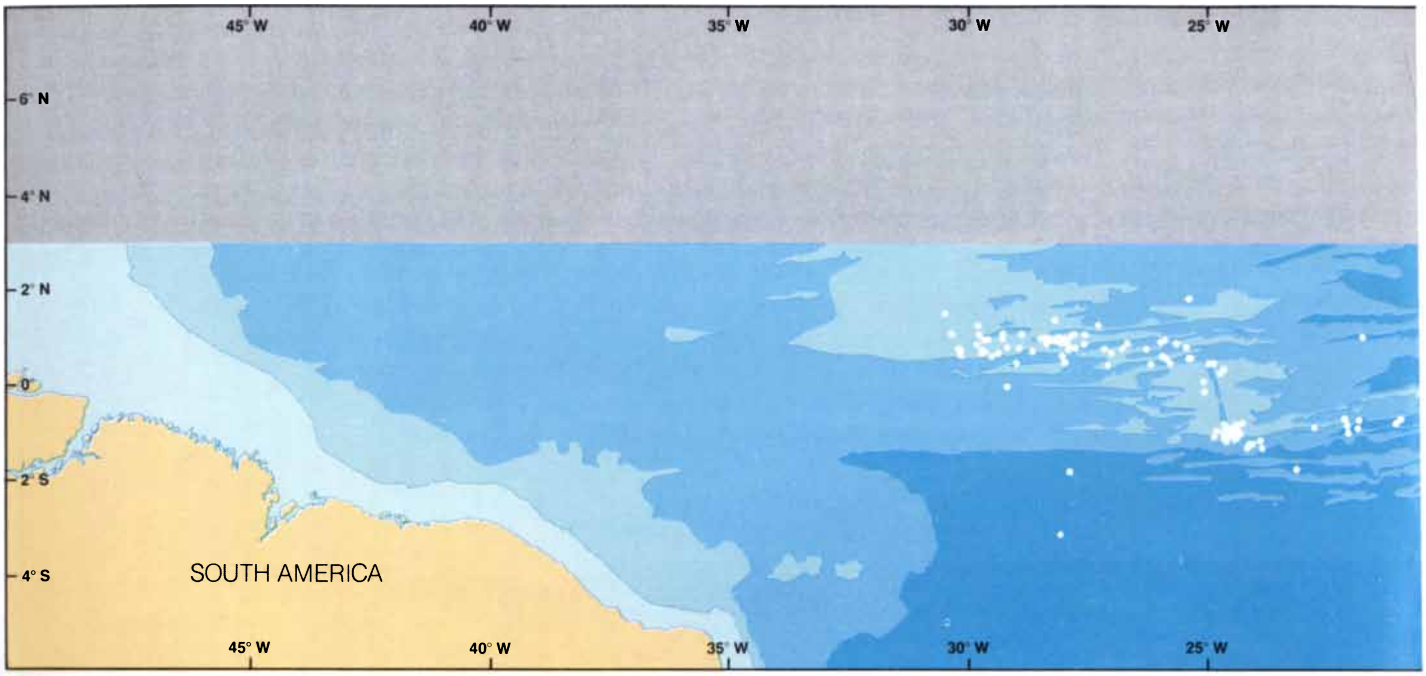




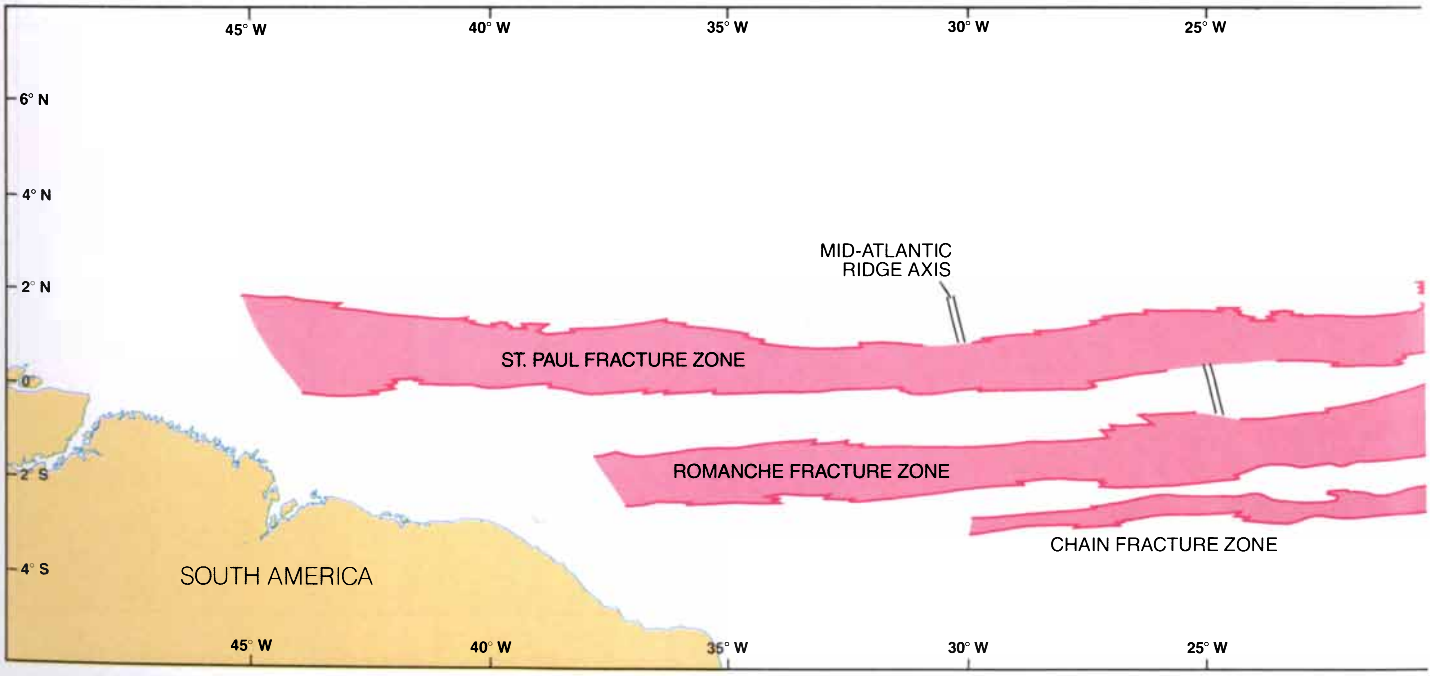
SEA-LEVEL MEASUREMENTS made by radar from a satellite yielded this image of a series of fracture zones dissecting the Mid-Atlantic Ridge. White patches mark areas such as the ridge itself, where the sea floor is relatively shallow. There the ocean's level rises. Dark patches mark increasing depths, notable among which are deep troughs found to be associated with oceanic fracture zones. There the ocean's level sinks. Many of the zones traverse the width of the equatorial Atlantic. The zones are identified in the accompanying map. A rectangle marks the area shown in illustrations on the next two pages. The image was provided by William F. Haxby of Lamont-Doherty.

dotite, particularly pyroxenes and spinels, vary slightly in chemical composition depending on the temperature and pressure—that is, the depth—at which they form. The mineral assemblages in the peridotites sampled at transform faults suggest they derive from depths in the earth exceeding 30 kilometers. (The boundary between the crust and the mantle is generally at a depth of from four to five kilometers.) These peridotites must somehow have risen that distance as essentially solid bodies.

In the course of such an ascent the peridotite will have reached levels in the crust, a few kilometers below the sea floor, where seawater is encountered. Indeed, seawater is particularly likely to penetrate into the crust at midocean ridges and transform faults because the crust in such places is greatly fractured and therefore is permeable. Some of the minerals in peridotite (notably olivine and pyroxenes) can react with the seawater, so that they are gradually changed into other minerals, notably the series of hydrated magnesium silicates called serpentines. As a result the density of the peridotite decreases dramatically, from about 3.2 to 2.6 or even less.



ELABORATE PLATE BOUNDARY consisting of segments of the Mid-Atlantic Ridge offset by a succession of transform faults can be inferred from two kinds of evidence: the topography of the ocean floor and the distribution of earthquake epicenters. The segments of the ridge are places where the sea floor is newly formed, so that it is hot and has expanded vertically; hence the ridge segments are rela-

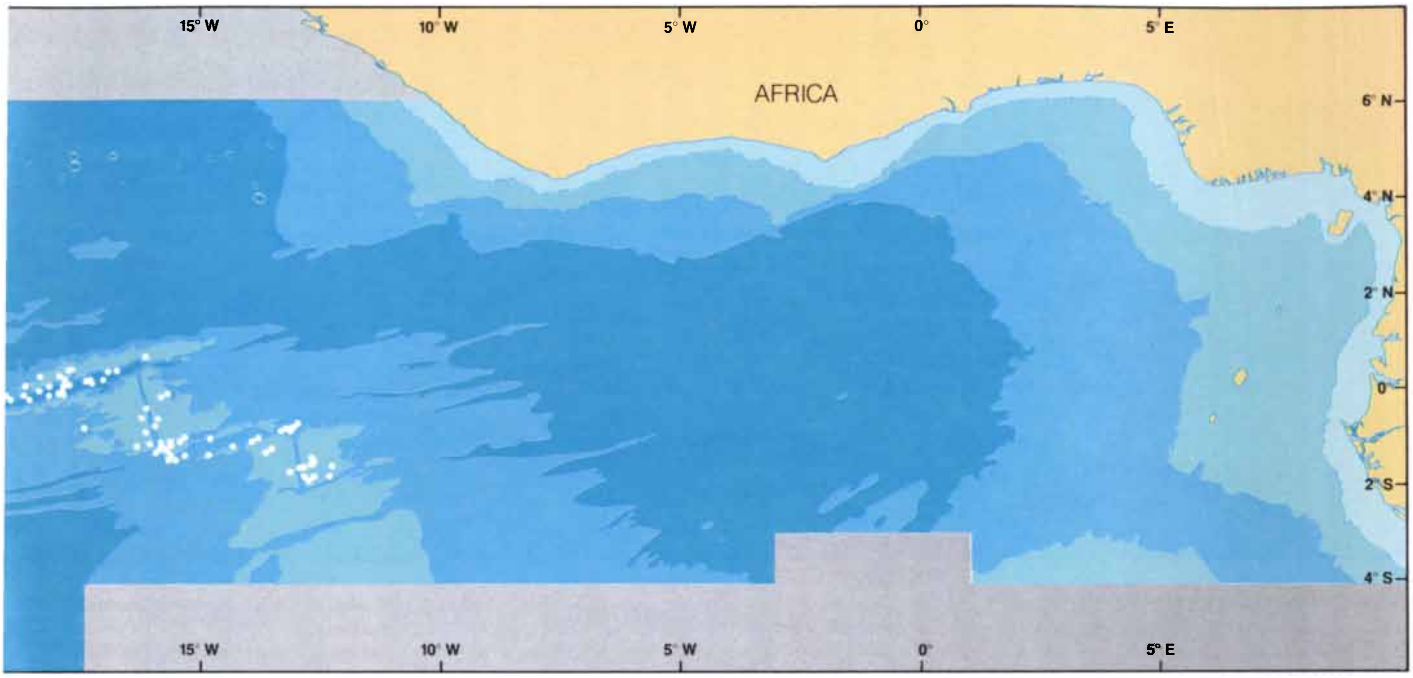


THREE MAJOR FRACTURE ZONES shown in red account for the long offsets in the Mid-Atlantic Ridge apparent in the illustration above. Each zone includes a transform fault that runs the length of one of the offsets. Then, beyond the limits of the offset, the zone extends across the width of the Atlantic as a trough flanked by anomalous sea-floor ridges. Near the coasts of Africa and South America

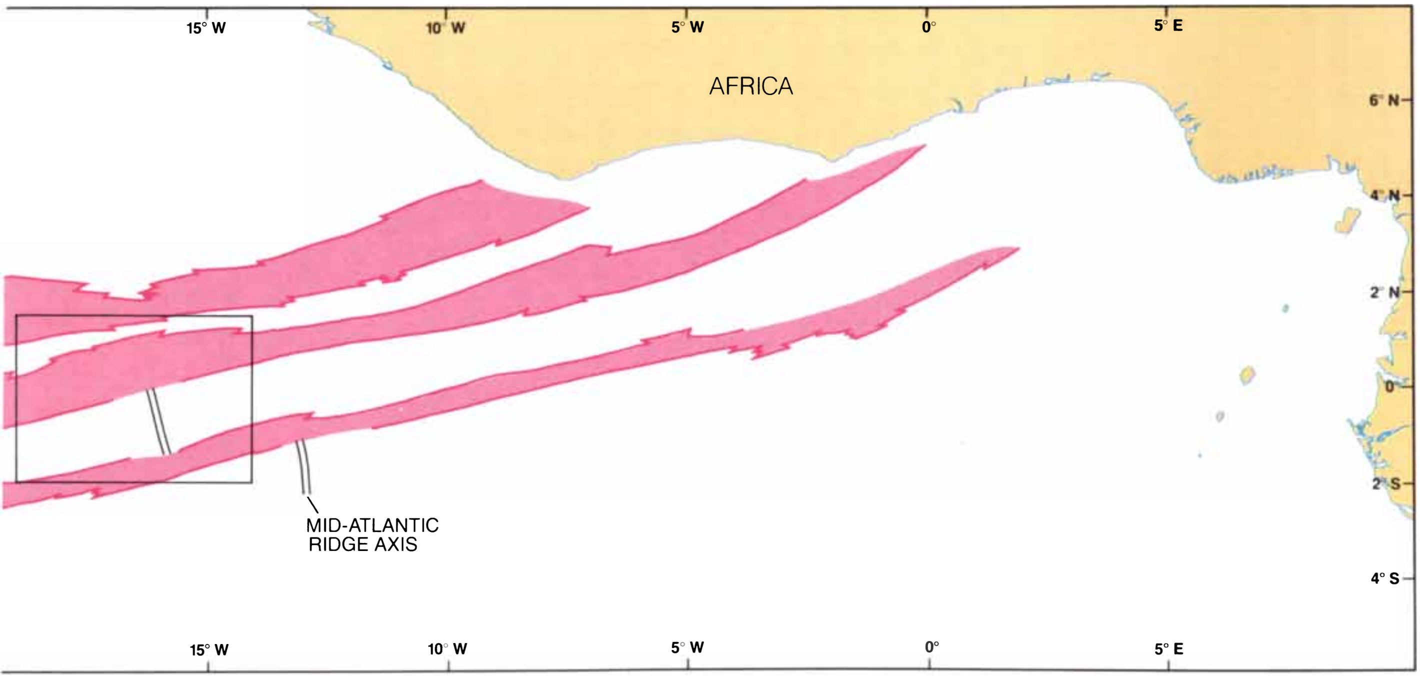
The decrease must tend to facilitate the ascent of the rock through the denser materials of the crust, including gabbro and basalt.

This is not to say the uplift of deep-seated rock at oceanic fracture zones has been inferred only from subtle analyses of rocks recovered from great depths in the ocean by dredging or with submersible research vessels. In at least one case the summits of a body derived from the mantle reach above sea level: the St. Peter-Paul islets rise from the deep ocean floor in the middle of the Atlantic just north of the Equator. In 1831 Charles Darwin, aboard the *Beagle,* landed briefly on their desolate, battered terrain. He saw that the islets were geologically unusual: they appeared not to be volcanic. It is now known that the islets consist of peridotites. They are a fragment of upper mantle that has been uplifted close to the intersection of the St. Paul transform with the axis of the Mid-Atlantic Ridge.

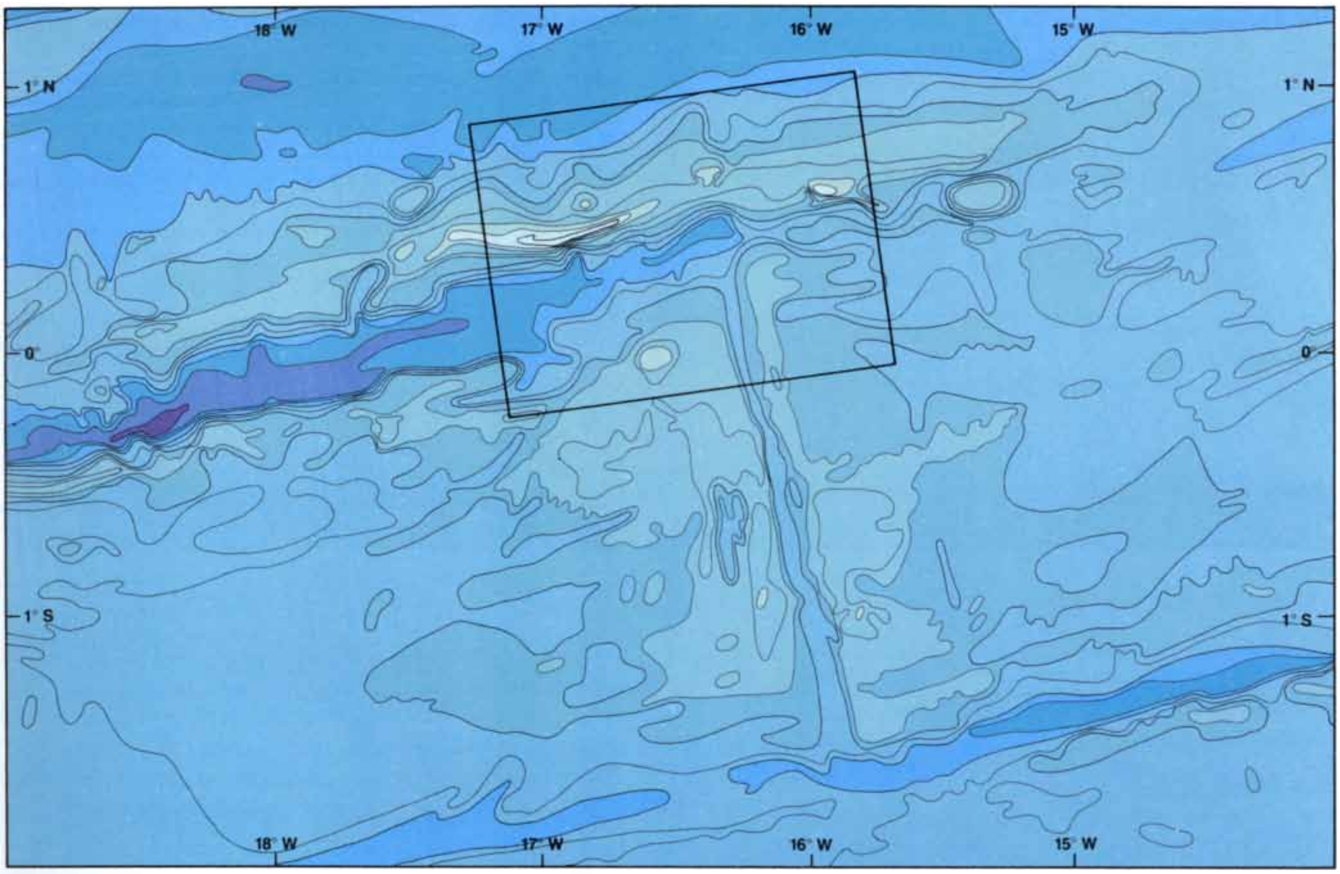
Moreover, it is now recognized that parts of other transverse ridges in the equatorial Atlantic have emerged as islands in recent times, geologically



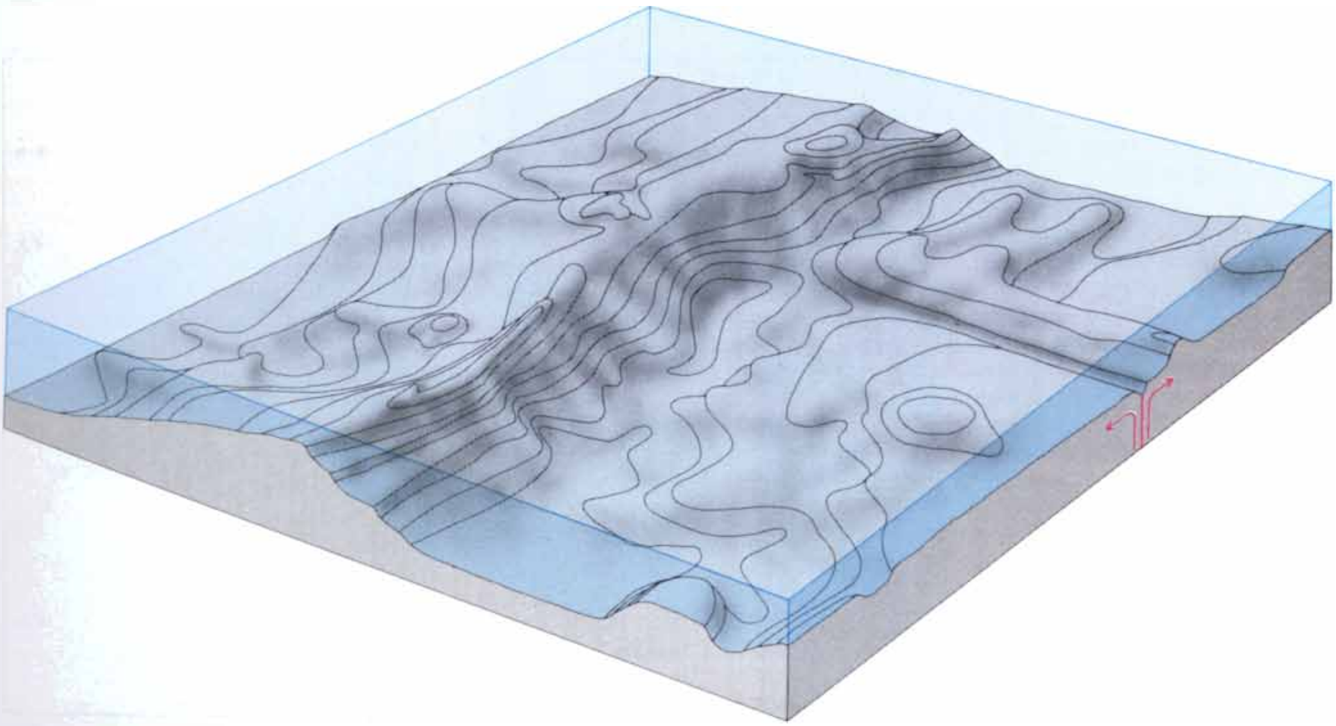
tively shallow parts of the Atlantic Ocean basin. The transform faults are places of sudden crustal slippage in the direction of sea-floor spreading; hence they (along with the ridge-axis segments) are places where earthquakes tend to occur. Contour lines on the map indicate depths of 1,000, 3,500 and 4,500 meters. White dots indicate the locations of earthquakes recorded from 1960 through 1970.



the zones are covered by continental sediment; nevertheless, seismic data confirm they are there. The three zones are at the heart of what is called the Equatorial Megashear Zone. Their positions were plotted by Marcus Gorini of Lamont-Doherty. The part of the Romanche Fracture Zone and the Mid-Atlantic Ridge enclosed by the rectangle is depicted in greater detail in the illustration on the next page.



ROMANCHE FRACTURE ZONE is the most pronounced of the fracture zones composing the Equatorial Megashear; its depth varies from less than one kilometer at the top of its ridges to more than 7.5 kilometers at the deepest parts of its trough. Here detailed topographic relief for part of the zone is mapped. Contour lines are at 500-meter intervals from 1,000 to 5,000 meters, with additional contours at 6,000 and 7,000 meters. The characteristic pattern is clear: a prominent trough is flanked by prominent ridges. A segment of the Mid-Atlantic Ridge meets the zone from the south. The part of the fracture zone enclosed by the rectangle appears in the illustration below.



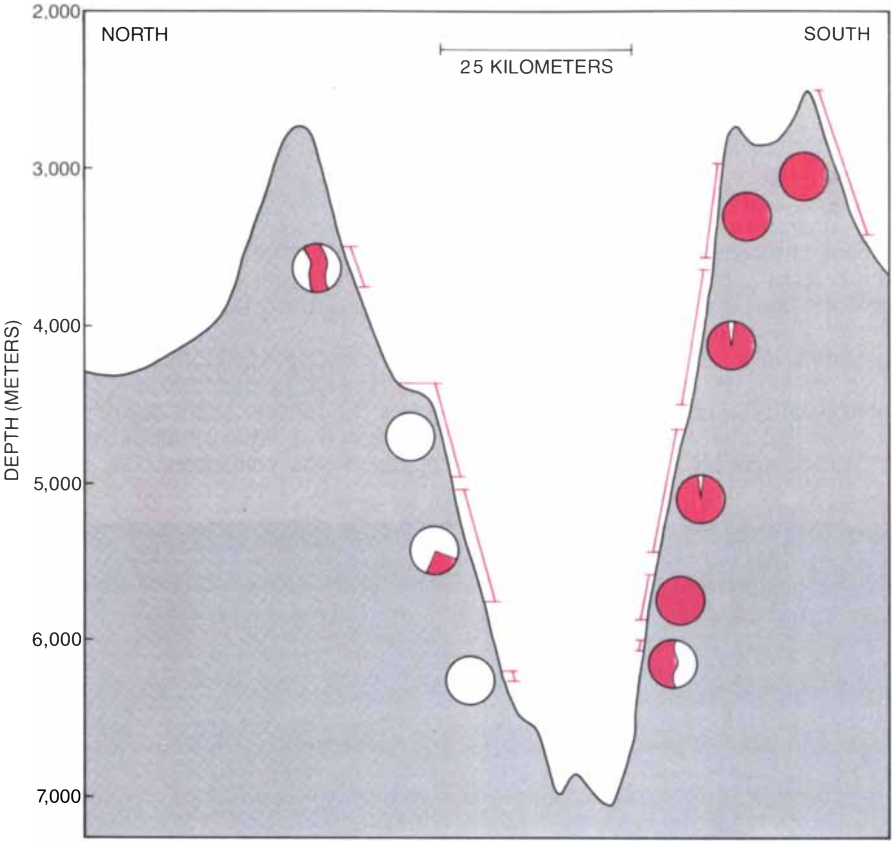
OBLIQUE VIEW of part of the Romanche Fracture Zone emphasizes the roughness of its topography by exaggerating the vertical scale. Contour lines are at 500-meter intervals from 1,000 to 5,000 meters. In plate-tectonic theory the elevation of the sea floor decreases with the age of the crust. The trough and ridges that characterize an oceanic fracture zone are thus anomalies in the age-depth relation.

speaking. At the Vema transform, 11 degrees north of the Equator, the axis of the Mid-Atlantic Ridge is offset by about 300 kilometers. Throughout the length of the offset an uplifted sliver of oceanic crust lies on the south side of the transform trough. The summit of the uplifted sliver is only 600 meters below sea level. A detailed sampling of the sliver shows it is capped by reef limestones similar to those ringing islands such as the Bahamas.

Several lines of evidence have helped us to deduce the environment in which the limestones formed. For one thing the limestones include shallow-water structures such as oolites: calcareous deposits formed when cold water flows onto warm, shallow banks. They include the fossil remains of organisms such as corals, calcareous algae and mollusks, which live in shallow water or in the zone between high tide and low tide. In addition the limestones show alterations typical of erosion on land. Finally, the proportions of the isotopes of oxygen are revealing. If a marine organism forms a shell of calcium carbonate, the oxygen atoms in the $CaCO_3$ will have a ratio of the isotopes oxygen 16 and oxygen 18 related to that of the oxygen dissolved in seawater. If at a later time the shell is exposed to rainwater, in which the ratio of the isotopes is slightly different, the ratio will change.

The evidence confirms that the limestones formed close to sea level as part of coral reefs and lagoons. Moreover, the evidence shows that the limestones were weathered and compositionally altered under the sky. This means that the shallowest parts of the Vema Transverse Ridge emerged as an island or islands at some time in the past. The evidence tells when. The identification of the fossil shells of foraminifera, a microscopic marine organism, suggests that the Vema Transverse Ridge was above sea level either continuously or intermittently from mid-Miocene to mid-Pliocene times, that is, from about 10 million to three million years ago. It has since sunk into the sea at an average rate of .3 millimeter per year.

A similar situation is found at the eastern intersection of the Romanche transform with the Mid-Atlantic Ridge. Here the Romanche Transverse Ridge shallows to little more than 1,000 meters below sea level for a distance of more than 100 kilometers. Again we were able to demonstrate that the summit of the ridge is capped by a fossil reef and that the summit emerged some five million years ago as an island or series of islands. They have subsided since then at the average rate of .2 millimeter per year.



PROFILE ACROSS ROMANCHE shows locations at which rocks were dredged by the authors. A small pie chart at each location shows the identity of the rocks. Some of the rocks (*white*) were unexceptional; they included basalt, the rock of which the oceanic crust consists. On the other hand, peridotite (*color*) proved to form 4.5 kilometers of the height of the southern wall of the Romanche trough. Peridotite is not normally part of the oceanic crust.

At the risk of oversimplifying, let us consider quite schematically what would happen at an oceanic transform fault if the direction of spreading of the oceanic plates that slide past each other along the length of the transform were to change even slightly from the orientation in which the transform makes a right angle with each segment of a ridge axis [*see illustrations on page 42*]. Suppose the angles increased. The plates would no longer slide past each other. Instead they would diverge. The transform zone would be subjected to extension; thus the zone would tend to open. For example, the zone might widen, and into the extended, weakened part of the oceanic crust some basaltic magma could rise. This possible formation has therefore been termed a "leaky" transform. Conversely, the angles could decrease. The plates would then converge on the transform zone. The crust in the zone would be deformed by lateral compression, and the resulting vertical movements could conceivably bring on the tectonic uplift of slivers of crust.

Is there any evidence that the direction of spreading of oceanic plates has changed over time? More than a decade ago Henry W. Menard, Jr., and Tanya M. Atwater of the Scripps Institution of Oceanography pointed out that magnetic anomalies in several parts of the earth's ocean basins do show changes in orientation. Magnetic anomalies are perhaps the most important evidence favoring the theory of plate tectonics. They arise because oceanic crust forms from cooling lava, and the magnetic field in the lava becomes aligned with the field of the earth at the time. The earth's field has reversed its polarity many times throughout geologic history; hence the field in the crust that formed at a midocean-ridge axis and then was carried off by sea-floor spreading should bear the record of a series of reversals. It does. Moreover, the pattern of reversals on one side of a midocean-ridge axis is the mirror image of the pattern on the other side of the axis. What Menard and Atwater saw is that the orientations of the magnetic "stripes" between reversals do not always parallel the orientations of the midocean ridges. The implication is that the direction of sea-floor spreading can change.

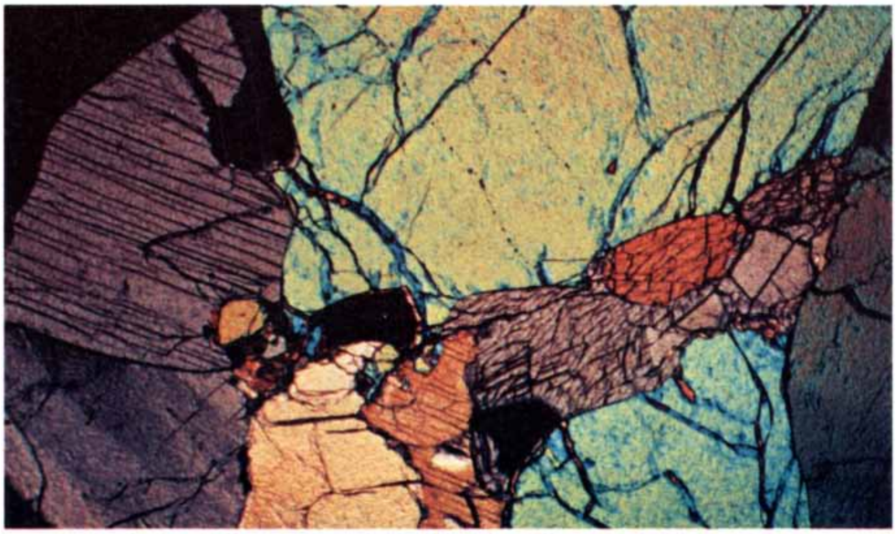
A second piece of evidence suggesting that the direction of sea-floor spreading can change is provided by the oceanic fracture zones themselves. The troughs and transverse ridges produced at transform faults and carried across the ocean basins by sea-floor spreading display sharp breaks in orientation. The breaks are clearly seen in the maps of the floor of the Atlantic produced with radar altimetry by William F. Haxby, a colleague of ours at Lamont-Doherty. Presumably each break was caused by the stresses arising when a plate changed its direction of motion.

Qualitatively, then, we could assert that the vertical tectonic movements giving rise to the anomalous shallows at transform zones are caused primarily by compressional stresses resulting from the reorientation of the transform fault in response to changes in a plate's direction of motion. The data that would make this assertion quantitative are not yet all available. One would need measurements of strain in the crust and also an extensive recovery of rock samples.

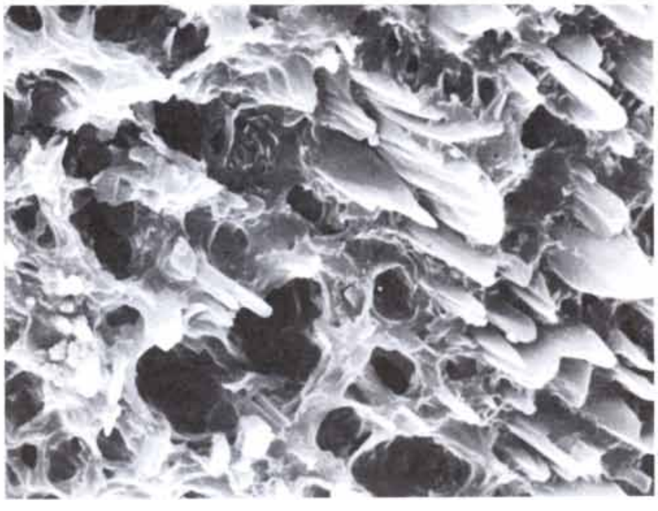
The closest we can come at present to making a quantitative analysis is at the Vema Fracture Zone. There in the early 1970's T. H. Van Andel of the Scripps Institution recognized the presence of a set of inactive transform valleys to the south of the transform that is known to be active today. Each inactive valley makes an angle of from 10 to 15 degrees with respect to the active transform. In addition each inactive valley can be traced only in oceanic crust more than 10 million years old.

These observations can be interpreted according to the following model. Before 10 million years ago a transform fault was active south of the present-day Vema transform. The tip of one of the Mid-Atlantic Ridge axis segments joined by the fault then propagated northward. Longitudinal propagation of that type has been documented in the East Pacific, in the Gulf of Aden, in the Red Sea and elsewhere by investigators including R. N. Hey of the Scripps Institution and Vincent Courtillot of the University of Paris. Next the Vema transform relocated so that the fault and the axis segments it joins assumed the geometry now observed. It is a geometry in which the fault meets the axis segments at right angles and the stresses in the region are relieved.

The model accounts for a curious finding we made at Vema: the uplifted crustal block on the south side of the



PERIDOTITE from the Oceanographer Fracture Zone in the North Atlantic was photographed from the submersible research vessel *Alvin* (*left*); its internal structure is suggested by that of a peridotite from a Red Sea fracture zone, photographed in thin section illuminated by polarized light (*right*). Peridotite consists of three minerals: pyroxene, olivine and spinel. A crystal of pyroxene (*brown in this illumination*) is at the upper left; crystals of olivine (*green and orange*) and spinel (*black*) fill the rest of the field. The field is about .5 centimeter across. Peridotite is thought to occupy the upper part of the mantle, the layer subjacent to the crust. Hence its presence in the elevated parts of fracture zones suggests that the elevations are slivers of crust and upper mantle lifted through vertical distances as great as 30 kilometers.



SCANNING ELECTRON MICROGRAPHS of peridotites from oceanic fracture zones show how they can be altered by contact with seawater. The peridotite at the left is from the Red Sea. It is unaltered. The surface of a crystal of pyroxene is stepped by a series of cleavages. The peridotite at the right is from Romanche. It is greatly altered: contact with seawater percolating downward through cracks in the oceanic crust has hydrated the pyroxene in it, forming a network of crystals of the mineral serpentine. The hydration and the porosity of the network have markedly reduced the density of the peridotite, perhaps facilitating its ascent from the upper mantle.

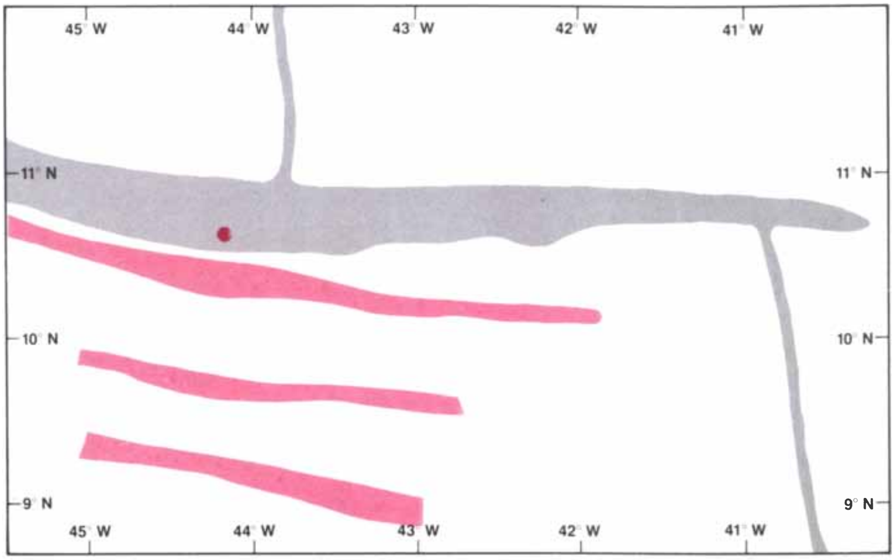
current transform fault is anomalously old. In particular some of the limestone recovered from the block proves to be more than 55 million years old. (It is limestone representing not a fossil reef but the solidification of calcareous sediment on the ocean floor.) Yet the rate of sea-floor spreading at Vema is little more than one centimeter per year. If the basaltic crust underlying the limestone had formed at the ridge-axis segment south of the current Vema transform and then had been moved, with the limestone on top of it, to its present location, the limestone should be no older than 30 million years.

We suggest that the crustal block capped by the limestone has undergone a process we call oscillatory spreading. Before the Vema transform relocated, the limestone was north of the fault and was moving eastward as part of the African plate. After the relocation (and within the present geometry) it is south of the fault and is moving westward as part of the American plate. Because of the reversal in its direction of travel, it has been moving for more than 55 million years but has yet to leave the area. Presumably the fault's transition from one configuration to the other was accompanied by the strong compressional stresses that lifted the crustal block the limestone is on.

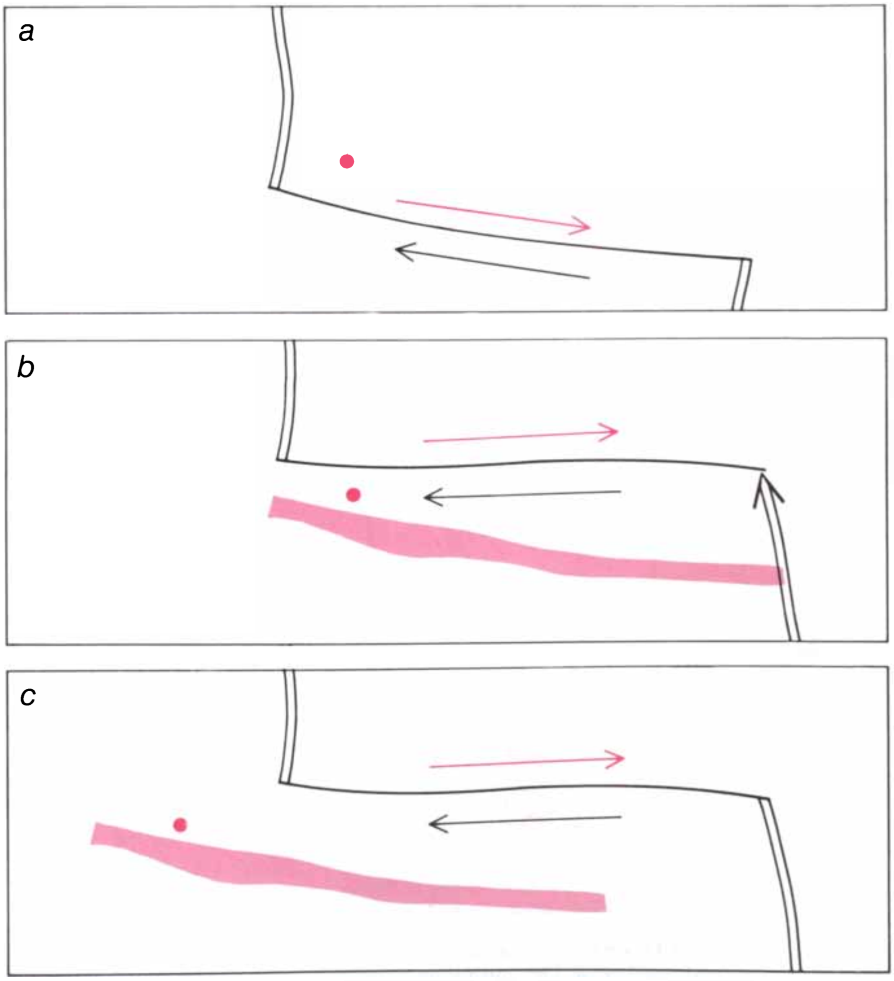
The observations and inferences at the Vema Fracture Zone demonstrate that the long transform faults dissecting a midocean ridge cannot be seen as quiescent features of the ocean floor. They are notably active features: they are subject to migration, to change in orientation and to deformation. One can see why. A length of several hundred kilometers means that a transform fault traverses relatively cold oceanic crust as it passes from one ridge-axis segment to another. The crust there behaves rather rigidly. If the offset were shorter, the crust it traverses would be hotter and more plastic. Readjustments would be easier. Perhaps it is little cause for wonder that topographic roughness and seismicity both increase with the length of the offset between two ridge-axis segments.

The transforms dissecting the Mid-Atlantic Ridge are generally much rougher than the ones dissecting the East Pacific Rise: the great ridge axis of the Pacific. Again one can see why. Sea-floor spreading is up to 10 times faster at the East Pacific Rise than it is in the Atlantic. Thus for equal lengths of offset the crust affected by a transform will be younger, hotter and less rigid in the Pacific than it is in the Atlantic.

Ultimately, of course, the instability of ridge-axis transforms must occur in response to the distribution of stresses within the oceanic plates, and the stresses must occur in response to the



MAP OF THE VEMA FRACTURE ZONE in the Equatorial Megashear shows the presence of numerous valleys (presumably inactive transform faults) south of the transform that is active today. The map displays the work of T. H. Van Andel of the Scripps Institution of Oceanography. In part of the Vema zone (*colored dot*) the rock proved to be limestone some 55 million years old. If the limestone had always been south of the Vema transform and had been carried to its present-day position by westward sea-floor spreading, it would be only half that age.



OSCILLATORY SPREADING is proposed as the reason for the anomalous age of the limestone sampled at Vema. At first the limestone (*colored dot*) was north of the Vema transform fault and so was moving eastward as part of the African plate (*a*). Then, however, the position of the transform shifted, so that the limestone is south of it (*b*). Today the limestone is moving westward as part of the American plate (*c*). A fossil valley marks the transform's former position.

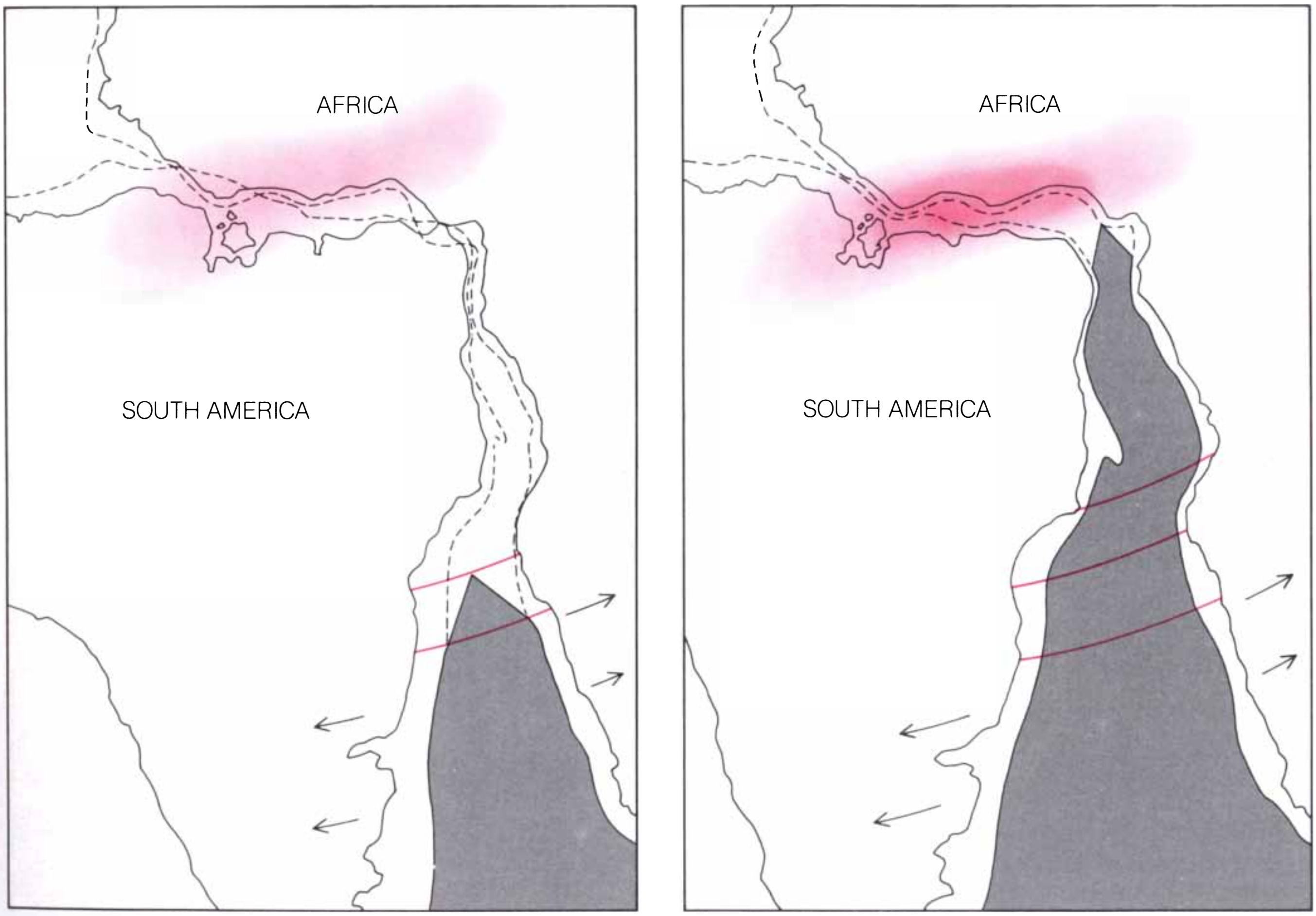
large-scale movements of the plates. The instability must therefore be extreme when a continent breaks up and an ocean basin opens. After all, continental lithosphere is thicker, colder and more rigid than oceanic lithosphere. The opening of the Atlantic is a case in point.

An analysis of the pattern of magnetic anomalies on the floor of the present-day Atlantic reveals that the Atlantic ocean basin broke open in discrete increments. Moreover, an analysis of the basin's geology suggests that in both the North and the South Atlantic the initial lines of opening were determined by the position of preexisting fracture systems traversing the continent that was soon to break apart. Presumably these zones were the wounds of tectonic events (collisions between plates, say) that occurred before the Atlantic developed. From the south to the north the major fracture systems were the Falkland Shear Zone and the Equatorial Megashear Zone, dominated by the Romanche Fracture Zone; then, in the North Atlantic, the Atlas Shear Zone and the Charlie Gibbs Shear Zone; and finally, in the Norwegian Greenland Sea, the Jan Mayen and the De Geer shear zones. The evidence suggesting the presence of the Equatorial Megashear Zone is particularly good. The remnants of the zone can be recognized in the Amazon trough on the American side and in the Benue trough on the African side.

The South Atlantic opened, then, from the south to the north, like a zipper, at a rate that varied from 10 to 20 centimeters per year, according to work by Philip D. Rabinowitz and John L. LaBreque of Lamont-Doherty. In essence the tip of a midocean-ridge axis moved northward, giving rise to oceanic crust that spread to each side. The motion began somewhere along the Falkland Shear Zone about 130 million years ago.

By 100 or 90 million years ago the tip of the axis had reached the Equatorial Megashear Zone [*see illustration below*]. The encounter must have been dramatic. For one thing, the Equatorial Megashear may well have represented what Courtillot calls a locked zone, that is, a part of a plate greatly resistant to being opened by a propagating rift. Then too, it is likely that the propagating axis tip impinged on the Megashear at an angle other than 90 degrees. The Megashear and the rift arose in different tectonic episodes, and so there is no reason to suppose they would be at right angles to each other. The oblique configuration would have been highly unstable. At one side of the ridge axis newly formed oceanic crust would have rammed into the Megashear, causing intense compression and probable uplift. At the other side of the axis the opposite would have happened: the stress would have been extensional. Perhaps the upwelling of basaltic magma into the weakened crust created a series of local oceanic spreading centers there.

Meanwhile the tip of a North Atlantic rift was apparently moving southward toward the Equator. The lateral distance from it to the southern rift would



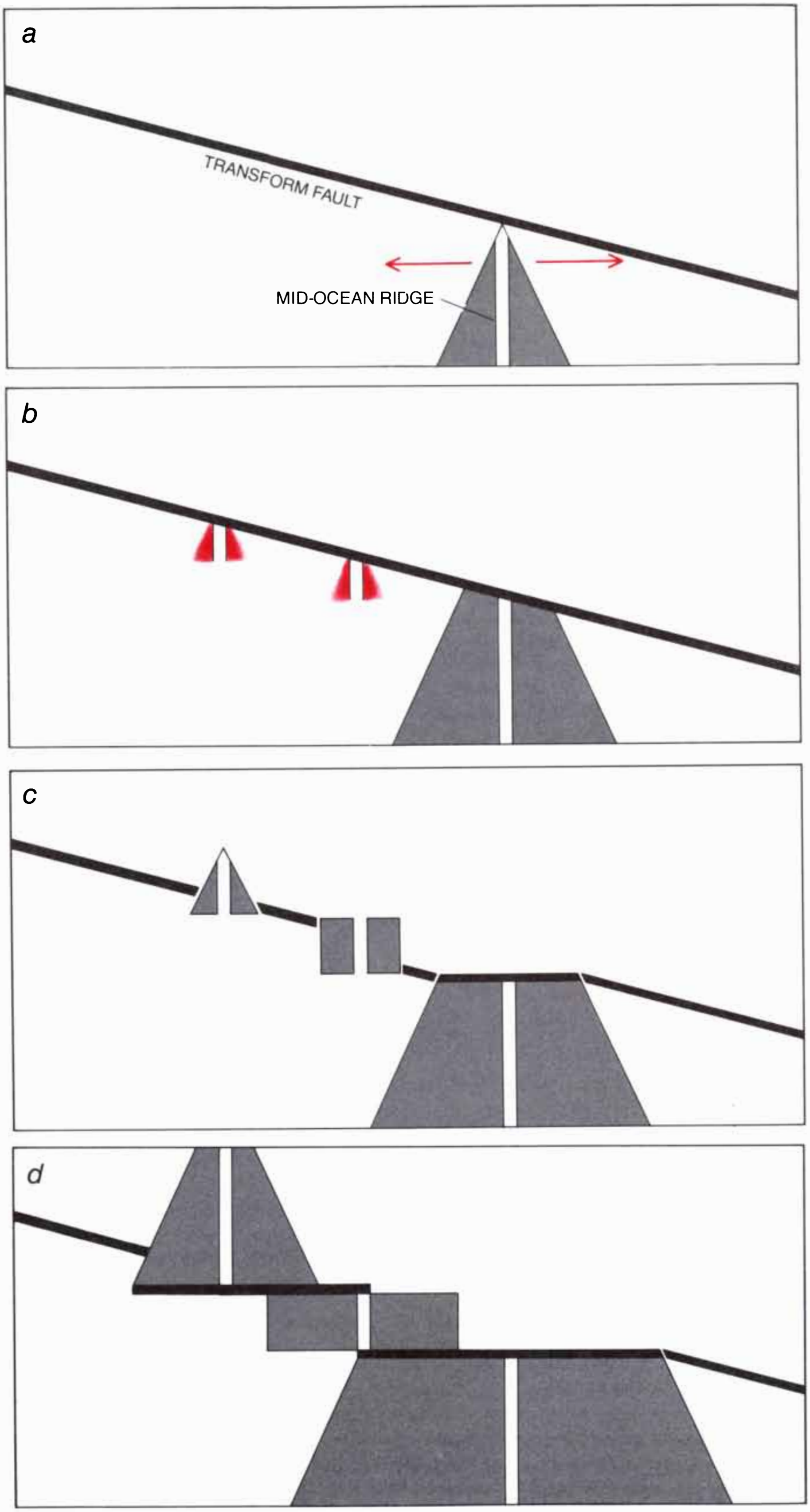
OPENING OF THE ATLANTIC was guided by preexisting fracture zones in the continental crust. The illustration shows two stages in the process. By 105 million years ago (*left*) the northern tip of a propagating rift had reached the first of a number of shear zones in the South Atlantic. Then, five million years later (*right*), the tip of the rift reached the Equatorial Megashear (*color*). At each successive encounter the northward advance of the rift was impeded until enough newly minted oceanic crust had spread from the rift to allow the fracture zone and the rift to reorient themselves in a direction at right angles to the spreading. Hypothetical details of the reorientation are shown in the illustration on the opposite page. The ancient positions of Africa and South America are inferred by repositioning their present-day outlines (*solid lines*) and their continental shelf, which is defined by the oceanic 1,000-meter contour (*broken lines*).

have been so great—more than 2,000 kilometers—that in order for the Atlantic to open at the Equator, a series of long-offset transforms would have had to develop. For at least 20 million years these transforms would have lain in continental lithosphere. Moreover, they probably would not have been at right angles to the rifts. Intense stresses must have arisen. The stresses would have continued until the quantity of newly minted oceanic crust enabled the rift segments and the fracture zone connecting them to readjust to a more nearly right-angled configuration, one that matched the direction of relative plate motion. On the basis of work by Dennis E. Hayes of Lamont-Doherty we surmise the new ocean basin would have had to be at least 1.5 times the width of the offset traversed by the fracture zone in order for the entire Equatorial Megashear to have made the readjustment. Only then could the zone have migrated and rotated with little or no impedance from continental lithosphere.

Surely the opening of the equatorial Atlantic was greatly influenced by the Equatorial Megashear Zone. One result is that the Mid-Atlantic Ridge axis shows the impressive series of long offsets observed at the Equator today. A second result is the general shape of the Atlantic. The rifts that opened the ocean tended to separate Europe and Africa from the Americas along a roughly north-south line. At the Equator, however, the Megashear Zone made the line east-west. Traces of the initial separation can be seen in the steplike east-west offsets of the equatorial Atlantic coastlines of both Africa and South America.

Throughout the further evolution of the Atlantic the equatorial transforms have been active, so that each change in the direction of sea-floor spreading has had dramatic consequences. In each case the readjustment of the geometry of the equatorial transform faults will have led to episodes of extension and compression, giving rise to volcanism and to the vertical motion of crustal slivers. On one resulting uplift along the St. Paul transform the birds observed by Darwin build their nests on the mantle of the earth. Elsewhere fossil coral reefs have sunk to the depths of the sea. It is likely that in the young and narrow equatorial Atlantic crustal blocks rose and fell quite often. Thus the geology of oceanic fracture zones probably influenced events as diverse as migrations of life between the continents and exchanges of water between the northern and southern ocean. The poetic intuition of William Carlos Williams was keen when he wrote:

> Oh most powerful connective, a bead
> to lie between continents through
> which a string passes . . .



HYPOTHETICAL SEQUENCE illustrates how a propagating rift might impinge on a preexisting shear zone. The illustration assumes that a propagating rift such as the one that opened the South Atlantic would tend to meet a preexisting shear at an oblique angle (*a*). Hence at one side of the rift the newly formed oceanic crust would be compressed against the shear; at the other side the crust would be stretched and would weaken. The weakened crust would admit basaltic magma; thus new oceanic spreading centers would arise (*b*). Ultimately they would produce enough oceanic crust for the fracture zone to relocate (*c, d*). In the illustration one of the new spreading centers resumes the propagation, so that the ridge axis comes to have offsets in it.

How Receptors Bring Proteins and Particles into Cells

Receptor-mediated endocytosis is a process whereby cells can take up specific large molecules. In most cases a receptor, having delivered its ligand, is recycled to the plasma membrane to bind more ligand

by Alice Dautry-Varsat and Harvey F. Lodish

The cells of a multicellular organism are surrounded by an aqueous medium, derived from the blood, that is like a very rich soup. It is an unusual soup in that it has many thousands of ingredients, most of them present at exceedingly low concentrations. Some of the ingredients are cellular building materials such as amino acids and nutrients such as vitamins, each of which is needed by a given cell in particular quantities and at particular times. Some ingredients are hormones delivering specific intercellular signals. Some are waste products or even toxic substances that particular cells are equipped to break down. Each cell must extract from the extracellular medium the substances it needs to internalize, rejecting the rest.

The sieve is the plasma membrane, which bounds the cell and controls its traffic with the medium and thus with every other cell of the organism. Like all biological membranes, the plasma membrane is mainly a double layer of phospholipid molecules in which many kinds of protein molecules are embedded. The proteins have a wide range of functions, one of which is to facilitate the selective uptake of specific water-soluble substances through the otherwise impermeable lipid bilayer. Ions (charged atoms) and small water-soluble molecules such as amino acids (the constituents of proteins) and sugars simply flo through or are pumped through specialized channels in the membrane; the channels are composed of proteins called permeases. Large molecules and particulate matter are brought into the cell by a quite different process, called endocytosis: a patch of the plasma membrane surrounds the material to be taken in, which is thus brought into the cell enclosed in a vesicle derived from the plasma membrane. There are three kinds of endocytosis.

In phagocytosis the binding of a very large particle or molecular complex to the surface of the cell triggers an expansion of the membrane around the object, which is incorporated into the cell in a baglike vesicle, an invaginated patch of the membrane, that can be several micrometers in diameter. Phagocytosis is the process whereby protozoans ingest bacteria and other food particles; in higher animals the immune-system cells called macrophages engulf intruders such as bacteria by phagocytosis.

Pinocytosis is a different process that results in the nonspecific uptake of extracellular fluid. A minute droplet of liquid is surrounded by a bit of invaginated plasma membrane and is internalized in a vesicle only about .1 micrometer in diameter, bringing with it whatever ions or small molecules happen to be in the droplet.

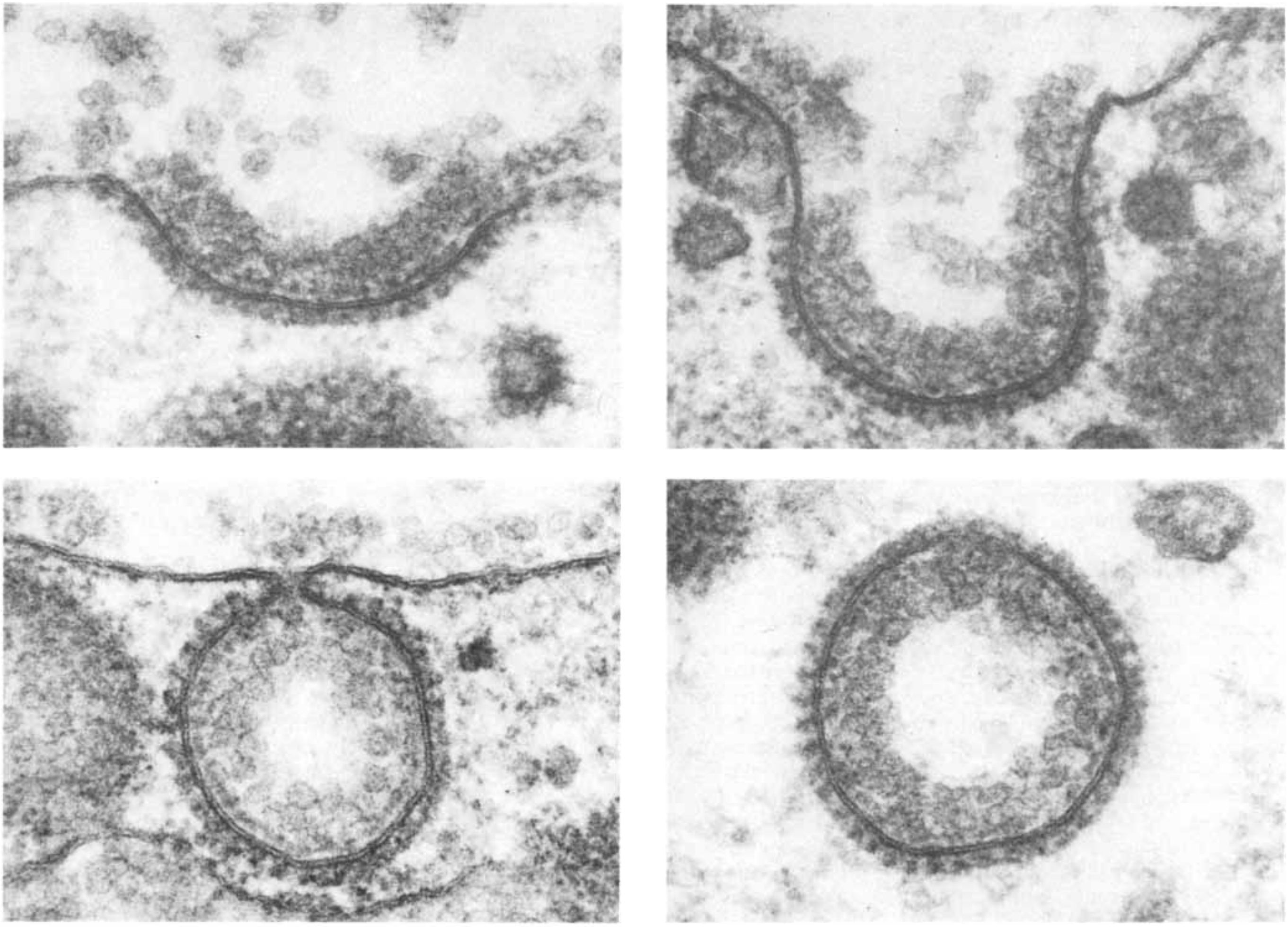
Receptor-mediated endocytosis, in contrast, is exquisitely specific. The receptors are membrane proteins, each of which has a binding site that fits a particular ligand: a protein or a small particle. The receptor in effect plucks one ingredient from the extracellular soup—even if it is present in a very low concentration and with a vast excess of unrelated molecules—and holds it fast. The binding triggers an invagination of the membrane to form a membrane-bounded vesicle enclosing the ligand and so to bring the ligand into the cell. In the past few years much has been learned about the mechanism of receptor-mediated endocytosis, and in particular about the remarkable events whereby the internalized receptor-ligand complex is dissociated, the ligand is dispatched to its intracellular destination and the receptor is recycled to the surface of the cell to bind more of the ligand.

A wide variety of ligands are brought into cells by receptor-mediated endocytosis, and once in a cell they are disposed of differently. The first example of protein uptake by this method was reported in 1964 by Thomas F. Roth and Keith R. Porter, who were then at Harvard University. It explained how insect and bird eggs acquire the large amount of protein stored in the yolk body to nourish the embryo: the proteins are imported into the developing oocyte after having been synthesized elsewhere. Roth and Porter showed that in the mosquito the precursor protein vitellogenin is synthesized in the female's liver, secreted into the blood and carried to the ovary, where it binds to receptors in some 300,000 tiny pits on the surface of the oocyte. The pits are internalized as vesicles, which fuse to form the yolk body. The vitellogenin is cleaved by protein-digesting enzymes (proteases) to yield two essential yolk proteins, lipovitellin and phosvitin.

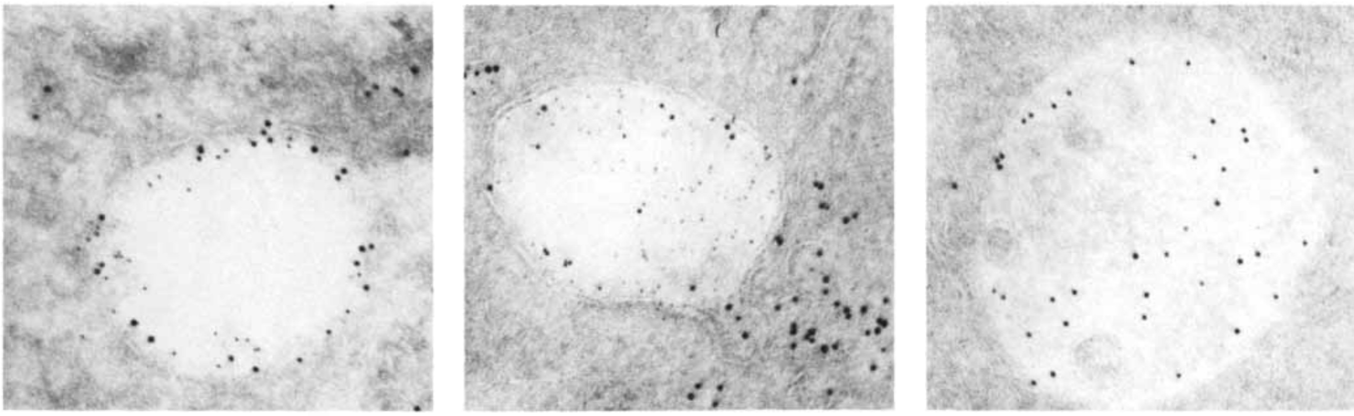
In mammals maternal immunity is transferred to the developing fetus by receptor-mediated endocytosis. Antibodies from the mother's blood bind to fetal cells lining the yolk sac, which have surface receptors that recognize a region common to all circulating antibodies of the immunoglobulin-G type. The antibodies are transported into the cells and are secreted into the fetal circulation.

Cholesterol is a major lipid component of the plasma membrane of all mammalian cells. Richard G. W. Anderson, Michael S. Brown and Joseph L. Goldstein of the University of Texas Health Science Center at Dallas have shown in elegant detail how the cholesterol is made available to the cell. Cholesterol is synthesized and stored in the liver and there packaged into a large, spherical low-density lipoprotein (LDL) particle, which serves as its transporter. Each LDL particle has a core of some 1,500 cholesterol molecules chemically bound to fatty-acid chains. The core is enveloped in a single-layer phospholipid membrane and more cholesterol; a binding protein called *apo-b* is embedded in the membrane and the particle is secreted into the circulation.

LDL particles bind to specific cell-



SUCCESSIVE STAGES OF ENDOCYTOSIS are seen in a series of electron micrographs of hen oocytes, the precursors of eggs. The shallow indentation in the oocyte's plasma membrane (*top left*) is a coated pit. The thick coating on the cytoplasmic (inner) side of the pit is a lattice of the protein clathrin. Large particles lining the membrane on the extracellular side of the pit are thought to be lipoprotein particles; they are presumably bound to specific receptors embedded in the oocyte's membrane. At later stages the pit deepens (*top right*) and rounds up to form a nascent coated vesicle (*bottom left*), and the vesicle is pinched off from the surface membrane, internalizing the lipoprotein ligand (*bottom right*). The micrographs were made by M. M. Perry and A. B. Gilbert of the Agricultural Research Council's Poultry Research Centre in Edinburgh, who treated cells with tannic acid and stained them with lead citrate. Enlargement is 135,000 diameters.



SEPARATION OF RECEPTOR AND LIGAND in rat liver cells is visualized by a double-label electron-microscope technique. Frozen, sectioned cells were treated sequentially with an antibody preparation specific for the ligand (an abnormal galactose-terminal glycoprotein) and then with one specific for its receptor, and the two antibodies were indirectly labeled with gold particles of differing size. The antibodies bound to their respective antigens, whose location was revealed by the electron-dense gold particles, which are seen as large and small dots in the micrographs. In an endosome (*left*), an endocytic vesicle near the cell surface, receptors (*large dots*) and ligand molecules (*small dots*) are seen together on the inner surface of the vesicle membrane. Deeper in the cell, in a vesicle called a CURL (*center*), ligand molecules have been released from the receptors and are free in the interior of the vesicle; the receptors are concentrated in tubular structures attached to the lower right side of the vesicle. These membranous tubules are thought to return the receptors to the cell surface. A micrograph made with the gold labeling reversed shows that a later-stage vesicle, a multivesicular body (*right*), contains ligand molecules but no receptors. The micrographs were made by Hans J. Geuze, Jan W. Slot and Ger J. A. M. Strous of the University of Utrecht in collaboration with Alan L. Schwartz and one of the authors (Lodish). The cell sections are enlarged about 90,000 diameters.

surface receptors that recognize the *apo-b,* are internalized in membrane-bounded vesicles and then transported, in a series of progressively larger vesicles, to the lysosome: a cell organelle equipped with a number of hydrolases, or digestive enzymes. The enzymes break down the LDL particles, degrading the binding protein, splitting the fatty acid from the cholesterol molecules and releasing the cholesterol and the fatty acids for incorporation into the cell membrane. Brown and Goldstein and their colleagues have shown that a lack of functional LDL receptors is responsible for familial hypercholesteremia, a genetic disease characterized by extremely high blood-cholesterol levels and consequent early atherosclerosis.

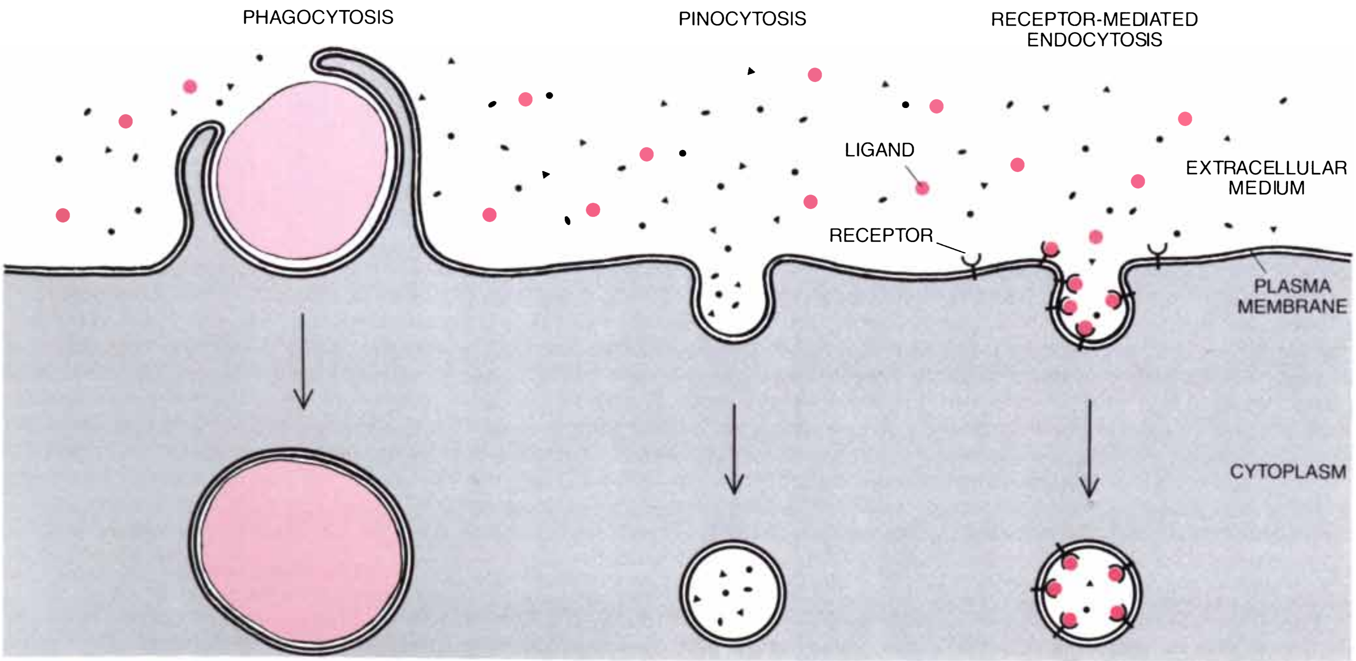
Iron is an essential constituent of all cells. It is brought to cells by a carrier protein called transferrin. The carrier binds ferric ions (Fe^{+++}) in the intestine (where iron is absorbed from foods) and in the liver (where iron is stored). A loaded ferrotransferrin molecule, carrying two iron ions, binds to a specific receptor on the surface of a cell, and the receptor-transferrin complex is internalized by endocytosis, making the iron available to the cell.

One of the major roles of receptor-mediated endocytosis is to internalize hormones and other proteins that deliver specific signals to certain cells. Insulin molecules, for example, bind to specific receptors on a target cell. The binding triggers several metabolic processes, including an increase in the uptake of glucose. It is thought the hormone response is terminated when the insulin-receptor complexes are internalized and the insulin is degraded in lysosomes. Potentially harmful materials are also disposed of in this way. For example, there are abnormal glycoproteins (proteins with attached sugar chains) that have the sugar galactose at the terminus of the chain instead of the usual sugar, sialic acid. Liver cells have receptors that recognize and bind galactose-terminal glycoproteins, which are then taken into the cell to be broken down in lysosomes.

The receptors that mediate endocytosis are proteins that span the thickness of the plasma membrane. Receptor proteins are amphipathic: they have two hydrophilic regions, which extend into the aqueous medium outside the cell and within the cytoplasm, and a central hydrophobic region that binds tightly to the fatty acids forming the core of the membrane. Not much is known about the structure of these receptors, but the receptor for transferrin is known to be a glycoprotein with a molecular weight of 180,000. Ian S. Trowbridge of the Salk Institute for Biological Studies, Robert Allen of the University of Colorado and Howard H. Sussman of Stanford University have shown that the transferrin receptor is made up of two identical polypeptide chains, each one about 800 amino acids long, that are linked by a single disulfide bond. In addition to at least three carbohydrate chains each polypeptide carries a fatty acid, palmitate, that may help to anchor the receptor in the membrane.

Each receptor binds two molecules of transferrin, presumably one molecule per polypeptide chain. Under physiological conditions the binding is followed quickly by internalization. The two steps can be separated for experimental purposes because binding does not require energy, whereas internalization does require it and can therefore be blocked by low temperature or by inhibitors of cellular energy production; under these conditions binding takes place normally and can be quantified. Like most receptors, the one for transferrin has a very high affinity for its ligand. Its dissociation constant (the concentration of ligand at which half of the receptors are occupied) is 5 nanomolar (5×10^{-9}M), or some 350 micrograms per liter. This means that transferrin is bound to its receptor even when the transferrin concentration is only about one hundred-thousandth of the total concentration of proteins in the blood.

Ligand-receptor complexes cluster at particular sites on the plasma membrane, namely in what are called coated pits. The pits, first observed by Roth and Porter, turn out to be present on almost all animal cells; typically they account for some 2 percent of the cell surface. The coat for which the pits are named, a thick proteinaceous layer on the inner side of the plasma membrane under each pit, is mainly clathrin, a fibrous protein that was identified by Barbara M. F. Pearse of the Medical Research



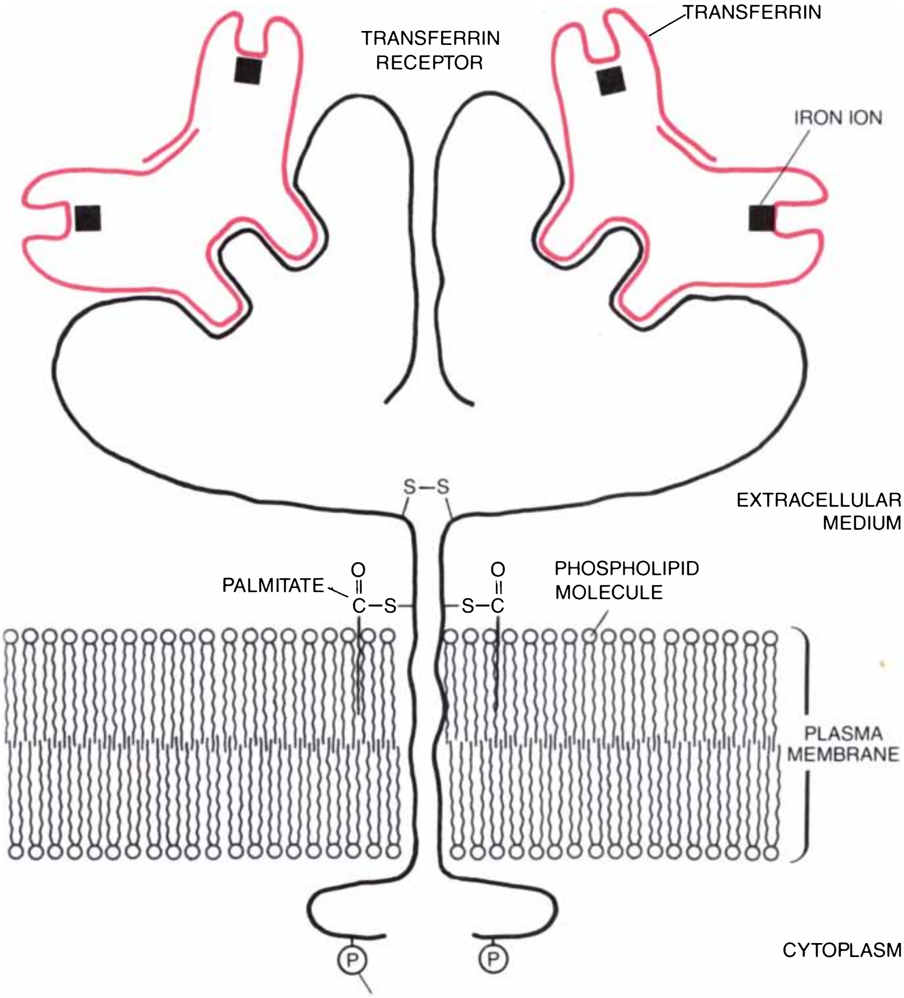
THREE KINDS OF ENDOCYTOSIS are diagrammed schematically. In phagocytosis (*left*) the cell's plasma membrane binds to a large particle such as a bacterium and expands to wrap around it and engulf it within the cell. In pinocytosis (*center*) a droplet of the liquid extracellular medium is surrounded by a patch of the membrane, which folds inward and "buds off" to form a membrane-bounded vesicle enclosing the droplet and any small molecules dissolved in it. Receptor-mediated endocytosis (*right*) is a mechanism for the selective uptake of large molecules or particles. A ligand binds to its specific receptor on the plasma membrane, triggering the internalization of the receptor-ligand complex in an invagination of the plasma membrane. The vesicle thus formed buds off inward, carrying the ligand into the cell.

Council Laboratory of Molecular Biology in England.

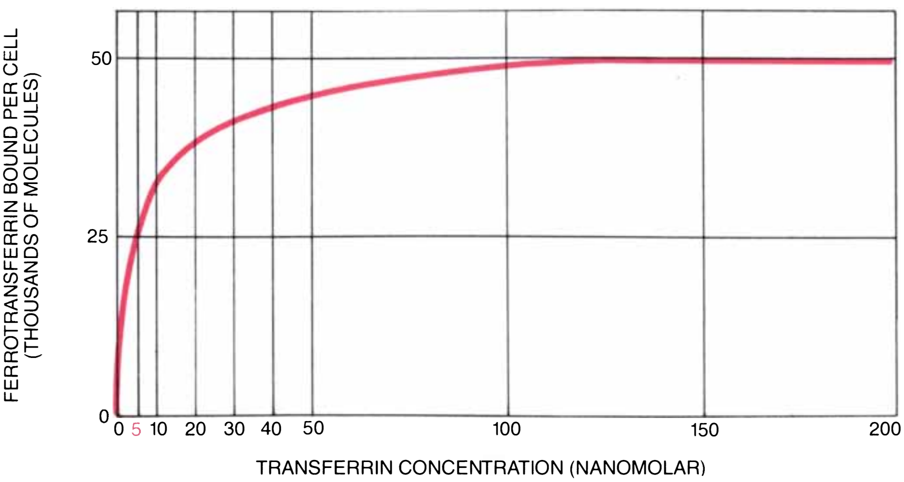
It is the coated-pit region of the membrane that invaginates to form a vesicle. Indeed, the process is continual. Coated pits keep folding inward to form vesicles and are continually regenerated on the cell surface. The vesicles are themselves coated, but now, of course, on their outer, or cytoplasmic, surface: the clathrin coat envelops the vesicle membrane in a fibrous network of pentagons and hexagons. (Apparently it is the polymerization of clathrin that actually forms the coated vesicle. When coated vesicles are purified free of membrane, the clathrin can be dissociated into three-armed subunit structures called triskelions. Under appropriate conditions the triskelions can aggregate to form the kind of basketlike structure that surrounds a coated vesicle.) As the coated vesicles move deeper into the cytoplasm they shed their clathrin coat and fuse with one another (or with a different kind of vesicle) to form larger, smooth-surfaced vesicles called endosomes.

Do receptors always reside in coated pits or is it only the binding of a ligand that makes them migrate there through the essentially fluid membrane? It depends on the receptor. Anderson has shown that the LDL receptors on fibroblasts (connective-tissue cells) are found primarily in coated pits even in the absence of bound LDL particles. On the other hand, receptors for transferrin, insulin and galactose-terminal glycoproteins are ordinarily distributed diffusely on the plasma membrane and seem to cluster in coated pits only when the ligand has bound and when the temperature is at least 37 degrees Celsius. A simple explanation would be that certain receptors have some property directing them to coated pits and holding them there; part of the receptor might, for example, have an affinity for a component of the pit region. Such a property might be elicited in other receptors only by the binding of its ligand, perhaps through a change in receptor conformation.

Although different ligands have different final destinations, most of them are dispatched to the lysosomes. There some 40 digestive enzymes, functioning in an acidic environment (a *p*H of from 4.5 to 5 in contrast to the neutral *p*H 7 to 7.4 of the cytoplasm), break down the ligands. The digestion products are either eliminated from the cell or (as in the case of the LDL components) exported into the cytoplasm as raw materials. It would be wasteful indeed if receptors shared the fate of their ligands, and indeed they do not. Some years ago it became clear that receptors, having delivered their ligands to the interior of the cell, are recycled to the cell surface to do it all over again. There were several lines of evidence. The making of new recep-



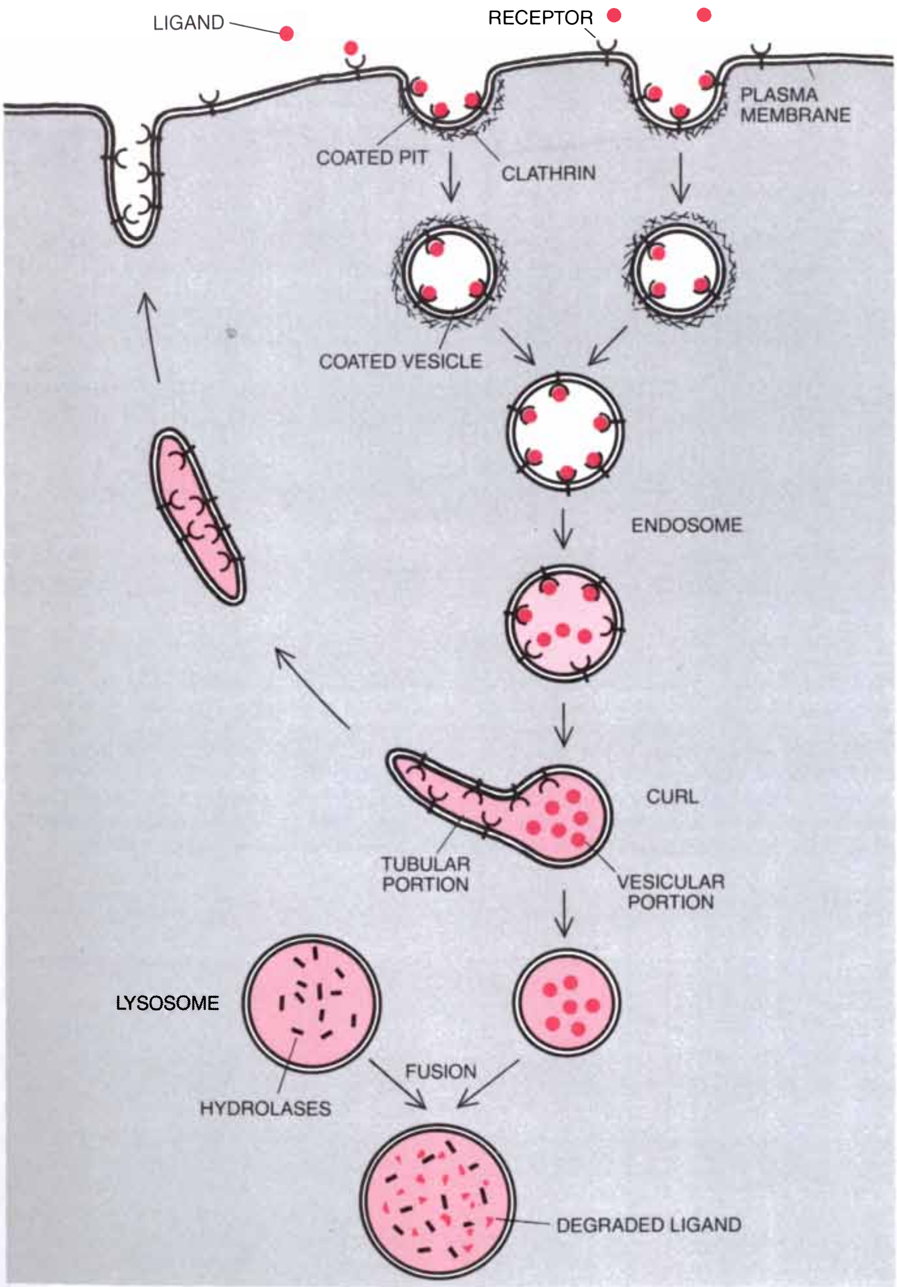
RECEPTOR FOR TRANSFERRIN, an iron-carrying protein, is embedded in the plasma membrane of a cell. The membrane is mainly a bilayer of phospholipid molecules whose fatty-acid tails make the membrane impermeable to water-soluble molecules. The actual structure of the transferrin receptor is not known; its shape in this highly schematic drawing reflects what is known. The receptor has two identical polypeptide chains linked by a disulfide bond (*S–S*); the bulk of the chains is exposed outside the cell; each chain carries a fatty acid, palmitate, that may help to anchor the receptor; the receptor binds two molecules of transferrin, presumably one molecule per chain. Each transferrin molecule can carry two ferric ions (Fe^{+++}).



AFFINITY OF RECEPTORS for iron-loaded ferrotransferrin is calculated. Liver-tumor cells are incubated in increasing concentrations of radioactively labeled transferrin at four degrees Celsius. The ligand binds at the surface but is not internalized, and the amount bound is quantified. The maximum binding level, 50,000 molecules per cell, shows a cell has some 50,000 transferrin receptor sites on its surface. The concentration at which half of them are occupied, 5 nanomolar, is the dissociation constant, a measure of the receptor's affinity for its ligand.

tors could be blocked with specific inhibitors of protein synthesis and yet receptor-bound ligand continued to be internalized. It could be shown that the lifetime of most receptors far exceeds the lifetime of their ligand and that each receptor for transferrin, LDL or the galactose-terminal glycoproteins internalizes a ligand every 10 or 15 minutes, and keeps doing so for many hours.

Do receptors cycle continuously, like a shuttle bus, or is their cycling induced by the binding of ligand (making them more like a taxicab)? Again the answer differs for different receptors. The LDL receptors seem to keep cycling through human fibroblasts whether or not they have bound ligand LDL. On the other hand, there is evidence that the receptors for galactose-terminal glycoproteins are internalized only when the ligand is bound. Apparently it is the receptors that are always localized in coated pits (even in the absence of ligand) that cycle continuously, because the coated pits keep invaginating; other receptors are seen to cycle only in the presence of ligand because it is only then that they move to coated pits.



PATHWAY OF RECEPTORS AND LIGANDS shown here was determined for galactose-terminal glycoproteins but is thought to apply in the case of other ligands and receptors as well. Ligand binds to receptors diffusely and then collects in coated pits, which invaginate and are internalized as coated vesicles whose fusion gives rise to endosomes and then to a CURL (so designated because it is a "compartment of uncoupling of receptor and ligand"). In the acidic CURL environment (*color*) ligand is dissociated from receptors. Ligand accumulates in the vesicular lumen of the CURL and the receptors are concentrated in the membrane of an attached tubular structure, which then becomes separated from the CURL. The vesicular part moves deeper into the cell and fuses with a lysosome, to which it delivers the ligand for degradation. The membranous tubular structure is thought to recycle receptors to the plasma membrane.

At what stage are receptors separated from their ligands, and how? Hans J. Geuze, Jan W. Slot and Ger J. A. M. Strous of the University of Utrecht, collaborating with Alan L. Schwartz and one us (Lodish), applied a novel electron-microscope technique in an effort to find out where the separation takes place. An antibody directed against a galactose-terminal glycoprotein and one directed against the galactose receptor were made and the two antibodies were labeled with gold particles of differing size. Rats injected with the glycoproteins were sacrificed at varying times after the injection, the livers were fixed and sectioned and the sections were treated with the two antibodies, which bound to the ligand and to the receptor; the gold particles were visible as dense dots in electron micrographs.

In sections made in the early stages of endocytosis both the receptor and its ligand are seen to be closely associated with the membrane of vesicles lying just under the surface of the cell. The lumen (the fluid-filled interior) of the vesicles does not contain any free ligand. Apparently the ligand is still bound to receptors embedded in the vesicle membrane. Farther from the cell surface, however, one sees larger vesicles, and these have free ligand in the lumen; the larger the vesicle is, the more free ligand seems to be present. In these larger vesicles the receptors seem to be still associated with membrane, but they are no longer distributed randomly over the entire vesicle membrane. Rather, they either are clustered near an end of the vesicle where the vesicle seems to be adjacent to or actually fused with thin, membranous tubules, or they are in the tubules themselves.

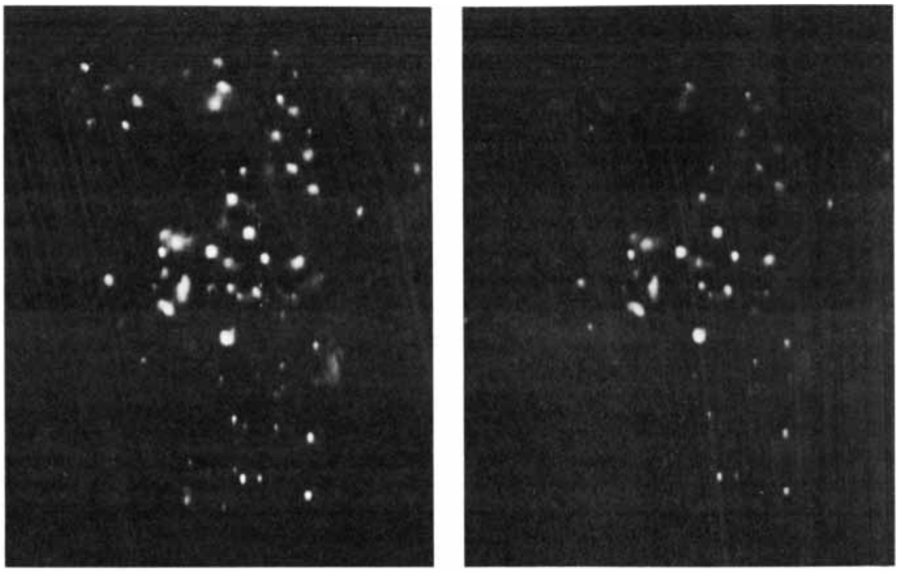
We call this vesicle-plus-tubule structure a CURL (for "compartment of uncoupling of receptor and ligand"). It seems to be the organelle in which receptors and ligands are separated and redistributed, with the ligands accumulating in the vesicular part and the receptors accumulating in the membranous tubules. Tissue sections made after endocytosis has proceeded for a longer time show that the CURL vesicles eventually fuse with lysosomes, where the ligand is degraded. Before the fusion takes place, however, the tubular portions loaded with receptors have been detached from the CURL vesicles, and so the receptors escape lysosomal degradation. The tubular structures seem to serve as intermediates that somehow get the receptors back to the cell surface. Whether they are vesicles that move to the surface or are part of a fixed tubular system, something like the endoplasmic reticulum, extending to the surface is still not known.

On the surface of a cell a receptor binds its ligand tightly. What causes a receptor and its ligand to dissociate within the cell? A low *p*H promotes such dissociation, and a few years ago it was reported that the recycling of receptors can be inhibited by agents that raise the *p*H of acidic vesicles within the cell. Benjamin Tycko and Frederick R. Maxfield of the New York University Medical Center showed directly that endosomes (and presumably CURL's) are normally acidic. They studied the endocytosis in human fibroblasts of alpha-2-macroglobulin, a protein that binds proteases and is itself bound by cell-surface receptors and internalized, thereby clearing the proteases from the circulation. Tycko and Maxfield coupled the fluorescent dye fluorescein to the alpha-2-macroglobulin. The excitation spectrum of fluorescein is sensitive to *p*H, that is, the intensity of its fluorescence varies with *p*H as well as with the wavelength of the light to which the dye is exposed.

The fibroblasts were incubated with the fluorescein-labeled ligand just long enough for the ligand to be internalized and reach prelysosomal vesicles such as endosomes or CURL's. Then the cells were chilled, illuminated at two different wavelengths and observed with an image-intensifier camera mounted on a fluorescence microscope. Calculations based on the relative intensity of fluorescence at the two excitation wavelengths showed that the *p*H of the vesicles containing the ligand was about 5.0.

Experiments with purified receptors and ligands show that whereas receptors bind their ligand tightly at a neutral *p*H, the ligands are rapidly dissociated when the *p*H goes below 5.5. It is clear that when a receptor-ligand complex enters an acidic vesicle, the ligand is separated from the receptor and becomes soluble in the vesicle's lumen; the receptor remains bound to the vesicular membrane and is eventually recycled. How the endosomes (or CURL's) are acidified is not yet established, but it is known that there is an enzyme in the endosomal membrane that exploits the energy stored in adenosine triphosphate (ATP) to pump protons into the lumen. An excess of protons means a low *p*H.

Unlike other ligands, ferrotransferrin is neither degraded nor stored after internalization. Instead it gives up its load of iron ions, which remain inside the cell, and is itself quickly secreted from the cell as iron-free apotransferrin. Why is the transferrin protein excreted from the cell and why does it leave its iron behind? Together with Aaron Ciechanover we studied the endocytosis of transferrin in cultured human liver-tumor cells and found the answers in the differential effect of *p*H on the binding of ferrotransferrin and of apotransferrin to the transferrin receptor. Richard D. Klausner of the National Cancer Institute and his colleagues, working with cultured precursors of red blood cells, reached similar conclusions.

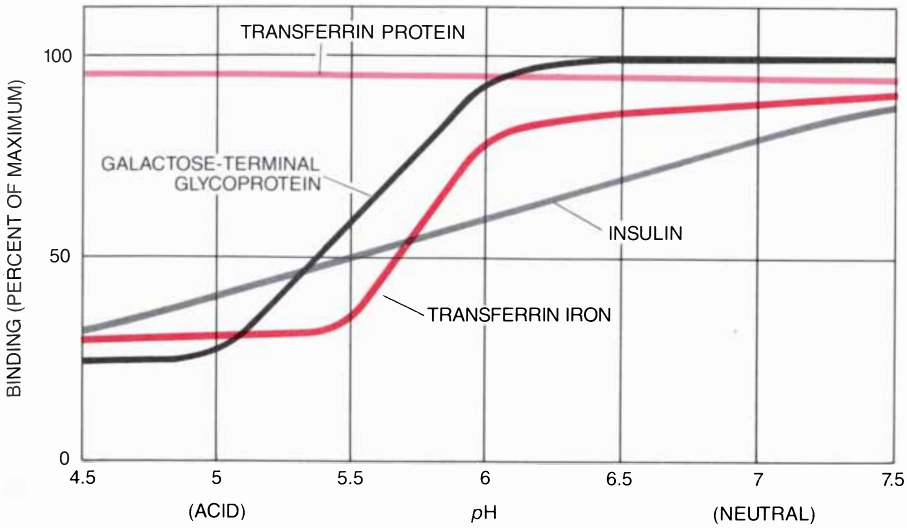


ACIDITY of endocytic vesicles was demonstrated by Benjamin Tycko and Frederick R. Maxfield of the New York University Medical Center. They labeled a ligand, the protein alpha-2-macroglobulin, with fluorescein, a *p*H-sensitive fluorescent dye. Cultured fibroblasts (connective-tissue cells) were incubated with the ligand just long enough for the ligand to reach endosomes or CURL's. Then the cells were illuminated with 450-nanometer violet light or with 490-nanometer blue light and were observed with an image-intensifier video camera mounted on a fluorescence microscope. The ratio of the intensities of fluorescence under illumination at the two wavelengths provides a measure of the *p*H within the vesicles harboring the labeled ligand. The 450-nanometer fluorescence (*left*) is somewhat brighter than the 490-nanometer fluorescence (*right*), indicating that the *p*H is acidic; calibration against intensity ratios measured in solutions of known *p*H shows that the *p*H in these vesicles is about 5.0.

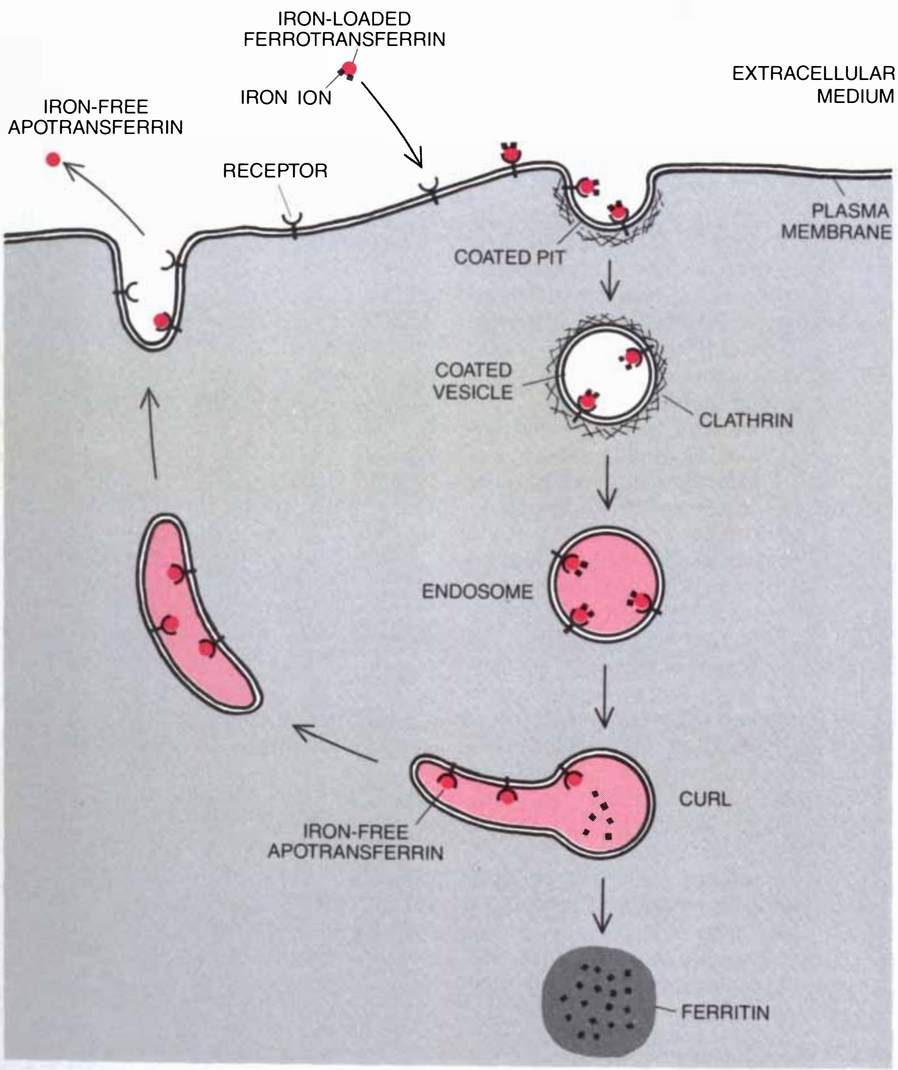
In our laboratory at the Massachusetts Institute of Technology we found that ferrotransferrin binds avidly to its receptor at the neutral *p*H of the extracellular environment. It is endocytosed and transferred to an acidic prelysosomal vesicle just as other ligands are. Observations some years ago had established that iron ions become dissociated from their transferrin carrier at an acidic *p*H (5.5 or less), and we found that to be the case even when the transferrin is bound to its receptor. Remarkably, the iron-free apotransferrin remains bound to its receptor even at *p*H 5, in striking contrast to such ligands as LDL, insulin and galactose-terminal glycoproteins. Indeed, the affinity of iron-free apotransferrin for its receptor at *p*H 5 is the same as the affinity of iron-loaded ferrotransferrin at a neutral *p*H 7. On the other hand, the apotransferrin has no measurable affinity for its receptor at a neutral *p*H; if an apotransferrin-receptor complex is suddenly transferred from an acidic to a neutral environment, the bound ligand dissociates within seconds.

The course of transferrin in endocytosis can now be traced. Ferrotransferrin, with its two iron ions, binds to the cell-surface receptor and the complex is internalized and transported to an endosome and thence to a CURL. When the complex reaches an acidic vesicle, the iron is released. It is transported (how is not known) to ferritin, an iron-storage protein in the cytoplasmic fluid. The transferrin receptor is recycled to the cell surface, just as other receptors are after they have been freed of their ligand. The transferrin receptor is not freed of its ligand, however; it carries the iron-free apotransferrin with it to the outside of the cell, where the neutral *p*H of the extracellular medium causes the apotransferrin to dissociate from the receptor. The iron-free apotransferrin enters the bloodstream, ready to load more iron; the receptor takes its place in the plasma membrane, ready to bind another ferrotransferrin molecule—for which, as we mentioned above, it has high affinity in the neutral extracellular environment.

All of this happens very quickly. We can show that each of the liver-tumor cells we work with has some 150,000 transferrin-binding sites on its surface or at some stage of recycling within the cell. A cell takes up transferrin for several hours at a rate of 19,000 iron atoms, or 9,500 transferrin molecules, per minute. Dividing 150,000 by 9,500 gives the cycling time: some 16 minutes from the binding of ferrotransferrin to the secretion of apotransferrin. We have measured the time required for each step of the cycle. On the average it takes four



EFFECT OF ACIDIC ENVIRONMENT on the binding of various ligands differs. Whereas most ligands dissociate from their receptor as the *p*H is lowered below 6, the binding of transferrin to its receptor remains unaffected. When the binding of iron carried by ferrotransferrin is measured separately from that of the transferrin, the iron is seen to dissociate from the transferrin-receptor complex below *p*H 6 while transferrin protein remains bound to the receptor.



TRANSFERRIN CYCLE is traced. After endocytosis of the transferrin-receptor complex, iron is released from the transferrin in the acidic environment of the CURL and is transferred to the iron-storing protein ferritin. The iron-free apotransferrin remains bound to its receptor and is recycled with it to the cell surface. When the receptor-apotransferrin complex encounters the neutral *p*H of the extracellular medium, the iron-free apotransferrin is released.

minutes for a transferrin receptor on the surface to bind a ferrotransferrin ligand. The ferrotransferrin-receptor complex is internalized in about five minutes. It takes another seven minutes for the iron to dissociate from the transferrin and the apotransferrin-receptor complex to return to the surface, after which the apotransferrin is released from the receptor in only 16 seconds. The total elapsed time is about 16 minutes.

The entire process of endocytosis involves a succession of rapid rearrangements and fusions of biological membranes. Certain membrane proteins are selectively trapped in a coated pit, whereas others are specifically excluded. Every time a receptor-ligand complex is internalized a piece of the plasma membrane "buds off" from the surface to form a coated vesicle. The newly generated vesicles fuse, giving rise to endosomes and CURL's. The CURL membrane is dissociated, some of it apparently fusing with the tubules that return receptors to the surface and some of it fusing with a lysosome to deliver ligands for degradation. In turn the lysosome may bud off parts of its own membrane as small vesicles that return to the cell surface and disgorge digestion products outside the cell.

Cells may internalize a large part of their plasma membrane (as much as 50 percent of it per hour), not only in receptor-mediated endocytosis but also in pinocytosis or phagocytosis. It would be wasteful for the cell to keep resynthesizing membrane components. Presumably the bits of membrane or their components are instead restored to the surface or to the organelle (such as the lysosome) to which they belong. The phospholipid composition of the plasma membrane and other membranes is somewhat different, and all biological membranes have very different complements of integral membrane proteins, of which the receptors we have been discussing constitute only one class. Somehow the plasma membrane and every other biological membrane must maintain their individuality in the face of continual invagination, fusion and recycling events.

There must be mechanisms for the continual sorting of membrane components, but nothing is yet known about such mechanisms. What accomplishes the unequal partitioning of membrane proteins so that, for example, the plasma membrane gets the permeases it should have and the lysosome gets the right proton-pumping enzyme? How are vesicles targeted from one organelle to another? Are there specific sorting signals or receptorlike molecules on the outer surface of each vesicle? More detailed understanding of receptor-mediated endocytosis will require answers to this kind of question.

INCIDENTALLY, THERE ARE ALSO A FEW PRACTICAL REASONS TO OWN ONE.

A lot of people buy cars because they like the way they look.

We're not against that.

But we do think it would be a shame if people bought the new Cougar only for its beautiful lines.

Because an extraordinary amount of advanced design and engineering has gone into Cougar's styling.

Its aerodynamic shape, for example, is as practical as it is beautiful, able not only to slip cleanly through onrushing air, but to use that air to help the car hug the road.

Get it together—buckle up.

Its EEC-IV electronic engine control system is the most advanced in the world, and is capable of controlling six different functions at once to keep the engine operating at peak levels of both power and fuel efficiency.*

Even the Cougar's shock absorbers are specially designed with pressurized nitrogen to respond to jarring road obstructions consistently better than conventional shocks.

And while most cars in its class have not undergone significant styling changes in several years, this Cougar was newly and aerodynamically shaped for 1983.

Obviously, we're pleased if you are attracted to Cougar's looks.

But after a careful consideration of its many practical features, we hope you won't let its looks blind you to its real beauty.

Call us toll-free at 1-800-MERCFAX for a copy of our 1984 Cougar catalog.

*27 EST. HWY. [20] EPA EST. MPG. Compare to the estimated MPG of other cars. Your mileage may differ depending on speed, weather and trip length. Actual highway mileage less.

MERCURY. A MORE ENLIGHTENED APPROACH.

LINCOLN-MERCURY DIVISION Ford

MERCURY COUGAR

SCIENCE AND THE CITIZEN

Man or Mouse?

Ever since William Harvey published his treatise *On the Movement of the Heart and Blood in Animals* in 1628, thus inaugurating the modern era of experimental physiology, experiments on laboratory animals have been the basis of most advances in medical knowledge. Today millions of rats and mice and smaller numbers of rabbits, dogs, cats and other animals are sacrificed every year in the U.S. for basic studies in biology; for the development of vaccines, new surgical procedures and drug therapies; for nutritional studies; for testing potential carcinogens and for establishing the efficacy and safety of new pharmaceutical agents. All such activities are now threatened by a growing new movement to sharply restrict or even to prohibit the use of animals in research.

Organizations opposed to animal experimentation are currently proliferating. Legislation limiting the availability of animals for research has been passed in many states and localities. Bills some investigators think would inhibit research are pending in Congress. Fringe groups in the animal-protection movement have broken into laboratories to "liberate" animals that were the subjects of research, in one instance dogs carrying experimental heart pacemakers and in another instance rats serving in a study of Alzheimer's disease.

The most significant trend is not, however, the increase in actions aimed at animal experimentation but a change in the movement's conceptual foundation. Instead of concern for animal "welfare" there is now an assertion of animal "rights." All sentient beings, it is held, have an inalienable legal right to life and protection from harm; it is not for human beings to decide that certain other species may properly be subjected to experimentation, no matter what the benefits may be for human beings; to condone animal experimentation is "speciesism," which is as reprehensible as racism and sexism.

A number of the proliferating activist groups embrace the animal-rights concept and would abolish all animal experimentation. Most of the old-line humane societies committed to animal welfare would allow some experimentation but would restrict it and supervise it more closely. The "abolitionists," however, have been putting increasing pressure on the "restrictionists," and they have received impressive public support. In a survey conducted by *The Boston Globe* only a third of the respondents answered in the affirmative when they were asked: "Do you believe scientists should be allowed to experiment with live animals?" In a survey done by the magazine *Glamour* 59 percent of the respondents said they would be willing to use a drug that had not been tested in animals even though it might not be safe.

Even the most restrained animal-welfare organizations maintain that many experiments done with animals are unnecessary or repetitive and that animals are often abused. Some groups that would allow "biomedical" research are strongly opposed to "behavioral" research, which they consider to be capricious and unrelated to human needs. They argue that alternative methods are available and cite in particular the development of mathematical models and of tests that can be done with bacteria or with animal and human cells cultured in laboratory glassware; they allege that some investigators prefer animal studies because they either are too lazy to learn alternative methods or are fundamentally sadistic.

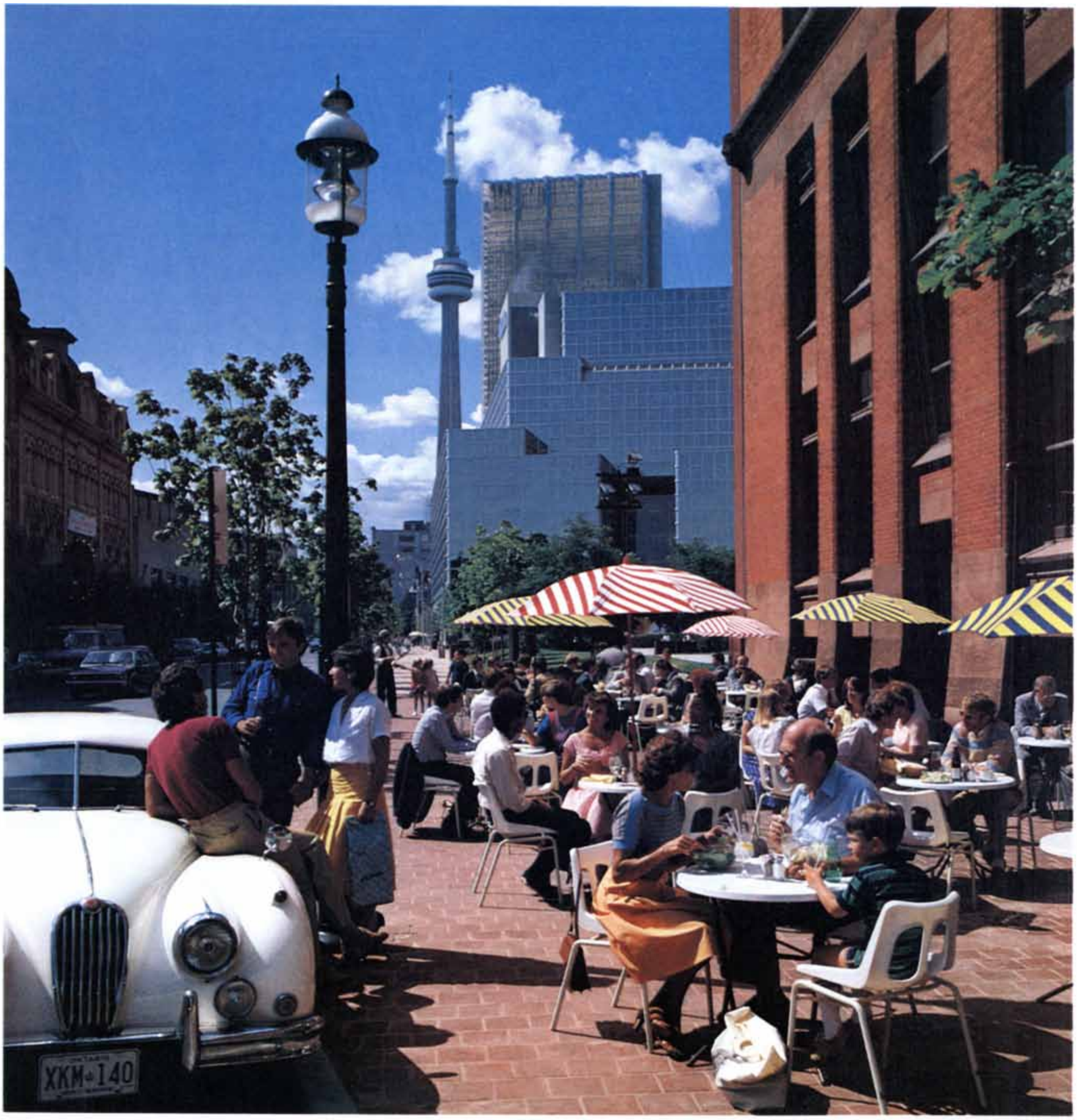
Investigators have been slow to respond to the challenge, much as biologists and students of evolution have tended to ignore the challenges of creationists. Biomedical workers have felt the necessity and value of animal experimentation is so clearly manifest as to need no defense. The National Society for Medical Research, based in Washington, was founded in 1946 to "assure the continued opportunity for progress in the study of human and animal diseases." Its activities have largely been confined to lobbying against restrictive legislation and to promoting improved standards for animal experimentation and the development of "adjunct" methods to replace some animal studies. Now the Foundation for Biomedical Research, with headquarters in Waltham, Mass., has been organized to undertake a broad program of public education on the value and necessity of animal research.

Investigators have cited a long list of advances in basic biological knowledge and in medical treatment that have been made possible only through animal experimentation. They have maintained that alternative in vitro tests are exploited as much as possible, if only because they are quicker, more precise and much less expensive, but that no such test can replace whole-animal studies in every case. They have conceded that some animals are unnecessarily abused but insist that responsible investigators and their institutions are eager to prevent such abuse and that there are adequate regulations and mechanisms to prevent it, given enough financial support.

Standards for the housing, care and treatment of all laboratory animals other than rodents are set by the Animal Welfare Act of 1966. Laboratories are subject to inspection by the Animal and Plant Health Inspection Service (APHIS) of the Department of Agriculture, but inspectors may not interfere with the actual conduct of research. Since 1963 the National Institutes of Health have established more specific guidelines for the care and use of laboratory animals in institutions receiving N.I.H. support. Such an institution must give assurance of compliance with the guidelines. It can show it is certified by the American Association for the Accreditation of Laboratory Animal Care or submit a statement from its own animal-care committee either certifying full compliance or listing deficiencies and showing progress toward remedying them; N.I.H. inspectors conduct spot checks for compliance.

Estimates of the total number of animals used for research and testing in the U.S. vary widely. The most authoritative estimate, from a 1978 study done by the Institute of Laboratory Animal Resources of the National Academy of Sciences, is some 20 million animals, a 40 percent decline from the 1968 total. More than 95 percent of the animals were rats, mice and other rodents. Dogs and cats may have accounted for 1 or 2 percent of the total and nonhuman primates for less than 1 percent. It has been estimated that about 40 percent of the animals are used in basic research, 8 percent for either teaching or experimental surgery, 26 percent for demonstrating the beneficial effects of particular products, 20 percent to test products for deleterious effects and 6 percent for other purposes.

Laboratory rats and mice are bred for research by licensed suppliers, as are most laboratory primates and some cats and dogs. Other cats and dogs come from public pounds and shelters. About 15 million dogs are destroyed every year in pounds, according to one estimate by a humane society; fewer than 400,000 dogs are used for research. "Pound release" has nonetheless become the major target of many animal-welfare groups that prefer the killing (by humane methods) of unclaimed strays to their release for research. As a result of such opposition New York State prohibited the sale of pound animals to research institutions in 1980. Similar legislation was passed in New Hampshire and Massachusetts last year and in Maine this year; in Massachusetts even the importation of animals from out-of-state pounds will be prohibited in 1986. California's legislature is considering two pound-release laws, each



Toronto

Vintage Vantage

And well worth a closer look. Gleaming New World shapes vie side by side with the hallmarks of a bygone era. Come catch a glimpse of Toronto! Lend an ear to our concerts and cabarets or savour the flavours, the creature comforts, the urban delights and the cosmopolitan pleasures. It's a new way of looking at the world – and yours to discover! For more information, write: Ontario Travel, Queen's Park, Toronto M7A 2E5 Ontario/Canada, or call TOLL FREE 1-800-828-8585. From New York State, call 1-800-462-8404.

ONTARIO yours to discover!®

How to move with and take your

Introducing the IBM *Portable* Personal Computer.

It's a complete PC. In a case. With a handle. And a welcome addition to the family.

The IBM *Portable* Personal Computer is the first IBM PC system you can pick up and take with you. Across town or across the hall. Or put away easily for another day.

It's a powerful system, with 256KB of user memory (expandable to 512KB) and a slimline double-sided 5¼" diskette drive (and room for another). Plus a built-in

modern times PC with you.

9″ monitor with easy-to-read amber characters. Text and graphics capability. And an 83-key keyboard.

All fitted into a sturdy, transportable case that's easy to handle. And park.

The IBM® of portable personal computers. Make no mistake about it, this is a true IBM PC.

Which means it is part of the same dependable family as the IBM Personal Computer, the IBM PC/XT and the new IBM PC*jr.* And that means you can use many IBM Personal Computer Software programs to help you reach your goals.

All this and five expansion slots, ready to accept expanded memory, printers and other useful IBM Personal Computer options. Which should keep you rolling far into the future.

Pick one up at a store near you. You can see the new IBM *Portable* Personal Computer at any authorized IBM PC dealer or IBM Product Center.

To find the store nearest you, call 1-800-447-4700. In Alaska or Hawaii, call 1-800-447-0890.

IBM®

of which has already been passed in one house. One bill, supported by many investigators, would tighten procedures to ensure that no animals identified as pets are made available for research. The other one would not only prohibit pound release but also make it illegal to subject dogs or cats to "pain which causes suffering."

In both houses of Congress bills to renew the basic authorization of N.I.H. programs call for an 18-month study of animal experimentation in the U.S. to determine the current level of such experimentation and the extent to which alternative methods might be substituted. Biomedical-research organizations favor this provision. The House bill would go further. It would make adherence to the N.I.H. guidelines a legal requirement rather than an N.I.H. administrative requirement, and it would specify the composition of institutional animal-care committees, which would have to include a member from outside the institution.

A broader bill, introduced by Senator Robert J. Dole (R., Kans.), would amend the basic Animal Welfare Act. It would make the N.I.H. guidelines into firm "standards" and would apply them to all research institutions rather than just those receiving N.I.H. funding. It would also expose research protocols and procedures, which are currently exempt from APHIS surveillance, to review for compliance with animal-welfare standards.

Cosmic Catalysis

In the first few moments after the beginning of the big bang only gravity, among the four fundamental forces now recognized in nature, had become distinct. The other three forces, or interactions, namely the electromagnetic force and the weak and strong nuclear interactions, were bound together into a unified force that brought about a continual interchange of identity among elementary particles. One of the most remarkable effects of the unified force is that it can lead to the decay of the proton, a particle whose stability, until only a few years ago, had been assumed to be absolute. For all practical purposes the proton must still be considered a highly stable particle: according to the "grand unified" theories that describe the unified force, proton decay does not readily proceed unless the temperature exceeds 10^{30} degrees Kelvin. The reaction rate is so low under the conditions generally prevailing in the universe that physicists estimate the life of the proton in a vacuum is at least 6.5×10^{31} years. It has now been shown, however, that in the presence of certain exotic particles the decay of the proton can be catalyzed. The seed of this destruction is the particle called the magnetic monopole.

The existence of the monopole is predicted by most versions of grand unified theory. Such a particle has no electric charge, but it does carry a single magnetic charge. Theorists differ over its internal composition and over the population of monopoles in the universe at large; it is generally agreed, however, that the monopole has an enormous mass for an elementary particle, perhaps 10^{16} to 10^{17} times the mass of the proton. A particle that heavy could not be created anywhere in the universe today. Only in the first 10^{-35} second after the beginning of the big bang was the energy density in the universe high enough to enable heavy magnetic monopoles to materialize freely. If monopoles still exist, they must be the stable byproducts of that early epoch.

The idea that monopoles catalyze proton decay was not at all obvious in the initial formulations of grand unified theory. A general theoretical condition for the decay was set forth by Stephen L. Adler of the Institute for Advanced Study in Princeton, N.J., John S. Bell of the European Laboratory for Particle Physics (CERN) in Geneva and Roman W. Jackiw of the Massachusetts Institute of Technology. In 1981 Valery Rubikov of the Institute of Nuclear Research in Moscow suggested that monopoles could meet this condition; the decay mechanism was also discussed independently a year later by Frank Wilczek, now at the University of California at Santa Barbara. The catalytic properties of the monopole turn out to be a direct consequence of its stable internal structure.

In the simplest version of grand unified theory the monopole is a soliton, a multidimensional kink in the fields of unified force that were present everywhere in the universe soon after its birth. As the universe expanded and began to cool, the unified-force fields disappeared everywhere except along the monopolar kinks. In quantum theory every field of force is generated by the exchange of vector bosons: the particles that mediate the force. Accordingly the monopole can be thought of as a composite made up of roughly 100 vector bosons associated with the unified force. In effect the inside of a monopole is a hot coal that bears the fire of the early universe; if such a knot of vector bosons were to pass near a proton, the proton would be subjected to the unified force and would decay almost instantly.

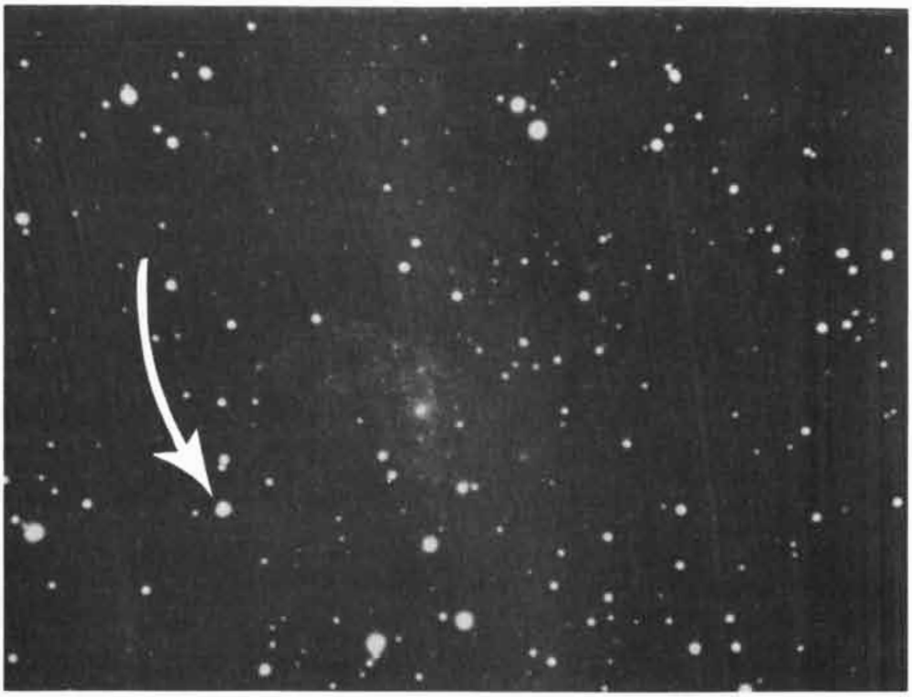
When the possibility of monopole catalysis was first recognized, there seemed little reason to regard it as an effective mechanism for proton decay. Grand unified theory also predicts that in spite of its large mass the monopole must be incredibly small: some 10^{14} times smaller than the proton. According to Curtis Callan of Princeton University, if the effects of the monopole could be estimated solely from its size, its influence on proton decay would be entirely negligible.

Callan's calculations show, however, that the magnetic field of the monopole can interact strongly with the angular momentum of the proton. In quantum mechanics the angular momentum of a particle is analyzed as a wave. One of the simple components of that wave, namely the one that corresponds to zero angular momentum, is sucked into the core of a monopole whenever the proton passes close enough to it to be affected by its magnetic field. Inside the monopole the wave component interacts directly with the vector bosons of the unified force; the proton is converted into, say, a positron and a neutral pi meson, and a prodigious amount of energy is released. The monopole remains unaffected by the interaction, and so it can continue to catalyze the decay of more protons.

The interaction of the monopole with the proton is still only a theoretical possibility; indeed, neither the monopole nor the decay of the proton has been seen experimentally. Nevertheless, the theoretical development opens grand unified theory to new and exotic observational tests. Experimenters have attempted to establish progressively larger upper bounds for the lifetime of the proton and progressively smaller upper bounds for the density of monopoles in the universe. Since the intrinsic rate of proton decay increases enormously in the presence of monopoles, any upper bound on the rate of proton decay can also be understood as an upper bound on the density of monopoles.

Michael S. Turner of the University of Chicago points out that the most stringent upper bounds on the monopole density can be obtained from the measurement of radiation emitted by astrophysical objects. According to Turner, free monopoles moving through space could be stopped and captured by a dense object such as a star. The weight of the monopole would cause it to sink immediately to the center of the star; there, as Turner puts it, each monopole would begin "gobbling up nucleons," or in other words causing protons and neutrons to decay, at a rate that depends on the density of the matter at the center of the star. For each nucleon decay roughly a billion electron volts of energy would be released, which would eventually be emitted by the star as thermal radiation. The amount of radiation depends on the density of the star and on the number of monopoles that had collected there in the star's lifetime. The number of collected monopoles depends in turn on the age of the star and on its cross-sectional area, as well as on the density of monopoles in space.

The lowest upper bound on the monopole density can therefore be obtained



by measuring the X-ray emission of a neutron star, the densest object known. The estimate is conservative in the sense that all the thermal radiation emitted by the star is assumed to be caused by nucleon decays that are catalyzed by monopoles. Observations of the pulsating neutron star designated 1929 + 10, made by the Einstein orbiting X-ray telescope, set an upper limit on the monopole flux of 10^{-21} monopole per square centimeter per year, the equivalent of one monopole per year passing through an area the size of Chicago. The estimate can be reduced by a factor of a million if it is assumed that the neutron star began its life as an ordinary star; in that way its emissions could arise from monopoles collected over a time much longer than its life as a neutron star. On the other hand, if the flux is as high as one "Chicago" per year, the cosmic rain of monopoles will in 10^{18} years cause the earth to evaporate.

Multifactorial Mortality

The death of a human being is generally thought of as having a single cause such as cancer, heart disease or accident, and the statistics on mortality tabulated and published by the National Center for Health Statistics (NCHS) have long reflected this assumption. For the most part NCHS reports have attributed each death recorded in the U.S. to one medical cause. The total number of deaths resulting from a particular disease condition is taken to be a measure of the threat to public health presented by that condition. Data recently published by the center show, however, that most lives are ended not by a single condition but by a combination of several conditions. The data suggest that an accurate picture of mortality in a large population can be had only by examining the multiple causes of death in relation to one another.

The new data, which are published in the NCHS *Monthly Vital Statistics Report,* are a count of all the causes of death listed on death certificates in the U.S. in 1978. That year was chosen because it was the last one in which the World Health Organization disease-classification scheme that had been introduced in 1968 was employed. The scheme is revised every 10 years, so that the data for 1978 are compatible with those for the preceding decade.

The multiple-cause data collected from the 50 states of the U.S. show that in 1978 almost 75 percent of the death certificates filed listed more than one medical condition. More than 30 percent of the certificates had two conditions, 25 percent had three and 15 percent had four or more. The conditions recorded on the death certificate, however, are not necessarily unrelated. Indeed, the form of the standard death certificate guarantees that in many instances the illnesses will be closely related.

The standard death certificate, which is completed by the attending physician or by a physician who happens to be present when the person dies, has two sections where disease conditions can be entered. Part I asks the physician to record a causal sequence of three medical conditions that led directly to the death of the patient. There is space for an underlying cause of death, an intervening cause and an immediate cause. The underlying cause from Part I of the death certificate is the cause given in most mortality statistics, and most conclusions about the prevalence of disease in the U.S. are based on it. Part II asks for illnesses the patient had at the time of death that may have contributed to death but did not play a role in the direct causal sequence.

If Part I is completed, the illnesses given as causes of death will be connected as components of a causal chain. According to Harry Rosenberg, chief of the mortality-statistics branch of the NCHS, typical entries in Part I might include as the underlying cause of death chronic ischemic heart disease, as the intervening cause acute myocardial infarction and as the immediate cause rupture of the myocardium (the muscular tissue of the heart). Given in Part II as an additional cause might be pneumonia or influenza.

The recent report includes all the conditions included in Part I or Part II of the standard death certificate. As a result it is now known how many times a particular illness was implicated in a death, whether or not it was the underlying cause. In addition workers at the NCHS have computed the ratio of the number of times an illness was entered to the number of times it was entered as the underlying cause of death. The higher the ratio is, the less likely the illness is to be considered the fundamental reason for the cessation of life.

Among common illnesses there is considerable variation in the ratio. The ratio for malignant neoplasms is 1.1 and for hypertensive heart and renal disease 1.3. Such a low ratio implies that if the condition is entered on the death certificate, it is almost certain to be entered as the underlying cause. For anemia, on the other hand, the ratio is 8.2, for arteriosclerosis 6.5 and for pneumonia 3.3. Such conditions are reported as the underlying cause in only a fraction of the cases where they are recorded on the death certificate.

Why should the variation in the ratio be so great? According to Richard A. Kaslow, chief of the epidemiology and biometry section of the National Institute of Allergy and Infectious Diseases, the specificity of the disease is a significant factor. The more specific the disease is and the more careful the diagnosis is, the more probable it is that the illness will be considered the underlying cause of death. For example, leukemia, which has a ratio of 1.3, is a small group of very specific disorders and not merely an array of symptoms. Furthermore, the diagnosis of leukemia is made with considerable care on the basis of precise diagnostic tests.

At least two other factors can contribute to lowering the ratio. If a disease has dramatic and devastating consequences, as leukemia does, it will tend to be entered as the underlying cause of death. According to Kaslow, there is also a "self-fulfilling quality" in the process of determining and entering the underlying cause of death. If a patient has a condition, such as ischemic heart disease, that is known to have been fatal in many previous cases, the disease will often be entered as the underlying cause independent of its position in the causal sequence.

Disorders that are less specific than leukemia, such as anemia, generally have higher ratios. Anemia, with its ratio of 8.2, can be manifested in many diseases, including leukemia. In most instances more effort goes into finding the underlying cause of the anemia than goes into investigating the condition itself. In addition anemia by itself is not often fatal because the patient can be kept alive with blood transfusions.

The ratio of reported-to-underlying causes is beginning to play a role in the investigation of the prevalence of certain diseases. One such disease is septicemia, or "blood poisoning," which results from the presence in the circulatory system of toxic microorganisms. Of the frequently reported causes of death in the U.S. septicemia is the most rapidly increasing; in 1978 it was entered on some 51,000 death certificates. Even so, it has a ratio of 6.6, which suggests that it can result from several underlying illnesses.

Kaslow and his colleagues at the Institute of Allergy and Infectious Diseases have begun work aimed at finding out why septicemia is being reported more often as a cause of death. Among the possible explanations are the utilization of therapeutic agents for cancer that suppress the immune response, the increased administration of antibiotics (which leads to the development of resistant strains of bacteria) and the increased utilization of invasive diagnostic and therapeutic methods in an aging population the members of which are quite susceptible to infection. Even when these explanations are combined, however, they cannot account for all the deaths where septicemia is a contributing cause. Finding the additional factors will require correlating all the reports of septicemia on death certificates with the underlying cause reported on the certificate and examining the

medical histories of selected victims of septicemia to discover where and how the condition was initiated.

The data on multiple causes of death could be a valuable starting point for studies of particular syndromes such as septicemia. More generally the information could help to give an accurate picture of the events leading to death in the U.S. population. Rosenberg states: "The medical community feels that the underlying cause of death, which has been the traditional index, might not be the best indicator of how widespread a disease is in the population. For example, diabetes, which has a ratio of 3.9, is more prevalent as a condition associated with death than it would appear if only the underlying cause of death were considered. This is particularly true in a country [such as] the U.S., where the population is aging and hence more people are dying with more than one medical condition. When this is so, the concept of the underlying cause becomes less valid and must be augmented with information on multiple causes of death."

Spot Check

A few things are known about sunspots. They are relatively cool areas on the surface of the sun (about 3,900 degrees Kelvin, compared with an average surface temperature of about 5,600 degrees), which explains why they are dark. The intense magnetic fields invariably associated with sunspots probably cause the coolness by inhibiting the convective transfer of heat from the interior of the sun to its surface. The number of spots visible on the solar disk varies according to an 11-year cycle, during which period the area of spot generation moves from latitudes of 35 degrees toward the equator. Why the spots follow this cycle, however, and what accounts for their strong magnetism are questions for which satisfactory answers have yet to be found in spite of decades of efforts by theorists.

Robert Howard of the Mount Wilson and Las Campanas Observatories and Peter A. Gilman of the High Altitude Observatory in Boulder, Colo., now report a tantalizing new observation that will have to be put through the theoretical mills. From white-light photographs of the sun made at Mount Wilson, Howard and Gilman measured the displacement of individual sunspots from one day to the next and found that the sun's rate of rotation is not constant: it increases near the maximum and the minimum of the solar cycle, particularly the minimum. Over the 62-year period from 1921 through 1982 the increase in rotational velocity was sometimes as much as 3 to 4 percent.

Earlier studies had suggested that there might be a correlation between rotation and the sunspot cycle, but those studies had all estimated the displacement not of individual sunspots but of sunspot groups. According to Howard and Gilman, that method is flawed because the growth and decay of spots within a group may change the overall average position of the group in ways having nothing to do with rotation. Howard and Gilman benefited from the heroic efforts of their research assistant, Pamela I. Gilman (not related to Peter Gilman), who spent more than two years measuring the position of each sunspot on 23,000 photographs with the help of a hand-held cursor that enabled her to record the position in a digital, computer-readable form. A computer programmed with a special pattern-recognition algorithm then picked out spots that recurred on successive days, and the displacement of those spots was used to calculate the sun's rotation.

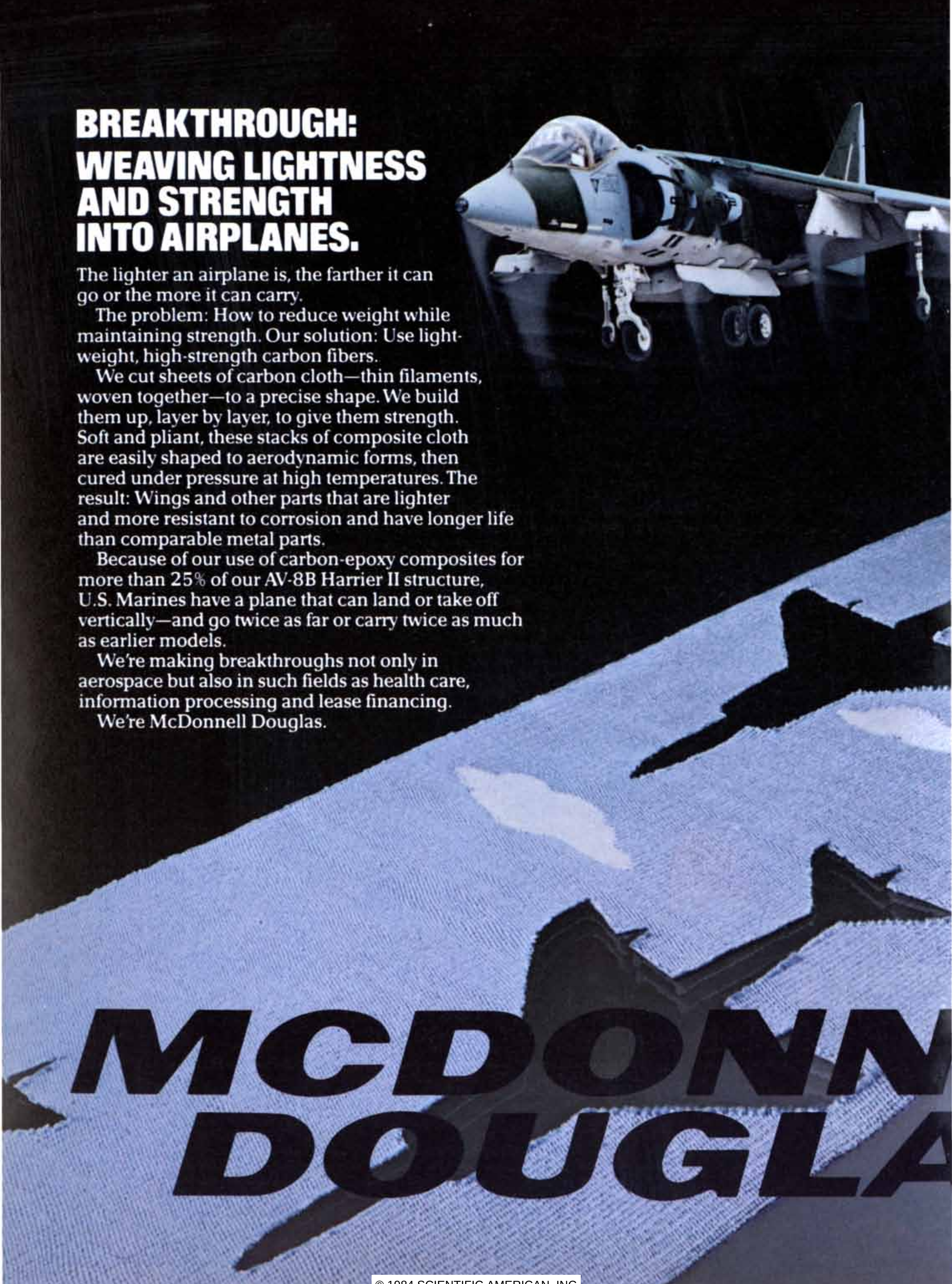
The investigators compared their calculations with results obtained from an independent method of measuring solar rotation: Doppler spectroscopy. As the sun turns from east to west, spectral lines emitted by gas on its eastern limb, which is moving toward the earth, are shifted toward the blue end of the spectrum; light coming from the western limb, which is moving away, is shifted toward the red. The amount of the Doppler shift is proportional to the speed of rotation. Doppler measurements made at Mount Wilson since 1967 follow a pattern similar to that of the sunspot data, with major speed peaks in 1971 (just after a solar-cycle maximum), in 1974–76 (just before a minimum) and in 1979 (near another maximum).

Comparison of the two sets of data also lent support to a widely held hypothesis: the average rotation speed of sunspots was found to be higher by about 3 percent than that of the gaseous surface of the sun (as measured by Doppler spectroscopy). According to current theory, the magnetic fields that give rise to sunspots are not confined to the surface; they are "anchored" at a deeper level in a layer of denser, heavier fluid. That layer moves faster than the solar surface, and as it does it drags along with it the sunspots' magnetic fields, which are so strong that they can move through the gas like sticks through water. In this view the sunspot-speed peaks near the minimum and the maximum of the solar cycle reflect an acceleration of the anchoring layer.

On the other hand, the peaks in the Doppler data reflect accelerations at the surface. Since the overall angular momentum of the sun cannot change, and since no area on the surface is seen to slow down at the same time, a deceleration must be taking place somewhere inside the sun, possibly at a deeper level than that at which the sunspots are anchored. In effect, angular momentum must be transferred from the sun's interior to its surface. Howard and Gilman offer a "highly speculative" hypothesis for how that might happen. Angular momentum might propagate upward from deep within the sun as slow-moving waves, accelerating first the sunspot layer and later the surface, where the waves would be reflected back toward the interior. Such reflection would explain why changes in spot rotation seemed to lag behind Doppler-measured surface changes from 1973 through 1982: that might have been a period when angular-momentum waves were being reflected from the surface. The hypothesis will gain credibility if at some time in the future spot-rotation peaks are found to precede the Doppler peaks, corresponding to a period in which the waves are propagating upward through the sun's interior.

The correlation between the sunspot cycle and changes in the sun's rotation suggests that the two phenomena have the same underlying cause, but theorists are a long way from understanding that connection. They have had some success so far in explaining the solar cycle as a consequence of the sun's differential rotation: the fact that the sun spins faster at the equator, where the period of rotation is about 25 days, than it does near the poles, where the period is about 34 days. According to these models, the north-south lines of the sun's magnetic field are embedded in the convective zone, the layer below the surface that extends about a fourth of the way to the center and in which hot gas rises toward the surface. Because the sun is rotating faster at the equator, the field lines are stretched in equatorial regions and concentrated in tubes of magnetic flux. Sunspots form where flux tubes rise to the surface.

The problem with these solar dynamo models, according to Gilman, is that they make no attempt to explain how differential rotation arises in the first place and in fact contradict theories that do. Gilman's own model portrays the sun's fast equator as a consequence of a global convective circulation, but it does a poor job of reproducing the solar cycle. The conflict between theories that explain the sun's magnetic phenomena and those that focus on its dynamics has persisted for nearly a decade, Gil-

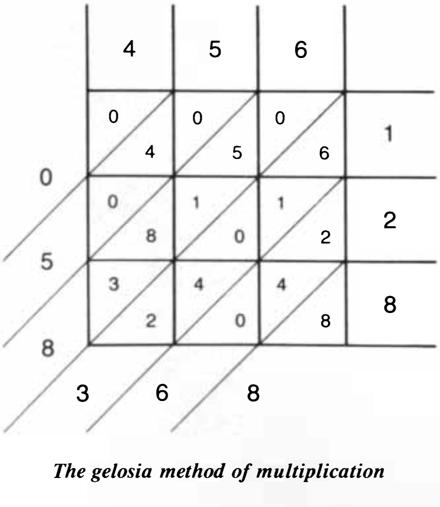

BREAKTHROUGH:
WEAVING LIGHTNESS AND STRENGTH INTO AIRPLANES.
The lighter an airplane is, the farther it can go or the more it can carry.
The problem: How to reduce weight while maintaining strength. Our solution: Use light-weight, high-strength carbon fibers.
We cut sheets of carbon cloth—thin filaments, woven together—to a precise shape. We build them up, layer by layer, to give them strength. Soft and pliant, these stacks of composite cloth are easily shaped to aerodynamic forms, then cured under pressure at high temperatures. The result: Wings and other parts that are lighter and more resistant to corrosion and have longer life than comparable metal parts.
Because of our use of carbon-epoxy composites for more than 25% of our AV-8B Harrier II structure, U.S. Marines have a plane that can land or take off vertically—and go twice as far or carry twice as much as earlier models.
We're making breakthroughs not only in aerospace but also in such fields as health care, information processing and lease financing.
We're McDonnell Douglas.

man says, and a unifying resolution remains elusive.

Dem Bones

John Napier, Baron of Merchiston, Protestant polemicist and inventor of martial devices including burning mirrors, artillery and an armored, weapon-carrying chariot, devoted much of his leisure to easing "the difficulty and prolixity of calculation." In 1614 he published the invention for which he is best remembered: a table of logarithms, for aid in trigonometric calculations. Then in 1617, the year of his death, he described what would turn out to be an ancestor of the modern computer: a machine intended to ease the labor "of those who prefer to work with the natural numbers as they stand." The machine is the Rabdologia, a set of rods inscribed with numbers, for aid in performing multiplications. The rods are remembered today as "Napier's bones." (The better ones were actually made out of bone.) The Rabdologia was to give rise to a remarkable sequence of calculating machines extending through the 17th century. The sequence is traced by M. R. Williams of the University of Calgary, who writes in *Annals of the History of Computing*. Williams himself has enlarged the history of the sequence by finding a manuscript describing in detail a 17th-century machine whose design was thought to have been lost.

Williams begins by describing the inspiration for Napier's bones. It was surely an ancient method of multiplication that became known as the gelosia method (because of its resemblance to an Italian window grating) when it found its way from India into Italy in the 14th century. The method requires that the digits of one of the numbers to be multiplied be placed across the top of a grid of boxes and the digits of the other number be placed along the right-hand edge of the grid (*see illustration below*). Then diagonal lines are drawn through the grid so that they bisect each box. The products of the digits are placed inside the boxes: the tens digit above the diagonal, the units digit below it. Each digit in the desired product is then the sum of all the digits between successive diagonals, with numbers carried if necessary. (In the illustration 456 times 128 yields 058368.)



The gelosia method of multiplication

Napier's bones, writes Williams, "are simply a collection of strips of all possible columns of [a] gelosia table.... To perform the multiplication of 456 by 128, one simply selects the strips headed 4, 5 and 6, places them side by side, and reads the partial products of 456 times 1, 456 times 2 and 456 times 8.... These partial products are then added to obtain the final result." Napier must have regretted the need for this final addition; thus he devised a Promptuary of Multiplication. It was a box of numerous copies of elaborations of gelosia strips and numerous copies of strips with holes strategically cut in them. The user selected the strips he needed, made a grid by arranging them on top of the box and added the digits he saw through the holes along each successive diagonal. That completed the multiplication.

Wilhelm Schickard, a German mathematician and astronomer, modified the scheme; the result was "the first really workable mechanical adding machine." The invention came in response to a request from a fellow astronomer, Johannes Kepler. In 1620 Kepler arrived in Tübingen to take part in the defense of his mother, who had been imprisoned as a witch. There he discussed Napier's work with Schickard. Schickard was "inspired to consider the design of a machine that would incorporate both a set of Napier's bones and a mechanism to add the partial products they produced in order to automate the process of finding the product of two numbers." By 1623 Schickard had succeeded. He wrote to Kepler: "You would burst out laughing if you were present to see how it carries by itself from one column of tens to the next."

Guided by Schickard's sketches and a knowledge of 17th-century clockmaking, Bruno von Freytag Loringhoff of the University of Tübingen has reconstructed the device. Its upper part is a set of Napier's bones inscribed on cylinders so that a bone is chosen by turning a dial. A set of slides exposes particular constellations of digits; thus a one-digit multiple is obtained. The bottom part of the device is an accumulator: a set of wheels by which the multiples are added. By means of intermediate wheels the turning of each accumulator wheel past the digit 9 causes a one-digit rotation in the wheel to its left. The device had a drawback: the force required to carry a digit through a sequence of places would undoubtedly ruin the wheels. Accordingly Schickard intended the operator of the device to have available a set of brass rings. Each ring, notes Williams, would signify that "a carry had been propagated off the end of the accumulator."

The next innovator was Samuel Morland, who served Oliver Cromwell but was actually "a spy for the exiled King Charles." In the mid-1660's he too mechanized Napier's bones. In Morland's scheme the digits on the bones were placed along the perimeter of circular disks. A number at the center of each disk showed which bone it was. To do a multiplication a set of disks would be mounted in his device. Next a key would be turned until a pointer indicated the desired single-digit multiplicand. The key would rotate the disks, so that certain combinations of numbers would appear in a sequence of windows. Adding the numbers would yield the digits of the product. Complex multiplications would again be done by adding partial products on an "accumulator."

In 1891 Henri Genaille, a French civil engineer, made a final modification of Napier's bones. Inspired by a problem set by a recreational mathematician, Édouard Lucas, he made a set of bones that incorporate the carrying of digits. Each bone provides a sector for each digit from 0 to 9, and each sector has along its left margin a network of arrows. When a set of bones are aligned for a particular multiplication, the arrows lead the calculator through the digits of the product. Genaille went on to devise a set of bones for division.

Williams' own discovery concerns René Grillet, clockmaker to Louis XIV. Grillet is known to have published, in 1678, an account of a machine based, he said, on the "rulers of Neper." The only thing the account explains is where to buy the device. In 1977 Williams came on a further explanation: a handwritten manuscript, 18 pages long, tucked into a book owned by a French mathematician, Michel Chasles (1793–1880). The upper part of Grillet's machine proves to be a set of accumulator dials much like those devised by Morland; the lower part of the machine proves to be a set of Napier's bones. There is, however, a problem. Chasles was gullible. Indeed, he once was sold a book including "manuscript letters from such famous people as Julius Caesar, Nostradamus and Cleopatra." The letters were written in French. The seller, who was paid 165,000 francs, was Édouard Lucas, the recreational mathematician. "The fact that Lucas was capable of producing one set of forged documents and that he obviously had an intimate knowledge of the operation of Napier's bones tends to cast some doubt on the authenticity of the Grillet manuscript." Still, "if Lucas was once again having some fun at Chasles's expense, he made a much better job of it than he did with the letter from Cleopatra."

Engineering.
The Buick Electra.

ADVANCED TECHNOLOGY CENTERS USA ADVERTISING SUPPLEMENT/SCIENTIFIC AMERICAN/MAY 1984

In 8.5 seconds, Skyhawk will change your mind about Buicks.

Let's face it. Some people think of a Buick as being, well, traditional. But an 8.5 second zero-to-sixty time at our proving grounds is hardly what you would call staid.

This is a Buick? Ah, yes. This *is* a new Buick Skyhawk T TYPE. It's smartly outfitted, as you can see. And full of surprises.

It sports an available multi-port fuel-injected, turbocharged 1.8 litre engine that makes it both responsive and practical.

It has front-wheel drive.

And high-rate suspension linked with quick steering produces the precise handling you just might not expect from Buick.

Visit a Buick dealer and buckle yourself into a Skyhawk. Then brace yourself for a change of mind. And a pleasant one at that.

Official Car of the XXIIIrd Olympiad
Los Angeles 1984

BUICK

Wouldn't you really rather have a Buick?

"If you do things the way you have always done them, you will get what you have always gotten before."

At least one American automaker is heeding this dictum of scientific inquiry, producing a new luxury car this year that exemplifies new ways of thinking about technology, process verification and marketing implementation.

The company is Buick, and the car is the Electra/Park Avenue—an entirely new vehicle, yet with a nameplate redolent of a 25-year luxury car tradition. The objective was to build a car with advanced, state-of-the-art technology that would not only fulfill expectations for loyal Electra sedan buyers, but go beyond to appeal to younger, sophisticated buyers who have looked to Audi, BMW and Mercedes as their standard of performance and luxury.

"From its inception, this car has been treated in more new and exciting ways than any car in Buick history," says David Sharpe, Buick Vehicle chief engineer. "Even the concept itself is new—a single car that can meet the expectation of two groups that have traditionally not seen eye to eye."

Sharpe explains that the older luxury car buyers want a substantially built car with room and comfort. The younger market looks to luxury too, but wants performance, handling and high technology, and all in a more space-efficient package.

Buick's new Electra and Park Avenue models serve both markets. The car is modern, with an aerodynamic wedge shape, front-wheel drive, a fuel-injected engine, four-wheel independent suspension and numerous other features normally found only in European luxury cars.

"We took 600 pounds out of the weight of the previous Electra," Sharpe says, "yet the interior room is within one inch as spacious as before. And with the relatively flat floor achieved with front-wheel drive, the car is actually roomier."

Then, to serve the divergent customers, Buick designers created three variations on the same theme: first, the Electra with a high standard of luxury; then the top-line Park Avenue with a traditional luxury look; and finally, the T Type, a performance-oriented model that speaks directly to the European car buyers. All provide luxury, yet with refinements in their accent.

According to Dr. Glen Urban, marketing professor at the Massachusetts Institute of Technology, "Luxury is the key word to describe both the product and the development process, for, from the outset, it was clear that Buick was giving its scientists and engineers the luxury of time to innovate, test new precepts and verify assumptions."

Dr. Urban's participation is characteristic of the new thinking that went into the Electra. He was asked to conduct a three-year research project by John Dabels, Buick general director of marketing and sales planning, who himself had studied at M.I.T. as a Sloan fellow under Urban. Urban had been conducting pioneering work on consumable-goods modeling. His paper on this subject, in fact, earned him the highest recognition from the *Journal of Marketing* and created a stir in the field.

"No one had ever established a viable model for the marketing of durable goods," Dabels explains. "We felt that the Electra was such an exciting new product—with distinctive possibilities—that this would be the perfect opportunity to merge the science of marketing with our advancements in technology."

Urban's study included a wide range of methodology, from traditional focus groups to in-depth, two-hour interviews with a sample of nearly 600 prequalified individuals.

A multiple modular system, Urban's approach envelops the complexity of the buying decision in ways that makes possible a forecast of sales and, with the use of conjoint analysis, determination of the impact a new, competitive product introduction will have on a vehicle line's sales.

"Many of the study's findings raised some eyebrows at Buick," Dabels says. "Implicit in the work, for example, was the assumption that a luxury car like Electra will not compete so much with other cars as against a hierarchy of other uses for discretionary spending, such as vacations, tuition, home improvements and even video cameras.

"I believe Urban's work here is going to have a profound impact on the durable-goods industry, starting, of course, with the way it will change the marketing of the Electra."

Another luxury the Electra project inspired was a new, cross-disciplined way of looking at the product. An internal group was formed called EPATAF (Electra/Park Avenue Task Force). In a sense, it was an epitaph to the traditional ways of product management.

"We wanted a holistic approach," Dabels says. "Instead of having each discipline take a piece of Electra, we brought all the specialties together in informal brainstorming sessions, with engineers, researchers, service and quality specialists, and marketing and advertising people contributing."

Sharpe feels the benefit of EPATAF for his engineers was in broadening the perception of their role. "Engineers have a tendency to build to specifications as if customers actually came at a car with micrometers," he says. "We need to understand that meeting specifications alone will not necessarily satisfy customers. We need to experience what we're creating more through the user's perspective."

For example, Sharpe points out that every measurement on the seam of an instrument panel might be perfect, yet the way it looks from the driver's seat, or even the touch of the surface materials, can alienate the customer in almost subliminal ways. A few such irritants and the customer is lost.

"By combining data and insights from all disciplines," he says, "we are far more likely to have caught many of these minor points that

"Electra would be the perfect opportunity to merge the science of marketing with our advancements in technology."

would have been missed if we had taken a traditional engineering outlook alone."

Perhaps the greatest luxury the Electra program inspired—and the one that the customer will most appreciate—was time for the most extensive quality testing and product verification process ever conducted.

Stan Fowler, a Buick engineer who was part of the development team from its inception four years ago, explains the process:

"Computer-aided design and manufacturing saved time at the preliminary stages of product development, giving us more time in the later stages of development to dramatically increase our product verification process," he says.

Traditionally, structure design has involved a great deal of costly, time-consuming iteration. Parts and components were fabricated, tested, redesigned, refabricated and retested through several iterations.

In recent new-car development programs, General Motors has led the industry by using analytical techniques such as computer finite-element modeling to refine a design before experimental parts are produced.

"With Electra," Fowler says, "we used finite-element modeling for more components than on any previous car project."

With Electra, Buick engineers went even further. Using structural scale modeling, they were able to build ⅜-scale, half-scale and full-scale structurally representative components of plastic.

"The plastic fabrication process takes a few hours, compared with a few days working with metal," Fowler says. "By running stress tests and barrier crashes on plastic components and using formulas to obtain statistically significant results, we saved a tremendous amount of time."

With the addition of laser-scanner technology, Buick cut 50 percent from the normal time for experimentation-learning and turnaround.

The time saved was well used. A major quality testing program, more extensive than any before, was employed. Instead of going from a few handmade prototypes directly into full-scale production, a new step was added.

"Nearly 300 Electras were built on a production line very early on," Sharpe says. "This did two things for us. First, we gained experience in manufacturing and could therefore eliminate most start-up errors. Second, we could get millions of miles more testing on production vehicles."

Virtually every engineer involved with Electra got a car to use, evaluate and suggest changes on. "I, myself, have put 10,000 miles on a Park Avenue already," Sharpe says. "Believe me, when you live with a car that long, you get to know every minute potential customer like and dislike.

"Equally important, we've gained this experience early enough to make a lot of refinements before production."

The large fleet of prototype Electras was put to the test around the world. The new car was tested again and again at Buick test centers in America. Cold weather performance was tested under severe conditions in Kapuskasing in northern Ontario. Fuel-injection performance was monitored at Bosch Center, near Stuttgart, West Germany. And to test Electra to the ultimate in wind, sun, sand and heat, prototypes were taken to the Australian outback.

According to Randy Harlan, Buick's supervisor of the fuel-injection group, "We went to Australia because we couldn't get harsh enough weather in the United States at the time of year we wanted to test in. By going to Australia during our winter, which was their summer, we got an extra six months of experience in over 100-degree heat. During our summer we did the testing in Phoenix, Arizona, and Death Valley, California."

Testing in Australia was fraught with problems, such as having to import unleaded fuel, to decorate the Electras with warning placards that read "LEFT-HAND DRIVE" and to add "roo bars."

"A roo bar," Harlan says, "is the Australian equivalent of a cowcatcher. In the evening kangaroos come to the side of the road for water and food. Apparently, headlights stun them, so we had to be especially careful."

The extensive testing and product verification process was deemed so important that it actually delayed the introduction of the Electra. The car was supposed to have been introduced with other new 1984 Buicks in September. Instead, its introduction was set back to April.

General Motors president F. James McDonald explained the delay in saying, "Only after we're satisfied with the results of special verification testing for the early production runs of these new cars will we offer them to the public.

"To put it simply, the product will tell us when it is ready."

Electra is ready. The attention to detail with this car, much of it so subtle as to escape the average layperson's eye, shows a commitment to refinement clearly in keeping with its Mercedes rivals.

The underhood is a fine example. According to Don Runkle, Buick's assistant chief engineer, "As much time was spent designing the underhood area as the car interior."

The hood itself is hinged forward to give service technicians greater access to the engine compartment. This makes possible a three-inch-lower side line and thus far easier reach to anything under the hood.

"Instead of the spaghetti factory you normally find under modern car hoods," Runkle explains, "the Elec-


BUICK

The new Buick Electra. We put it to the test. Now it's your turn.

The inevitable question is, *why?*

Why take a car that will ultimately be judged for its luxury and comfort, and submit it to inhuman extremes in the Australian Outback, the wide-open German Autobahns or the sub-zero cold of Canada?

Because simply being luxurious isn't enough for a luxury car anymore. This Buick Electra is trimmer, lighter, more fun to drive than its predecessors, and it begs to be judged by a tougher, more demanding set of criteria. Ours. And yours.

First, ours.

A BOULEVARD DANDY, ELECTRA IS NOT.

The Australian Outback confirmed the ability of Electra's new computer-controlled engine to withstand sustained 108°F temperatures, day after punishing day. A comforting thought if you're stuck in rush-hour traffic in mid-July, or taking a vacation drive across Death Valley.

AT SPEED IN GERMANY.

We decided long ago the new Electra should respond to the throttle. So we enlisted the aid of Bosch, acknowledged world experts in fuel injection, to help us evaluate Electra's fuel injection system, one of the world's most sophisticated multi-port fuel injection systems. Then, to see how it would do in the real world, we drove the Electra on the Autobahn, where there are vast sections with no speed limit. Not that you're going to do a lot of wide-open running, as our test engineers did. But the knowledge is reassuring when you have an 18-wheeler to pass.

Then, we tried the oxygen-thin altitudes and steep grades of the Austrian Alps. Electra's fuel injection system is designed to handle such conditions. And it did. The secret here is a mass air-flow sensor, a tiny

The ultimate Electra: Buick Park Avenue.

electronic device that measures the mass of the air as it enters the engine, and allows for tiny rapid-fire corrections in the fuel charge–up to 80 times a second, for optimal performance.

A MOST UN-TRADITIONAL ELECTRA, TO BE SURE.

The Alps, with their slippery, twisting roads, also highlighted two more of Electra's noteworthy features: front-wheel drive and fully independent rear suspension. The traction of front-wheel drive, together with the ability of the rear wheels to handle bumps independently, endows the new Electra with impressive roadholding. Combined with Electra's traditional ride smoothness, the new Electra is an absolute pleasure on long drives.

NOT JUST HIGH TECH–*HIGH TOUCH.*

Happily, the new Electra is not *all* new.

Inside, per Electra tradition, you will find generous room for six adults, and true *comfort* for six adults. Indeed, Electra's interior dimensions are within an inch or so of its predecessors. So it's the same great cross-country luxury tourer it's always been.

Nor did we lose sight of the fact Electra is an elegant car. The supple feel of the seats, the rich upholstery and thick carpeting are there in force. Not to mention high levels of convenience–like an optional "memory seat" that remembers where you've set it, even after the parking lot attendant has readjusted it to his liking.

MORE THAN A CAR, A COMMITMENT.

The new Electra is so special, it is produced in one of the world's most advanced automotive assembly plants. For example, it enables every Electra to get two extra coats of clear enamel after the exterior color has been applied. For added lustre and protection.

The attention to detail extends all the way to the dealership. There, your new car will undergo a thorough inspection by both the Service Department and the salesman. Only then, when you go over the same detailed checklist, will you be asked to accept delivery.

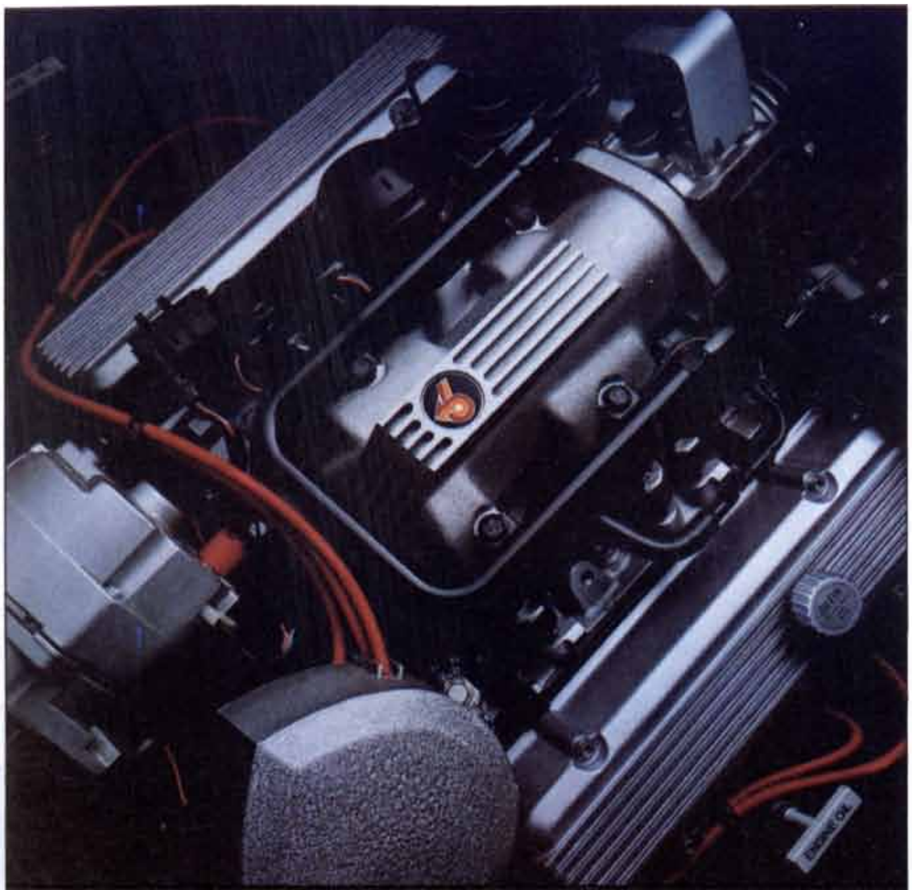
WE'VE HAD OUR TURN –NOW IT'S YOURS.

Official Car of the XXIIIrd Olympiad
Los Angeles 1984

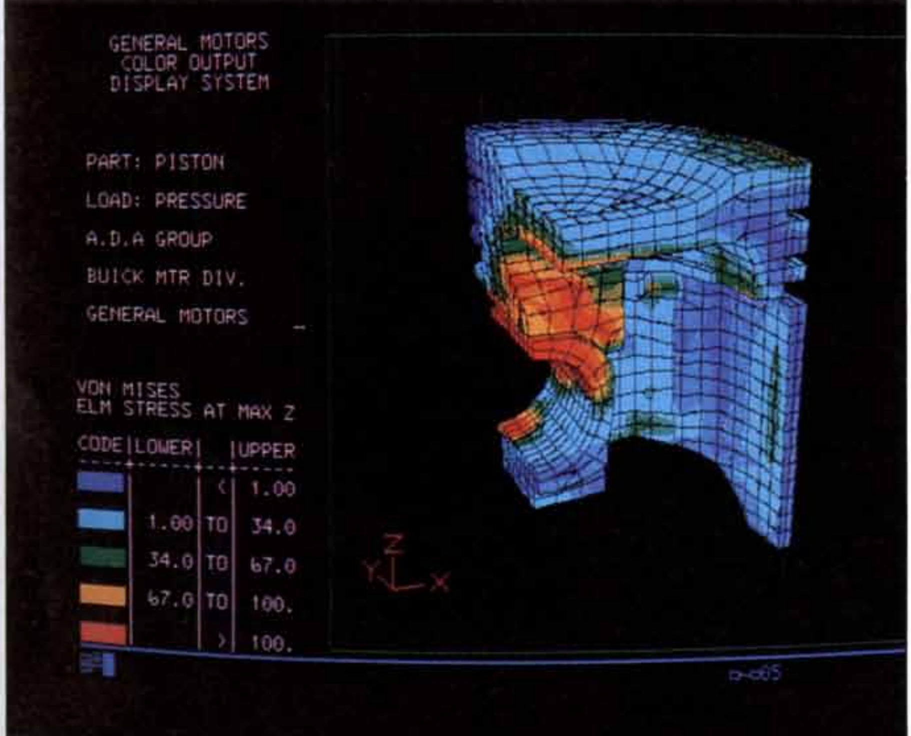
Buckle yourself in and put the new Electra to your own personal road test. To guide you, we've written a 20-page book that will assist you in your evaluation. The book is free, and we'd really like your opinion. Ask for it at your Buick dealer.

BUICK

Wouldn't you really rather have a Buick?



A Buick engineer calls Electra's V6 "probably the most advanced engine General Motors has ever created."



Engine technology: color codes illustrate stresses encountered on the piston assembly due to combustion pressure.

tra's hood area is clean, uncomplicated and aesthetically appealing. No one in the world has a better-looking underhood, not even the Porsche 928, and certainly not any Japanese car."

The underhood was designed to be as functional as it is aesthetic. By properly organizing the engine compartment, Electra designers achieved a car that was easier to build at the factory and to service in the shop.

"It's harder to make mistakes in an ordered environment," Runkle says. "It's clear where the plug lines go, and where the vacuum lines must be.

"An even greater benefit is that customers will take care of the engine a little more. Items like oil and transmission-fluid dipsticks are clearly marked and designed for easy access."

Numerous other subtle refinements on Electra may never be noticed, yet add up to an overall impression of workmanship. Window glass, for example, is flush mounted, which creates cleaner lines and reduces wind resistance. Exterior moldings have insulators and end caps, adding to good appearance while reducing any chance of paint damage. And pillars and body panels are filled with foam to eliminate noise.

In the higher levels of engineering, Electra exceeds all levels of engineers' expectations. Each engineer who worked on this car, in fact, is excited by a special aspect. For Runkle, it is the new 3.8-liter fuel-injected engine.

Runkle is a performance car specialist. Before coming to Buick, he helped design the Camaro and is now also in charge of Buick's "specialty" vehicles—he oversees limited-production cars and off-road racing components, one of which, a special turbocharged V6 engine, will race at the Indianapolis 500 this year.

While Electra offers a 4.3-liter diesel engine and a conventionally carbureted 3.0-liter V6, the new 3.8-liter fuel-injected engine is Runkle's enthusiastic favorite. He says, "It is probably the most advanced engine GM has ever created."

The multiport fuel-injection system represents a quantum leap in

> **"To put it simply, the product will tell us when it is ready."**

fuel delivery technology. Buick engineers worked with Bosch of Germany in developing the system. Six injectors, one at each intake point, are activated in pairs with each revolution. A fuel rail feeds the injectors via a high-pressure pump system.

The development of this fuel delivery system is another example of new thinking at Buick. Although the company had several years' experience in fuel injection, Bosch of Germany had some technology in this field that Buick felt was advanced. Rather than lag behind, General Motors entered into a joint development agreement with Bosch.

"It's called leapfrogging technologically," Runkle says. "By buying the highest-level technology, we are saving developmental time and, in effect, speeding up reaching the high point of our learning curve. And with our own edge of advanced control systems, we could combine Bosch and Buick knowledge for true product superiority.

"With the multiport system we were able to eliminate a number of engine control parts by incorporating them into the fuel rail," Runkle says. "The result is a less complex mechanism with lower emissions, better fuel economy and better starting characteristics than any engine on the market."

Unique to this engine is its hot film mass air flow sensor, the first of its kind. This device measures the density of air flowing into the fuel-injection system for precise air-fuel mixtures and optimum combustion. The air flow sensor and numerous other sensors such as the oxygen sensor, coolant temperature sensor, detonation sensor and throttle-position sensor all provide critical input to the on-board computer, called the Electronic Control Module (ECM). Computerization, sensor technology and the fuel delivery system reach a level of sophistication in the 3.8-liter engine that no other engine offers anywhere in the industry.

Electra's power is transmitted through a new, four-speed automatic transmission with overdrive fourth gear. The engine has new first-gear ratios for fast initial acceleration, and, for an overdrive ratio of .07 to 1, highway fuel economy is impressive. In fact, Electra should achieve an estimated 22 miles per gallon in city driving and 35 miles per gallon on the highway. In comparison, a subcompact car such as the Honda Civic gets 29 miles per gallon in the city—only a seven-mile-per-gallon difference separating it from Electra's six-passenger spaciousness.

In a true luxury car, ride quality is almost as vital an aspect as having a steering wheel. Electra went to four-wheel independent suspension using MacPherson struts and automatic levelers for the rear suspension. These features, combined with rack-and-pinion steering and special tires for the TType, make performance a strong appeal to European-oriented buyers.

"The car handles terrifically," Runkle says. "The TType, for example, doesn't take a back seat by any means to the BMW 528e or the Audi 5000S. We've been running those cars in tests against Electra, and it compares favorably.

"The TType is far faster than the 5000S. It performs as well as the BMW. And handling is better than either one."

One advantage that most of the European competitors lack is front-wheel drive. With all the drive components up front, a far more efficient overall design is possible. Electra, for instance, is 25 inches shorter than previous models, yet has virtually the same interior dimensions. The front-wheel-drive system is simpler, with the traditional heavy differential and drive shaft eliminated.

With all of this, Electra has numerous extra touches that delight even the most non-engineering-minded person. The power window switches are concave and convex. The license plate pocket requires no screws. Serviceability: the rear taillights can be removed by undoing only four easily reached screws. For the precision-minded, there are many such points to fascinate.

One is the MacPherson strut suspension, adding a new look to the coil spring configuration. Rather than shaping the springs like cylinders, the form of conventional springs, Buick engineers shaped theirs like barrels—narrow at the top and bottom. In this shape, a spring can compress into itself, requiring less space and making possible a lower hood, which improves aerodynamics and fuel efficiency. This is a fine example of a subtly new concept with major implications.

In conjunction with the effort put into sophisticating the product design, an equally great effort was made in providing suitable production facilities. There is an old saying in the auto industry that the art of engineering is building one perfect car, but the art of manufacturing is building a hundred thousand more. Indeed, the challenge of mass production of the Electra was even greater than usual, because it would be compared head-on with far more expensive, limited-production cars such as Mercedes and BMW. Electra had to be ready to meet the comparative challenge.

Two new assembly plants were built, using the latest machine and computer technology, to manufacture the Electra. The nearly identical facilities, in Wentzville, Missouri, and Orion, Michigan, each had 150 robots installed for the beginning of production. Computer-aided manufacturing using robotics eliminates the tedious and therefore error-prone tasks from welding to component assembly.

"Robotics are used in environments that are unpleasant for humans," Sharpe explains. "Logically, that means increased quality and greater cost effectiveness because people simply cannot perform well in bad work situations.

"The computer and robotic aids have another advantage in that they free up manufacturing workers to do what people do best—to think. Every employee is being enlisted with Electra as never before to be a full participant in the quality process. And they are being given the tools: both salaried and hourly people have been trained in statistical process control techniques to chart machine behavior and monitor production parameters before a single part goes out of specifications."

And as a final, extra measure of assurance, Buick created a new customer delivery standard called NVI (New Vehicle Inspection) Delivery. Every Electra is inspected before delivery not once, but twice.

While all of this seems like enough, Buick researchers know from repeated market surveys that the real test of a new product comes after it has been delivered to the customer. The speed and efficiency of servicing the vehicle play a major role in ultimate satisfaction and repeat sales.

Every Electra has a built-in diagnostic computer to ensure rapid, accurate service. The car can be hooked up directly to yet another Electra innovation, the Diagnostic Data Recorder (DDR). With the DDR, a service technician anywhere can make a direct computer diagnosis. The DDR can be placed in the car to record patterns on problems that occur only intermittently.

Should the cause of a problem still elude the service technician, then he or she need only pick up a phone and call an 800 number to receive expert advice from a Buick Technical Assistance Center. In sum, the individual servicing this high-technology vehicle need never be stumped for a diagnosis. Both computer analysis and expertise are always at hand.

Will so many advanced features in product design, verification, manufacturing and serviceability be appreciated?

"Each owner may appreciate different aspects, to add up to an overall impression of quality," Dabels says. "One of the things we learned from Dr. Urban's study is that customers decide based not on one impression, but on the cumulative impact of many cues or specific positive indicators.

"The research also told us that with a luxury car like Electra, word of mouth communication is a strong decider. That means a single individual disenchanted over a couple of negative cues could undo all efforts to market such a car. From the outset, we were determined that not one detail, no matter how seemingly minute, would get by us. All the cues will add up on the plus side for Electra."

Buick broke with yet another pattern of current automakers with Electra. Many automakers, afraid that their new products might be too radically different to appeal to the older fans of rear-wheel-drive luxury cars, have kept old designs to sell alongside new ones. They are hedging their bet. Buick decided to go completely to the advanced design Electra and set aside old technology.

Dabels says: "If you look at the life cycle of a car, you'll see that it's heavily influenced by its initial sales to perceptive buyers who come back to buy again and again. By making a full commitment to what we believe in, to the Electra, we intend to get the initial sales up as high as possible at the start, to build the broad base for the future.

"We believe that even our most tradition-bound buyers of pre-1985 Electras are becoming more sophisticated along with the rest of us. They don't live in a vacuum. And if you're 55 years old and have a 30-year-old son or daughter who drives a BMW, you can say, "Look at this car," because it's got all of the sophistication and technology of theirs, and more."

Dabels contends that all buyers will accept innovation as long as they do not lose any of the features they have enjoyed in the past, such as big car ride and comfort. "Nothing was sacrificed, and much was gained," Dabels says. "Even though this new car is two feet shorter and 600 pounds lighter, you can sit behind the wheel and experience everything that the Electra loyalists want; all the comfort, interior room and substantial solidness of a much heavier car. It's got better performance than today's V8, significantly better. Better fuel economy. All true luxury."

Dabels is certain of Electra's impact on those who have long considered American luxury cars. He expects inroads with the younger, import luxury car enthusiasts. Inroads, but not miracles.

"We don't believe that everyone who owns a BMW right now is going to go out and buy an Electra. Our research shows that about 15 percent of all Americans are hard-core import buyers. Yet even these people, if they can achieve some level of objectivity in looking at an American car, will be impressed."

Buick believes so thoroughly in making inroads through communicating with intelligent comparison buyers that the division is making this information as accessible as possible. A new showroom computer is part of a system being tried with Electra, called EPIC: Electronic Product Information Center.

"EPIC is still in the verification stage," Dabels says. "But it has the potential to make the buyer's task easier. The customer can make comparisons of Electra next to its competitors, point by point. The system even has a mode for locating Electras in stock at other dealerships with specific equipment desired. It can speed up delivery significantly."

For many potential customers who have had the opportunity to compare and experience the new Electra and Park Avenue, the results are highly favorable. In fact, customer satisfaction is rated higher on this car than ever before. And with Buick having broken all sales records in 1983 and doing even better this year, the market appears ready to consider Electra.

For the people inside Buick—engineers, scientists, marketers, and so many other specialists—the Electra is even more. It represents new levels of shared goals, of encouraged innovation and of a commitment to quality and product verification never before attempted. There is an overwhelming enthusiasm at Buick about the Electra experience.

"Engineering. The Buick Electra" was researched and written by Jim Plegue and Albert Lee. Graphics by Sherin & Matejka, Inc., and Buick Motor Division.

Buick Century. Technology that makes you feel good inside.

In creating the Buick Century, it seems that we created something of a paradox.

With its sleek aerodynamics, advanced front-wheel-drive technology and computer-selected springs, the Century is a very nimble road car indeed.

But because it's a Buick, the Century is one road car that won't let the feel of the road interfere with your personal comfort and tranquility.

For an automobile that offers both driving pleasure and riding pleasure, visit your Buick dealer and buckle yourself into the Century.

It will make you feel very comfortable with today's technology.

Official Car of the XXIIIrd Olympiad
Los Angeles 1984

BUICK

Wouldn't you really rather have a Buick?

Right from day one, owning a Buick is special.

We'd like the drive home in your new Buick to be a memorable, even magical experience.

Hence, the New Vehicle Inspection and Delivery Checklist you see here. It's meant to be signed by you and your dealer when you agree that your new Buick works the way you expect it to. And that you know exactly *how* it works.

This checklist is just one of many checkpoints your Buick passes through on its way from us to you.

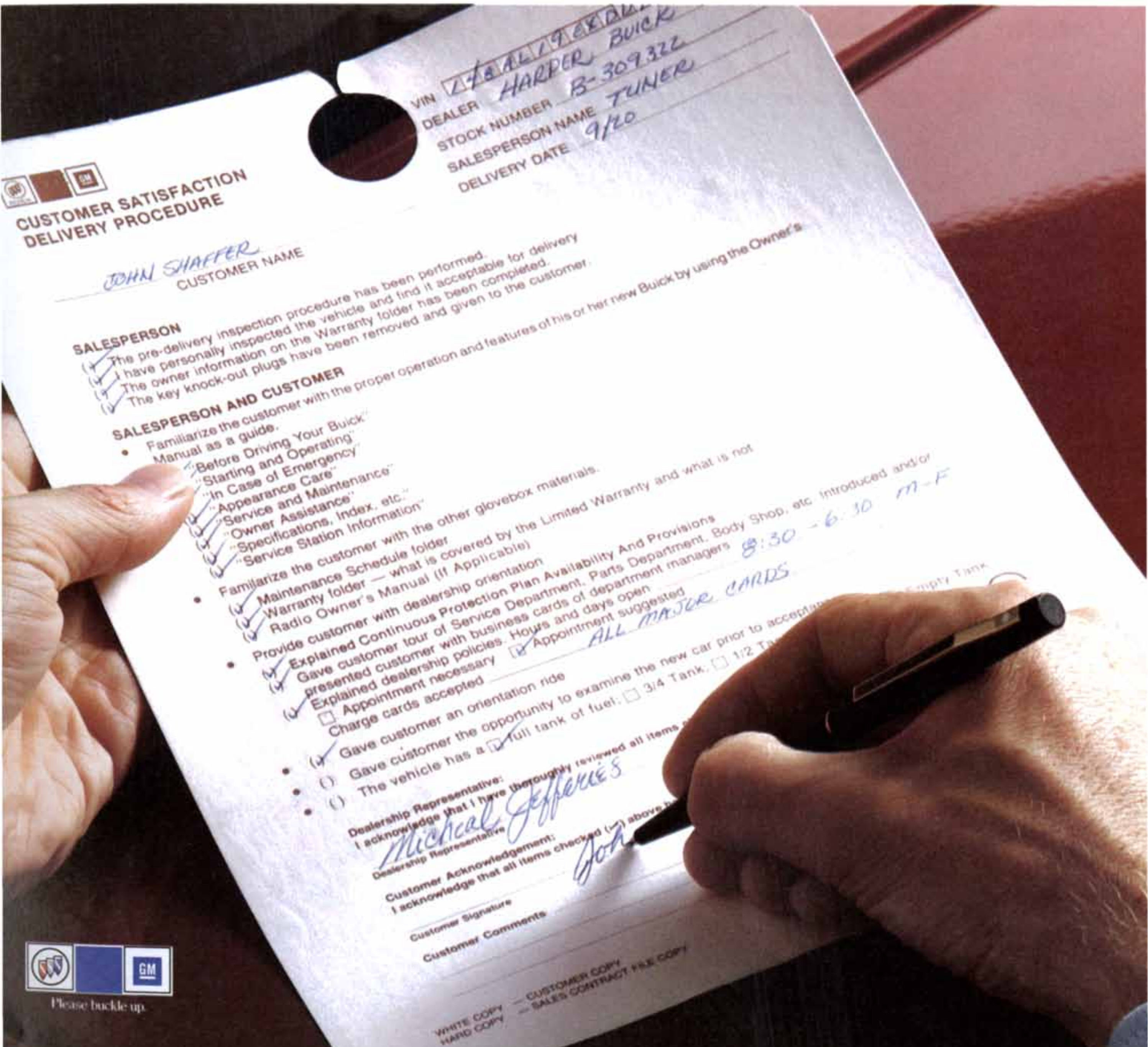
It's part of an entire program dedicated to keeping you, a Buick owner, a satisfied Buick owner. When you drive away in a new Buick, we're determined that you feel as special as the car you just bought.

Right from your first day together. To the day you trade it for your next one.

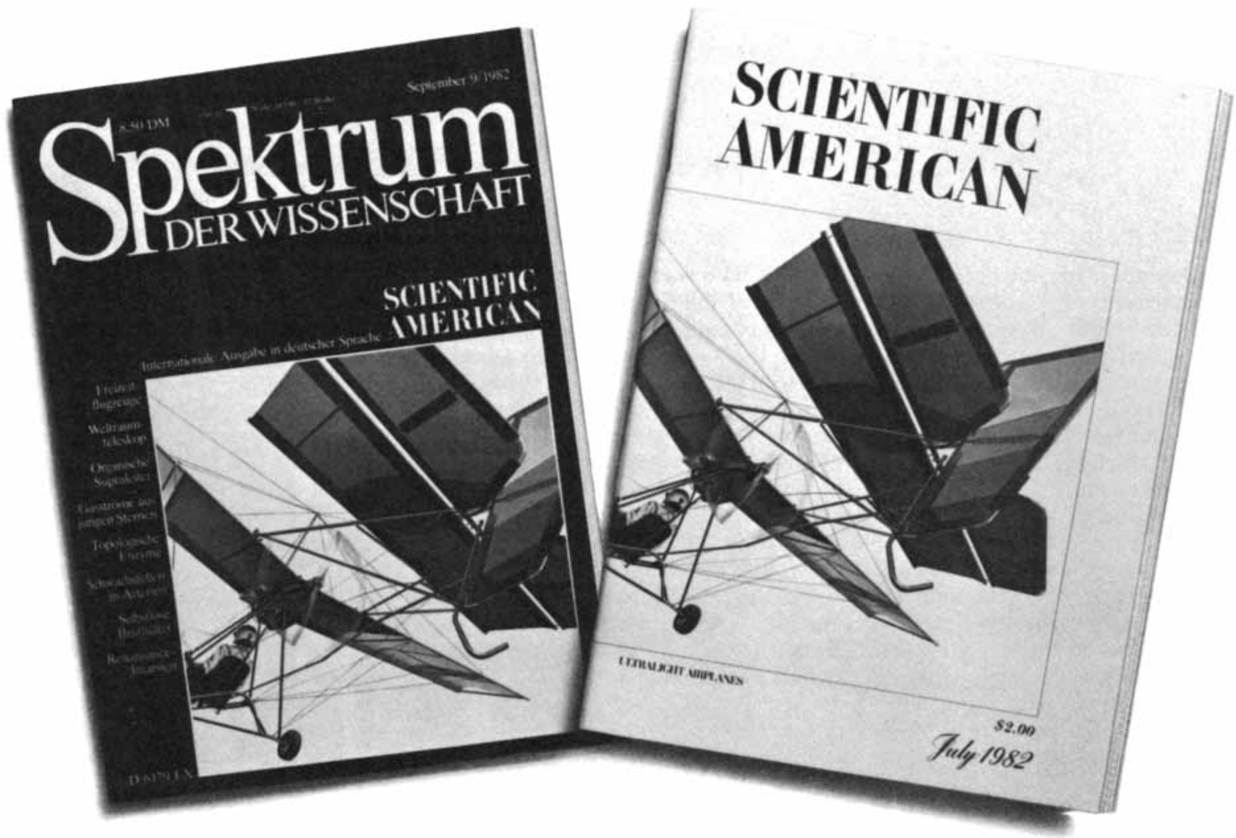
Official Car of the XXIIIrd Olympiad
Los Angeles 1984

BUICK

Wouldn't you really rather have a Buick?



HOW TO REACH THE PEOPLE WHO MAKE THE FUTURE HAPPEN IN GERMANY



SPEKTRUM DER WISSENSCHAFT is the German edition of SCIENTIFIC AMERICAN which attracts a young audience of affluent professionals who have come to the top because of their technical expertise. Almost half of them hold top management job titles. The fortunes of major corporations in Germany increasingly depend on technically sophisticated people. SPEKTRUM DER WISSENSCHAFT keeps these people in touch with the advances in science that drive the growth of Germany's industry.

SCIENTIFIC AMERICAN speaks the languages of more than half the world's population. We are the one publication in the world today providing efficient coverage of technology-based management. We reach men and women in industry, whose qualification to make technical decisions places them in key positions in their country's government and industry.

SCIENTIFIC AMERICAN in eight languages: English, Spanish, Italian, French, German, Japanese, Chinese and now Russian has gathered in its audience three million people who make the future happen around the world.

For more information on our German-language edition contact:

Claus G. Firchow
SPEKTRUM DER WISSENSCHAFT
Moenchhofstrasse 15
6900 Heidelberg
FEDERAL REPUBLIC OF GERMANY
Telephone 011-49-6221-46037

or in New York
John Kirby
V.P./Advertising Director
SCIENTIFIC AMERICAN
415 Madison Avenue
New York, New York 10017
Telephone 212-754-0262

Turing Machines

At its logical base every digital computer embodies one of these pencil-and-paper devices invented by the British mathematician A. M. Turing. The machines mark off the limits of computability

by John E. Hopcroft

In 1900 David Hilbert, the preeminent mathematician of his time, challenged the world of mathematics with a list of unsolved problems, presented in Paris before the International Congress of Mathematicians. The 23rd problem on the list was to discover a method for establishing the truth or falsity of any statement in a language of formal logic called the predicate calculus. Thirty-six years were to pass before the problem was settled, and its resolution marked an extraordinary and unexpected turn in mathematics. At the University of Cambridge a young fellow of mathematics in King's College named Alan Mathison Turing had become familiar with Hilbert's 23rd problem through a series of lectures given by M. H. A. Newman. Turing pondered the problem during long afternoon runs in the English countryside, and it was after one of these runs that the answer came to him. Hilbert's problem was impossible to solve.

The publication in which Turing announced his result has had a significance far beyond the immediate problem it addressed. In attacking Hilbert's problem, Turing was forced to ask how the concept of method might be given a precise definition. Beginning with the intuitive idea that a method is an algorithm—a procedure that can be mechanically carried out without creative intervention—he showed how the idea can be refined into a detailed model of the process of computation in which any algorithm is broken down into a sequence of simple, atomic steps. The resulting model of computation is the logical construct called a Turing machine.

The simplest way to describe the Turing machine is in terms of mechanical parts such as wheels, punched tape and a scanner that can move back and forth over the tape. The machinery is not essential—at the most fundamental level Turing's device is the embodiment of a method of mathematical reasoning—but it would be misleading to dispense entirely with the mechanical metaphor. That metaphor was suggestive to Turing himself, who was a pioneer in the development of the digital computer. Moreover, the claims of the computer scientist on the Turing machine as a conceptual tool are now at least as strong as the claims of the logician. Its significance for the theory of computing is fundamental: given a large but finite amount of time, the Turing machine is capable of any computation that can be done by any modern digital computer, no matter how powerful.

The universal capability of the Turing machine does not imply that it would be a practical computer. Any real computer can work many times faster than the Turing machine, because in the design of the real computer clarity of operation is willingly sacrificed for speed and efficiency. Nevertheless, for the theoretical study of the ultimate problem-solving capacity of the real computer the Turing machine has become indispensable. For example, it has enabled mathematicians and computer scientists to prove there are many problems in addition to Hilbert's problem that cannot be solved, no matter how fast or how powerful a computer is applied to their solution.

Turing Machine Operation

What is a Turing machine and how does it work? Andrew Hodges, in his recent biography of Turing, compares it to an ordinary typewriter. Like the typewriter, the Turing machine incorporates a movable printing head that prints discrete symbols, drawn from a finite alphabet, one at a time on a printing surface. To simplify the movements of the printing head, the printing surface is assumed to be a tape marked off into discrete frames or segments. The printing head of the Turing machine therefore needs to move in only one dimension, to the left or to the right, and the actions of the machine need not take account of such irrelevant complexities as the length of the printed line or the width of the space between two lines. Only one printed symbol is allowed in each frame of the tape, but there is no limit imposed on the length of the tape or, consequently, on the length of the string of symbols that can be printed on it.

The movable printing head of the Turing machine can carry out two other functions as well as printing. Like many typewriters manufactured in the past decade, it can remove or erase one symbol at a time from the printing surface. Unlike the typewriter, the printing head can also "read," or register the symbolic content of each tape frame one frame at a time. In this way the symbols on the tape can serve as input to the machine and play a role in determining its subsequent action.

A typewriter can assume one of several states, or modes of operation, in the course of its activity. In its "home" state it prints lowercase letters and numerals, whereas in its "shift" state it prints uppercase letters and special symbols. Similarly, a Turing machine can assume any one of a finite number of states. Each state presumably constitutes a different setup or configuration of the machine, but because the Turing machine is a relatively abstract device, usually no attempt is made to give a more concrete, mechanical description of the states. It will suffice to describe each state of the machine in terms of the effects that state has on the activity of the machine.

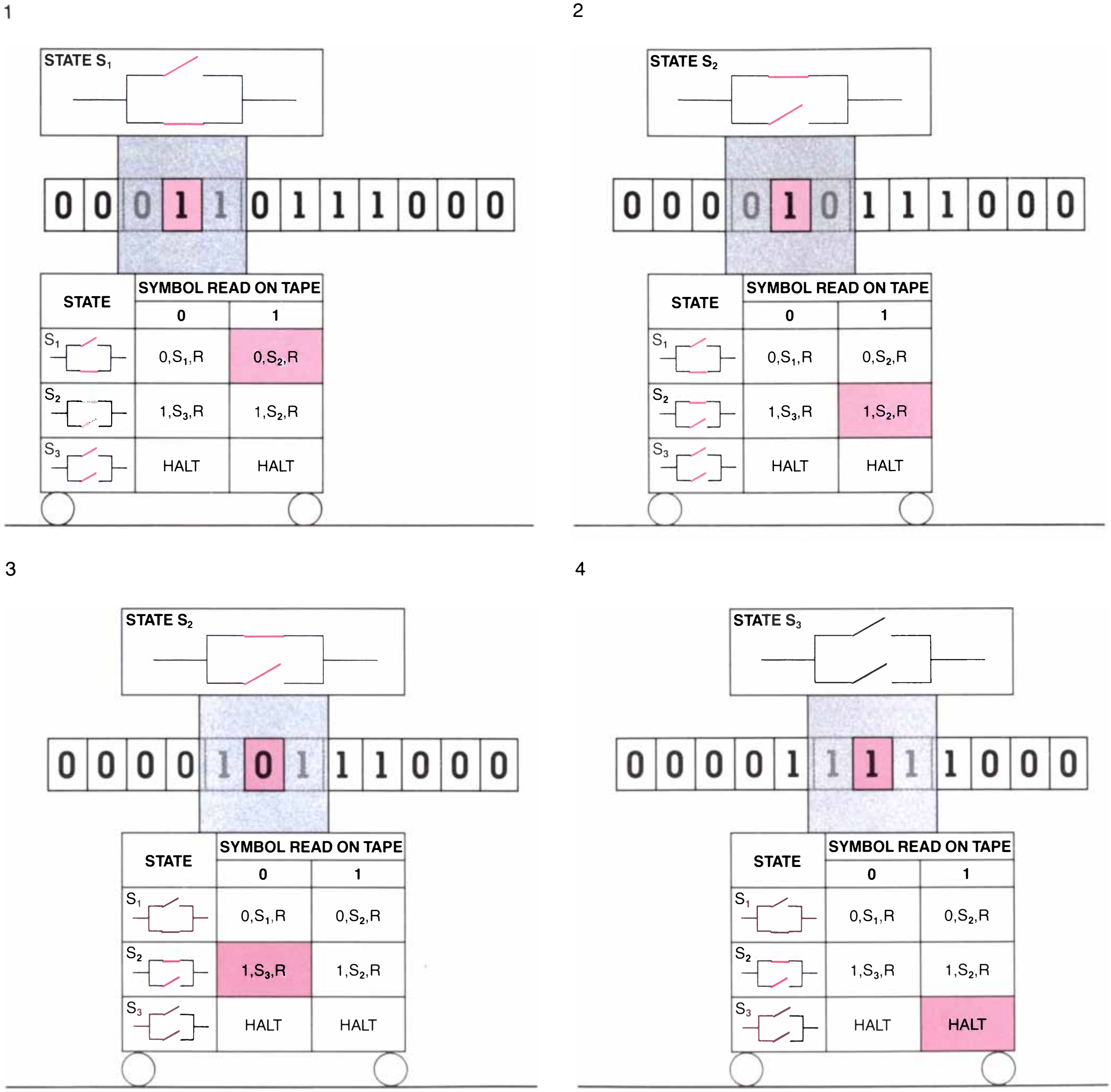
The activity of a Turing machine is made up entirely of discrete, instantaneous steps, and each step is determined by two initial conditions: the current state of the machine and the symbol that occupies the tape frame currently being scanned. Given some pair of initial conditions, the machine receives a three-part instruction for its next operating step. The first part of the instruction designates the symbol the machine is to leave in the tape frame being scanned. For example, if the instruction specifies that the symbol 1 is to be left in the frame, the machine prints the symbol if the frame is blank, leaves the symbol alone if a 1 already occupies the frame or erases the symbol if it is not a 1 and replaces it with a 1.

The second part of the instruction

specifies the next state of the machine. As with the specification of the symbol in the first part of the instruction, the designation of a state may require the machine to change its state or to remain in its current one. The third part of the instruction specifies whether the printing head is to scan one frame to the left or to the right along the tape.

The entire instruction can be abbreviated by listing the three values of the variables—the tape symbol, the machine state and the motion of the printing head—in a fixed order. For example, if a given pair of initial conditions are to make the machine leave the symbol 1 on the scanned tape frame, cause the machine to assume state S_2 and move the printing head one frame to the left, the instruction is abbreviated (1,S_2,L).

The best way to understand how a Turing machine works is to try to build one. In this context building a Turing machine means to construct a table of instructions that specify the action of the Turing machine for every possible pair made up of one state and one tape



ADDITION OF 2 AND 3 is accomplished by a Turing machine in four steps. Each number to be added is represented on a tape in unary notation: as a string of 1's bounded by 0's at both ends. The machine can register the content of one tape frame at a time (*colored frames*) by shuttling to the left or to the right across the tape in a series of discrete moves. The goal of the machine is to generate a string of five consecutive 1's and halt. The table of instructions, shown in the lower part of each machine in the diagram, is a fixed set of moves for all possible initial conditions and gives a procedure for adding any two numbers. Following the instructions, the machine removes the 0 separating two strings of 1's and shifts the left string one frame to the right to join the right string. The number of initial conditions available in the table of instructions must be large enough to meet all the contingencies that might arise on the tape; that number can be increased by increasing the number of internal states, or configurations, that are built into the Turing machine. For every possible combination of tape symbol and machine state the table of instructions must either halt the machine or specify three variables. First, it must give the symbol that is to be left in the frame of the tape currently being registered; second, it must specify the state the machine is to assume before it registers another tape frame, and third, it must indicate whether the machine is to move across the tape one frame to the left or to the right.

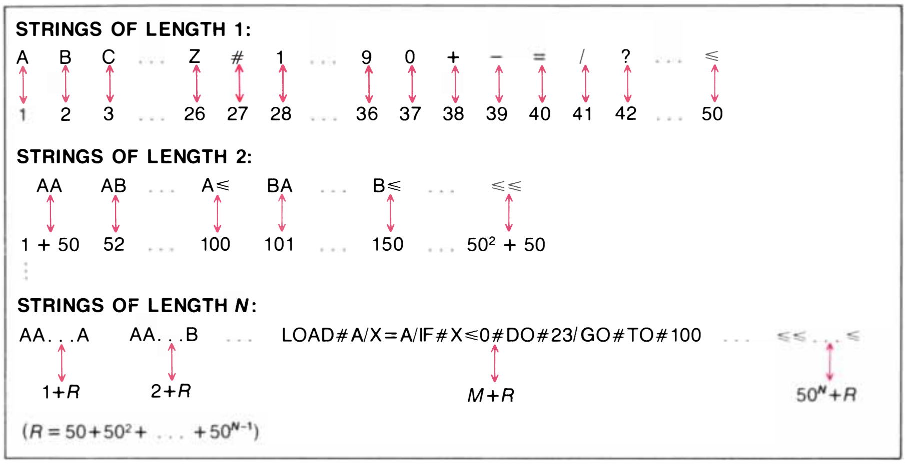
symbol. In practice the construction of the table also means granting enough possible states to the machine for it to do the task at hand.

The Adding Machine

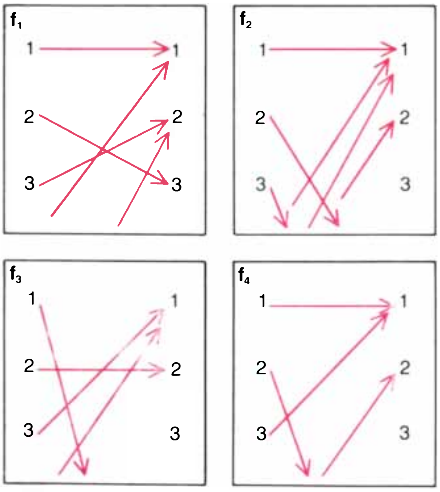
Consider how one might go about designing a Turing machine that adds two numbers and halts. It is customary, although by no means essential, to allow only two symbols to be printed on the tape, say 0 and 1. Any given number N can then be represented on the tape by a string of N 1's. If two numbers M and N are to be printed on the tape, they can be represented by a string of M 1's, followed on the right by a 0, followed in turn on the right by a string of N 1's. Assume the Turing machine is in its initial state S_1 and its printing head is scanning the leftmost 1 in the string of M 1's. To construct a Turing machine that adds, what one wants is to generate a table of instructions that will cause the Turing machine eventually to print a string of $M + N$ 1's and then halt.

One simple way to accomplish the task is to shift the leftmost string of M 1's one frame to the right. If the shift is done properly, the 0 no longer separates the two strings, and the single string that results has length $M + N$. It is not possible for the Turing machine to shift the string of 1's all at once; for example, the printing head can advance only one frame at a time. One way to get the result is to create a series of instructions for a Turing machine with three states. In the first state the printing head scans the tape from left to right, one frame at a time, until it reaches the leftmost 1. It changes the 1 to a 0, enters the second state and continues moving to the right. In the second state, when the head finds a 1 in the current frame, it takes no action—it does not change either the tape symbol or the machine state—except to move one more frame to the right. In this way the head scans past the $M - 1$ 1's remaining in the first string. When a 0 is finally found, the instructions cause the head to change the 0 to a 1 and halt.

The reader is now invited to construct a Turing machine that will find the product of any two given numbers. The conventions for the input are the same as they were for the machine that adds: the two integers to be multiplied are represented by two consecutive strings of 1's, separated by a 0. The output, a string of 1's whose length must be equal to the product of the first two strings, can be set off as a third string to the right of the first two. One working design is given in the illustration on page 91. It is fair to say at the outset that the construction is tricky and requires some careful bookkeeping. It also seems fair to give a hint, which the puzzle fancier can avoid by not reading the next paragraph.



CHARACTER STRINGS of any finite length can be paired one for one with the positive integers, provided each character in the string is selected from some finite set. According to the definition formulated by the German mathematician Georg Cantor, such a pairing implies the two sets are equivalent in size, even though both are infinite sets. Any infinite set whose elements can be paired one for one with the positive integers is called a countable set. Because every computer program and every possible table of instructions for a Turing machine can be encoded as a finite string of symbols, the pairing shows that the number of possible computer programs and the number of possible Turing machines are both countably infinite numbers.



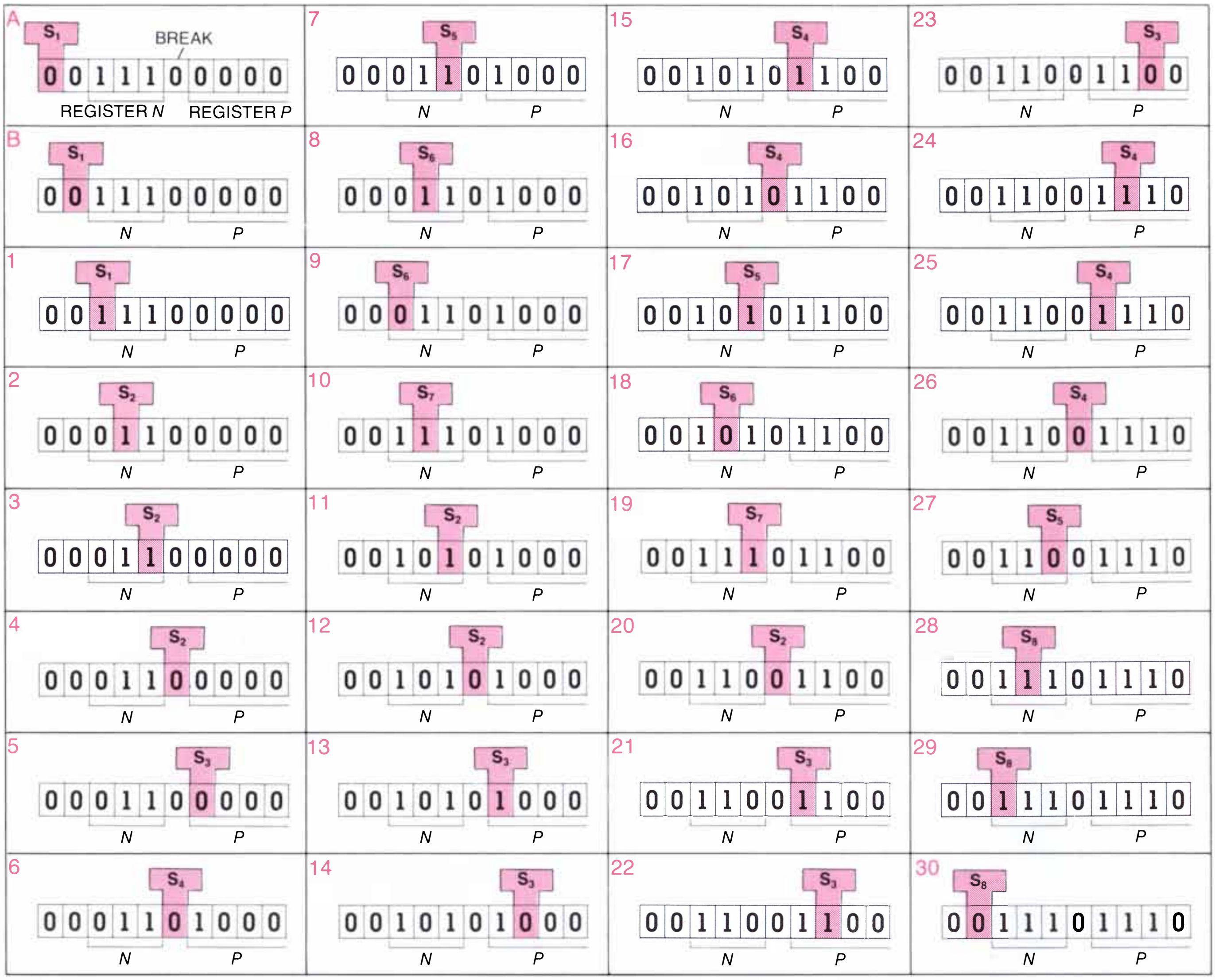
	1	2	3	4	5	6	...
f_1	[1]	3	2	1	7	2	...
f_2	1	[5]	9	1	2	1	...
f_3	9	2	[1]	6	1	6	...
f_4	1	8	1	[6]	2	4	...
f_5	1	1	2	4	[1]	1	...
f_6	1	7	1	8	9	[6]	...
⋮							⋮
F	2	4	2	5	2	5	...

SET OF MATHEMATICAL FUNCTIONS of the positive whole numbers that take integer values is a noncountable set; that is, there are too many such functions for them to be paired one for one with the infinite set of whole numbers. Suppose, on the contrary, all the functions could be named according to their subscripts and then listed in a fixed order. The value of the function corresponding to each integer could then be given as an infinite row of digits. Since a function is determined by its infinite row of digits, consider a function constructed by changing the first digit in the first row, the second digit in the second row and so on. The row of digits that define the constructed function (*color*) must differ from any row of digits in the original list by at least one digit. Hence the supposition that all the functions could be listed leads to the contradiction that a new function not in the original list can always be constructed.

Building Complex Programs

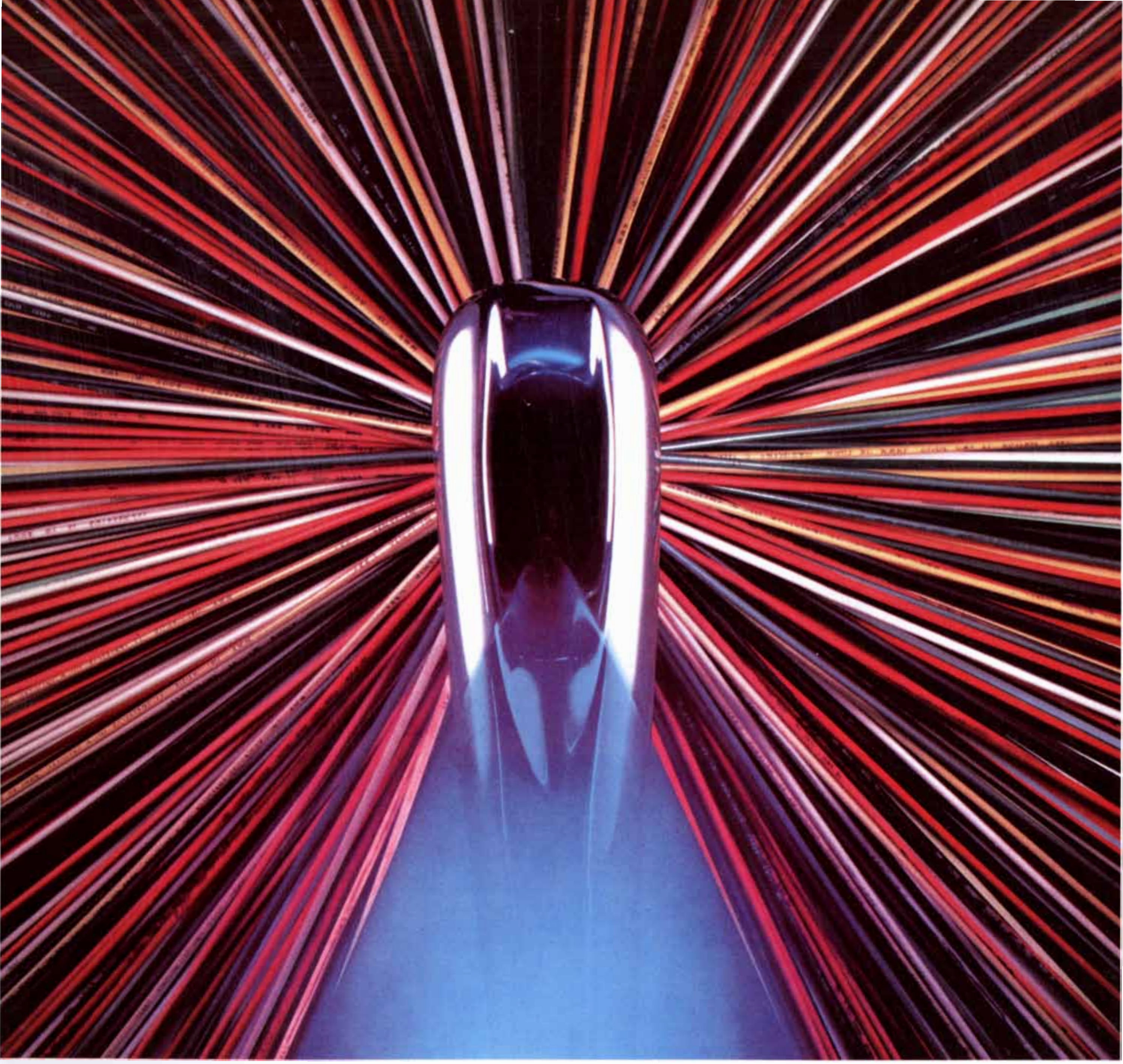
Multiplication is much easier to do if one first develops a routine for copying a string of 1's immediately to the right of some given point. It is always possible to copy N 1's with an N-state machine; the states can effectively count the 1's in the string. Since the number of states must not grow indefinitely large, however, it is desirable to find a way to copy the string without counting its elements. One way is to send a marker across the string from left to right; the marker is a 0 that successively takes the place of each 1 in the string. For each advance of the marker from one frame to the next the printing head is instructed to scan past the right end of the string, skip over a 0 that indicates spacing and then change the next 0 it finds to a 1. If the printing head shuttles back and forth once for each advance of the marker, a new string of N 1's is written immediately to the right of the first string [*see illustration on opposite page*].

With a little practice one quickly learns how to build Turing machines that can run simple computational routines and how to combine those ma-



STATE	TAPE SYMBOL 0	COMMENT	SAMPLE MOVES	TAPE SYMBOL 1	COMMENT	SAMPLE MOVES
S_1	0, S_1, R	Begin with right shift to left end of register *N*.	A, B	0, S_2, R	Mark left end of register *N* with 0.	1
S_2	0, S_3, R	Right shift across break.	4, 12, 20	1, S_2, R	Right shift across register *N* from marker to break.	2, 3, 11
S_3	1, S_4, L	Copy 1 at right end of register *P*.	5, 14, 23	1, S_3, R	Right shift across 1's in register *P*.	13, 21, 22
S_4	0, S_5, L	Left shift across break.	6, 16, 26	1, S_4, L	Left shift across register *P*.	15, 24, 25
S_5	1, S_8, L	0 marker detected to the immediate left of break, indicating that a complete copy of register *N* has been added to register *P*.	27	1, S_6, L	Left shift from break one unit into register *N*. State cannot remain S_5; if it did, the eventual encounter with the 0 marker would cause the Turing machine to halt prematurely.	7, 17
S_6	1, S_7, R	Begin move of 0 marker one unit to right.	9, 18	1, S_6, L	Continue left shift across register *N* to 0 marker.	8
S_7		Condition is not encountered, hence any dummy instruction or none at all may be inserted.		0, S_2, R	Complete move of 0 marker one unit to right; repeat cycle.	10, 19
S_8	0, —, —	Copy routine completed; stop or wait for new instruction.	30	1, S_8, L	Left shift to left end of register *N*.	28, 29

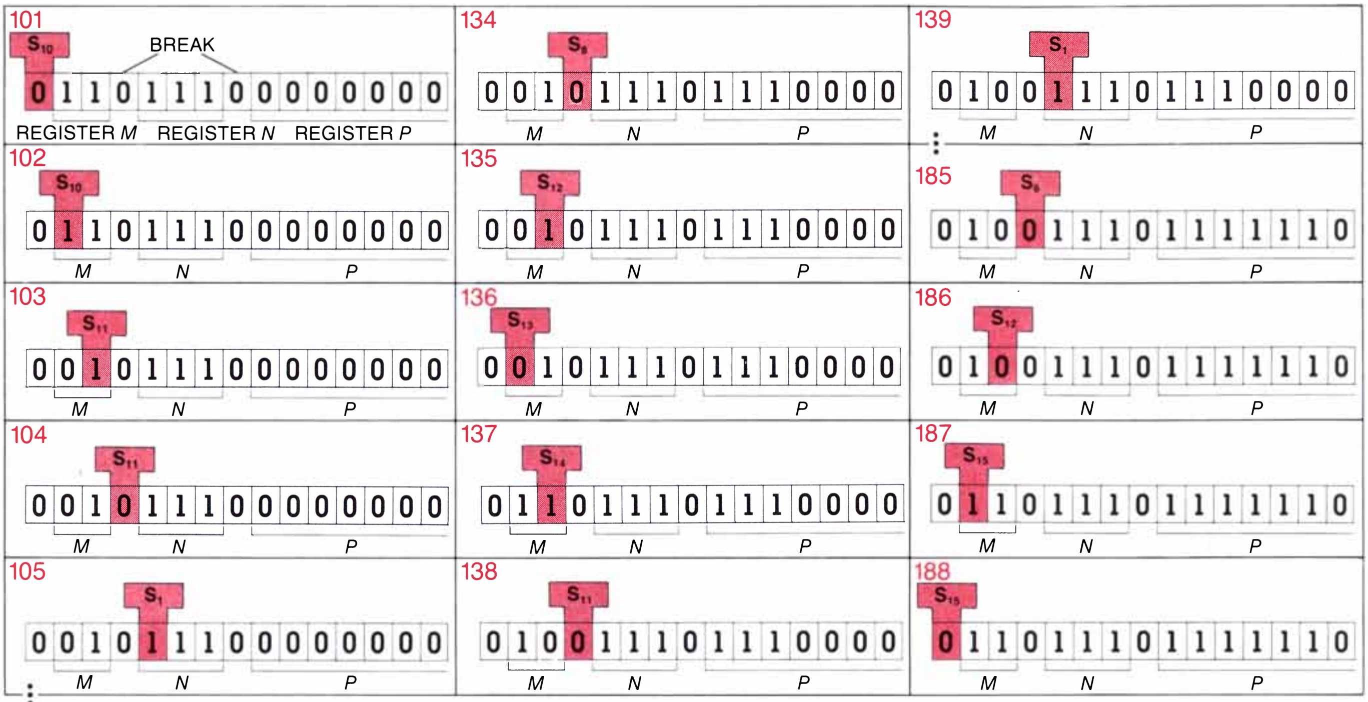
COPYING TURING MACHINE is a component of more elaborate devices. Given any string of 1's marked on a tape, the machine writes a second string with the same number of 1's to the right of the 0 that marks the end of the first string. Here the machine copies three 1's.



chines in order to carry out more complex calculations. For example, a polynomial expression can be evaluated by combining the routines for adding, copying and multiplying. Even more versatile are short, elementary routines for symbol manipulation, such as "Move the printing head to the right until it encounters a 0" and "Move the marker in the leftmost string of 1's one frame to the right." Variations of these short routines are exploited both in the Turing machine that copies a string of 1's and in the machine that multiplies.

The Universal Turing Machine

If one is inclined to try building, say, the Turing machine that multiplies, one soon begins to appreciate the difficulties that must be faced in the design of a useful computer program. Most small Turing machines, namely the ones with only a few possible states, do not carry out any useful or even sensible task. Many of them get caught in infinite loops and shuttle back and forth on a tape indefinitely without halting. From among the machines that do perform reasonable tasks, one must choose a combination of machines that work together efficiently. The initial impression can be that the simplest tasks are fiendishly difficult and that realistic computation is hopeless. Such difficulties can be frustrating, but they are merely techni-



STATE	TAPE SYMBOL 0	COMMENT	SAMPLE MOVES	TAPE SYMBOL 1	COMMENT	SAMPLE MOVES
S_{10}	0, S_{10}, R	Begin with right shift to left end of register *M*.	101	0, S_{11}, R	Mark left end of register *M* with 0.	102
S_{11}	0, S_1, R	Right shift across break between register *M* and register *N*; begin copy routine *(see illustration on page 89)*.	104, 138	1, S_{11}, R	Right shift across register *M* from marker to break.	103
S_8	0, S_{12}, L	Left shift across break between register *N* and register *M*; end of copy routine.	134, 185	1, S_8, L	Left shift to left end of register *N*.	Not shown
S_{12}	1, S_{15}, L	0 marker detected to the immediate left of break between register *N* and register *M*, indicating that the product of the numbers stored in registers *M* and *N* has been stored in register *P*.	186	1, S_{13}, L	Left shift from break between register *N* and register *M* one unit into register *M*. State cannot remain S_{12}; if it did, the eventual encounter with the 0 marker would cause the Turing machine to halt prematurely.	135
S_{13}	1, S_{14}, R	Begin move of 0 marker in register *M* one unit to right.	136	1, S_{13}, R	Continue left shift across register *M* to 0 marker.	Not shown
S_{14}		Condition is not encountered; hence any dummy instruction or none at all may be inserted.		0, S_{11}, R	Complete move of 0 marker in register *M* one unit to right; repeat cycle.	137
S_{15}	0, —, —	Multiplication routine completed; stop or wait for new instruction.	188	1, S_{15}, L	Left shift to left end of register *M*.	187

MULTIPLICATION can be done with a Turing machine by embedding in it another Turing machine that can copy a string of 1's. In the example the machine requires 88 cycles to find the product of 2 and 3; the breaks in the numbering are machine cycles that jump to states defined for the machine that copies a string of 1's (*see illustration on page 89*). In the last cycle the product of 2 and 3 is displayed as a string of six 1's in the section of the tape called the *P* register, immediately after the 0 that separates it from the two multipliers.

ONE SYMBOL

1 2 3 9 π

TWO SYMBOLS

11 99 $9!\ (=1\times2\times3\times4\times5\times6\times7\times8\times9=362{,}880)$ $9^9\ (=387{,}420{,}489)$

THREE SYMBOLS

TEN 9+9 $3\uparrow4\ (=3\times[3\times(3\times3)]=3^4=81)$ $9^{9^9}\ (=9^{387{,}420{,}489}\cong10^{369{,}700{,}000})$

FOUR SYMBOLS

NINE 10^{99} $3\uparrow\uparrow4\ (=3\uparrow[3\uparrow(3\uparrow3)]=3^{3^{3^3}}=3^{3^{27}}=3^{7{,}625{,}597{,}484{,}987}$
$\cong10^{3{,}638{,}000{,}000{,}000})$

FIVE SYMBOLS

EIGHT $3\uparrow\uparrow\uparrow4\ (=3\uparrow\uparrow[3\uparrow\uparrow(3\uparrow\uparrow3)]=3\uparrow\uparrow[3\uparrow\uparrow3^{3^3}]=3\uparrow\uparrow\left.3^{3^{\cdot^{\cdot^{\cdot^{3}}}}}\right\}3^{3^3}\text{ LEVELS}$

$$=\left.\left.3^{3^{\cdot^{\cdot^{\cdot^{3}}}}}\right\}3^{3^{\cdot^{\cdot^{\cdot^{3}}}}}\text{ LEVELS}\right\}3^{3^3}\text{ LEVELS})$$

⋮

58 SYMBOLS

1#MORE#THAN#THE#LARGEST#NUMBER#EXPRESSIBLE#WITH#58#SYMBOLS

RICHARD PARADOX, named after the French mathematician Jules Richard, arises if one supposes the positive integers can be ordered, or listed, according to the number of symbols needed to specify them. Isolated integers of remarkable size can be designated with the help of special symbols such as the arrow notation introduced by Donald E. Knuth of Stanford University. Nevertheless, according to its own description, the number allegedly specified above with 58 symbols is larger than any number that can be specified with 58 symbols, which is a paradox. A similar paradox arises if one tries to find a given Turing machine in a list of Turing machines arranged according to the length of the string of symbols needed to encode each machine.

cal; with a few well-chosen routines the power of the Turing machine for solving problems increases explosively, and one is struck not by the weakness of the machine but by its potential. As Turing was able to show, it is possible to combine simple Turing machines into a machine that can carry out any task that can be explicitly described.

The electronic computer, which in part owes its existence to Turing's conceptual machines, is now probably the most convincing demonstration of those machines' computational power. In the course of his work Turing pointed out that any Turing machine M can be encoded on a tape as a sequence of 0's and 1's. The fundamental reason the encoding can be done is that every Turing machine is uniquely defined by its table of instructions; that table must be finite in length because both the machine states and the alphabet of tape symbols are finite in number.

Turing showed that the operation of Turing machine M on any sequence of tape symbols X can be simulated by another Turing machine called the universal machine. The symbols on the tape registered by the universal Turing machine are grouped into two major sections: at the left is the encoded description of Turing machine M and at the right is the sequence of symbols X that would be encountered by M as it scanned its own tape. The universal machine is then constructed so that its printing head shuttles back and forth between the left and right sections of the tape. Through an elaborate system of markers the universal machine keeps track of the encoded state of M that is being consulted. Turing proved that the effect of the universal machine on the sequence of symbols X is exactly what the effect of M would be on the same sequence of tape symbols.

The successful simulation of Turing machine M by the universal Turing machine depends only on the fact that M is a machine that can be described exhaustively by a finite number of symbols. In principle, however, every digital computer can be described in the same way. The computer has a large but finite number of internal states, and its response to input data is entirely determined by the finite set of statements that make up its programming. Hence a complete description of any digital computer can be encoded on a tape as a sequence of 0's and 1's, and any input data can be encoded on the tape to the right of the description of the computer. By alternately consulting the description of the computer and the string of input data on the tape, the universal Turing machine can simulate the action of the computer on the input data step by step.

Given enough memory to serve as a tape for symbol manipulation, any real computer can play the role of the universal Turing machine. For example, if a home microcomputer were programmed to function as a universal Turing machine, and if a description of a large, "mainframe" computer were encoded on its input data, the microcomputer would simulate the action of the large computer on any string of data symbols. In this sense every digital computer can compute exactly the same class of mathematical functions, namely all the functions that are computable by some Turing machine. The existence of only one such class of functions strongly supports Turing's formal definition of computability: A mathematical function is computable if it can be computed by some Turing machine. Turing argued persuasively that his definition is equivalent to any reasonable interpretation of

NUMBER OF STATES	MAXIMUM NUMBER OF PRINTED 1's	LOWER LIMIT FOR VALUE OF σ
3	$\sigma(3)$	6
4	$\sigma(4)$	12
5	$\sigma(5)$	17
6	$\sigma(6)$	35
7	$\sigma(7)$	22,961
8	$\sigma(8)$	$3^{92}\cong7.9\times10^{43}$
9	$\sigma(9)$	$3^{92}+1$
10	$\sigma(10)$	$\left.\left.\left.a^{a^{\cdot^{\cdot^{a}}}}\right\}a^{a^{\cdot^{\cdot^{a}}}}\right\}\cdots\right\}a^a$

"BUSY BEAVER" PROBLEM is to find the maximum number of 1's that can be printed by an *N*-state Turing machine that begins its operation on a tape initially filled with 0's and eventually comes to a halt. The number, which depends on *N*, is the value of a function designated $\sigma(N)$. In 1962 Tibor Rado of Ohio State University proved the function grows too fast to be computable. Lower bounds on the function estimated for small values of *N* are shown. The lower bound for $\sigma(10)$ can be expressed by a number *a* whose value is approximately $\sqrt{8}$; $\sigma(10)$ is an exponentiated stack of *a*'s, where the number of *a*'s in the stack is expressed by another stack of *a*'s. The process of defining the height of one stack by another is carried out 10 times.

The TeleVideo IBM PC
The best hardware for

TeleVideo versus IBM. Make a few simple comparisons and you'll find there is no comparison.

RUNS IBM SOFTWARE.

With the TeleVideo® IBM Compatible line—PC, XT and portable computers—you'll get the most out of all the most popular software written for the IBM® PC—more than 3,000 programs.

Because every TeleVideo Personal Computer offers the highest level of IBM compatibility on the market

THE BEST HARDWARE FOR THE BEST PRICE.

Features	Tele-PC	IBM PC	Tele-XT	IBM XT
Monitor	YES	OPTIONAL	YES	OPTIONAL
Screen Size	14″	12″	14″	12″
Tilt Screen	YES	NO	YES	NO
Quiet Operation	YES (NO FAN)	NO	YES	NO
Memory	128K	128K OPTION	256K	256K OPTION
Graphics Display (640 x 200 resolution)	YES	OPTIONAL	YES	OPTIONAL
Printer Port	YES	OPTIONAL	YES	OPTIONAL
Communication Port	YES	OPTIONAL	YES	YES
MS™-DOS/BASIC®	YES	OPTIONAL	YES	OPTIONAL
System Expansion Slot	YES	YES	YES	YES
RGB and Video Port	YES	OPTIONAL	YES	OPTIONAL
Typical System Price	**$2995**	**$3843**	**$4995**	**$5754**

compatibles. the best software.

and has the standard—not optional—features you need to take full advantage of every job your software can do.

Study the chart at the left. It proves that TeleVideo—not IBM—offers the best hardware for the best price.

Note that TeleVideo's ergonomic superiority over IBM extends from fully sculpted keys and a comfortable palm rest to a 14-inch, no glare screen that tilts at a touch.

THE BEST MICROCHIPS.

What is perhaps most impressive about the TeleVideo IBM PC Compatible can be found deep within its circuitry. We use the same 8088 central processing unit that runs an IBM PC. But we also employ new VLSI (Very Large Scale Integration) microchips that are designed and built exclusively for TeleVideo.

These interface more efficiently with the powerful 8088 and yield numerous benefits.

For example, our tiny custom chips do the work of many of the larger, more expensive circuit boards in an IBM PC. So we can offer a computer system that comes in one attractive, integrated case, is ready to run and occupies less desk space. A computer that edges out IBM's added-cost component system for reliability, ease of service and purchase simplicity.

Fewer circuit boards to cool also allowed us to eliminate the noisy, irritating fan IBM and most other PCs force you to put up with. And TeleVideo compatibles accept any IBM hardware options without modification.

THE BEST PORTABLE FOR THE BEST PRICE.

Features	TPC II	COMPAQ
High Capacity Storage	YES	NO
2nd Disk Drive	YES	OPTIONAL
Quiet Operation (No Fan)	YES	NO
Ergonomic Display	YES	NO
Communication Port	YES	OPTIONAL
International Power Supply	YES	NO
MS™-DOS 2.11	YES	NO
Graphics Display	YES	YES
Typical System Price	**$2995**	**$3710**

THE BEST LINE.

But the Tele-PC is only one element of the TeleVideo IBM PC Compatible line.

The TeleVideo XT is the best hardware for users of popular IBM XT software who would appreciate an extra 10 megabytes of storage capacity along with the advantages listed on the preceding chart.

As the chart above demonstrates, our portable IBM compatible computer, the TPC II, is far and away better hardware than COMPAQ.™ Better hardware—standard—at a better price.

THE BEST MANUFACTURER.

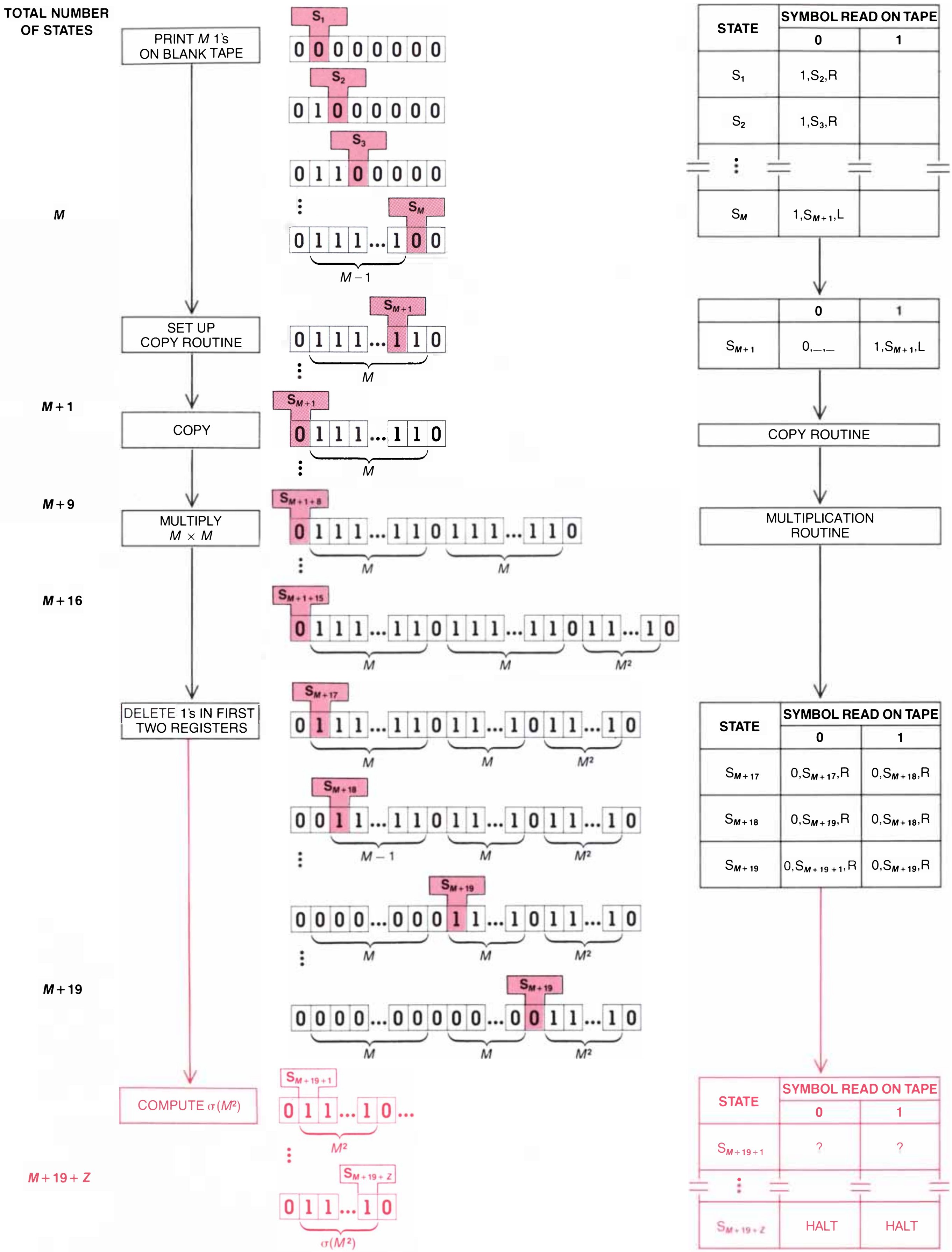
The TeleVideo IBM PC Compatible line is made by the world leader in multi-user computer systems and the number one independent manufacturer of terminals.

Our compatibles are available at participating ComputerLand and Entré (call 800-HI-ENTRE) dealers or you may call 800-538-8725 for the dealer nearest you. In California, call 800-345-8008.

Before you invest, make a few simple comparisons. You'll find that TeleVideo—not IBM or COMPAQ—has the best hardware for the best software. At the best price.

IBM is a registered trademark of International Business Machines. MS is a trademark of MicroSoft Corporation. GW Basic is a registered trademark of MicroSoft Corporation. COMPAQ is a trademark of COMPAQ Computer Corporation.

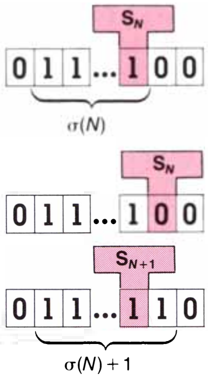
TeleVideo®
Personal Computers
TeleVideo Systems, Inc.



PROOF THAT $\sigma(N)$ IS NONCOMPUTABLE begins with an estimate of the number of states needed to generate a string of M^2 1's on a tape initially filled with 0's. The Turing machines for copying and multiplying, shown in the illustrations on pages 89 and 91, are combined to form a machine that has $M + 16$ states. Three additional states delete the two leftmost strings of 1's, leaving a string of M^2 1's. It is assumed there is a Z-state machine that, given any string of N 1's, will generate a string of $\sigma(N)$ 1's (*type and tables in color*).

the intuitive concept of computability. It is probably worth mentioning, however, that it is pointless to ask for a rigorous mathematical proof that a formal definition such as Turing's fully captures some originally intuitive notion.

First notice that σ is uniformly increasing; that is, if $X>Y$, $\sigma(X)>\sigma(Y)$. No matter how many 1's are printed by an N-state Turing machine that halts, an $(N+1)$-state Turing machine can always be constructed that will add one 1 to the string of 1's and then halt.



STATE	SYMBOL READ ON TAPE	
	0	1
S_N	$1,S_{N+1},L$	$1,S_N,R$
S_{N+1}	HALT	HALT

Suppose there is a Z-state Turing machine that computes $\sigma(M^2)$.
By the definition of σ and the calculation shown in the illustration on the opposite page,

$$\sigma(M+19+Z) \geqslant \sigma(M^2). \quad (1).$$

If M is large enough, however, statement (1) contradicts the fact that σ is uniformly increasing.
To demonstrate this, let $M = Z+20$, or $Z = M-20$.
Then $M+19+Z = M+19+M-20 = 2M-1$.
By elementary algebra $2M-1 = M^2-(M-1)^2$, and so $M+19+Z = M^2-(M-1)^2$.
Since $(M-1)^2 > 1$, $M^2-1 > M^2-(M-1)^2$.
Hence, because σ is uniformly increasing,

$$\sigma(M^2-1) > \sigma(M^2-[M-1]^2) = \sigma(M+19+Z) \geqslant \sigma(M^2).$$

Since $\sigma(M^2-1) > \sigma(M^2)$, however, σ cannot be uniformly increasing, and so the supposition that there is a Z-state Turing machine that computes $\sigma(M^2)$ leads to a contradiction.

FINAL STEPS IN PROOF that $\sigma(N)$ is noncomputable derive a contradiction from the assumption there is some Z-state Turing machine that, given a string of M^2 1's, computes $\sigma(M^2)$.

The Hilbert Program

In order to understand why Turing was so intent on defining computability it is necessary to have some sense of the history of mathematical logic before 1936. Rigorous mathematics as it is known today is a relatively recent development. The first serious attempt to reduce mathematical statements to statements in formal logic was begun by Gottlob Frege in 1879, with the publication of his *Begriffsschrift* ("The Notation of Concepts"). The problem posed by Hilbert in 1900 was therefore of direct significance to mathematics: if Frege's scheme could be carried to completion and if a method could be found for determining the truth or falsity of any statement in formal logic, then that method could also determine the truth or falsity of any mathematical statement, no matter how complex. If such a method could be found, mathematical conjectures such as Pierre de Fermat's "last theorem," which had resisted proof or disproof for centuries, would immediately be disposed of. An affirmative answer to Hilbert's bold challenge would reduce all mathematics to mechanical computation.

Two major developments in logic in the early decades of this century threw much of Hilbert's program into disarray. In 1901 Bertrand Russell discovered an irrefutable paradox in the elementary theory of sets, a theory that was essential to Frege's program of reducing mathematics to logic. Russell communicated his discovery to Frege just as the second volume of Frege's last major work, *Grundgesetze der Arithmetic* ("The Fundamental Laws of Arithmetic"), was about to be published. Frege ended the volume with a dispirited note: "A scientist can hardly meet with anything more undesirable than to have the foundation give way just as the work is finished. I was put in this position by a letter from Mr. Bertrand Russell when the work was nearly through the press." In spite of the flaws Frege was unable to remove from his work, Russell and Alfred North Whitehead were later able to salvage Frege's program and circumvent the paradox in set theory.

A second major discovery in logic was made by Kurt Gödel of the Institute for Advanced Study in Princeton, N.J. Implicit in Hilbert's program was the assumption that there must exist some method for distinguishing true statements from false ones in formal logic; the problem was to find the method. Gödel showed that the assumption is unjustified. In 1931 he proved that any consistent system of formal logic powerful enough to formulate statements in the theory of numbers must include true statements that cannot be proved. Because consistent axiomatic systems such as the one devised by Russell and Whitehead cannot encompass all the true statements in the subject matter they seek to formalize, such systems are said to be incomplete.

The Logic of Computability

Gödel's work effectively put an end to Hilbert's program. There can be no method for deciding whether some arbitrary statement in mathematics is true or false. If there were, the method would constitute a proof of all the true statements, and Gödel had demonstrated that within a consistent axiomatic system such a proof is impossible. The attention of logicians shifted from the concept of truth to the concept of provability. In this context there remained a simple analogue to Hilbert's question that had not been settled: Does a single method exist whereby all the provable statements in mathematics can be proved from a set of logical axioms?

The preeminent investigator of the logic of provability in the years immediately following Gödel's proof was Alonzo Church of Princeton University. Church and two of his students, Stephen C. Kleene and J. Barkley Rosser, developed a consistent formal language called the lambda calculus; it is useful for reasoning about mathematical functions, such as the square root, the logarithm and any more complicated functions that might be defined. (Lambda, the Greek letter corresponding to the Roman letter L, was chosen by Church to suggest that his formal system is a language.) The modern language for computer programming called Lisp (for list processing) is modeled on the lambda calculus. Kleene showed that large classes of mathematical functions, including all the functions employed by Gödel in his proof, could be expressed in the lambda calculus.

The next major step in this line of thought was taken by Church. He argued that if a mathematical function can be computed at all—meaning that it can be evaluated for every number in its domain of definition—then the function can be defined in the lambda calculus. Church's work showed that if there were such a thing as a mathematical function expressible in the lambda calculus that is not computable, there would be no method for determining whether or not a given mathematical statement is provable, let alone true. The final surviving hypothesis in Hilbert's program would be disproved. In April, 1936, Church published a logical formula that is not computable in his system.

Turing, working independently of Church, had also grasped the technical connection between Hilbert's problem and the idea of a computable function. In attacking the problem, however, he proceeded in a far more direct and concrete manner than Church. What was needed was a simple but precise model of the process of computation, and Turing's machines were designed to meet that need. Once their properties were specified, however, Turing made a brilliant connection between the idea of a computable function and the results of mathematical work done some 50 years earlier by the German mathematician Georg Cantor. Cantor had argued that although there is no largest whole number, any infinite set of objects that can be counted, or paired with the positive whole numbers, is a set of the same size as the set of whole numbers. Since any Turing machine can be expressed as a character string of finite length, all possible Turing machines and with them all computable functions can be listed in numerical or alphabetical order; hence they can be paired one for one with the whole numbers [*see upper illustration on page 88*]. There is, of course, no fixed upper limit to the size of a Turing machine, and so there is no limit to the number of possible Turing machines. Nevertheless, Cantor's analysis shows that the set of all possible computable functions is the same size as the set of all whole numbers; both sets are called countable sets.

Cantor had also shown there are infinite sets that are not countable; they are larger than the set of whole numbers in the sense that they cannot be paired one for one with the whole numbers. One example of such a noncountable set is the set of all the functions of the positive whole numbers that take on integer values. A careful analysis shows there must be more such functions than there are whole numbers. The implication is that not all functions are computable: there are not enough computer programs to compute every possible function.

The Halting Problem

Which functions are noncomputable? Unfortunately the proofs by Church and by Turing do not readily yield examples of noncomputable functions. In the past 20 years, however, computer scientists have exploited Turing machines to construct several such functions. One of the early examples of a noncomputable function was devised by Tibor Rado of Ohio State University in 1962. Consider all the Turing machines that have some fixed number of states N. Suppose all the machines begin their operations on a blank tape, or in other words a tape on which a 0 is marked in every frame. Imagine for the moment that all the Turing machines that never halt are excluded from this set. Among the remainder of the Turing machines, pick the machine or the group of machines that print the largest number of 1's in succession on the blank tape before they halt. That number of 1's, for each value of N, is the value of Rado's function; it is usually designated $\sigma(N)$. The detailed proof that $\sigma(N)$ is not computable proceeds by assuming it can be computed and then deriving a contradiction. The argument is straightforward, but the technical details are rather intricate; they can be found in the illustrations on the preceding two pages.

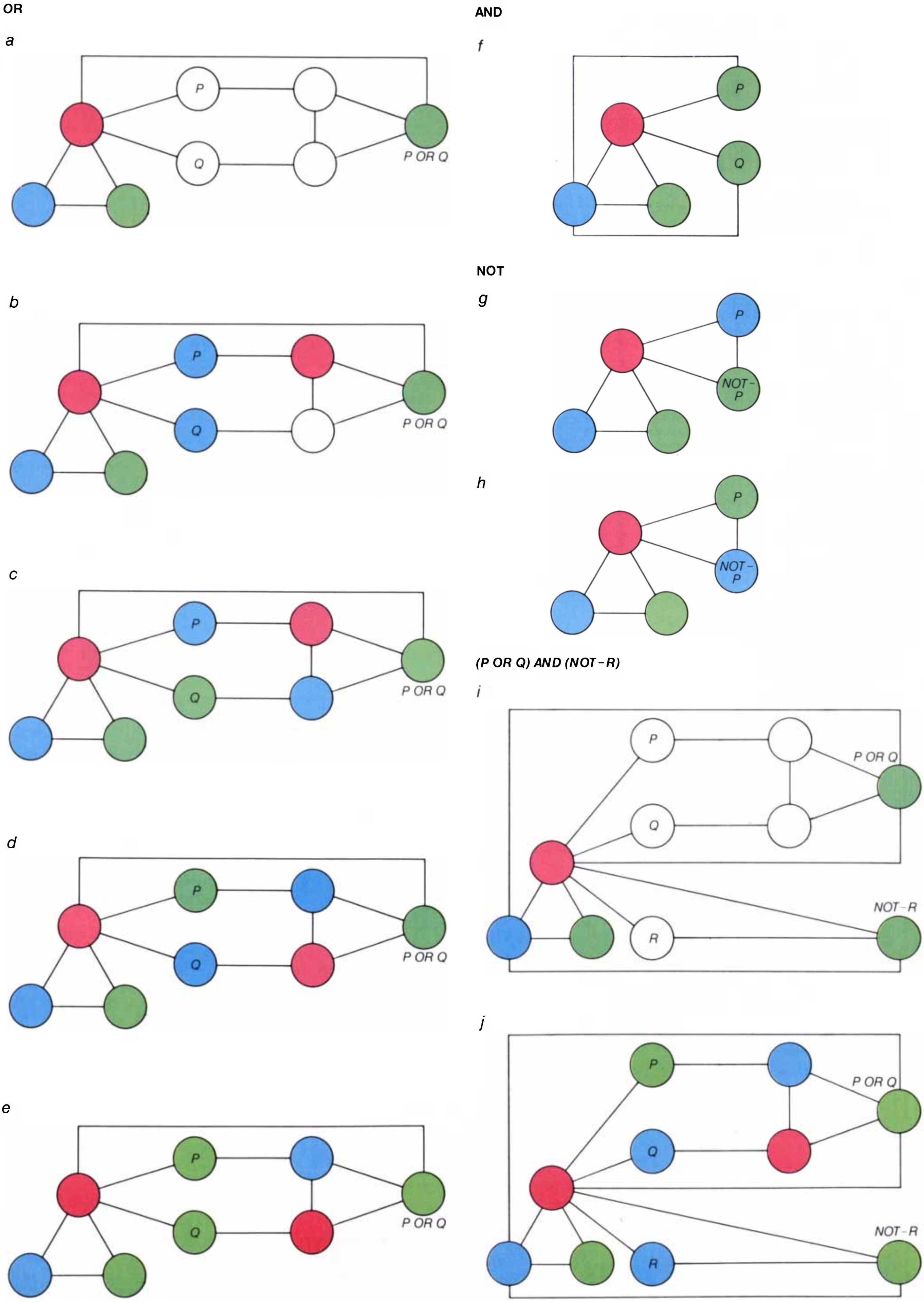
One might think the construction defective because it assumes the N-state Turing machines that do not halt can be sorted out in advance. The objection is a serious one. Consider, therefore, how one might attempt to compute $\sigma(N)$ by brute force. List all the N-state Turing machines in some numerical order, simulate each one on a universal Turing machine and select the machine or the group of machines that print the most 1's. Although this method of computation seems to avoid the objection, the difficulty with Turing machines that do not halt reappears in an intractable form. Some of the N-state machines that do not halt can be eliminated by simple algorithms, but there are other machines for which no such decision can be made. If one cannot determine that a particular machine does not halt, the machine cannot be eliminated from the list of N-state machines and the simulation must continue. Since the machine may actually never halt, there is no guarantee the computation of $\sigma(N)$ can be carried through to completion.

Although the early impact of the Turing machine was in logic, it has also played a dominant role in computer science since the early 1960's. In 1965 Juris Hartmanis and Richard E. Stearns, then at the General Electric Research Laboratories in Schenectady, N.Y., showed that the Turing machine can help to establish tight bounds on the complexity of computations. Subsequent investigators began to classify problems according to the way in which the running time, or equivalently the number of computational steps, varies with the size of the problem. For example, suppose some number of points N are interconnected by lines to form a graph of vertexes and lines. The problem is to color the vertexes in such a way that no two vertexes connected by a line have the same color. Suppose furthermore the fastest method known for solving the problem requires a time that varies as some power of N, say N^2. The problem is then said to be in the class of problems that can be solved in polynomial time, designated P. The class P has increased in importance as many computer scientists have come to regard all problems that are not in P as intractable.

Modern Complexity Theory

Note that a problem is assigned to class P only if no instance of the problem requires more than polynomial time to solve. In other words, the method of solution for the problem is deterministic, in the sense that it guarantees a solution in a time less than some fixed power of the size of the problem, N. A nondeterministic Turing machine can also be defined: it is allowed to solve a problem by guessing the answer and then verifying the guess. For example, to determine whether an integer is composite, the nondeterministic machine guesses a divisor, divides and, if the division is exact, verifies that the number is composite. The deterministic machine, on

TWO PROBLEMS ARE EQUIVALENT if the solution to one problem immediately gives the solution to the other. Here the problem of determining the conditions under which a complex Boolean formula, or logical proposition, is true has been mapped, or transformed, into another problem, that of coloring the vertexes of a graph with three colors in such a way that no two vertexes connected by a line are the same color. The truth or falsity of any Boolean formula depends on the truth or falsity of the atomic, or simple, propositions that make it up and on the ways the truth values are combined by the connectives "or," "and" and "not." For every possible Boolean formula a simple graph of lines and vertexes can be constructed, which can be colored with three colors if and only if truth values can be given to the atomic propositions in such a way that the complex formula is true. For example, consider the complex formula *P or Q*, which is made up of the atomic proposition *P* and the atomic proposition *Q*. The formula *P or Q* is true only if proposition *P* is true, proposition *Q* is true or both propositions are true. These conditions are reflected in the coloring of the graph at the upper left (*a*), in which green represents true and blue represents false. The coloring can be completed in such a way that the rightmost vertex is green (true) only if either one or both of the labeled vertexes are colored green (*b–e*). Similarly, the complex formula *P and Q* is true only if the atomic propositions *P* and *Q* are both true, and that state of affairs can be reflected in a graph in which the vertex representing *P* and the vertex representing *Q* are both colored green (*f*). Finally, the formula *not-P* is true only if the vertex labeled *P* is colored blue (*g*), and the formula *P* is true only if the vertex labeled *not-P* is colored blue (*h*). At the lower right of the illustration is the graph that corresponds to the more complex Boolean formula (*P or Q*) *and* (*not-R*); the upper coloring (*i*) represents the constraints imposed by the connectives "or," "and" and "not." At the bottom (*j*) one of the four ways to color the graph is shown; it corresponds to one way of satisfying the Boolean formula.


OR
a
P
Q
P OR Q
b
P
Q
P OR Q
c
P
Q
P OR Q
d
P
Q
P OR Q
e
P
Q
P OR Q
AND
f
P
Q
NOT
g
P
NOT-P
h
P
NOT-P
(P OR Q) AND (NOT-R)
i
P
Q
P OR Q
NOT-R
R
j
P
Q
P OR Q
NOT-R
R

the other hand, must search systematically for a divisor.

The time needed for solving a problem with a nondeterministic machine is measured by the length of the shortest computation, and so the nondeterministic machine would seem to have an enormous advantage over the deterministic one. Ordinary experience suggests it is easier to verify a solution than it is to find it in the first place. Nevertheless, no one has been able to prove that the problems solvable in polynomial time with a nondeterministic machine—the class of problems designated *NP*—are intrinsically any more difficult than the problems in the class *P*. Whether or not the class *P* is distinct from the class *NP*, which is called the *P-NP* problem, has become one of the major open questions in mathematics.

Important progress on the *P-NP* problem was made in 1970 by Stephen A. Cook of the University of Toronto. Cook was investigating the problem of determining the conditions under which a complex logical proposition is true. For example, the complex proposition formed when two simple propositions are linked by the word "or" is true if either one of the simple propositions is true or if both are true. In general it is quite difficult to determine the range of truth conditions for simple propositions that satisfy a complex proposition, or in other words make the complex proposition true. Cook was able to show that the problem, which is called the satisfiability problem, is as difficult as any other problem in class *NP*. There is an efficient algorithm for solving the satisfiability problem only if there is an efficient algorithm for solving every other problem in the class *NP*. Any problem having this property with respect to an entire class of problems is said to be complete for that class.

A year went by before most investigators grasped the significance of Cook's result. In 1971 Richard M. Karp of the University of California at Berkeley began to ask what other natural problems might play the same role as the satisfiability problem with respect to the class *NP*. Karp discovered that many important problems in operations research, including the problem of coloring a graph with three colors, are also as difficult as any problem that can be assigned to the class *NP*; that is, they are *NP*-complete. It can be shown directly, by mapping one problem into the domain of the other, that the graph-coloring problem and the satisfiability problem are equivalent [*see illustration on preceding page*].

It has now been proved in a similar way that several hundred problems, previously thought to be distinct, are actually notational variants of one another. All these problems are equivalent to the satisfiability problem and so all are *NP*-complete. Several other collections of such complete problems have been discovered, both for the class *P* and with respect to the classes of intractable problems for which the number of steps required for a solution on a Turing machine grows exponentially with the size of the problem. A direct assault on the *P-NP* problem, however, still seems premature. Some of the difficulty can be appreciated from recent developments in the theory of computation.

Relative Computability

The idea that a function is computable by a Turing machine can be extended by making computability depend on strings of symbols the machine may encounter on its tape. If a string found on the tape belongs to a previously specified set of strings *A*, the Turing machine can be instructed to advance to a special state that continues to compute the function in question. If the string does not belong to the set *A*, the machine delivers the decision that the function is not computable. The function is said to be computable relative to the set *A*. If the set *A* were made up of all the strings encoding Turing machines that eventually halt when they are given a blank tape as input, Rado's function $\sigma(N)$ would be computable relative to *A*.

In 1974 Theodore P. Baker of Cornell University, John Gill of Stanford University and Robert M. Solovay of Berkeley asked whether the nonequivalence of class *P* and class *NP* could be proved for relative computations. In the course of this work they made a startling discovery. They could specify two sets, *A* and *B*, for which the relations between class *P* and class *NP* are contradictory. In other words, for computations relative to set *A*, both *P* and *NP* are equivalent, whereas for computations relative to set *B*, *P* and *NP* are not equivalent classes. Moreover, it was discovered that for any formal system there are relative computations for which one can assume either that *P* and *NP* are equivalent or that they are not, without detriment to the consistency of the system. Other investigators have since found that many other problems can be relativized in such a way that either of two possible outcomes is true.

This perplexing state of affairs is obviously unsatisfactory as it stands. No problem that has been relativized in two conflicting ways has yet been solved, and this fact is generally taken as evidence that solutions to such problems are beyond the current state of mathematics. Nevertheless, one must remember that the mere formulation of these seemingly intractable problems is made possible by the simple solution to an impenetrable problem of an earlier generation. The next major advance may seem as simple in retrospect as A. M. Turing's imaginative machines.

ADVANCED TECHNOLOGY CENTERS 1984:

ECONOMIC DEVELOPMENT AND INDUSTRIAL CHANGE

The United States has passed through recession and the beginnings of readjustment. Declining demand, increased foreign competition and reduced capital investment still limit U.S. business and industry. As a result, the public sector is attempting to direct material and intellectual resources to areas with potential for additional employment and growth.

"For our economy to grow in a world of rapid industrial change, we must be flexible," said Commerce Secretary Malcolm Baldridge in a Washington address in March. "Our industrial managers cannot play 'catch-up'; they must leapfrog their competitors, and deal with the ingenuity of other nations."

Baldridge believes a "national industrial policy" is the wrong approach for the public sector. "It's protectionist legislation at heart," he said, "and it will lose jobs over time because it tries to enforce the status quo in a dynamic and ever more flexible world trade situation." Other industrial leaders agree in general; John Latona, vice president of Houdaille Industries, says, "America does not need a Cambridge don or a 'philosopher king' to decree which industries shall be supported by MITI-type industrial policy and which shall not." Says Brookings Institution senior fellow Robert Lawrence: "Who has the vision to know where the next resource dollar should be applied? That vision resides in the marketplace—or at best, as is increasingly the fact, in the states."

The states have long since individually developed "industrial policy" initiatives and programs "targeting" specific industries and geographic areas. Area economic or industrial development are other terms for industrial policy. An individual state, free from the expense of territorial defense and able to clearly assess its innate features and benefits, can encourage industrial growth productively. After study and experience, smaller geographic units such as cities and counties may be effective as well. Assistance has come from collective research conducted under the auspices of the National Governors' Association (NGA).

In 1983, the NGA Task Force on Technological Innovation reported on "Technology & Growth: State Initiatives in Technological Innovation." James B. Hunt, Jr., governor of North Carolina, chaired the task force; Governor Dick Thornburgh of Pennsylvania was vice chairman. The NGA report concluded: "States should seek to identify and report on the venture capital needs of the business community by involving business people and entrepreneurs in ongoing assessments of obstacles to innovative financing, current infrastructure and capital requirements, and initiatives for further state development." States, the report said, should facilitate contacts between business and private capital and, in general, act in part as information conduits. States should develop closer linkages between seed funding programs, incubator facilities and patent-aid programs to encourage startup and commercialization of new technology-based ventures.

Cover photo: This SmithKline Beckman Corporation facility, housing the headquarters of SmithKline Chemicals and a chemical production plant, is a sharp contrast of old and new. In the background are the remains of production facilities of the former Allenwood Steel Company, founded in about 1830. In the foreground, SmithKline's 1982-built state-of-the-art production facilities.

Space shuttle photograph courtesy of NASA.

Traditionally, states and localities have competed to attract industry from other areas; they often took for granted their existing industrial base. Lures including tax incentives, revenue bonds, buildings and land, training grants and developed industrial parks were commonly available only to the newly relocating firms. Contemporary practice credits the existing industrial base with both the motivation and the resources to expand and diversify, and successful new state initiatives are more and more directed to building on assets already in place.

The Office of Technology Assessment (OTA) is the arm of Congress that collects and publishes information regarding new industrial growth and change. In a recent background paper titled "Encouraging High-Technology Development" OTA identified six general categories of advanced-technology state undertakings:

Technology transfer: improved linkages between universities and industry;

Entrepreneurship: technical and management assistance;

Human capital: training and education;

Financial capital: tax aid and venture capital formation;

Physical capital: infrastructure improvements and research-science parks;

Information: assembly and dissemination.

The OTA report adds that advanced-technology industrial development is aided by the proximity of a strong research university, a skilled labor pool, formed capital, similar firms or corporate headquarters, transportation, and climate and similar life-quality amenities. "All are desirable or necessary factors," the report states, "but they are not always enough."

One excellent new source for use by top economic officials in the private and public sectors is the *Directory of Incentives for Business Investment and Development in the United States: A State-by-State Guide.* The directory is the product of research conducted jointly by the National Association of State Development Agencies, the National Council for Urban Economic Development and the Urban Institute. The *Directory* is published by the Urban Institute Press, 2100 M St. NW, Washington 20037.

During the first years of the 1980's, states looked on advanced-technology manufacture as a panacea for the problems caused by stagnation among what were considered "basic" industries: steel, equipment for heavy manufacturing, automotive. The debate over national industrial policy has more clearly defined American industry, allowing a rational approach at state and national levels. "All industry is technology," says Motorola chairman Bob Galvin. "'Advanced' is a matter of degree.

"If you quantify it, basic industry accounts for about 8 percent of U.S. output. You can say it's 6 percent or 10 percent—but that's the order of magnitude. But basic industry is also a market for other industry—consuming perhaps 12 percent of the output from that sector." AFL-CIO Industrial Union president Howard Samuel agrees: "Basic industry and high-tech are one and the same," he states. "It's much easier to determine the areas where they have much in common than the areas that distinguish them." Samuel cites the emergence of a newly vital, technology-using automotive industry in Michigan as one example.

State initiatives over the past several years have had the desired effect, and the professional-technical and top executive managers charged to evaluate sites for expansion are in a position to "feed back" with regard to the factors that have influenced their own decisions.

Studies made by the U.S. Congress, the National Governors' Association, the National Association of State Development Agencies, major accounting firms, the states themselves and various consultants closely agree on the factors that make regions attractive to high-technology industry.

Many states have begun to question the national and even local benefit gained by merely moving an established enterprise from one area to another—promoting "runaway" industry—often at great cost to taxpayers in both old and new locations.

To persuade an expanding company to move is not zero-sum. Expanding companies have many reasons to set up new production, research or regional administrative operations in new locations.

When all the studies and reports have been analyzed, one person says at last: "This is where we will go or stay."

THE MANAGEMENT VIEW

Usually such a decision is inscribed over the signature of the chief executive. What influences that decision, however, may differ among companies of varying size and maturity. The process reflects company dynamics, markets, raw materials, labor,

finance, transportation and many other factors that are becoming better defined.

"We obtained capital by producing a business plan. The plan says we'll double production every three years. That means we'll need new people, regularly, and I have to know we'll be able to get them. And all our people will be expected to work very hard *and* keep up to date in their fields. We'd better be somewhere that can happen," says an engineer in management.

Established and growing companies can and do rationally evaluate and decide on new sites. These firms are the proper target of state and regional development agencies. Their busy executives need all the aid they can find.

The smaller growing company may not have much interest in the traditional incentives. "Startup companies don't make much money for the first five years," says one company president, "so tax incentives don't help much. Rent incentives and job training would be useful to the mature company, but not to us."

States and localities in recent years have sought to create a climate that will, as much as possible, attract expanding advanced-technology companies. More important, they also strive to cultivate the fertile ground in which entirely new enterprises will take root. Those attributes that will lure a new industry inward should also persuade a native one to stay. Whether those conditions also help or cause new ones to begin is an open question.

North Carolina's Research Triangle Park, now nearly 30 years old, has been a standard for state-initiated advanced-technology industrial change and growth. Edward Moore is a beneficiary. Moore is cofounder and president of Wilmore Electronics in Durham, within the "Triangle." Moore obtained a doctorate in electrical engineering at Duke; while there, he helped to develop an idea for manufacturing computer power supplies for computer systems, along with a group of professors and other graduate students.

"Lots of companies like Wilmore end up where they are because there's an educational institution nearby," says Moore. "It was really the university situation that got things started. But without the Research Triangle, we would not have our airline connections or our pool of trained and skilled technical people. We don't have to educate local suppliers to our needs. We probably have the most comprehensive secondary education system in the nation. The state has been helpful in all kinds of ways. We're very pleased."

In order to promote further industrial growth, Michigan's bureaus of community development and economic development, formerly separate, have merged into the Office of Community and Business Development.

Director Carol Hoffman explains the restructuring: "We found we were duplicating services; interested people didn't know who to call. Now we have a very active outreach program where each geographic area of the state has direct contact with economic development specialists exclusively assigned to their region. Few of these communities are organized to handle economic development on their own.

"The merging," Hoffman continues, "has made us even more receptive to statewide needs and opportunities. Our 'Commerce Roadshow,' as it is affectionately called, is very well received by the individual communities."

Marketing specialist Mary Soper comments that Michigan's excellent research universities have helped the state to retain industry and jobs. "We find that people who are educated in Michigan want to stay here."

Vermont's Agency of Development and Community Affairs in Montpelier assists and advises advanced-technology industry in the state—both established firms such as IBM, which employs about 10,000 people in Essex Junction, and newer companies such as New England Digital Corporation. The Vermont Industrial Development Authority (VIDA) provides appropriate financial assistance.

New England Digital executive vice president Sydney Alonso worked with VIDA to move his fledgling company from Norwich to a former warehouse facility in White River Junction. "VIDA was extremely helpful," Alonso recalls. "We manufacture the Synclavier II™ digital music system, a production instrument for motion-picture scores and commercial use. It's very high-tech, and so a lot of things come together to make a proper environment for a firm like ours.

"Our key people are from Dartmouth, and all the people here would rather not be anywhere else," Alonso says. "We hear from people who'd like to come here from the cities—even from Route 128 and California, for that matter.

"One thing one really notices about Vermonters, at least those in this area. They are *really* hard-working people."

Cliff Williams runs International Sensor Systems from a former cow pasture in Aurora, Nebraska. His company is involved with hybrid thick-film technology used in computer disk drives and in solar cells. "I had some real qualms about moving here," Williams recalls, although he graduated from the University of Nebraska.

"That summer of 1972, I returned to visit friends and family. I got a call from Harold Edgerton, inventor of the stroboscope, a founder of EG&G and a native Nebraskan. He'd learned of me through a local banker.

"I became impressed with the attitude of the people—they were aggressively eager to learn, different from what I'd known back East. So I told Edgerton that, if he'd be on

my board, I'd start up in Aurora. He agreed, and here we are. Later, I joined the dean of engineering's advisory board at the university. Now we have three or four good courses in solid-state technology. That ties in with our personnel needs; we teach the corporate and technology structure."

Until recently, Oak Ridge National Laboratory in Tennessee was cloaked in secrecy as the Federal government researched classified atomic-energy applications. Now the laboratory regularly acquires patents, awards and recognition internationally. Passage of the Stevenson-Wydler Act of 1980 mandated that one-half of 1 percent of total Federal laboratory funding be used to transfer laboratory technologies to private industry.

Tennessee governor Lamar Alexander has not overlooked the opportunities such transfer offers, nor has Tennessee's established advanced-technology industry. A "technology corridor" now extends from Oak Ridge to McGhee Tyson Airport, with an interstate highway offshoot to Knoxville 10 miles to the east. Within the "corridor" are the laboratory, the University of Tennessee at Knoxville, the Tennessee Valley Authority, Boeing, Martin-Marietta, Westinghouse, EG&G ORTEC, Bechtel, Lockheed, Fluor and others.

The Tennessee Technology Foundation, a private, nonprofit corporation, coordinates industry with state and education resources. "We already have a critical mass of 30,000 technology-rich people in the 'corridor,'" says Dr. John Cruthers. "We see this in payrolls, we see it in enrollments for continuing education. We've inventoried over 80 spin-offs from ORNL and the university.

"The industry is here. More important, the people are here. We've got an approved $250 million bill for upgrading public education—so the people will continue to be here, for our next generation of knowledge-based industrial growth."

Like Tennessee, New Mexico benefits from "spillovers" caused by a national research laboratory—Sandia, in Albuquerque. More than 40 percent of Sandia's staff of 7,000 have advanced degrees in physics, chemistry and electrical engineering, creating a rare high-quality professional labor pool for surrounding industry. Separately, Albuquerque-based University of New Mexico is a resource for advanced-technology industry. The Albuquerque Vocational-Technical Institute offers a laser-electro-optics program to train and place technicians; it, too, assists industry in R&D as well as production processes.

Advanced-technology benefits from Federally funded projects are not limited to national laboratories. Montgomery County, in Maryland due north of the District of Columbia, has a high concentration of medical science installations, including the National Institutes of Health, the Food and Drug Administration, the National Library of Medicine and the Naval Medical Center. Supporting educational institutions such as Johns Hopkins are nearby. The county has created a 230-acre science complex, Shady Grove Medical Park, centered on existing hospital facilities, which expands the area's medical-technology resources. Although the state did not initiate the Shady Grove development, its economic and community development professionals have long believed that technology-rich people are what constitute a progressive center for advanced-technology industry. The Technology Extension Service makes professional people available and known to one another.

The park will soon house the county-initiated Center for Advanced Research in Biotechnology. Otsuka Pharmaceutical Company will construct a $7 million R&D center there. Microbiological Associates' headquarters and R&D operation will be constructed there as well. The county expects its center to "contribute to diverse sectors of the economy—agriculture, pharmaceuticals, commodity chemicals and health-care products."

Colorado's high-tech history began not with startups but with Hewlett-Packard's expansion beyond its Palo Alto base. "Most of Colorado's high-tech and medical-tech industry grew in a balanced economy," says Governor Richard D. Lamm. "Our economy has shifted away from the sustaining industries, agriculture, mining and tourism, toward the industries of the future. We compete for the future jobs while stabilizing the industries that have served us so well for so long...."

MiniScribe Corporation is one of Colorado's "industries of the future." It makes Winchester disk drives and was founded by Terry Johnson, a former storage technology professional, in 1980. "It was just logical to start the business where he lived," says engineering vice president Robert Ganter. "He converted his basement into the start-up operation."

Colorado is not considered the center of disk technology—a benefit, according to Ganter. "You don't hear all the reasons why this or that can't be done. Unlike the Valley, we have excellent staff stability. And there is good infrastructure in the front range now, with some 600 new technology companies in place. There are all sorts of people to do circuit boards, special machining.

"We recruit at the University of Denver, the University of Colorado and Colorado State University," says Ganter. "Governor Lamm isn't a drum beater, but he does promote the state. I can't tell you how many people we've hired who first walked in as tourists."

In Virginia, proximity to Washington together with ample labor and careful government planning have led to the development of an extensive advanced-technology sector. The college of engineering at Virginia Polytechnic Institute & State University in Blacksburg ranks as the nation's seventh largest; there are more than 10,000 engineering students attending colleges and universities within the commonwealth. Governor Charles Robb's Task Force on Science and Technology has announced a cooperative government-university-industry

CAPITAL FINANCING

Many of the states in the Northeast and Midwest have enormous numbers of well-trained and willing workers and also a new awareness that to grow they must compete with the rest of the world and the rest of the country.

More than most, these states know the cost of "runaway" industry. They have in recent years taken many steps to so shape their industrial climates that the industry they already have will stay home, to encourage the development of new industry and to make themselves attractive to expansion industry from other states and countries.

Many states now have some form of risk-capital encouragement designed to generate seed money for new enterprises, and Michigan is no exception. The legislature in 1982 authorized the use of the Michigan State Retirement Funds, currently some $350 million, as venture capital for equity investment. The hope is to generate additional matching funds from private venture capital firms. State funding will be limited to 40 percent of the total required.

Connecticut was one of the first states to take a risk-capital approach to development. The decade-old Connecticut Product Development Corporation (CPDC) provides existing companies with capital financing of up to 60 percent of the costs of new product development. However, in 1982 the state established the Connecticut Innovation Development program, which makes venture capital available.

Indiana typifies a different approach. The state's Corporation for Innovation Development (CID) is a private, for-profit corporation established by the legislature in 1981. CID, which has raised some $10 million in capital through the sale of equity, functions as a private venture capital firm. But it is empowered to grant up to $5 million in tax credits for private investments in its venture capital fund. Also, CID must donate 5 percent of its net income to Indiana state universities for R&D aimed at new technology applications.

Center for Innovative Technology to facilitate advanced-technology research.

Dr. Coleman Raphael, chairman of Atlantic Research Corporation in Alexandria, comments, "One of the highest-rated public school systems is here... it's a great pro-business state."

ETA Systems of St. Paul, Minnesota, a developer of supercomputers, spun off from Control Data Corporation, with staff and financing, in 1983. CDC owns 40 percent of the company. ETA conducted a nationwide site search before choosing St. Paul. Says president and chief executive officer Lloyd Thorndyke, "The design work would have been done here in any case. We were looking for another manufacturing site. We looked at California and at Texas, where Microelectronics & Computer Technology Corporation (MCT) had gone, because of resources offered there by the electronics industry and the two big universities." MCT was a Minnesota initiative, with input from Minnesota companies including CDC. "MCT had done a full national search," says Thorndyke, "and we had that data bank to work with. But in the final decision it was the locally available technology that convinced us."

Governor Dick Thornburgh of Pennsylvania has placed his state among the nation's leaders in attracting new advanced-technology industry. "The resources for building with new technology have always been here," Thornburgh explained at a Philadelphia reception in January. "We have 150,000 scientists and engineers at work in the commonwealth, perhaps because four of the country's top 50 graduate research universities are located here."

Pennsylvania's innovative Ben Franklin Partnership is designed to increase advanced-technology jobs by helping to commercialize research and development. James D. Pickard, Pennsylvania secretary of commerce, administers the partnership. "We prefer to speak of 'advanced technology' instead of 'high technology,'" Pickard says. "Although it may seem a minor distinction, 'high technology' frequently refers to small, new firms that 'spin off' into wholly new fields. 'Advanced technology' involves the scores of established major international corporations in Pennsylvania that 'spin in' innovative departments, divisions and subsidiaries, applying technology to the strengths and capabilities they already have."

As states and localities turn to advanced technology for the jobs of today and tomorrow, many organizations prepare studies and reports. These are designed to help government and industry understand advanced technology, examine ways in which it can be attracted and nurtured, detail what has actually been done and review the effectiveness of specific programs.

One of the more prolific producers of studies has been the Joint Economic Committee of Congress, under the direction of economist Dr. Robert Premus. In 1982, Dr. Premus produced a seminal paper, "Location of High-Technology Firms and Regional Economic Development," which was based on a survey of 691 companies. This report ranked the factors that entrepreneurial high-technology companies consider important; a key factor is "infrastructure," the necessary services and facilities already in place.

Dr. Premus's work, updated in a forthcoming paper called "Urban Growth and Technological Innovation," finds that in general high-technology firms will set up where support exists, support that may include an array of services ranging from a heliport to a reliable commuter railroad, from a cable TV system to a deep-water harbor to dependable and adequate electrical power.

Indeed, the local power company is a most important part of the infrastructure, not only because it provides energy but because it is an industrial development resource. Unlike "footloose" (Dr. Premus's word) high-technology companies, utility companies are tied to their service areas. They grow only when their areas develop. As industrial factors themselves, they have developed a feel for the area, land availability, local custom, taxes and the like.

Occasionally the utility may do more state promotion than the state. New Jersey's Public Service Electric and Gas Company puts it plainly.

"We're part of New Jersey. The vitality of our company is related to the economy of the state."

PSE&G has included Rutgers, Princeton, the New Jersey Institute of Technology and Stevens Institute in its planning for efficiency, environmental impact and new technology. "We instituted the Public Service Research Corporation, a subsidiary of PSE&G, back in 1977," says Stephen Mallard, PSE&G senior vice president for planning and research. "New Jersey has become a focus for technology from all over the world; we published one of our documents in Japanese. It's our job to use this technology, and to fuel it."

Michigan's Consumers Power Company has been most active in assisting several recent Michigan operations. The utility's director of economic development serves on the site selection committee for the state's new Molecular Biology Institute. The utility has helped GMF Robotics, a joint venture between General Motors and Fanuc, Ltd., of Japan that is currently site-seeking in America. The utility was also instrumental in attracting a new Weyerhaeuser company plant to Michigan.

INTERNATIONAL TECHNOLOGY PARTNERSHIPS

Queen's University of Belfast, Northern Ireland, first known as Queen's College, opened its doors in 1849—when Silicon Valley was still farm country. Now Queen's University's doors open to a microelectronics design laboratory and similar advanced-technology facilities, operated in conjunction with the Wolfson Signal Processing Unit.

The unit, now seven years old, coordinates industry programs in Northern Ireland with the university's staff and equipment. Commercially oriented and deriving its income from project work from industry, the unit has produced specialized items such as a 32 analogue channel, eight digital channel data logger for use in the Antarctic and a microwave radar alarm for protection of grounded aircraft.

The province's second university, the New University of Ulster at Coleraine, has developed equally close links with industry, particularly in surface physics research, and the microelectronics program based at the university promotes greater computer awareness and activity in primary and secondary schools.

Northern Ireland Polytechnic, through its Innovation and Resource Center, offers a troubleshooting team that, for example, designs specialized hardware or tests prototype products for clients.

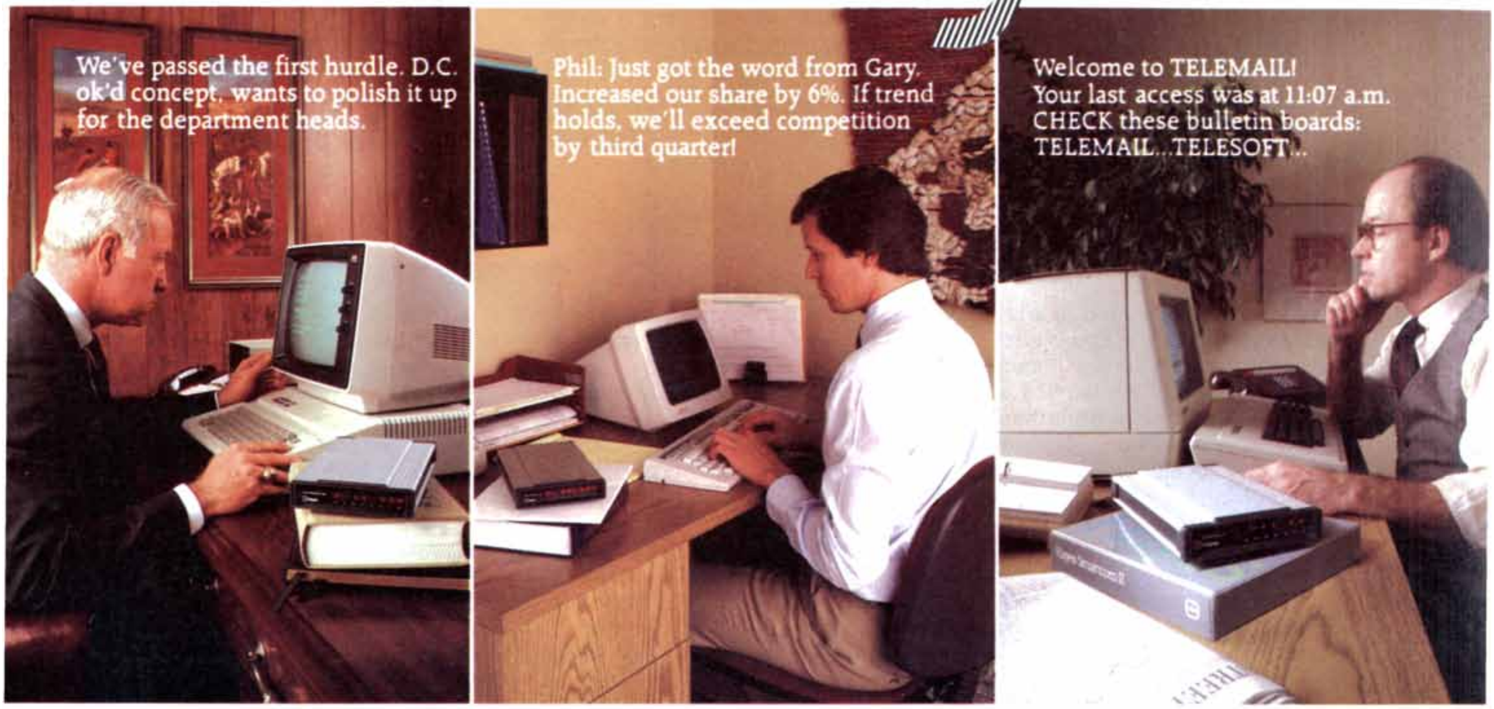
Supported by these universities, the electronics industry in Northern Ireland is sophisticated and vigorous. AVX, the world's largest manufacturer of multilayer ceramic capacitors, is so pleased with their Northern Ireland location that they announced a major expansion in October, 1983. During 1983, Standard Telephone and Cable in County Antrim received a $10-million radio-pager order from U.S.-based Tandy Corporation—an indication of the plant's advanced technology.

In recent years, Northern Ireland has also become a major world center for R&D in the field of medical technology. The first mobile coronary care unit was developed in the province, as was the first portable cardiac defibrillator.

The Industrial Development Board for Northern Ireland (IDB) has recognized that a partnership must exist between an expanding firm and its national host, a partnership that must continue past the site selection decision. The IDB offers available labor skilled in medical and industrial as well as information technologies, and it is broadly empowered to provide financial incentives. But Northern Ireland's most important attribute may well be the "intellectual capital" represented by its respected universities and skilled workers.

This second annual report on Advanced Technology Centers was written by Peter J. Brennan and produced by Development Counsellors International, Ltd., of New York. Graphic design: Sherin & Matejka, Inc.

Hayes™ Your computer's telephone.



Whether they're collaborating on competitive strategies or accessing data bases for valuable information, more and more corporate managers are communicating with each other. With networks, data bases and mail services all over the country. And for thousands of personal computer owners, we are providing the communications link.

Hayes Smartmodem. Think of it as your computer's telephone. Just plug Smartmodem into any phone jack, and watch it send messages to and from your computer. Over ordinary phone lines, and without using the telephone receiver.

Operating with rotary, Touch-Tone® and key-set systems, Smartmodem will dial, answer, and disconnect calls. Automatically.

Choose your speed; choose your price. The lower-priced Smartmodem 300™ is ideal for inter-office communications and local data swaps, transmitting at 300 bits per second. For longer distances and larger volumes, Smartmodem 1200™ communicates at speeds up to four times faster.

Both are compatible with many computers (*any* computer with an RS-232C port), and a variety of communications programs. However, if you have an IBM® Personal Computer, Hayes also provides the perfect companion software.

Smartcom II™ communications software. For the IBM PC.

We've spent a lot of time developing it, so you can spend less time using it.

Smartcom II makes quick work of the communications process—reducing to mere keystrokes what would take a lengthy series of entries on other programs. Because Smartcom II is menu driven, you simply respond to the screen's prompts. There are no complicated procedures to learn or troublesome instructions to unravel. Success comes swiftly. Simply!

And now, a complete plug-in telecomputing system for the IBM PC. Smartmodem 1200B™ All the great features that have made our freestanding Smartmodem a hit. On one, convenient plug-in board. And packaged with its own software. Smartcom II. Together, they're all you need to communicate—*fast!*

Backed by the experience and reputation of Hayes. Our record speaks for itself. Five years of solid leadership in the microcomputer industry. Nationwide availability through retail computer stores. Trouble-free factory service and call-in assistance. A limited two-year warranty on all hardware. And the most efficient telecomputing system available. Anywhere.

To appreciate the capabilities of Hayes Smartmodem and Smartcom II fully, see your dealer for a demonstration. Then get beyond "desktop" computing. Get a telephone for your computer.

Hayes Microcomputer Products, Inc.
5923 Peachtree Industrial Boulevard,
Norcross, Georgia 30092.
404/441-1617.

Hayes®

Smartmodem 300, Smartmodem 1200, Smartmodem 1200B and Smartcom II are trademarks of Hayes Microcomputer Products, Inc. Telemail Electronic Mail Service is a service mark of GTE Telenet Communications Corp. TeleSoft is a trademark of TeleSoft, Inc. Touch-Tone is a registered service mark of American Telephone and Telegraph. IBM is a registered trademark of International Business Machines, Corp. ©1983 Hayes Microcomputer Products, Inc.

Parasites That Change the Behavior of Their Host

In doing so they make the host more vulnerable to predation by their next host. Among such parasites are certain thorny-headed worms, which infest pill bugs that are later eaten by songbirds

by Janice Moore

One of the most familiar literary devices in science fiction is alien parasites that invade a human host, forcing him to do their bidding as they multiply and spread to other hapless earthlings. Yet the notion that a parasite can alter the behavior of another organism is not mere fiction. The phenomenon is not even rare. One need only look in a lake, a field or a forest to find it.

For a long time most ecologists considering the animal component of living communities have preferred to concentrate on free-living animals and to leave the study of parasites to parasitologists. Nevertheless, it was well known that many parasites have complex life cycles, spending their early life in one animal (the intermediate host) and reaching maturity in another (the definitive host). Gradually evidence began to accumulate that some parasites do not wait passively for a chance to reach their final destination but have ways of increasing the likelihood that the first host will fall prey to the second. They may do so by simply changing the intermediate host's size or color, making it more conspicuous. In other instances, however, they change the animal's behavior in a way that makes it more vulnerable. For example, as early as 1931 Eloise B. Cram of the U.S. Department of Agriculture noted that grasshoppers infested with the larvae of the nematode *Tetrameres americana* might be easy prey for chickens, in which the worm lives its adult life; the larvae encyst in the grasshopper's muscles and make it less active.

Some parasites, it was later found, change the behavior of their intermediate host by invading its central nervous system. Gid, a disease that makes ruminants such as sheep stagger in circles and become separated from the herd, is caused by an invasion of the animal's brain or spinal column by the larva of the canine tapeworm *Taenia multiceps.* Wolves and wild dogs, which prey on such ruminants, are the tapeworm's definitive hosts.

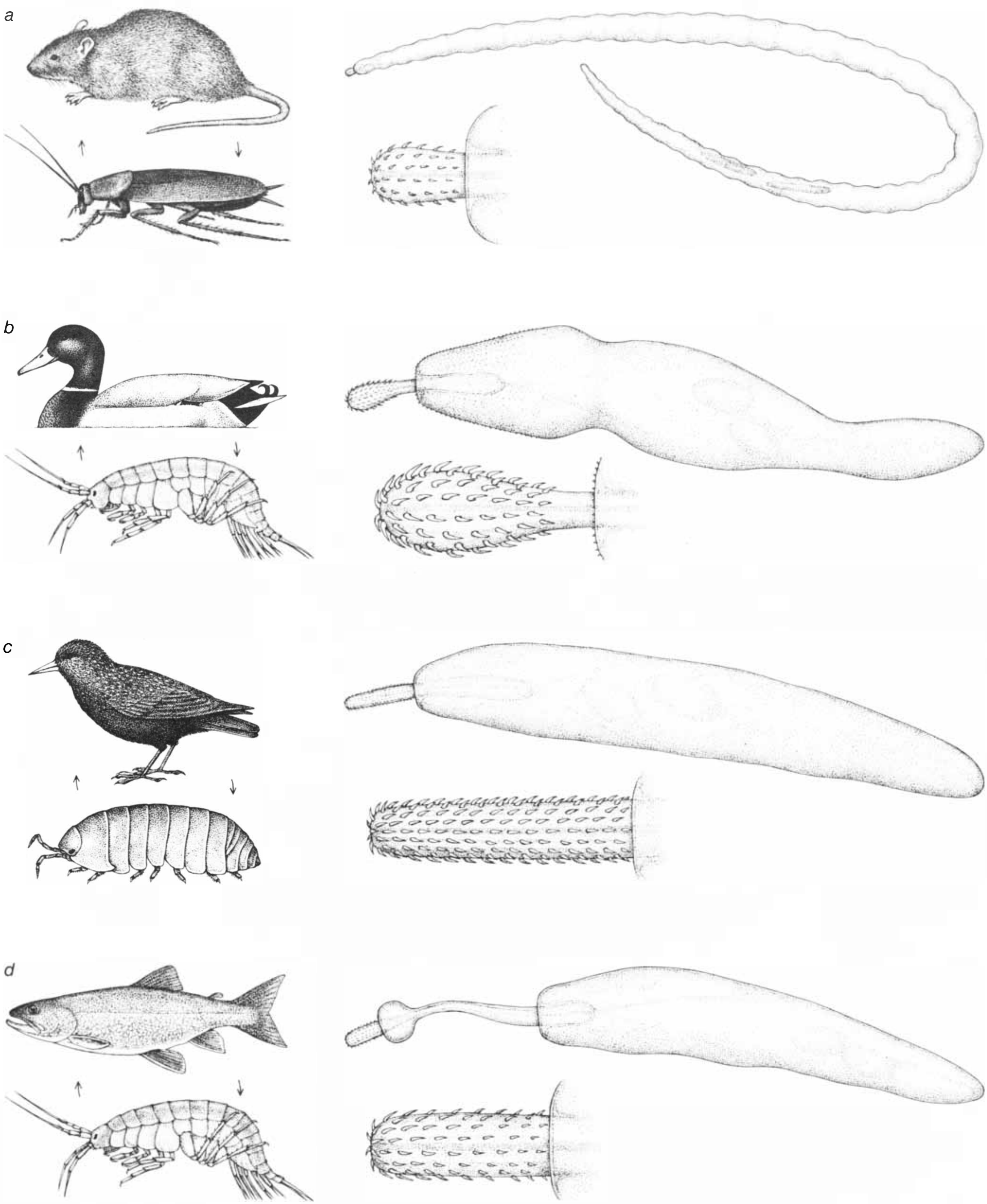
The lancet fluke *Dicrocoelium dendriticum* interferes with the behavior of its intermediate host in a more specific way. *D. dendriticum* matures in sheep but spends part of its early life in ants; it must thus overcome the fact that sheep do not ordinarily eat ants. When a group of immature worms invade an ant, one of them encysts in the subesophageal ganglion, the part of the nervous system that controls the insect's mouth parts and locomotion. Wilhelm Hohorst and his co-workers at Hoechst A.G. in Frankfurt, West Germany, showed that an infested ant crawls to the top of a plant and, if the temperature is low enough, becomes locked onto the plant by its mandibles. It is then likely to be eaten by a grazing sheep.

At least one group of parasites induce specific behavioral changes without damaging either the muscular or the nervous tissue of their host: the members of the phylum Acanthocephala, commonly called the thorny-headed worms. Acanthocephalans do not invade muscles or the nervous system, so that the mechanism underlying their behavioral effects is probably biochemical; little, however, is known about it. The ability of these organisms to elicit what would otherwise be normal behavior patterns from their hosts at inappropriate times has in recent years attracted the interest of a number of investigators. The resulting research, including my own, has demonstrated that this ability is present in members of all three classes of acanthocephalans. It may well be universal in the phylum.

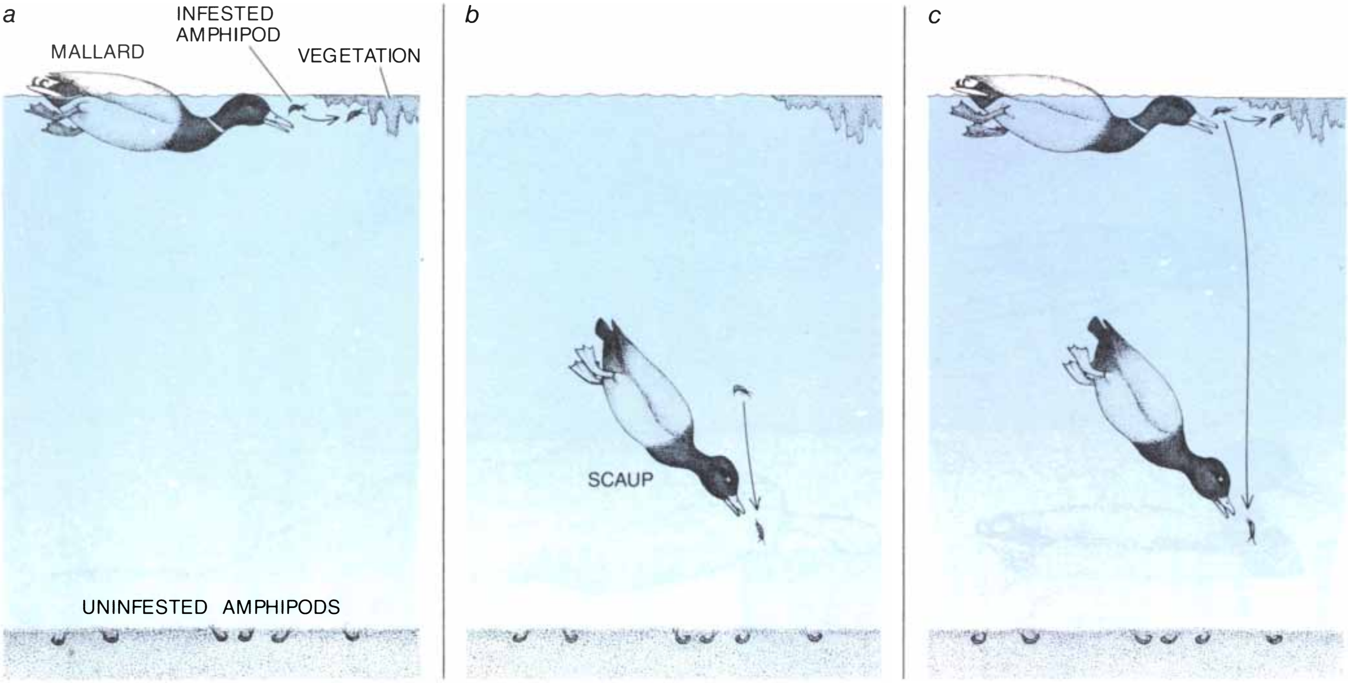
The number of acanthocephalan species is not precisely known; the highest estimate puts the total at less than 1,200. The few species for which life cycles are known all live as adults in the small intestine of vertebrates, particularly birds and fishes. There the female releases eggs that are excreted by the vertebrate and subsequently eaten by an intermediate host: an arthropod such as an insect or a crustacean. The egg hatches in the arthropod's intestine, and the larva burrows through the intestinal wall to the body cavity, where it develops to the stage called a cystacanth, which can infest vertebrates. When an infested arthropod is eaten by an appropriate vertebrate, the cystacanth takes up residence in the small intestine and develops to a sexually mature adult.

Morphologically, acanthocephalans are marvelously adapted to their parasitic way of life. The adult body is essentially a little bag of reproductive organs attached to a proboscis that is covered with backward-curved hooks. This thorny proboscis enables the organism to attach itself to the intestinal wall of the vertebrate. Acanthocephalans themselves have no intestine or any developmental remnant of one; they can survive without a digestive tract because they live in an environment that is rich in digested nutrients, which they simply absorb through their skin. How acanthocephalans evolved these morphological characteristics, and what their closest living relatives are, remain subjects of speculation—largely because no other organism looks very much like an acanthocephalan.

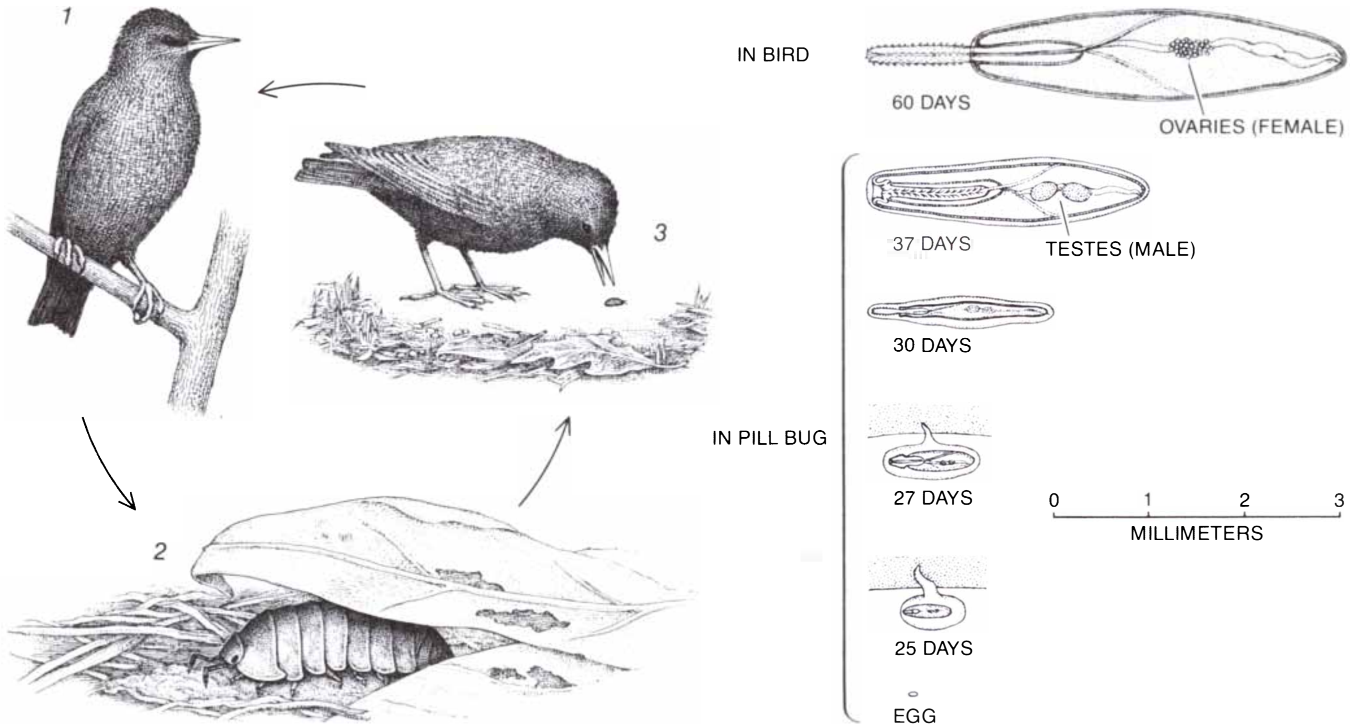
Given the worm's life cycle it is reasonable to suppose it may have evolved, through natural selection, traits that increase its chances of reaching a definitive host. In the 1970's William M. Bethel and John C. Holmes, working at the University of Alberta in Edmonton, did a series of experiments designed to test specifically whether acanthocephalans change the response of their intermediate host to environmental stimuli. They examined three species of worm that have amphipods (small aquatic crusta-



THORNY-HEADED WORMS, or acanthocephalans, are depicted along with their intermediate host (an arthropod), in which the larva develops, and their definitive host (a vertebrate), in which the parasite reaches sexual maturity. All four species depicted here have been found to change the behavior of arthropods such as crustaceans or insects. *Moniliformis moniliformis* (*a*) grows in cockroaches and then infests rats that eat the roaches. *Polymorphus paradoxus* (*b*) is transmitted from amphipods (small aquatic crustaceans) to mallards and other predators. *Plagiorhynchus cylindraceus* (*c*), whose behavioral impacts on pill bugs were studied by the author, matures in starlings and other songbirds that eat pill bugs. *Pomphorhynchus laevis* (*d*) moves from amphipods to fishes such as trout. The thorny probosces with which adult worms attach themselves to the vertebrate intestinal wall are drawn to a common scale; in life the proboscis of *Plagiorhynchus* is about a millimeter long. The worms themselves vary considerably in size: *Moniliformis* can be as much as 30 centimeters (12 inches) long, and the other species are less than a twentieth that size. They have reproductive organs but no digestive tract.



DIFFERENT BEHAVIORAL IMPACTS on amphipods of three acanthocephalan species were demonstrated in a laboratory experiment conducted by William M. Bethel and John C. Holmes of the University of Alberta. Uninfested amphipods avoid light; when they are disturbed, they burrow into the mud at the bottom. Amphipods infested with *Polymorphus paradoxus* (*a*) move toward light, and when they are threatened, they tend to cling to floating vegetation or skim along the surface; there they are eaten by dabbling ducks such as mallards. Although crustaceans infested with *Polymorphus marilis* (*b*) also prefer lighted areas, they do not go to the surface; they are preyed on by diving ducks such as scaups. Amphipods harboring *Corynosoma constrictum* (*c*) do swim to the surface, but some of them dive when they are disturbed; their predators are dabbling as well as diving ducks. The worms mature in the duck's intestine.

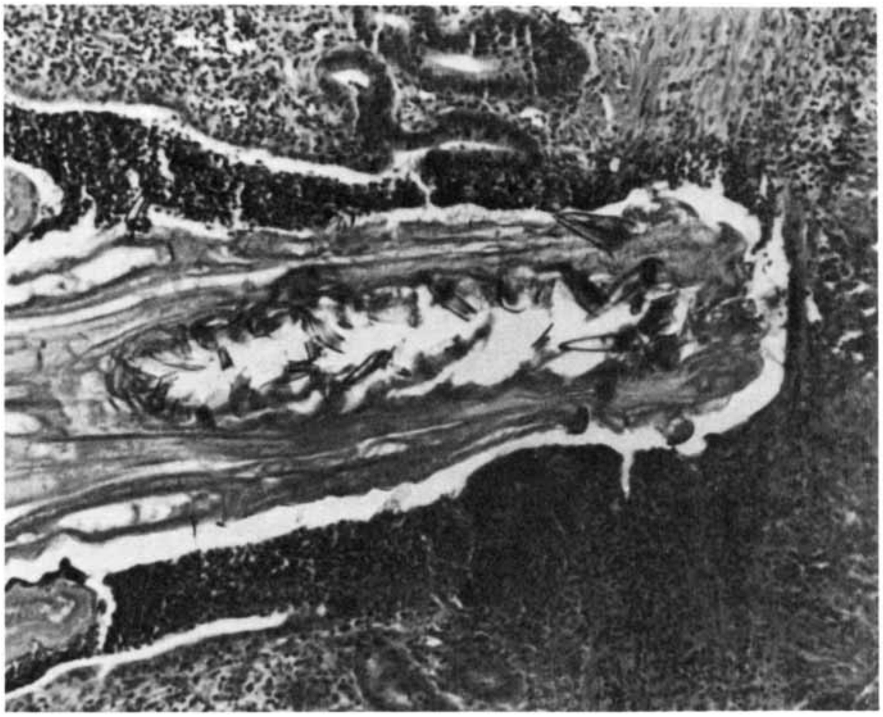


LIFE CYCLE OF PLAGIORHYNCHUS CYLINDRACEUS begins in the small intestine of a starling, where the adult female lays eggs that are excreted by the bird (*1*). When the bird feces are eaten by a pill bug or some other suitable isopod (*2*), the eggs hatch within a few hours. The larva, initially no more than a tenth of a millimeter long and armed with tiny spines, burrows through the wall of the isopod's gut and eventually drops into the body cavity, remaining attached to the intestine by a stalk. In the body cavity the larva greatly increases in size and develops adult organs. From 60 to 65 days after hatching it is about four millimeters long and can infest a bird. Through a mechanism that is not yet understood the larva changes the behavior of a pill bug; for example, infested pill bugs are less likely to seek shelter from predators under leaf litter. The thorny proboscis usually remains inverted inside the body of the worm until the pill bug is eaten by a bird (*3*), at which time the proboscis evaginates and attaches to the bird's intestine. The parasite continues to grow, eventually becoming as much as 15 millimeters long. *P. cylindraceus*' life cycle was studied by Gerald D. Schmidt and O. Wilford Olsen.

ceans) as intermediate hosts but that reach adulthood in different vertebrates. They found each species induces different behavioral changes that increase the likelihood of the amphipod's being consumed by the appropriate type of definitive host.

Uninfested amphipods move away from light and are rarely at the surface of a pond or lake. When they are disturbed, they dive and burrow into the mud at the bottom. In contrast, Bethel and Holmes showed that amphipods infested with cystacanths of *Polymorphus paradoxus,* which lives as an adult in the small intestine of mallards, beavers and muskrats, move toward light. When the infested amphipods are disturbed, they skim along the surface, sometimes clinging to vegetation or other floating objects. This makes them likelier to be eaten by surface-feeding predators—such as mallards, beavers and muskrats. Crustaceans infested with *Polymorphus marilis,* on the other hand, move toward light but do not go all the way to the surface; the definitive hosts for this parasite are diving ducks such as the lesser scaups. Finally, amphipods harboring *Corynosoma constrictum* move toward light, but more than half of them dive when they are disturbed. *C. constrictum* matures in both diving and surface-feeding ducks.

In tests conducted with mallards and muskrats in a laboratory tank Bethel and Holmes confirmed that the observed changes in behavior affect feeding patterns. Compared with uninfested control animals the surface-feeding mallards ate a large number of amphipods containing *P. paradoxus,* a smaller but still significant number of amphipods infested with *C. constrictum* and no amphipods containing *P. marilis,* whose intermediate hosts avoid the surface. Muskrats ate a significant number of amphipods containing *P. paradoxus* and none containing *P. marilis.* The investigators also found that the changes in amphipod behavior only occur once the acanthocephalan has reached the cystacanth stage, when it can infest vertebrates. If the amphipod were eaten at an earlier stage, the immature parasite would not survive.



THORNY PROBOSCIS of *P. cylindraceus* is seen embedded in the intestinal tissue of a starling in this photomicrograph. The damage caused by the parasite is generally local, confined to within 90 micrometers of the proboscis, which in life is about one-quarter millimeter wide.

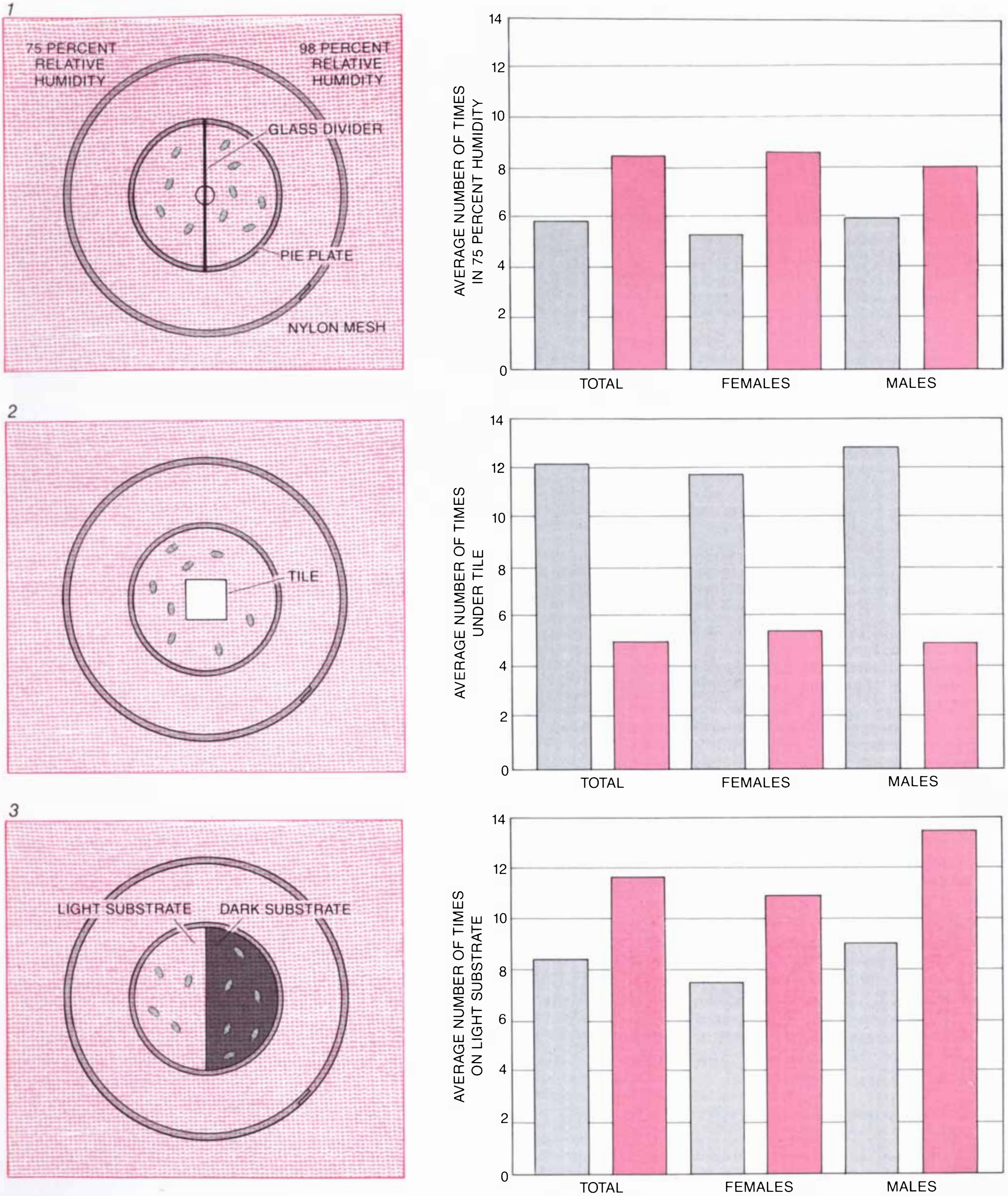
By 1979 seven species of acanthocephalans, all in the class Palaeacanthocephala, had been shown to alter the behavior of five species of aquatic crustaceans, both amphipods and isopods. Laboratory predation tests on six of these parasites had in every case confirmed that infested crustaceans were likelier to be eaten than uninfested controls. I wanted to extend this work to a terrestrial palaeacanthocephalan and for the first time to examine a parasite's effects on predation in the field as well as in the laboratory. In addition, by looking at an acanthocephalan thought to cause disease in its definitive host (none of the species studied previously were known pathogens) I hoped to determine whether the host might develop a way of avoiding the parasite.

Plagiorhynchus cylindraceus, cited in anecdotal reports and some textbooks as being highly pathogenic for songbirds, seemed to be the worm I was looking for. As an adult it often lives in starlings, and its intermediate hosts are terrestrial isopods such as the common pill bug. There was good reason to believe *P. cylindraceus* might alter pill bug behavior: in an eight-millimeter pill bug its cystacanth may grow to almost three millimeters in length and one millimeter in width. Moreover, Gerald D. Schmidt of the University of Northern Colorado had reported that female pill bugs infested with this parasite do not develop ovaries. Finally, field surveys had shown that more than 40 percent of birds were infested with *P. cylindraceus* in areas where the worm's prevalence in isopods was less than 1 percent. Since isopods are not known to be a significant part of a bird's diet, these results were hard to explain without assuming that the acanthocephalan had some way of making isopods more attractive to birds.

To examine pill bug behavior in the laboratory I made a simple chamber out of two pie plates. The bottom plate, containing a saturated salt solution that generated the desired relative humidity, was covered with a nylon mesh on which the pill bugs could move around. The inverted top plate closed the chamber, and the junction was sealed with weather stripping. All the experiments were done at the same temperature of 24 to 25 degrees Celsius (75 to 77 degrees Fahrenheit).

I then fed the pill bugs pieces of carrot covered with *P. cylindraceus* eggs. After waiting three months for cystacanths to develop I mixed infested animals with uninfested controls. The experiments were conducted blind: infestation does not change the isopods' dark color, and so I did not know which animals contained *P. cylindraceus* until I dissected them after observing their behavior in the experimental apparatus. In order to keep track of more than one pill bug at a time I marked the pill bugs with paint of different colors. In a series of experiments I then examined the impact of *P. cylindraceus* infestation on four types of behavior that may be important in determining whether a pill bug falls prey to a bird: the isopod's reactions to humidity, shelter, light and color of substrate (the surface on which it rests).

First I gave the pill bugs a choice between relative humidities of 98 percent and 75 percent, which I generated on opposite sides of a divided chamber with different salt solutions. The pill bugs could move freely under the glass divider because it did not quite touch the nylon mesh. At one-minute intervals over a period of about half an hour I



CHANGES IN PILL BUG BEHAVIOR were observed by the author when she compared pill bugs infested with *P. cylindraceus* (*color*) with uninfested controls (*gray*) in a series of laboratory experiments. The "behavioral chamber" consisted of a pie plate with a nylon mesh stretched over it and a second plate inverted on top of the first. After an acclimation period groups of 10 individually marked pill bugs were allowed to move about freely on the nylon mesh for half an hour, and each isopod's location was recorded at intervals of one minute. Pill bugs were first allowed to choose between relative humidities of 98 percent and 75 percent (*1*). Infested isopods were on the low-humidity side of the divider for an average of 8.5 out of 30 observations, uninfested controls for fewer than six observations. An even more striking difference was found in the isopods' response to a raised tile that simulated a shelter (*2*): uninfested pill bugs were under the tile, which covered 9.5 percent of the mesh, for more than 12 out of 30 observations, whereas infested pill bugs were under the shelter only slightly more often than one would expect from random motion. Finally, infested isopods were found more often on light substrate (*3*), where they were notably conspicuous. Pill bugs moving in open areas or on light substrate are more likely to be eaten by birds.

checked the location of each isopod. Infested pill bugs were found on the low-humidity side much oftener than the uninfested controls.

Ordinarily pill bugs will desiccate and eventually die when the relative humidity is less than 98 percent. Since infested pill bugs display less aversion to lower humidity, *P. cylindraceus* may either increase their ability to withstand dryness or impair their ability to perceive humidity, or both. The mechanism by which terrestrial isopods perceive moisture is not known, so that it is impossible to say how the parasite might alter it. Conceivably *P. cylindraceus* could enable an isopod to withstand drier environments by somehow decreasing the permeability of its skin, thereby reducing its water loss, but the results of experiments I did to test this hypothesis were inconclusive. One conclusion, however, seems reasonable: pill bugs that frequent dry areas might spend more time in exposed locations, where they would run a greater risk of being preyed on by birds.

Next I decided to test directly the isopods' response to shelter. I placed a tile on four pebbles in the center of the nylon mesh and recorded the location of each animal, again at one-minute intervals. The tile covered 9.5 percent of the area of the nylon mesh, so that one would expect a randomly moving pill bug to be under the shelter during approximately three of 30 observations. Uninfested pill bugs were under the tile much more often, suggesting they somehow perceived and favored the shelter, but infested animals were not, suggesting they were ignoring the tile. Again the physiological mechanism underlying the change is not known. Strangely enough, it does not seem to be caused by a different response to light; when in a separate experiment I darkened half of the chamber and allowed the pill bugs to move around, I found no difference between the behavior of infested animals and that of uninfested ones. Whatever the mechanism, a random response to shelter is clearly no way for an animal to avoid predators.

Finally, I tested the pill bugs' reaction to substrates of different colors. Unlike some of their aquatic relatives, parasitized terrestrial isopods have normal pigmentation, but a change in their choice of substrate could make them just as conspicuous as a change in their own color. A dark pill bug is much more visible on a sidewalk, for example, than it is on dark soil. I re-created this choice in the laboratory with aquarium gravel, black on one side of the pie plate and white on the other. The results were quite clear: infested isopods were on the white gravel far more often than uninfested ones. Since a dry substrate is often lighter than a wet one, the changes in humidity response and choice of substrate may even be mutually enhancing.

The laboratory experiments thus showed that *P. cylindraceus* alters the behavior of pill bugs in ways that make them likelier to encounter birds. An infested isopod spends more time away from shelter and on a light substrate. The next step was to find out whether these behavioral changes improve the parasite's chances of being transmitted to a starling.

With the assistance of Danny Bell and Mike Freehling I conducted field studies showing that they do. It would have been difficult to answer the question by observing individual adult starlings in the wild and then trapping and dissecting them. Instead I chose a more docile group of experimental subjects: nestlings. I provided nest boxes for starlings on the University of New Mexico campus. During the breeding season I tied pipe cleaners around the young starlings' neck for periods of no more than two hours, tight enough to keep them from swallowing prey but not tight enough to choke them. By collecting the prey I could estimate the rate at which parents were delivering pill bugs to their young (about one every 10 hours) and also could make sure that the nestlings were not acquiring any *P. cylindraceus* worms from prey other than pill bugs. At the same time I collected 250 pill bugs from the vicinity of the nest boxes, where the adult starlings foraged. From this sample I determined that approximately .4 percent of the pill bug population in the area were harboring *P. cylindraceus* larvae.

When I dissected 22 of the nestlings, however, I found at least one *P. cylindraceus* in seven of them, or 32 percent. Given the prevalence of the parasite in the pill bug population, the rate of pill bug delivery and the fact that none of the nestlings was more than 18 days old, only about half that number of nestlings should have contained worms if the parents were foraging randomly, that is, if the behavioral changes were not making the infested pill bugs more attractive prey. Moreover, the pill bug delivery rate was in most cases less than one every 10 hours, and most of the nestlings were less than 18 days old. If I had used averages for these factors rather than maximums, the parasite transmission rate estimated for random foraging would have been even lower. The actual transmission rate would have been higher if I had not counted as single infestations instances in which a nestling harbored more than one worm. Thus the above estimate of the difference between the observed transmission rate and the rate attributable to random foraging is conservative, and the evidence that *P. cylindraceus* influences pill bug predation by starlings is probably stronger than I have suggested.

Similar results were obtained in three subsequent field studies in which I dissected a total of 96 nestlings and 3,000 pill bugs. To buttress this evidence and to observe more directly the starling's response to the changed behavior of infested pill bugs I also did a laboratory predation test. I covered half of the bottom of a container with moist black sand and the other half with drier white sand. Then I put equal numbers of infested and uninfested pill bugs in the container. After allowing the pill bugs to wander around overnight I presented

PARASITE SPECIES	INTERMEDIATE HOST	BEHAVIOR CHANGE	DEFINITIVE HOST
ARCHIACANTHOCEPHALA			
MONILIFORMIS MONILIFORMIS	*PERIPLANETA AMERICANA* (COCKROACH)	HYPERACTIVE; MOVES TOWARD LIGHT.	RAT
PALAEACANTHOCEPHALA			
POLYMORPHUS PARADOXUS	*GAMMARUS LACUSTRIS* (AMPHIPOD)	MOVES TOWARD LIGHT; CLINGS TO VEGETATION AT SURFACE EVEN WHEN DISTURBED.	SURFACE FEEDERS: MALLARD, MUSKRAT, BEAVER
POLYMORPHUS MARILIS	*GAMMARUS LACUSTRIS*	PREFERS LIGHTED AREAS BUT DOES NOT GO TO SURFACE.	DIVING DUCKS: SCAUP
CORYNOSOMA CONSTRICTUM	*HYALELLA AZTECA* (AMPHIPOD)	MOVES TOWARD LIGHT AND GOES TO SURFACE; SOME ANIMALS DO NOT DIVE WHEN DISTURBED.	SURFACE FEEDERS AND DIVING DUCKS: MALLARD AND SCAUP
POLYMORPHUS MINUTUS	*GAMMARUS LACUSTRIS*	MOVES TOWARD LIGHT.	SURFACE-FEEDING DUCKS
PLAGIORHYNCHUS CYLINDRACEUS	*ARMADILLIDIUM VULGARE* (PILL BUG)	SPENDS MORE TIME IN LOWER HUMIDITY, AWAY FROM SHELTER AND ON LIGHTER SUBSTRATE.	SONGBIRDS
POMPHORHYNCHUS LAEVIS	*GAMMARUS PULEX* (AMPHIPOD)	SPENDS MORE TIME IN OPEN WATER; MOVES TOWARD LIGHT AND RESTS ON SURFACE VEGETATION; SWIMS IN A SPIRAL.	FISH: TROUT
ACANTHOCEPHALUS DIRUS	*ASELLUS INTERMEDIUS* (AQUATIC ISOPOD)	MORE ACTIVE.	FISH: CREEK CHUB
ACANTHOCEPHALUS JACKSONI	*LIRCEUS LINEATUS* (AQUATIC ISOPOD)	MORE ACTIVE; CRAWLS OVER RATHER THAN UNDER OBSTACLES.	FISH
EOACANTHOCEPHALA			
NEOECHINORHYNCHUS CYLINDRATUS	*PHYSOCYPRIA PUSTULOSA* (OSTRACOD)	GOES TO SURFACE.	FISH: MOSQUITO FISH
OCTOSPINIFEROIDES CHANDLERI	*CYPRIDOPSIS VIDUA* AND *PHYSOCYPRIA PUSTULOSA* (OSTRACODS)	GOES TO SURFACE; PREFERS LIGHTED AREAS.	FISH

ELEVEN SPECIES OF ACANTHOCEPHALAN, including members of all three classes of the phylum, have been found to alter the behavior of their intermediate hosts. In every case the change appears to make the intermediate host easier prey for the definitive host.

the container to a captive starling and watched until it had eaten approximately half of them. The black pill bugs were very conspicuous against the white sand, and in repeated experiments the birds consistently took most of their prey from that area. In doing so they ate significantly more infested pill bugs than uninfested ones.

The fact that *P. cylindraceus* is good at increasing the probability that it will get transmitted from a pill bug to a bird raises an interesting question: Given that the parasite is supposed to be pathogenic, why does a starling eat infested pill bugs and feed them to its young? It may be that the bird cannot distinguish infested pill bugs from uninfested ones. (The two look alike to human beings.) But then one might ask why a starling does not totally avoid pill bugs, which are ordinarily an insignificant part of its diet. The answer, I believe, is that although *P. cylindraceus* may have adverse effects on some birds, it is not usually, contrary to anecdotal reports, a pathogen.

Weight loss is a general indication of ill health, but I found no relation between weight and the presence or number of *P. cylindraceus* worms, either in the nestlings I examined or in 103 adult starlings I trapped. Furthermore, when infested bird tissue is examined under the microscope, there is no sign of damage more than 90 micrometers from the acanthocephalan's proboscis. If the parasite causes disease only rarely, or if its adverse effects are usually minimal, it may not be worthwhile to a starling to forgo eating pill bugs. Along with the fact that deviant behavior makes infested pill bugs easier prey, that would explain why *P. cylindraceus* is present in so many starlings.

The pill bug, of course, pays a higher price for harboring *P. cylindraceus.* Not only is it likelier to be eaten but also, if it is female, it develops no ovaries and therefore does not reproduce. The low incidence of the parasite in natural pill bug populations may mean that the isopods are avoiding the parasite, or it may simply reflect the fact that infested pill bugs are being eaten. To determine whether pill bugs, particularly females, make any effort to avoid consuming acanthocephalan eggs, I dried and weighed bird feces and gave two pieces to each of 26 females. One piece was covered with an aqueous suspension of *P. cylindraceus* eggs, the other was simply rehydrated. After a week I dried and weighed the remaining pieces. The pill bugs showed a tendency to shun the "infested" feces, but the result was not statistically significant, and in any case the tendency did not seem dramatic given the penalties involved. Perhaps pill bugs, as omnivores, are not capable of being very discriminating. It may even be that the benefits of eating nutrient-rich bird feces outweigh the risk of being eaten by birds.

The ability to alter host behavior is not limited to palaeacanthocephalans such as *P. cylindraceus* but extends to the other two classes of acanthocephalans. David J. DeMont and Kenneth C. Corkum of Louisiana State University recently found the ability in two eoacanthocephalans, both of which mature in fishes and have as intermediate hosts ostracods (another type of aquatic crustacean). The two parasites probably make ostracods more vulnerable by increasing their attraction to light.

Cockroaches infested with *Moniliformis moniliformis,* a member of the third class of acanthocephalans (Archiacanthocephala), are also attracted to light. Furthermore, they are hyperactive, as I discovered when I placed infested animals in activity wheels connected to devices that record the frequency of motion. Hyperactive cockroaches that move toward light are less likely to remain hidden and so are probably more likely to be caught by rats, which are the definitive hosts for *M. moniliformis.*

The effects these acanthocephalans and others have on their hosts are not as magical as they might sound. For one thing, the parasites do not induce novel behavior patterns but merely elicit existing patterns at disastrously inappropriate times. Nevertheless, this is quite a feat, and a general physiological explanation of how an acanthocephalan accomplishes it while floating in the body cavity of the host has yet to be found.

The realization that parasites can change host behavior has intriguing implications. Biologists observing certain animals in the field must now take into account the possibility that the observed behavior may have been "rigged." A parasite may even influence the evolution of its host, as many predators influence the evolution of their prey. Although this phenomenon is not evident in the case of *P. cylindraceus,* a truly pathogenic parasite might cause the host to evolve ways of avoiding the parasite or of resisting its pathogenic effects; any animal with these abilities would have an advantage over other members of its species. William D. Hamilton and Marlene Zuk of the University of Michigan have recently suggested that resistance to parasitic disease is expressed in the physical appearance and courtship display of songbirds and may affect the choice of a mate. More information is needed to confirm such a connection, but it is clear from this and similar research that biologists are beginning to understand the extent to which parasites can influence the behavior of other organisms.

The Inflationary Universe

A new theory of cosmology suggests that the observable universe is embedded in a much larger region of space that had an extraordinary growth spurt a fraction of a second after the primordial big bang

by Alan H. Guth and Paul J. Steinhardt

In the past few years certain flaws in the standard big-bang theory of cosmology have led to the development of a new model of the very early history of the universe. The model, known as the inflationary universe, agrees precisely with the generally accepted description of the observed universe for all times after the first 10^{-30} second. For this first fraction of a second, however, the story is dramatically different. According to the inflationary model, the universe had a brief period of extraordinarily rapid inflation, or expansion, during which its diameter increased by a factor perhaps 10^{50} times larger than had been thought. In the course of this stupendous growth spurt all the matter and energy in the universe could have been created from virtually nothing. The inflationary process also has important implications for the present universe. If the new model is correct, the observed universe is only a very small fraction of the entire universe.

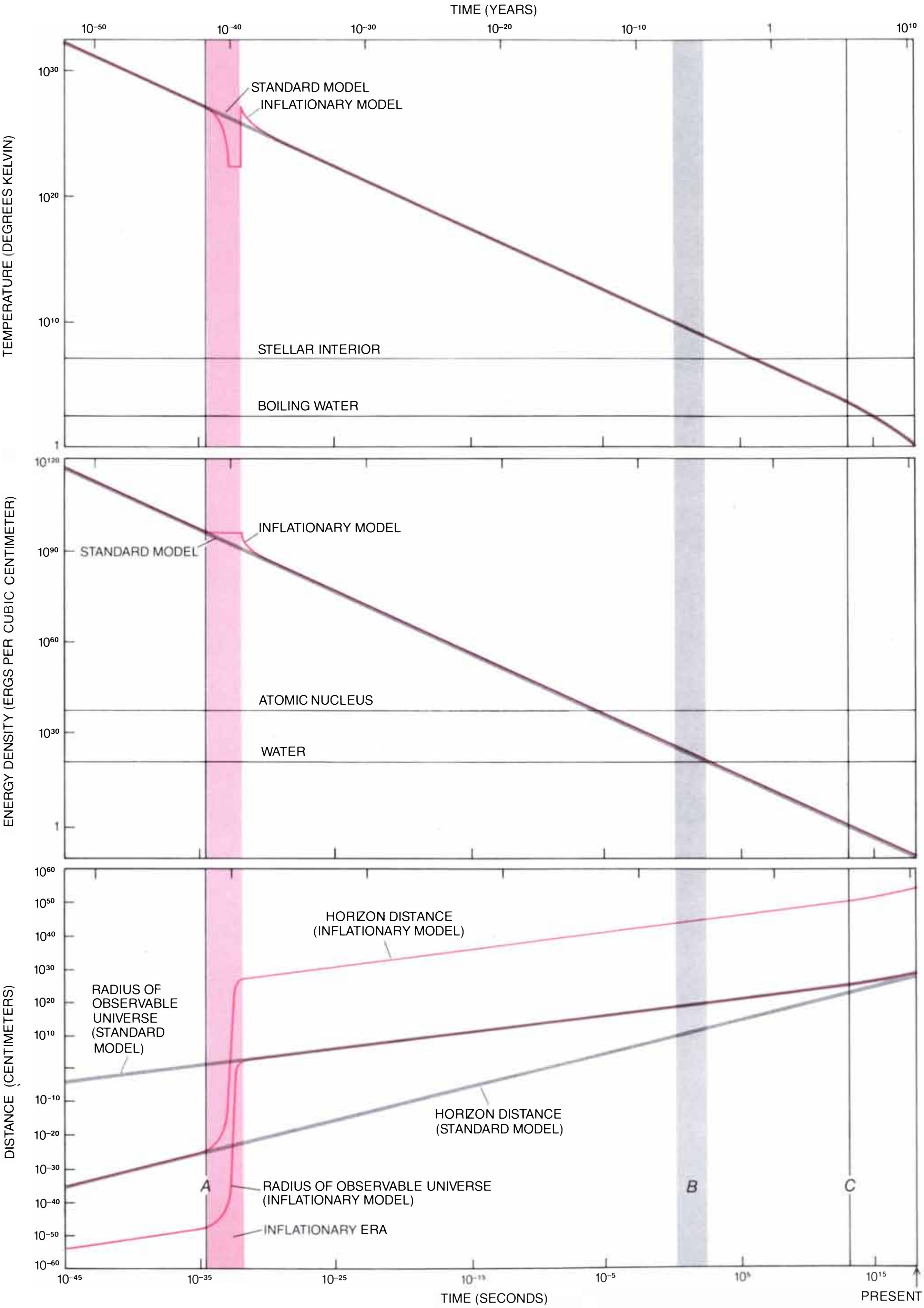
The inflationary model has many features in common with the standard big-bang model. In both models the universe began between 10 and 15 billion years ago as a primeval fireball of extreme density and temperature, and it has been expanding and cooling ever since. This picture has been successful in explaining many aspects of the observed universe, including the red-shifting of the light from distant galaxies, the cosmic microwave background radiation and the primordial abundances of the lightest elements. All these predictions have to do only with events that presumably took place after the first second, when the two models coincide.

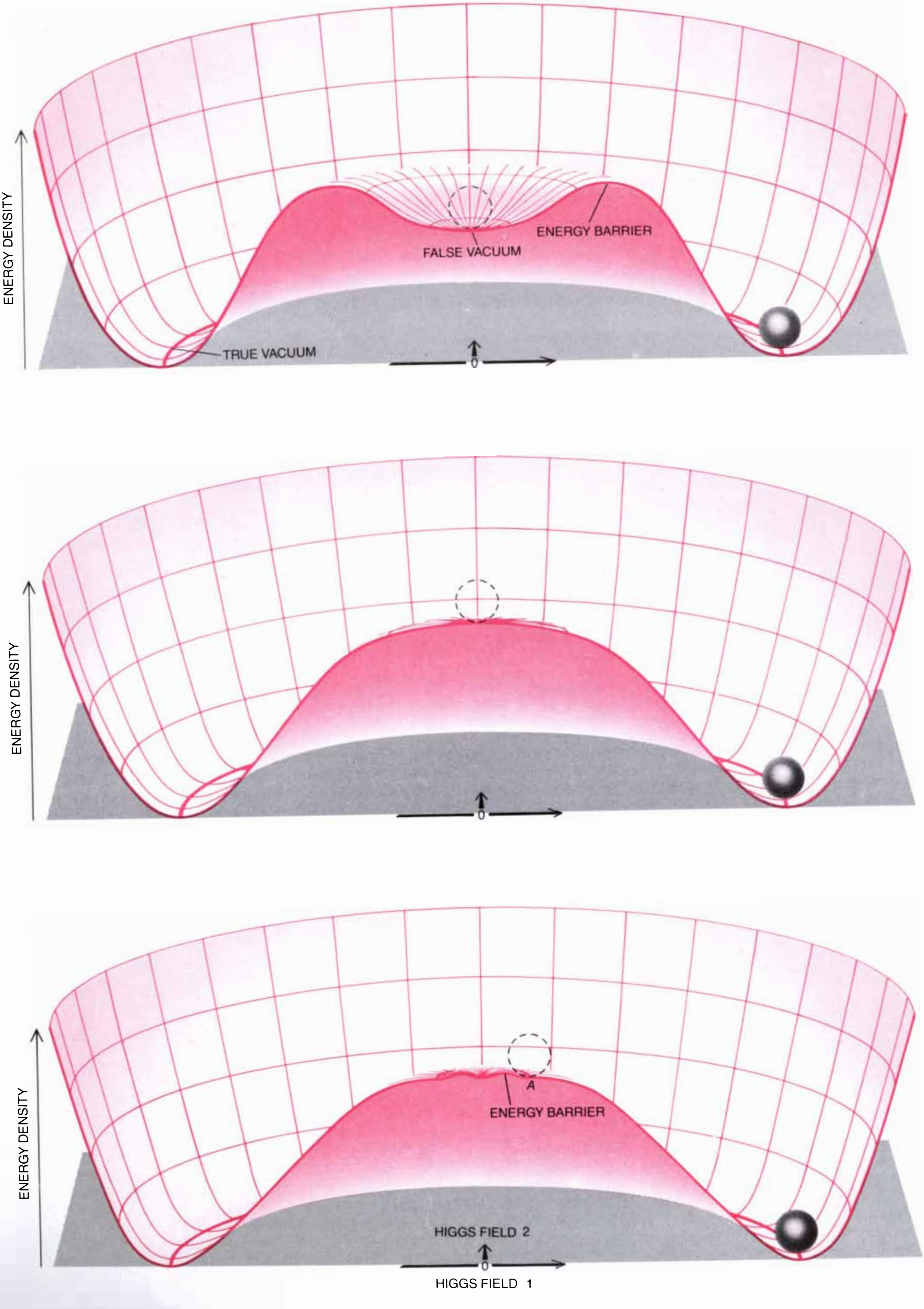
Until about five years ago there were few serious attempts to describe the universe during its first second. The temperature in this period is thought to have been higher than 10 billion degrees Kelvin, and little was known about the properties of matter under such conditions. Relying on recent developments in the physics of elementary particles, however, cosmologists are now attempting to understand the history of the universe back to 10^{-45} second after its beginning. (At even earlier times the energy density would have been so great that Einstein's general theory of relativity would have to be replaced by a quantum theory of gravity, which so far does not exist.) When the standard big-bang model is extended to these earlier times, various problems arise. First, it becomes clear that the model requires a number of stringent, unexplained assumptions about the initial conditions of the universe. In addition most of the new theories of elementary particles imply that the standard model would lead to a tremendous overproduction of the exotic particles called magnetic monopoles (each of which corresponds to an isolated north or south magnetic pole).

The inflationary universe was invented to overcome these problems. The equations that describe the period of inflation have a very attractive feature: from almost any initial conditions the universe evolves to precisely the state that had to be assumed as the initial one in the standard model. Moreover, the predicted density of magnetic monopoles becomes small enough to be consistent with observations. In the context of the recent developments in elementary-particle theory the inflationary model seems to be a natural solution to many of the problems of the standard big-bang picture.

The standard big-bang model is based on several assumptions. First, it is assumed that the fundamental laws of physics do not change with time and that the effects of gravitation are correctly described by Einstein's general theory of relativity. It is also assumed that the early universe was filled with an almost perfectly uniform, expanding, intensely hot gas of elementary particles in thermal equilibrium. The gas filled all of space, and the gas and space expanded together at the same rate. When they are averaged over large regions, the densities of matter and energy have remained nearly uniform from place to place as

INFLATIONARY MODEL of the universe is represented by the colored curves in this set of graphs, showing how several properties of the observed universe could have changed with time starting at 10^{-45} second after the big bang. The gray curves represent the standard big-bang model, which is coincident with the inflationary model for all times after 10^{-30} second. For comparison the graph of temperature (*top*) also includes an indication of the boiling point of water (373 degrees Kelvin) and the temperature at the center of a typical star (10 million degrees K.). Similarly the graph of energy density (*middle*) indicates the energy density of water (10^{21} ergs per cubic centimeter) and of an atomic nucleus (10^{36} ergs per cubic centimeter). On the graph of spatial dimensions (*bottom*) each cosmological model is represented by two curves. One curve shows the region of space that evolves to become the observed universe and the other shows the horizon distance: the total distance a light signal could have traveled since the beginning of the universe. One problem of the standard model, known as the horizon problem, arises from the fact that the horizon distance is much smaller than the radius of the observable universe for most of its history. In the inflationary model the horizon distance greatly exceeds the radius of the observable universe at all times. On the time axis several significant events are marked. *A* indicates the time of the phase transition predicted in the standard big-bang model by grand unified theories of the interactions of elementary particles; at the high temperature prevailing before this time the various nongravitational forces acting between particles are thought to have been related to one another by a symmetry that was spontaneously broken when the temperature fell to a critical value of about 10^{27} degrees. A key feature of the inflationary model is the prolongation of the phase transition, which extends through a period called the inflationary era (*color band*); during this era the universe expands by an extraordinary factor, perhaps 10^{50} or more. Meanwhile the temperature plunges, but it is stabilized at about 10^{22} degrees by quantum effects that arise in the context of general relativity. The gray band labeled *B* indicates the period when the lightest atomic nuclei were formed and *C* indicates the time when the universe became transparent to electromagnetic radiation.


TIME (YEARS)
10⁻⁵⁰
10⁻⁴⁰
10⁻³⁰
10⁻²⁰
10⁻¹⁰
1
10¹⁰
TEMPERATURE (DEGREES KELVIN)
STANDARD MODEL
INFLATIONARY MODEL
STELLAR INTERIOR
BOILING WATER
ENERGY DENSITY (ERGS PER CUBIC CENTIMETER)
INFLATIONARY MODEL
STANDARD MODEL
ATOMIC NUCLEUS
WATER
DISTANCE (CENTIMETERS)
HORIZON DISTANCE (INFLATIONARY MODEL)
RADIUS OF OBSERVABLE UNIVERSE (STANDARD MODEL)
HORIZON DISTANCE (STANDARD MODEL)
RADIUS OF OBSERVABLE UNIVERSE (INFLATIONARY MODEL)
INFLATIONARY ERA
A
B
C
TIME (SECONDS)
PRESENT


ENERGY DENSITY
FALSE VACUUM
ENERGY BARRIER
TRUE VACUUM
0
ENERGY DENSITY
0
ENERGY DENSITY
A
ENERGY BARRIER
HIGGS FIELD 2
0
HIGGS FIELD 1

the universe has evolved. It is further assumed that any changes in the state of the matter and the radiation have been so smooth that they have had a negligible effect on the thermodynamic history of the universe. The violation of the last assumption is a key to the inflationary-universe model.

The big-bang model leads to three important, experimentally testable predictions. First, the model predicts that as the universe expands, galaxies recede from one another with a velocity proportional to the distance between them. In the 1920's Edwin P. Hubble inferred just such an expansion law from his study of the red shifts of distant galaxies. Second, the big-bang model predicts that there should be a background of microwave radiation bathing the universe as a remnant of the intense heat of its origin. The universe became transparent to this radiation several hundred thousand years after the big bang. Ever since then the matter has been clumping into stars, galaxies and the like, but the radiation has simply continued to expand and red-shift, and in effect to cool. In 1964 Arno A. Penzias and Robert W. Wilson of the Bell Telephone Laboratories discovered a background of microwave radiation received uniformly from all directions with an effective temperature of about three degrees K. Third, the model leads to successful predictions of the formation of light atomic nuclei from protons and neutrons during the first minutes after the big bang. Successful predictions can be obtained in this way for the abundance of helium 4, deuterium, helium 3 and lithium 7. (Heavier nuclei are thought to have been produced much later in the interior of stars.)

Unlike the successes of the big-bang model, all of which pertain to events a second or more after the big bang, the problems all concern times when the universe was much less than a second old. One set of problems has to do with the special conditions the model requires as the universe emerged from the big bang.

The first problem is the difficulty of explaining the large-scale uniformity of the observed universe. The large-scale uniformity is most evident in the microwave background radiation, which is known to be uniform in temperature to about one part in 10,000. In the standard model the universe evolves much too quickly to allow this uniformity to be achieved by the usual processes whereby a system approaches thermal equilibrium. The reason is that no information or physical process can propagate faster than a light signal. At any given time there is a maximum distance, known as the horizon distance, that a light signal could have traveled since the beginning of the universe. In the standard model the sources of the microwave background radiation observed from opposite directions in the sky were separated from each other by more than 90 times the horizon distance when the radiation was emitted. Since the regions could not have communicated, it is difficult to see how they could have evolved conditions so nearly identical.

The puzzle of explaining why the universe appears to be uniform over distances that are large compared with the horizon distance is known as the horizon problem. It is not a genuine inconsistency of the standard model; if the uniformity is assumed in the initial conditions, the universe will evolve uniformly. The problem is that one of the most salient features of the observed universe—its large-scale uniformity—cannot be explained by the standard model; it must be assumed as an initial condition.

Even with the assumption of large-scale uniformity, the standard big-bang model requires yet another assumption to explain the nonuniformity observed on smaller scales. To account for the clumping of matter into galaxies, clusters of galaxies, superclusters of clusters and so on, a spectrum of primordial inhomogeneities must be assumed as part of the initial conditions. The fact that the spectrum of inhomogeneities has no explanation is a drawback in itself, but the problem becomes even more pronounced when the model is extended back to 10^{-45} second after the big bang. The incipient clumps of matter develop rapidly with time as a result of their gravitational self-attraction, and so a model that begins at a very early time must begin with very small inhomogeneities. To begin at 10^{-45} second the matter must start in a peculiar state of extraordinary but not quite perfect uniformity. A normal gas in thermal equilibrium would be far too inhomogeneous, owing to the random motion of particles. This peculiarity of the initial state of matter required by the standard model is called the smoothness problem.

Another subtle problem of the standard model concerns the energy density of the universe. According to general relativity, the space of the universe can in principle be curved, and the nature of the curvature depends on the energy density. If the energy density exceeds a certain critical value, which depends on the expansion rate, the universe is said to be closed: space curves back on itself to form a finite volume with no boundary. (A familiar analogy is the surface of a sphere, which is finite in area and has no boundary.) If the energy density is

ENERGY DENSITY of the universe is represented in these three-dimensional diagrams as a function of two Higgs fields, members of a special set of fields postulated in grand unified theories to account for spontaneous symmetry breaking. Each surface shown in cross section is rotationally symmetric about a vertical axis, which corresponds to a state in which both Higgs fields have a value of zero. In the absence of thermal excitations this state of unbroken symmetry, known as the false vacuum, would have an energy density of about 10^{95} ergs per cubic centimeter, or some 10^{59} times the energy density of an atomic nucleus. The rotational symmetry is broken whenever one of the Higgs fields acquires a nonzero value (or both of them do). Here the theory has been formulated in such a way that the states of lowest energy density, known as the true-vacuum states, are states of broken symmetry, forming a circle in the horizontal plane at the bottom of each diagram. In this analogy the evolution of the universe can be traced by imagining a ball rolling on the surface. The ball's distance from the central axis represents the combined values of the Higgs fields and its height above the horizontal surface represents the energy density of the universe. When the Higgs fields both have a value of zero, the ball is poised at the axis of symmetry; when the Higgs fields have a value that corresponds to the lowest possible energy density, the ball is lying somewhere in the trough that defines the broken-symmetry, or true-vacuum, states. In the original form of the inflationary-universe model it was assumed that the energy-density function had the shape in the diagram at the top. The inflationary episode would then take place while the universe was in the false-vacuum state. If the laws of classical physics applied, this state would be absolutely stable, because there would be no energy available to carry the Higgs fields over the intervening energy barrier. According to the laws of quantum physics, however, the fields in small regions of space can "tunnel" through the energy barrier, forming bubbles of the broken-symmetry phase, which would then start to grow. In the new inflationary model (*middle diagram*) there is no energy barrier; instead the false vacuum is at the top of a rather flat plateau. Under these circumstances the transition from the false vacuum to the broken-symmetry phase occurs by means of a slow-rollover mechanism: the Higgs fields are pushed from their initial value of zero by thermal or quantum fluctuations, and they proceed toward their true-vacuum values much as a ball would roll off a plateau of the same shape. The accelerated expansion of the universe takes place during the early stages of the rollover, while the energy density remains roughly constant. A single domain of broken-symmetry phase could then grow large enough to encompass the entire observable universe. When the Higgs fields reached the bottom of the trough, they would oscillate about the lowest energy-density value, causing a reheating of the universe. In a variant of the new inflationary model (*bottom diagram*) the false vacuum is surrounded by a small energy barrier. As in the original inflationary model, the false vacuum decays by the random formation of bubbles, created by the tunneling of the Higgs fields through the energy barrier. Because the energy barrier is small in this case, the Higgs fields tunnel only as far as the circle labeled *A*. Since the slope is quite flat at *A*, the Higgs fields evolve very slowly toward their true-vacuum values. The accelerated expansion of the universe continues as long as the Higgs fields remain near *A*, and a single bubble could grow large enough to encompass the observable universe.

less than the critical density, the universe is open: space curves but does not turn back on itself, and the volume is infinite. If the energy density is just equal to the critical density, the universe is flat: space is described by the familiar Euclidean geometry (again with infinite volume).

The ratio of the energy density of the universe to the critical density is a quantity cosmologists designate by the Greek letter Ω (omega). The value $\Omega = 1$ (corresponding to a flat universe) represents a state of unstable equilibrium. If Ω was ever exactly equal to 1, it would remain exactly equal to 1 forever. If Ω differed slightly from 1 an instant after the big bang, however, the deviation from 1 would grow rapidly with time. Given this instability, it is surprising that Ω is measured today as being between .1 and 2. (Cosmologists are still not sure whether the universe is open, closed or flat.) In order for Ω to be in this rather narrow range today, its value a second after the big bang had to equal 1 to within one part in 10^{15}. The standard model offers no explanation of why Ω began so close to 1 but merely assumes the fact as an initial condition. This shortcoming of the standard model, called the flatness problem, was first pointed out in 1979 by Robert H. Dicke and P. James E. Peebles of Princeton University.

The successes and drawbacks of the big-bang model we have considered so far involve cosmology, astrophysics and nuclear physics. As the big-bang model is traced backward in time, however, one reaches an epoch for which these branches of physics are no longer adequate. In this epoch all matter is decomposed into its elementary-particle constituents. In an attempt to understand this epoch cosmologists have made use of recent progress in the theory of elementary particles. Indeed, one of the important developments of the past decade has been the fusing of interests in particle physics, astrophysics and cosmology. The result for the big-bang model appears to be at least one more success and at least one more failure.

Perhaps the most important development in the theory of elementary particles over the past decade has been the notion of grand unified theories, the prototype of which was proposed in 1974 by Howard M. Georgi and Sheldon Lee Glashow of Harvard University. The theories are difficult to verify experimentally because their most distinctive predictions apply to energies far higher than those that can be reached with particle accelerators. Nevertheless, the theories have some experimental support, and they unify the understanding of elementary-particle interactions so elegantly that many physicists find them extremely attractive.

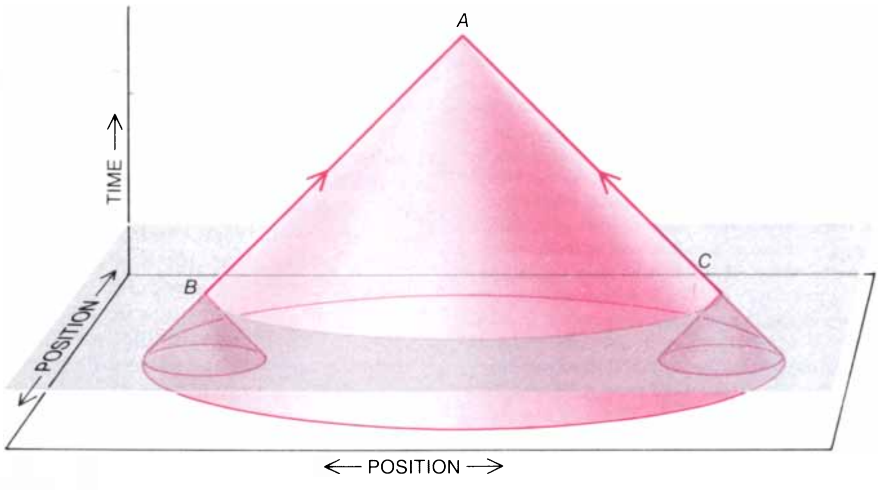
The basic idea of a grand unified theory is that what were perceived to be three independent forces—the strong, the weak and the electromagnetic—are actually parts of a single unified force. In the theory a symmetry relates one force to another. Since experimentally the forces are very different in strength and character, the theory is constructed so that the symmetry is spontaneously broken in the present universe.

A spontaneously broken symmetry is one that is present in the underlying theory describing a system but is hidden in the equilibrium state of the system. For example, a liquid described by physical laws that are rotationally symmetric is itself rotationally symmetric: the distribution of molecules looks the same no matter how the liquid is turned. When the liquid freezes into a crystal, however, the atoms arrange themselves along crystallographic axes and the rotational symmetry is broken. One would expect that if the temperature of a system in a broken-symmetry state were raised, it could undergo a kind of phase transition to a state in which the symmetry is restored, just as a crystal can melt into a liquid. Grand unified theories predict such a transition at a critical temperature of roughly 10^{27} degrees.

One novel property of the grand unified theories has to do with the particles called baryons, a class whose most important members are the proton and the neutron. In all physical processes observed up to now the number of baryons minus the number of antibaryons does not change; in the language of particle physics the total baryon number of the system is said to be conserved. A consequence of such a conservation law is that the proton must be absolutely stable; because it is the lightest baryon, it cannot decay into another particle without changing the total baryon number. Experimentally the lifetime of the proton is known to exceed 10^{31} years.

Grand unified theories imply that baryon number is not exactly conserved. At low temperature, in the broken-symmetry phase, the conservation law is an excellent approximation, and the observed limit on the proton lifetime is consistent with at least many versions of grand unified theories. At high temperature, however, processes that change the baryon number of a system of particles are expected to be quite common.

One direct result of combining the big-bang model with grand unified theories is the successful prediction of the asymmetry of matter and antimatter in the universe. It is thought that all the stars, galaxies and dust observed in the universe are in the form of matter rather than antimatter; their nuclear particles are baryons rather than antibaryons. It follows that the total baryon number of the observed universe is about 10^{78}. Before the advent of grand unified theories, when baryon number was thought to be conserved, this net baryon number had to be postulated as yet another



HORIZON PROBLEM is a serious drawback of the standard big-bang theory. In this three-dimensional space-time diagram the scales have been drawn in a nonlinear way so that the trajectory of a light pulse is represented by a line at 45 degrees to the vertical axis. Our position in space and time is indicated by the point *A*. Since no signal can travel faster than the speed of light, we can receive signals only from the colored region, which is called our past light cone. Events outside the past light cone of a given point cannot influence an event at that point in any way. The gray horizontal plane shows the time at which the microwave background radiation was released. Radiation that is now reaching us from opposite directions was released at points *B* and *C*, and since then it has traveled along our past light cone to point *A*. The past light cone of point *B* has no intersection with the past light cone of point *C*, and therefore the two points were not subject to any common influences. The horizon problem is the difficulty of explaining how the radiation received from the two opposite directions came to be at the same temperature. In the standard model the large-scale uniformity of temperature evident in the microwave background radiation must be assumed as an initial condition of the universe.

initial condition of the universe. When grand unified theories and the big-bang picture are combined, however, the observed excess of matter over antimatter can be produced naturally by elementary-particle interactions at temperatures just below the critical temperature of the phase transition. Calculations in the grand unified theories depend on too many arbitrary parameters for a quantitative prediction, but the observed matter-antimatter asymmetry can be produced with a reasonable choice of values for the parameters.

A serious problem that results from combining grand unified theories with the big-bang picture is that a large number of defects are generally formed during the transition from the symmetric phase to the broken-symmetry phase. The defects are created when regions of symmetric phase undergo a transition to different broken-symmetry states. In an analogous situation, when a liquid crystallizes, different regions may begin to crystallize with different orientations of the crystallographic axes. The domains of different crystal orientation grow and coalesce, and it is energetically favorable for them to smooth the misalignment along their boundaries. The smoothing is often imperfect, however, and localized defects remain.

In the grand unified theories there are serious cosmological problems associated with pointlike defects, which correspond to magnetic monopoles, and surfacelike defects, called domain walls. Both are expected to be extremely stable and extremely massive. (The monopole can be shown to be about 10^{16} times as heavy as the proton.) A domain of correlated broken-symmetry phase cannot be much larger than the horizon distance at that time, and so the minimum number of defects created during the transition can be estimated. The result is that there would be so many defects after the transition that their mass would dominate the energy density of the universe and thereby speed up its subsequent evolution. The microwave background radiation would reach its present temperature of three degrees K. only 30,000 years after the big bang instead of 10 billion years, and all the successful predictions of the big-bang model would be lost. Thus any successful union of grand unified theories and the big-bang picture must incorporate some mechanism to drastically suppress the production of magnetic monopoles and domain walls.

The inflationary-universe model appears to provide a satisfactory solution to these problems. Before the model can be described, however, we must first explain a few more of the details of symmetry breaking and phase transitions in grand unified theories.

TYPE OF UNIVERSE	RATIO OF ENERGY DENSITY TO CRITICAL DENSITY (Ω)	SPATIAL GEOMETRY	VOLUME	TEMPORAL EVOLUTION
CLOSED	>1	POSITIVE CURVATURE (SPHERICAL)	FINITE	EXPANDS AND RECOLLAPSES
OPEN	<1	NEGATIVE CURVATURE (HYPERBOLIC)	INFINITE	EXPANDS FOREVER
FLAT	1	ZERO CURVATURE (EUCLIDEAN)	INFINITE	EXPANDS FOREVER, BUT EXPANSION RATE APPROACHES ZERO

THREE TYPES OF UNIVERSE, classified as closed, open and flat, can arise from the standard big-bang model (under the usual assumption that the equations of general relativity are not modified by the addition of a cosmological term). The distinction between the different geometries depends on the quantity designated Ω, the ratio of the energy density of the universe to some critical density, whose value depends in turn on the rate of expansion of the universe. The value of Ω today is known to lie between .1 and 2, which implies that its value a second after the big bang was equal to 1 to within one part in 10^{15}. The failure of the standard big-bang model to explain why Ω began so close to 1 is called the flatness problem.

All modern particle theories, including the grand unified theories, are examples of quantum field theories. The best-known field theory is the one that describes electromagnetism. According to the classical (nonquantum) theory of electromagnetism developed by James Clerk Maxwell in the 1860's, electric and magnetic fields have a well-defined value at every point in space, and their variation with time is described by a definite set of equations. Maxwell's theory was modified early in the 20th century in order to achieve consistency with the quantum theory. In the classical theory it is possible to increase the energy of an electromagnetic field by any amount, but in the quantum theory the increases in energy can come only in discrete lumps, the quanta, which in this case are called photons. The photons have both wavelike and particlelike properties, but in the lexicon of modern physics they are usually called particles. In general the formulation of a quantum field theory begins with a classical theory of fields, and it becomes a theory of particles when the rules of the quantum theory are applied.

As we have already mentioned, an essential ingredient of grand unified theories is the phenomenon of spontaneous symmetry breaking. The detailed mechanism of spontaneous symmetry breaking in grand unified theories is simpler in many ways than the analogous mechanism in crystals. In a grand unified theory spontaneous symmetry breaking is accomplished by including in the formulation of the theory a special set of fields known as Higgs fields (after Peter W. Higgs of the University of Edinburgh). The symmetry is unbroken when all the Higgs fields have a value of zero, but it is spontaneously broken whenever at least one of the Higgs fields acquires a nonzero value. Furthermore, it is possible to formulate the theory in such a way that a Higgs field has a nonzero value in the state of lowest energy density, which in this context is known as the true vacuum. At temperatures greater than about 10^{27} degrees thermal fluctuations drive the equilibrium value of the Higgs field to zero, resulting in a transition to the symmetric phase.

We have now assembled enough background information to describe the inflationary model of the universe, beginning with the form in which it was first proposed by one of us (Guth) in 1980. Any cosmological model must begin with some assumptions about the initial conditions, but for the inflationary model the initial conditions can be rather arbitrary. One must assume, however, that the early universe included at least some regions of gas that were hot compared with the critical temperature of the phase transition and that were also expanding. In such a hot region the Higgs field would have a value of zero. As the expansion caused the temperature to fall it would become thermodynamically favorable for the Higgs field to acquire a nonzero value, bringing the system to its broken-symmetry phase.

For some values of the unknown parameters of the grand unified theories this phase transition would occur very slowly compared with the cooling rate. As a result the system could cool to well below 10^{27} degrees with the value of the Higgs field remaining at zero. This phenomenon, known as supercooling, is quite common in condensed-matter physics; water, for example, can be supercooled to more than 20 degrees below its freezing point, and glasses are formed by rapidly supercooling a liquid to a temperature well below its freezing point.

As the region of gas continued to supercool, it would approach a peculiar state of matter known as a false vacuum. This state of matter has never been observed, but it has properties that are unambiguously predicted by quantum field theory. The temperature, and hence the thermal component of the energy density, would rapidly decrease and the energy density of the state would be concentrated entirely in the Higgs field. A zero value for the Higgs field implies a large energy density for the false vacuum. In the classical form of the theory such a state would be absolutely stable, even though it would not be the state of lowest energy density. States with a lower energy density would be separated from the false vacuum by an intervening energy barrier, and there would be no energy available to take the Higgs field over the barrier [*see top illustration on page 118*].

In the quantum version of the model the false vacuum is not absolutely stable. Under the rules of the quantum theory all the fields would be continually fluctuating. As was first described by Sidney R. Coleman of Harvard, a quantum fluctuation would occasionally cause the Higgs field in a small region of space to "tunnel" through the energy barrier, nucleating a "bubble" of the broken-symmetry phase. The bubble would then start to grow at a speed that would rapidly approach the speed of light, converting the false vacuum into the broken-symmetry phase. The rate at which bubbles form depends sensitively on the unknown parameters of the grand unified theory; in the inflationary model it is assumed that the rate would be extremely low.

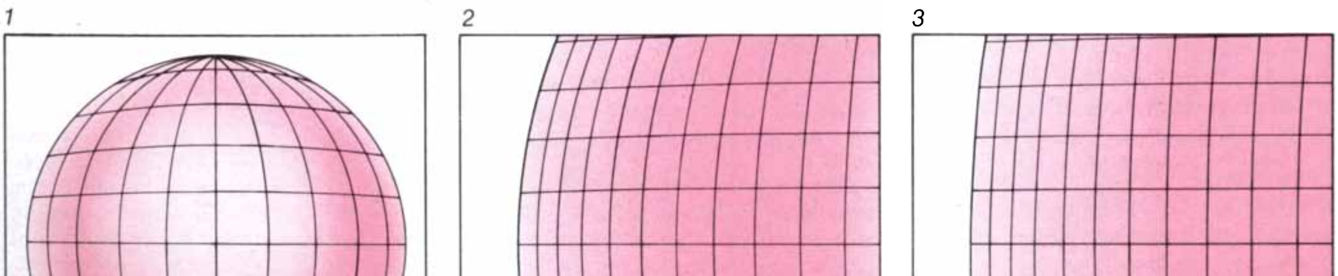
The most peculiar property of the false vacuum is probably its pressure, which is both large and negative. To understand why, consider again the process by which a bubble of true vacuum would grow into a region of false vacuum. The growth is favored energetically because the true vacuum has a lower energy density than the false vacuum. The growth also indicates, however, that the pressure of the true vacuum must be higher than the pressure of the false vacuum, forcing the bubble wall to grow outward. Because the pressure of the true vacuum is zero, the pressure of the false vacuum must be negative. A more detailed argument shows that the pressure of the false vacuum is equal to the negative value of its energy density (when the two quantities are measured in the same units).

The negative pressure would not result in mechanical forces within the false vacuum, because mechanical forces arise only from differences in pressure. Nevertheless, there would be gravitational effects. Under ordinary circumstances the expansion of the region of gas would be slowed by the mutual gravitational attraction of the matter within it. In Newtonian physics this attraction is proportional to the mass density, which in relativistic theories is equal to the energy density divided by the square of the speed of light. According to general relativity, the pressure also contributes to the attraction; to be specific, the gravitational force is proportional to the energy density plus three times the pressure. For the false vacuum the contribution made by the pressure would overwhelm the energy-density contribution and would have the opposite sign. Hence the bizarre notion of negative pressure leads to the even more bizarre effect of a gravitational force that is effectively repulsive. As a result the expansion of the region would be accelerated and the region would grow exponentially, doubling in diameter during each interval of about 10^{-34} second.

This period of accelerated expansion is called the inflationary era, and it is the key element of the inflationary model of the universe. According to the model, the inflationary era continued for 10^{-32} second or longer, and during this period the diameter of the universe increased by a factor of 10^{50} or more. It is assumed that after this colossal expansion the transition to the broken-symmetry phase finally took place. The energy density of the false vacuum was then released, resulting in a tremendous amount of particle production. The region was reheated to a temperature of almost 10^{27} degrees. (In the language of thermodynamics the energy released is called the latent heat; it is analogous to the energy released when water freezes.) From this point on the region would continue to expand and cool at the rate described by the standard big-bang model. A volume the size of the observable universe would lie well within such a region.

The horizon problem is avoided in a straightforward way. In the inflationary model the observed universe evolves from a region that is much smaller in diameter (by a factor of 10^{50} or more) than the corresponding region in the standard model. Before inflation begins the region is much smaller than the horizon distance, and it has time to homogenize and reach thermal equilibrium. This small homogeneous region is then inflated to become large enough to encompass the observed universe. Thus the sources of the microwave background radiation arriving today from all directions in the sky were once in close contact; they had time to reach a common temperature before the inflationary era began.

The flatness problem is also evaded in a simple and natural way. The equations describing the evolution of the universe



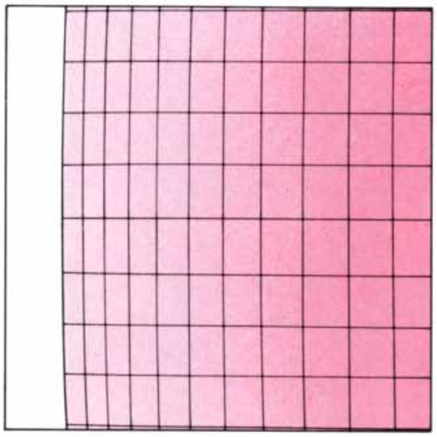
SOLUTION OF THE FLATNESS PROBLEM is illustrated by this sequence of perspective drawings of an inflating sphere. The illustration shows how a flat spatial geometry (which corresponds to a value of Ω equal to 1) can be produced by the inflationary scenario in a simple and natural way. In each successive frame the sphere is inflated by a factor of three (while the number of grid lines on the

during the inflationary era are different from those for the standard model, and it turns out that the ratio Ω is driven rapidly toward 1, no matter what value it had before inflation. This behavior is most easily understood by recalling that a value of $\Omega = 1$ corresponds to a space that is geometrically flat. The rapid expansion causes the space to become flatter just as the surface of a balloon becomes flatter when it is inflated. The mechanism driving Ω toward 1 is so effective that one is led to an almost rigorous prediction: The value of Ω today should be very accurately equal to 1. Many astronomers (although not all) think a value of 1 is consistent with current observations, but a more reliable determination of Ω would provide a crucial test of the inflationary model.

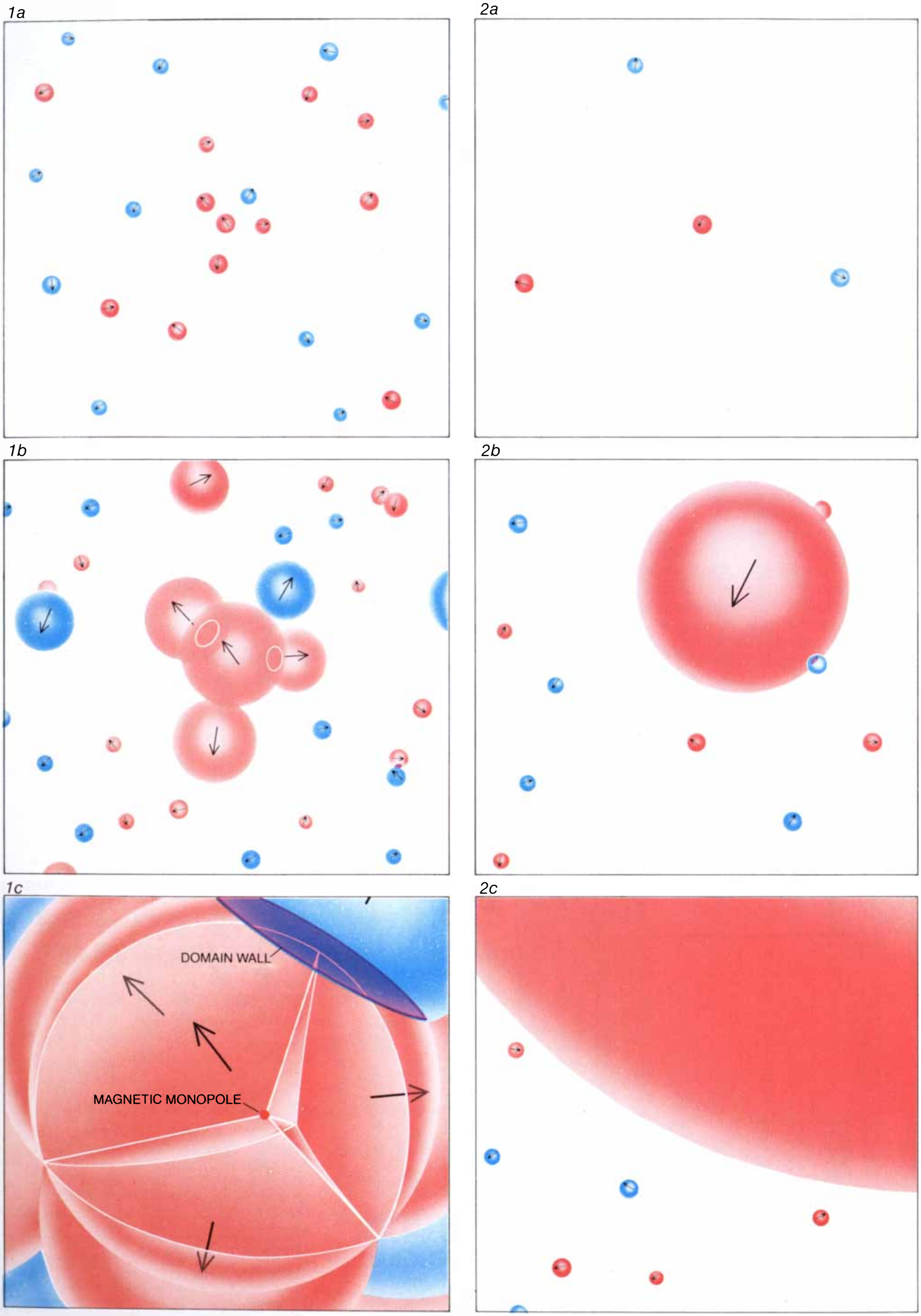
In the form in which the inflationary model was originally proposed it had a crucial flaw: under the circumstances described, the phase transition itself would create inhomogeneities much more extreme than those observed today. As we have already described, the phase transition would take place by the random nucleation of bubbles of the new phase. It can be shown that the bubbles would always remain in finite clusters disconnected from one another, and that each cluster would be dominated by a single largest bubble. Almost all the energy in the cluster would be initially concentrated in the surface of the largest bubble, and there is no apparent mechanism to redistribute energy uniformly. Such a configuration bears no resemblance to the observed universe.

For almost two years after the invention of the inflationary-universe model it remained a tantalizing but clearly imperfect solution to a number of important cosmological problems. Near the end of 1981, however, a new

4



surface is increased by the same factor). The curvature of the surface quickly becomes undetectable on the scale of the illustration.


1a
2a
1b
2b
1c
2c
DOMAIN WALL
MAGNETIC MONOPOLE

approach was developed by A. D. Linde of the P. N. Lebedev Physical Institute in Moscow and independently by Andreas Albrecht and one of us (Steinhardt) of the University of Pennsylvania. This approach, known as the new inflationary universe, avoids all the problems of the original model while maintaining all its successes.

The key to the new approach is to consider a special form of the energy-density function that describes the Higgs field [*see middle illustration on page 118*]. Quantum field theories with energy-density functions of this type were first studied by Coleman, working in collaboration with Erick J. Weinberg of Columbia University. In contrast to the more typical case shown in the top illustration on page 118, there is no energy barrier separating the false vacuum from the true vacuum; instead the false vacuum lies at the top of a rather flat plateau. In the context of grand unified theories such an energy-density function is achieved by a special choice of parameters. As we shall explain below, this energy-density function leads to a special type of phase transition that is sometimes called a slow-rollover transition.

The scenario begins just as it does in the original inflationary model. Again one must assume the early universe had regions that were hotter than about 10^{27} degrees and were also expanding. In these regions thermal fluctuations would drive the equilibrium value of the Higgs fields to zero and the symmetry would be unbroken. As the temperature fell it would become thermodynamically favorable for the system to undergo a phase transition in which at least one of the Higgs fields acquired a nonzero value, resulting in a broken-symmetry phase. As in the previous case, however, the rate of this phase transition would be extremely low compared with the rate of cooling. The system would supercool to a negligible temperature with the Higgs field remaining at zero, and the resulting state would again be considered a false vacuum.

The important difference in the new approach is the way in which the phase transition would take place. Quantum fluctuations or small residual thermal fluctuations would cause the Higgs field to deviate from zero. In the absence of an energy barrier the value of the Higgs field would begin to increase steadily; the rate of increase would be much like that of a ball rolling down a hill of the same shape as the curve of the energy-density function, under the influence of a frictional drag force. Since the energy-density curve is almost flat near the point where the Higgs field vanishes, the early stage of the evolution would be very slow. As long as the Higgs field remained close to zero, the energy density would be almost the same as it is in the false vacuum. As in the original scenario, the region would undergo accelerated expansion, doubling in diameter every 10^{-34} second or so. Now, however, the expansion would cease to accelerate when the value of the Higgs field reached the steeper part of the curve. By computing the time required for the Higgs field to evolve, the amount of inflation can be determined. An expansion factor of 10^{50} or more is quite plausible, but the actual factor depends on the details of the particle theory one adopts.

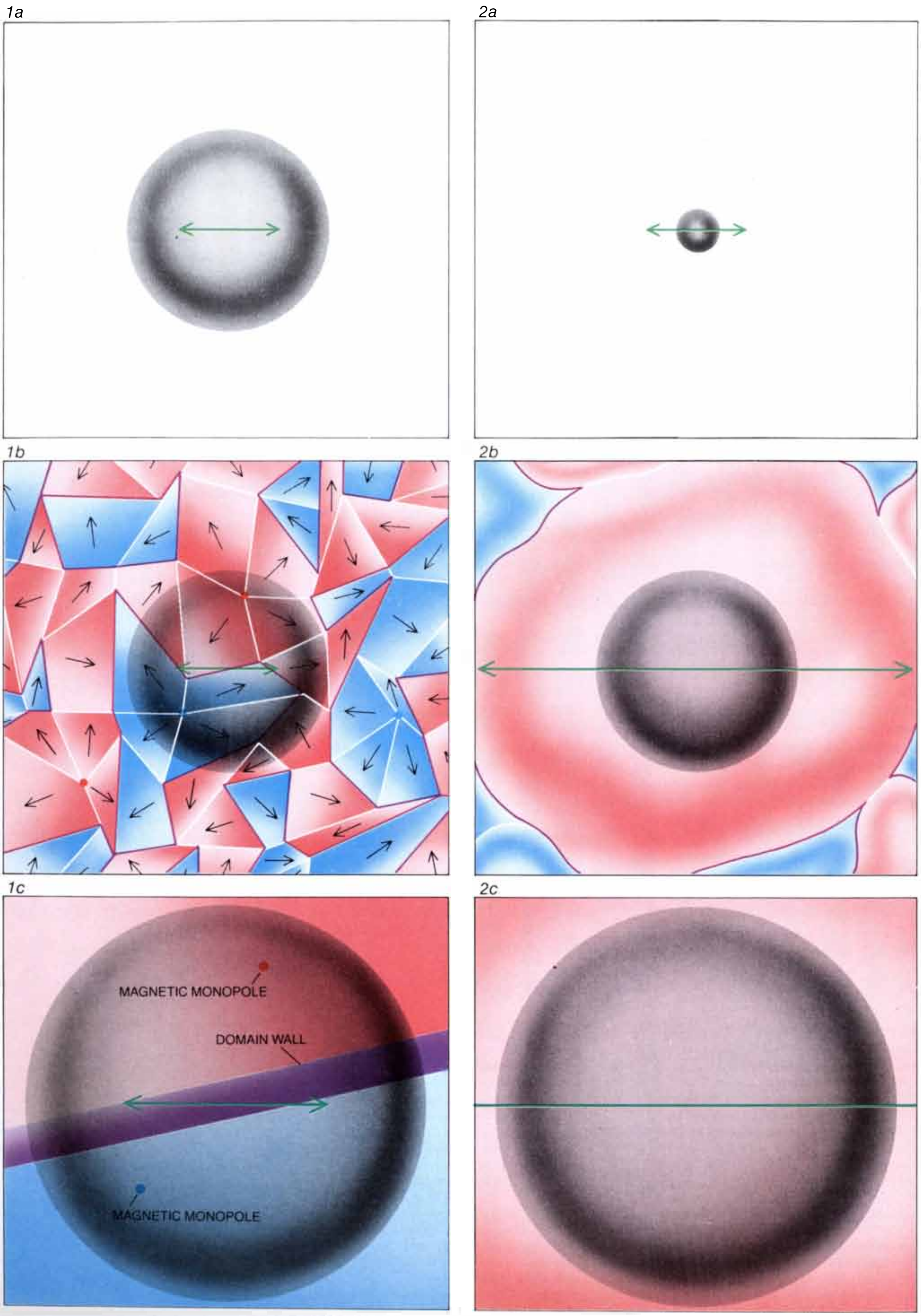
So far the description of the phase transition has been slightly oversimplified. There are actually many different broken-symmetry states, just as there are many possible orientations for the axes of a crystal. There are a number of Higgs fields, and the various broken-symmetry states are distinguished by the combination of Higgs fields that acquire nonzero values. Since the fluctuations that drive the Higgs fields from zero are random, different regions of the primordial universe would be driven toward different broken-symmetry states, each region forming a domain with an initial radius of roughly the horizon distance. At the start of the phase transition the horizon distance would be about 10^{-24} centimeter. Once the domain formed, with the Higgs fields deviating slightly from zero in a definite combination, it would evolve toward one of the stable broken-symmetry states and would inflate by a factor of 10^{50} or more. The size of the domain after inflation would then be greater than 10^{26} centimeters. The entire observable universe, which at that time would be only about 10 centimeters across, would be able to fit deep inside a single domain.

In the course of this enormous inflation any density of particles that might have been present initially would be diluted to virtually zero. The energy content of the region would then consist entirely of the energy stored in the Higgs field. How could this energy be released? Once the Higgs field evolved away from the flat part of the energy-density curve, it would start to oscillate rapidly about the true-vacuum value. Drawing on the relation between particles and fields implied by quantum field theory, this situation can also be described as a state with a high density of Higgs particles. The Higgs particles would be unstable, however: they would rapidly decay to lighter particles, which would interact with one another and possibly undergo subsequent decays. The system would quickly become a hot gas of elementary particles in thermal equilibrium, just as was assumed in the initial conditions for the standard model. The reheating temperature is calculable and is typically a factor of between two and 10 below the critical temperature of the phase transition. From this point on, the scenario coincides with that of the standard big-bang model, and so all the successes of the standard model are retained.

Note that the crucial flaw of the original inflationary model is deftly avoided. Roughly speaking, the isolated bubbles that were discussed in the original model are replaced here by the domains. The domains of the slow-rollover transition would be surrounded by other domains rather than by false vacuum, and they would tend not to be spherical. The term "bubble" is therefore avoided. The key difference is that in the new inflationary model each domain inflates in the course of its formation, producing a vast essentially homogeneous region within which the observable universe can fit.

Since the reheating temperature is near the critical temperature of the grand-unified-theory phase transition, the matter-antimatter asymmetry could be produced by particle interactions just after the phase transition. The production mechanism is the same as the one predicted by grand unified theories for the standard big-bang model. In contrast to the standard model, however, the inflationary model does not allow the possibility of assuming the observed net baryon number of the universe as an initial condition; the subsequent in-

EXPANDING BUBBLES of broken-symmetry phase form in an expanding region of symmetric phase in the two highly schematic time sequences on the opposite page. The sequence representing the standard big-bang model (*left*) covers a much shorter time span than the sequence representing the original form of the inflationary model (*right*). In both cases the Higgs fields have a value of zero in the region outside the bubbles, whereas at least one Higgs field has a nonzero value inside each bubble. In a grand unified theory the broken-symmetry states can in general be distinguished by parameters of two kinds: discrete and continuous. Here each bubble is labeled in two ways: by a color (blue or red) to indicate the discrete parameter and by an internal black arrow to indicate the value of the continuous parameter. In the standard model the bubbles would quickly coalesce and complete the transition from the symmetric phase to the broken-symmetry phase. A surfacelike defect called a domain wall would form at any boundary between regions with different values of the discrete parameter (*purple areas*). Within a region of uniform color a pointlike defect called a magnetic monopole would form at a center created by the intersection of many bubbles whenever the arrow representing the continuous parameter points everywhere away from the center. In the original form of the inflationary model the rapid expansion of the false-vacuum, or symmetric-phase, region would keep the bubbles from ever coalescing. Either hypothetical situation has consequences that are contrary to observation; the new inflationary model was developed to avoid both of them.


1a
2a
1b
2b
1c
2c
MAGNETIC MONOPOLE
DOMAIN WALL
MAGNETIC MONOPOLE

flation would dilute any initial baryon-number density to an imperceptible level. Thus the viability of the inflationary model depends crucially on the viability of particle theories, such as the grand unified theories, in which baryon number is not conserved.

One can now grasp the solutions to the cosmological problems discussed above. The horizon and flatness problems are resolved by the same mechanisms as in the original inflationary-universe model. In the new inflationary scenario the problem of monopoles and domain walls can also be solved. Such defects would form along the boundaries separating domains, but the domains would have been inflated to such an enormous size that the defects would lie far beyond any observable distance. (A few defects might be generated by thermal effects after the transition, but they are expected to be negligible in number.)

Thus with a few simple ideas the improved inflationary model of the universe leads to a successful resolution of several major problems that plague the standard big-bang picture: the horizon, flatness, magnetic-monopole and domain-wall problems. Unfortunately the necessary slow-rollover transition requires the fine tuning of parameters; calculations yield reasonable predictions only if the parameters are assigned values in a narrow range. Most theorists (including both of us) regard such fine tuning as implausible. The consequences of the scenario are so successful, however, that we are encouraged to go on in the hope we may discover realistic versions of grand unified theories in which such a slow-rollover transition occurs without fine tuning.

The successes already discussed offer persuasive evidence in favor of the new inflationary model. Moreover, it was recently discovered that the model may also resolve an additional cosmological problem not even considered at the time the model was developed: the smoothness problem. The generation of density inhomogeneities in the new inflationary universe was addressed in the summer of 1982 at the Nuffield Workshop on the Very Early Universe by a number of theorists, including James M. Bardeen of the University of Washington, Stephen W. Hawking of the University of Cambridge, So-Young Pi of Boston University, Michael S. Turner of the University of Chicago, A. A. Starobinsky of the L. D. Landau Institute of Theoretical Physics in Moscow and the two of us. It was found that the new inflationary model, unlike any previous cosmological model, leads to a definite prediction for the spectrum of inhomogeneities. Basically the process of inflation first smoothes out any primordial inhomogeneities that might have been present in the initial conditions. Then in the course of the phase transition inhomogeneities are generated by the quantum fluctuations of the Higgs field in a way that is completely determined by the underlying physics. The inhomogeneities are created on a very small scale of length, where quantum phenomena are important, and they are then enlarged to an astronomical scale by the process of inflation.

The predicted shape for the spectrum of inhomogeneities is essentially scale-invariant; that is, the magnitude of the inhomogeneities is approximately equal on all length scales of astrophysical significance. This prediction is comparatively insensitive to the details of the underlying grand unified theory. It turns out that a spectrum of precisely this shape was proposed in the early 1970's as a phenomenological model for galaxy formation by Edward R. Harrison of the University of Massachusetts at Amherst and Yakov B. Zel'dovich of the Institute of Physical Problems in Moscow, working independently. The details of galaxy formation are complex and are still not well understood, but many cosmologists think a scale-invariant spectrum of inhomogeneities is precisely what is needed to explain how the present structure of galaxies and galactic clusters evolved [see "The Large-Scale Structure of the Universe," by Joseph Silk, Alexander S. Szalay and Yakov B. Zel'dovich; SCIENTIFIC AMERICAN, October, 1983].

The new inflationary model also predicts the magnitude of the density inhomogeneities, but the prediction is quite sensitive to the details of the underlying particle theory. Unfortunately the magnitude that results from the simplest grand unified theory is far too large to be consistent with the observed uniformity of the cosmic microwave background. This inconsistency represents a problem, but it is not yet known whether the simplest grand unified theory is the correct one. In particular the simplest grand unified theory predicts a lifetime for the proton that appears to be lower than present experimental limits. On the other hand, one can construct more complicated grand unified theories that result in density inhomogeneities of the desired magnitude. Many investigators imagine that with the development of the correct particle theory the new inflationary model will add the resolution of the smoothness problem to its list of successes.

One promising line of research involves a class of quantum field theories with a new kind of symmetry called supersymmetry. Supersymmetry relates the properties of particles with integer angular momentum to those of particles with half-integer angular momentum; it thereby highly constrains the form of the theory. Many theorists think supersymmetry might be necessary to construct a consistent quantum theory of gravity, and to eventually unify gravity with the strong, the weak and the electromagnetic forces. A tantalizing property of models incorporating supersymmetry is that many of them give slow-rollover phase transitions without any fine tuning of parameters. The search is on to find a supersymmetry model that is realistic as far as particle physics is concerned and that also gives rise to inflation and to the correct magnitude for the density inhomogeneities.

In short, the inflationary model of the universe is an economical theory that accounts for many features of the observable universe lacking an explanation in the standard big-bang model. The beauty of the inflationary model is that the evolution of the universe becomes almost independent of the details of the initial conditions, about which little if anything is known. It follows, however, that if the inflationary model is correct, it will be difficult for anyone to ever discover observable consequences of the conditions existing before the inflationary phase transition. Similarly, the vast distance scales created by inflation would make it essentially impossible to observe the structure of the universe as a whole. Nevertheless, one can

NEW INFLATIONARY MODEL deftly evades the horizon, magnetic-monopole and domain-wall problems. In the two series of drawings on the opposite page, representing the standard big-bang model (*left*) and the new inflationary model (*right*), the gray sphere corresponds to the region of space that evolved to become the observed universe and the two-headed green arrow represents the horizon distance. (The relative scales shown here are suggestive only; the actual scales differ by factors that are too extreme to depict.) Three evolutionary stages are shown for each scenario: just before the phase transition (*top*), just after the phase transition (*middle*) and today (*bottom*). In the standard model the horizon distance is always smaller than the gray sphere, making the large-scale uniformity of the observed universe puzzling. Since in the standard model a domain of broken-symmetry phase created in the phase transition would have a radius comparable to the horizon distance, many monopoles and domain walls would be present in the observed universe. In the new inflationary model the horizon distance is always much larger than the gray sphere, and so the observed universe is expected to be uniform on a large scale and to have few, if any, monopoles and domain walls. Just before the phase transition the gray sphere in the inflationary model is much smaller than it is in the standard model; during the phase transition the gray sphere in the inflationary model expands by a factor of 10^{50} or more in radius to match the size of the corresponding sphere in the standard model.

still discuss these issues, and a number of remarkable scenarios seem possible.

The simplest possibility for the very early universe is that it actually began with a big bang, expanded rather uniformly until it cooled to the critical temperature of the phase transition and then proceeded according to the inflationary scenario. Extrapolating the big-bang model back to zero time brings the universe to a cosmological singularity, a condition of infinite temperature and density in which the known laws of physics do not apply. The instant of creation remains unexplained. A second possibility is that the universe began (again without explanation) in a random, chaotic state. The matter and temperature distributions would be nonuniform, with some parts expanding and other parts contracting. In this scenario certain small regions that were hot and expanding would undergo inflation, evolving into huge regions easily capable of encompassing the observable universe. Outside these regions there would remain chaos, gradually creeping into the regions that had inflated.

Recently there has been some serious speculation that the actual creation of the universe is describable by physical laws. In this view the universe would originate as a quantum fluctuation, starting from absolutely nothing. The idea was first proposed by Edward P. Tryon of Hunter College of the City University of New York in 1973, and it was put forward again in the context of the inflationary model by Alexander Vilenkin of Tufts University in 1982. In this context "nothing" might refer to empty space, but Vilenkin uses it to describe a state devoid of space, time and matter. Quantum fluctuations of the structure of space-time can be discussed only in the context of quantum gravity, and so these ideas must be considered highly speculative until a working theory of quantum gravity is formulated. Nevertheless, it is fascinating to contemplate that physical laws may determine not only the evolution of a given state of the universe but also the initial conditions of the observable universe.

As for the structure of the universe as a whole, the inflationary model allows for several possibilities. (In all cases the observable universe is a very small fraction of the universe as a whole; the edge of our domain is likely to lie 10^{35} or more light-years away.) The first possibility is that the domains meet one another and fill all space. The domains are then separated by domain walls, and in the interior of each wall is the symmetric phase of the grand unified theory. Protons or neutrons passing through such a wall would decay instantly. Domain walls would tend to straighten with time. After 10^{35} years or more smaller domains (possibly even our own) would disappear and larger domains would grow.

Alternatively, some versions of grand unified theories do not allow for the formation of sharp domain walls. In these theories it is possible for different broken-symmetry states in two neighboring domains to merge smoothly into each other. At the interface of two domains one would find discontinuities in the density and velocity of matter, and one would also find an occasional magnetic monopole.

A quite different possibility would result if the energy density of the Higgs fields were described by a curve such as the one in the bottom illustration on page 118. As in the other two cases, regions of space would supercool into the false-vacuum state and undergo accelerated expansion. As in the original inflationary model, the false-vacuum state would decay by the mechanism of random bubble formation: quantum fluctuations would cause at least one of the Higgs fields in a small region of space to tunnel through the energy barrier, to the value marked A in the illustration. In contrast to the original inflationary scenario, the Higgs field would then evolve very slowly (because of the flatness of the curve near A) to its true-vacuum value. The accelerated expansion would continue, and the single bubble would become large enough to encompass the observed universe. If the rate of bubble formation were low, bubble collisions would be rare. The fraction of space filled with bubbles would become closer to 1 as the system evolved, but space would be expanding so fast that the volume remaining in the false-vacuum state would increase with time. Bubble universes would continue to form forever, and there would be no way of knowing how much time had elapsed before our bubble was formed. This picture is much like the old steady-state cosmological model on the very large scale, and yet the interior of each bubble would evolve according to the big-bang model, improved by inflation.

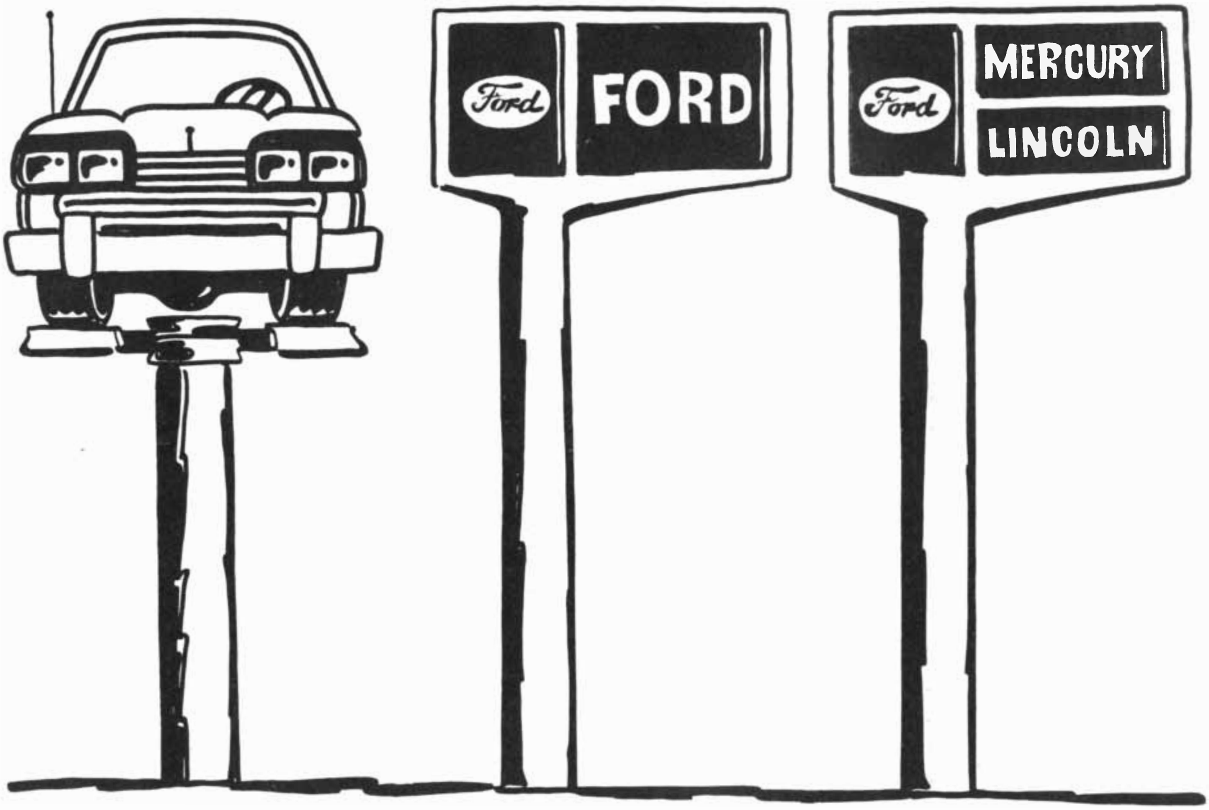
From a historical point of view probably the most revolutionary aspect of the inflationary model is the notion that all the matter and energy in the observable universe may have emerged from almost nothing. This claim stands in marked contrast to centuries of scientific tradition in which it was believed that something cannot come from nothing. The tradition, dating back at least as far as the Greek philosopher Parmenides in the fifth century B.C., has manifested itself in modern times in the formulation of a number of conservation laws, which state that certain physical quantities cannot be changed by any physical process. A decade or so ago the list of quantities thought to be conserved included energy, linear momentum, angular momentum, electric charge and baryon number.

Since the observed universe apparently has a huge baryon number and a huge energy, the idea of creation from nothing has seemed totally untenable to all but a few theorists. (The other conservation laws mentioned above present no such problems: the total electric charge and the angular momentum of the observed universe have values consistent with zero, whereas the total linear momentum depends on the velocity of the observer and so cannot be defined in absolute terms.) With the advent of grand unified theories, however, it now appears quite plausible that baryon number is not conserved. Hence only the conservation of energy needs further consideration.

The total energy of any system can be divided into a gravitational part and a nongravitational part. The gravitational part (that is, the energy of the gravitational field itself) is negligible under laboratory conditions, but cosmologically it can be quite important. The nongravitational part is not by itself conserved; in the standard big-bang model it decreases drastically as the early universe expands, and the rate of energy loss is proportional to the pressure of the hot gas. During the era of inflation, on the other hand, the region of interest is filled with a false vacuum that has a large negative pressure. In this case the nongravitational energy increases drastically. Essentially all the nongravitational energy of the universe is created as the false vacuum undergoes its accelerated expansion. This energy is released when the phase transition takes place, and it eventually evolves to become stars, planets, human beings and so forth. Accordingly, the inflationary model offers what is apparently the first plausible scientific explanation for the creation of essentially all the matter and energy in the observable universe.

Under these circumstances the gravitational part of the energy is somewhat ill-defined, but crudely speaking one can say that the gravitational energy is negative, and that it precisely cancels the nongravitational energy. The total energy is then zero and is consistent with the evolution of the universe from nothing.

If grand unified theories are correct in their prediction that baryon number is not conserved, there is no known conservation law that prevents the observed universe from evolving out of nothing. The inflationary model of the universe provides a possible mechanism by which the observed universe could have evolved from an infinitesimal region. It is then tempting to go one step further and speculate that the entire universe evolved from literally nothing.


Ford
FORD
Ford
MERCURY
LINCOLN

An Indian Hunters' Camp for 20,000 Years

A rock shelter in western Pennsylvania was periodically occupied from about 19,000 B.C. to A.D. 1000. The extreme age of the site bears on the question of when men entered the Western Hemisphere

by J. M. Adovasio and R. C. Carlisle

When did the people who first inhabited the New World arrive? Until recently scholarly opinion on the question was divided into two main camps. Some prehistorians suggested that migrations from Asia to North America began 30,000 or 40,000 years ago or even more. Others, citing a lack of firm evidence for such an early entry, placed the migrations at about 15,000 years ago or even less. Now, after almost a decade of field work at a sheltering overhang of rock in western Pennsylvania, the weight of evidence suggests that the answer lies somewhere between the two extremes.

The site, named Meadowcroft, is near the Ohio River 47 miles west of Pittsburgh. It is on the land of Albert Miller, a local historian with a strong interest in archaeology, and it was chosen for investigation more by accident than by design. In the early 1970's one of us (Adovasio) was looking for a place where students of archaeology at the University of Pittsburgh could be trained in excavation techniques. A colleague, the historian Phil R. Jack, suggested that Miller be approached for permission to work at a rock shelter on his land. Miller readily agreed, and in June, 1973, the work began. Here we shall describe the excavations and explain why the evidence of unexpectedly ancient use of the shelter by hunters and collectors has put a new complexion on the picture of early man in the New World.

The Meadowcroft rock shelter stands on the steep north bank of Cross Creek, a small stream that flows from east to west and enters the Ohio less than eight miles west of the site. Today the prevailing wind, blowing from the west, passes across the mouth of the shelter, providing almost constant ventilation. Created by the erosion of shales underlying the more resistant Morgantown-Connellsville Sandstone, the shelter is now 15 meters wide and six meters deep from the lip of its overhang to its back wall. It was once substantially larger, as is indicated by the numerous fallen blocks and smaller pieces of sandstone on and under the present shelter floor.

Before excavation began at Meadowcroft our field team made a detailed inventory of the kinds of plants growing around the shelter, inside it, on the slope above and on the slope below. When this had been done, all vegetation within 20 meters to the east and west of the midpoint of the shelter was stripped away to reveal the present topography. The area was mapped, and a basic grid system, consisting of two-meter squares, was established. (Where later work required a finer grid, the squares were subdivided into one-meter and even .5- and .25-meter squares.) A permanent elevation datum was then affixed to the back wall of the shelter, so that a series of points precisely describable in three dimensions were at the disposal of the excavators.

With these necessary preliminaries out of the way, the digging of the first trench began. It ran from south to north for approximately 12 meters, starting on the slope beyond the present drip line (where water dripping from the roof of the shelter hits the ground) and continuing on inside. The excavation revealed the presence of earlier drip lines and the sequence of their regression as parts of the roof fell away. This enabled us to calculate the amount of protected floor space available within the shelter at progressively earlier dates.

Over the next several seasons the extension of this initial trench to the east and west and the excavation of other trenches demonstrated to our surprise that the layers of sediment that had accumulated on the shelter floor were much thicker than the meter or so we had expected. Their total depth was in excess of five meters, and 11 distinct natural strata were evident. Within each of these strata, where natural microstrata were absent, the excavators proceeded by removing arbitrary one-centimeter, five-centimeter or 10-centimeter depths of sediments at a time. Both the principal strata and their natural subdivisions were defined by the sediments' texture, composition, degree of compaction, friability (tendency to crumble) and on occasion differences in color. We also observed that strata and microstrata of similar composition and degree of compaction often sounded the same when they were lightly tapped with the butt end of an excavation trowel. Thus in combination with more conventional field criteria even sound helped us to trace the course of individual levels within the rock shelter's deposits.

We assigned Roman numerals to the 11 strata, labeling the uppermost one Stratum XI and the lowermost, which contained no evidence of human activity, Stratum I. The lowest level that held evidence of human habitation, Stratum II, was further subdivided into IIa and IIb, the former being the lowest and oldest. The sediments from all the strata except Stratum IIa were sifted dry through a quarter-inch mesh to determine their contents, including seeds, nuts, bones and fragments thereof. The study of such faunal and floral remains helps us to determine what the natural environment at the site was like at various times in the past and also to reconstruct what its human inhabitants were hunting, collecting, eating or otherwise doing in the course of their stay. The sediments taken from Stratum IIa were sifted wet through an eighth-inch mesh. In addition to this relatively coarse screening process, samples were taken from every excavated square at the site (a total of some 147 square meters of surface both inside the modern drip line and outside it).

These samples, representative of every one-centimeter, five-centimeter or 10-centimeter downward cut in every stratum of the square, were treated in one of two ways. If the sample represented fill showing no sign of alteration by heat, or if it came from a cultural feature other than a fire pit or a hearth (such as a refuse pit or a storage pit), it was immersed in water and then passed through graded sieves. (The smallest sieve had the extremely fine mesh of 200 millimicrometers.) If the sample was from a fire pit or a hearth, the flotation was done not with water but with hydrogen peroxide. This minimized damage to any charred seeds, nuts or fragments of wood in the sample. In all the excavation removed a total of 230 cubic meters of sediment for screening.

Included in these totals are the yields from 12 sampling "columns." These were cut from the exposed trench faces and were subjected to geologic and other analyses. The columns extended from the site surface down to Stratum I and in two instances penetrated well into that stratum. Bulk samples were extracted from the columns at five- or 10-centimeter intervals. Each sample was subdivided, and the fractions were analyzed to determine such variables as the ratio of silt to clay and to particles as large as sand grains or larger, the amount of calcium carbonate in the material and the relative abundance of trace elements in each level. In addition to these geologic determinations the fractions were examined for microscopic plant and animal remains.

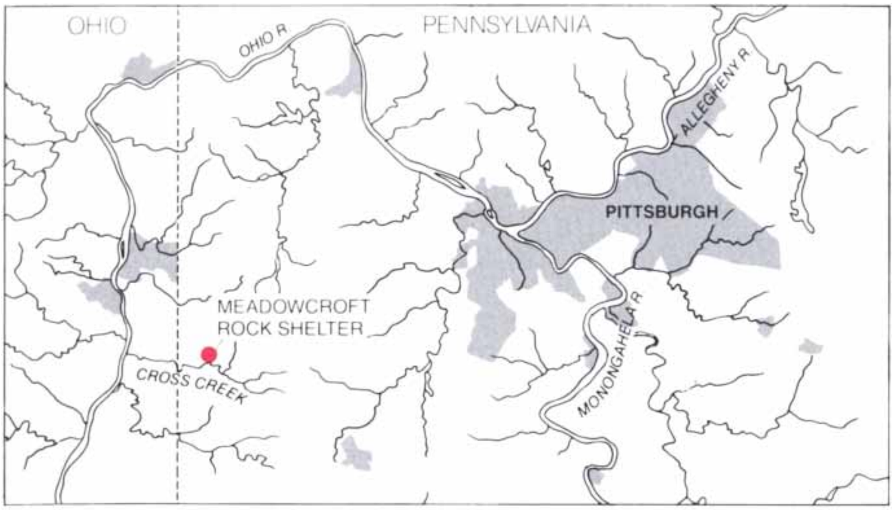
A further geologic project was the removal of rock samples from the cliff face itself. These were taken at 20- to 25-centimeter intervals from the base of the Morgantown-Connellsville Sandstone within which the rock shelter was formed to the top of that geologic formation, a distance of some 20 meters. The samples were used to prepare thin sections. The composition and the sand-grain size of the various sections were then compared with similar sample sections made from the sandstone that had fallen from the shelter roof in order to determine both the pattern and the mechanism of erosion that had shaped the shelter.

The greater part of the prehistoric record at Meadowcroft from upper Stratum II to upper Stratum IX belongs to the cultural period American archaeologists call Woodland. A long series of carbon-14 age determinations indicate that this period spanned some 2,250 to 2,500 years at the site, beginning in about 1115 B.C. To describe the more significant Woodland findings at the shelter it is convenient to reverse the order of their discovery and move from the past toward the present. Thus the earliest Woodland remains come from the upper part of Stratum III, where a large fire pit was found to contain sherds from several different pottery vessels. The wares, tempered with pulverized ironstone or sandstone and impressed with cord markings, are the earliest example of Indian pottery yet found in the upper Ohio River valley. Preserved squash seeds of the same age, that is, about 1115 B.C., indicate that horticulture, which is a key characteristic of the Woodland culture, probably reached this part of the upper Ohio more than 500 years earlier than had been supposed.

The continuity of horticulture is indicated by the presence of a squash seed in the lower part of the overlying level, Stratum IV, dating to about 865 B.C. The recovery of a carbonized cob of 16-row popcorn from the upper part of the same stratum, dating to between 375 and 340 B.C., establishes a new early date for the cultivation of maize in the upper Ohio area.

The arrival of horticulture roughly coincides with the end of the relatively intensive utilization of the Meadowcroft shelter. The immediate surround-

DEEP EXCAVATION at the Meadowcroft rock shelter has its walls marked with white tags of various sizes that designate the various strata at the site, locate such features as hearths, storage pits and work areas and identify points where samples were taken for such purposes as pollen analysis. The archaeological workers reached sterile soil (Stratum I) at a depth of more than five meters. They found that 10 overlying strata held evidence of visits to the site by people of three cultures: the Paleo-Indian, the Archaic and the Woodland.



MEADOWCROFT ROCK SHELTER stands on the north bank of Cross Creek, some 47 miles from Pittsburgh. Archaeology students began excavating there in the summer of 1973.

ings of the site lack the wide floodplains that were natural garden areas and village sites for the Woodland people. The shelter continued to be a hunters' haven and a station for the collection and processing of wild foods, but the evidence for such activities is generally scanty. The trend toward the abandonment of the shelter was doubtless accelerated by a major fall of roof rock at its eastern end sometime between A.D. 300 and 600. The fall reduced the shelter's available floor space by at least 20 percent. By the time the first European settlers made contact with the indigenous population in this part of Pennsylvania, early in the 18th century, the region around Meadowcroft appears to have been deserted.

The utilization of the Meadowcroft shelter by the Woodland people was preceded by some 7,000 years of visits by the hunters and gatherers of an earlier Indian culture: the Archaic. Moving again from older to younger levels, the Archaic Indians left evidence of their presence from the upper part of Stratum IIa, dating from about 8500 to about 8000 B.C., up to just below the upper part of Stratum III, dated at about the end of the second millennium B.C. The beginning of the Archaic period at Meadowcroft approximately coincides with the retreat of the last glacial ice sheet from North America and the beginning of the Holocene: the current geologic epoch.

From then until the arrival of the Woodland people the shelter and the surrounding region were periodically revisited by groups of Archaic Indians who practiced a hunting and collecting subsistence economy. It seems certain that Meadowcroft sheltered more human visitors as the Archaic period progressed, but it is not entirely clear whether this represents larger groups, a larger number of small groups or simply longer or more frequent visits.

Whatever the case, these people seasonally pursued an abundant fauna: deer, elk, wild turkey and diverse smaller game. That the Archaic people also exploited the plant resources of the area is documented by the presence in storage and refuse pits of, among other plant remains, nutshell fragments and the seeds and other parts of the hackberry and dwarf hackberry by the thousands in the storage and refuse pits. Also attributable to the Archaic hunters are a wide variety of stone projectile points, other stone artifacts and tools of bone and wood. Indeed, it may be said that Meadowcroft was never more heavily utilized than it was late in the Archaic period.

To have exposed a continuous record of prehistoric occupation spanning 10 millenniums at a site initially selected for teaching purposes was cause for considerable satisfaction. No one seriously expected that the part of Stratum IIa lying between the early Archaic levels above and the sterile Stratum I below would contain any surprises. In any event we concluded the 1973 field season by sending the first of some 100 charcoal samples, spanning the shelter's entire depositional sequence, to Robert Stuckenrath at the Radiation Biology Laboratory of the Smithsonian Institution for carbon-14 dating.

The first of the dates became available in the winter of 1973–74. They included those of samples Stuckenrath assigned to the 12th millennium B.C. At first we regarded dates this early with no small degree of skepticism. At that time we, like other conservative New World prehistorians, held the view that man had not reached the Americas until near the end of the Pleistocene epoch, some 15,000 years ago.

More samples from the site were sent to the Smithsonian for dating at the end of the 1974 field season and still more were sent over the next two years. By the time the latest dating results were received in the winter of 1978, no one connected with the Meadowcroft project any longer had reservations about the comparatively great age of Lower Stratum IIa. From bottom to top the sequence of carbon-14 dates was not only internally consistent but also consistent with the site stratigraphy. The earliest dates now showed a range from 9350 ± 700 B.C. (Middle Stratum IIa) to 17,650 ± 2400 B.C. (Lower Stratum IIa). Recently one even earlier carbon-14 age determination was made by a commercial radioactive-dating laboratory. The sample was found to date to 19,120 ± 475 B.C. It came from the deepest level in Stratum IIa, but it is apparently not associated with human activities.

The earliest inhabitants of the Western Hemisphere are assigned to a rather broadly defined cultural stage: the Paleo-Indian. Its most familiar characteristics are the hunting of big-game animals (most of which are now extinct) and the manufacture of elegantly fluted stone projectile points. The evidence that Paleo-Indians had reached Meadowcroft, in eastern North America, more than 19,000 years ago generated considerable controversy among American archaeologists. The crux of the controversy is precisely when the first emigrants from Asia crossed over the land bridge in what is now the Bering Strait. If some of them had reached the upper Ohio River by 17,000 B.C., it seems logical that, allowing some time for the migration east, they had crossed the land bridge at least 2,000 or 3,000 years earlier. Yet such a time of entry is thousands of years before the rough date customarily assigned to the event: 13,000 B.C. or later. This left those who still held to the customary view recourse to only one argument: The early dates at Meadowcroft must be mistaken, perhaps because the charcoal samples had been contaminated.

How could that be the case? One suggestion was that coal beds half a mile from the site were a source of "old carbon": particulate matter that had contaminated the carbon samples from Meadowcroft and made them appear to be older than they actually were. Alternatively, the ground water at the site could have been rich in dissolved "old carbon" from the coal deposit that could have had the same effect on the samples.

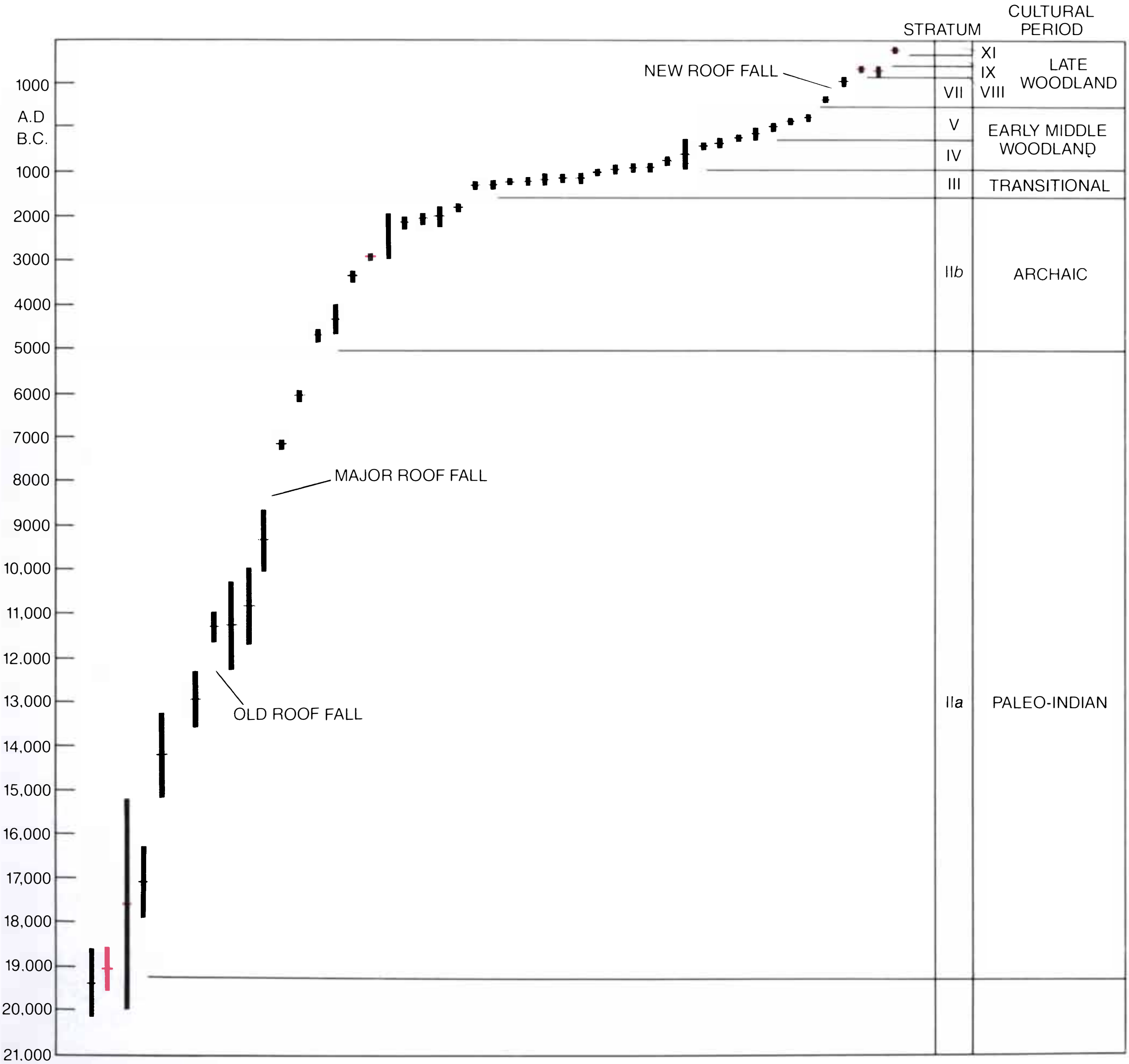
Actually, as part of the normal pretreatment procedure at the Smithsonian laboratory, each of the Meadowcroft samples had been microscopically examined for possible particulate contamination. None of the samples showed any signs of such contamination. As for the possibility of contamination by dis-

solved "old carbon," we supplied Stuckenrath with samples of wood of the notably coal-rich period generally known as the Carboniferous (in American usage the Pennsylvanian) from the vicinity of the rock shelter. This was the only remotely possible source of old-carbon contaminants in solution anywhere near the site. He was unable to dissolve the samples by boiling them either in water or in several other solvents. It is thus apparent that neither proposed source of potential contamination was active at Meadowcroft.

Critics of the site's oldest dates seem to have overlooked a further strong argument against their position. Surely any mechanism of contamination that involves percolating ground water would not have selectively affected only the oldest samples at the site. Yet the more recent dates from the site—for the terminal Paleo-Indian, the Archaic and the Woodland levels—have never been challenged. Indeed, as we have noted, they are remarkably consistent.

As a final word on the subject, it is worth mentioning that the one deep-level sample that was dated by a commercial laboratory was submitted "blind." Neither the site name nor the stratigraphic position (actually Stratum IIa) of the sample was stated. The laboratory was told only that the sample came from the general area of Cross Creek and that it was probably more than 9,000 years old. The processing and counting procedures were different from those employed at the Smithsonian laboratory. The sample was found to be the oldest of any from the site (19,120 ± 475 B.C.), which is consistent with its great depth in the rock-shelter deposit.

What are the earliest cultural remains recovered from Meadowcroft? In order to answer the question it is necessary to describe Stratum IIa and its various dates in detail. Overall the stratum consists of sediments derived principally from the weathering of the roof and walls of the shelter. It is divisible into three subunits of unequal thick-



CARBON-14 DATES obtained from charcoal samples unearthed at the rock shelter appear in this table, with the oldest samples at the left and the youngest at the right. A colored dash indicates the median date in each case; the length of the vertical bar shows plus-minus values. All dates except the one at the lower left (*color*) were determined at the Radiation Biology Laboratory of the Smithsonian Institution.


We built Laser XE to outperform the competition.
We gave it a turbo you can trust.
We built Chrysler Laser XE to win. We engineered its turbo to endure.
Laser's turbo is the sophisticated new wave. A water cooled bearing reduces critical turbo temperatures to prevent oil "coking" and bearing failure. Horsepower is boosted 43%. A multi-point injection system "spritzes" fuel in at 4 points and moves Laser like light. With 5-speed your time to 50 mph is 5.4 seconds. Camaro Z28, Trans Am, Toyota Supra and Nissan 300 ZX are in your remote-controlled side view mirrors.
We gave Laser XE world-class performance. In the slalom Laser beats all entries—from Trans Am to Mustang GT.
Laser does it when you equip it with turbo and performance handling package with nitrogen charged shocks.* Laser does it with front wheel drive, dual path suspension system and quick ratio power steering. In the slalom Laser finishes No. 1.
We gave it high-performance braking. Laser XE stops you where Trans Am doesn't.
We think total performance calls for performance braking. So we gave Laser XE semi-metallic brake pads, power brakes all around and optional wide 15" alloy wheels with Goodyear Eagle GT radials. Result: Laser stops quicker than Z28, Mustang GT, Toyota Supra, 300 ZX, Trans Am.
We gave it a brain—and a performance seat that performs.
Laser XE thinks with you. Its 19-feature electronic monitor even talks your language, while its color graphic displays make you a calculating driver. Laser XE's AM/FM stereo remembers what you like to hear and its self-diagnostic system is the nearest thing to an on-board mechanic. Your seat responds with cushions you pump up for thigh and lumbar support, and you can choose a 6-way power seat and Mark Cross leather.
We gave it our best: a 5 year/50,000 mile Protection Plan. Even your turbo is protected.
We believe a performer has to be a survivor, so we back your entire powertrain with 5/50 protection, with outer body rust-through protection for the same period.** See dealer for details. Buckle up for safety.
35 Est. Hwy
22 EPA Est. MPG†
Chrysler
The best built, best backed American cars.††
*Based on overall results of USAC tests against standard equipped models. Laser XE equipped with optional handling suspension. Turbo package and 15" road wheels and tires. **5 years or 50,000 miles, whichever comes first. Limited warranty. Deductible required. Excludes leases. †Use EPA est. mpg for comparison. Actual mileage may vary depending on speed, trip length and weather. Hwy. mileage probably less. ††Based on lowest percent of National Highway Traffic Safety Administration recalls for '82 and '83 models designed and built in North America.
Laser beats Nissan 300 ZX in the slalom.
Laser outperforms Trans Am in braking.
Laser is faster than Camaro Z28 from 0-50 mph.
50
CHRYSLER
LASER
"THE COMPETITION IS GOOD. WE HAD TO BE BETTER."
Lee A. Iacocca

ness. At the top of the uppermost subunit, Upper Stratum IIa, is a human occupation floor that yielded charcoal dating to 6060 ± 110 B.C. A slightly deeper occupation floor within the same subunit dates to 7125 ± 115 B.C. At the base of the subunit is a substantial quantity of fallen roof rock that marks the boundary between the upper subunit (early Archaic) and Middle Stratum IIa (Paleo-Indian). Directly under the roof rock another occupation floor yielded a still older date: 9350 ± 700 B.C. No further dates have been determined for the remainder of the second subunit, the bottom of which overlies another layer of fallen roof rock. A sample from an occupation floor directly under this layer, at the top of Lower Stratum IIa, is dated to 10,850 ± 870 B.C. Seven still lower samples from this deepest subunit range in age from 11,290 ± 1010 to 17,650 ± 2400 B.C.

To round off these dates, one may say that Upper Stratum IIa dates from 9000 to 6000 B.C. and is therefore entirely of Holocene age. Middle Stratum IIa in turn may be said to date from 11,000 to 9000 B.C. and is therefore of terminal Pleistocene age. Finally, the parts of Lower Stratum IIa that show evidence of human occupation may be said to date from 17,500 to 11,000 B.C.

The oldest evidence for the use of stone tools at Meadowcroft consists of 123 complete or fragmentary tools and several hundred waste flakes from the manufacture and maintenance of the tools. All were associated with occupation-floor features in Middle and Lower Stratum IIa. Some of the tools are bifacial (flaked on both sides). A few knives were flaked on only one side. Among them is a distinctive form of knife, the Mungai, named after another Cross Creek site, Mungai Farm. Also present were blades, unshaped flakes that had been utilized in various ways and fragments of the "cores" from which the tools had been struck.

The living floor at the top of Middle Stratum IIa yielded the first and oldest projectile point to be found at Meadowcroft. The type has been named Miller Lanceolate, in honor of our host at Meadowcroft, Albert Miller. Other projectile points of the same type have been found at the Mungai Farm site, the Pershina site (which will be further investigated this summer) and at other places within or near the Cross Creek drainage. So far these sites have not yielded Miller Lanceolate points in any stratigraphic context that would make it possible to date them directly. The Miller Lanceolate from Middle IIa, however, is one of the oldest projectile points yet unearthed in eastern North America. It may be locally ancestral to the widely distributed Paleo-Indian points named Clovis.

These deepest levels at Meadowcroft lay outside the shelter's present drip line, and so the organic material in them was poorly preserved. Fortunately enough material was recovered to yield a general picture of the environment at the time the early Paleo-Indians visited the area. For example, pollen and other plant remains indicate that both deciduous trees and conifers grew in the vicinity. As for animals, the hunters at the very least had at their disposal white-tailed deer and smaller game such as chipmunks and southern flying squirrels. It is most unlikely that there were no other animals in the Meadowcroft area, but none left any identifiable trace in the lowest levels at the site.

PROJECTILE POINT appears in situ, partially cleared of its surrounding sandy matrix, just above the centimeter rule in this photograph. The excavators named this point and similar ones found in the area Miller Lanceolate points for Albert Miller, who owns the land on which the Meadowcroft shelter is situated. This is one of the oldest Paleo-Indian points so far uncovered east of the Mississippi River: it is dated at between 10,850 ± 870 and 9350 ± 700 B.C.

Exciting as it is to have established the presence of hunting bands in western Pennsylvania some 16,000 to 19,000 years ago, the ultimate importance of the Meadowcroft–Cross Creek Archaeological Project will come from much more than the contents of the rock shelter's deepest strata. In addition to exposing the Archaic and Woodland levels at the shelter, our multidisciplinary work went on to embrace the entire Cross Creek watershed, an area of 55 square miles, beginning at the confluence of the stream with the Ohio River and extending east and southeast to the stream's headwaters. Between 1973 and 1978 field workers recorded 231 prehistoric sites in the watershed. Of these, 17 were extensively tested and two were fully excavated. The sites contain Paleo-Indian, Archaic and Woodland artifacts.

The archaeologists of the Meadowcroft–Cross Creek project have collected some 20,000 artifacts: both flaked and ground stone and tools made of bone and other perishable materials. The project zoologists have amassed nearly a million animal bones and the botanists have collected 1.4 million plant specimens. The geologic samples number in the thousands. Data from all these collections have been or are being prepared for computer analysis. They offer an unparalleled opportunity to examine the adaptations of different human populations to life in a single drainage system over a period of nearly 20 millenniums.

The questions that may be answered before the project's work is done are many and pointed. As one example, having to do with the terminal Archaic, one may wonder what effects the post-Pleistocene climate had on the people of the Cross Creek area. Was this climatic change responsible for the increase in the number of Archaic sites and the appearance of "base camps" at the rock shelter and elsewhere? As another example, is the diminishing collection of wild plants in the Woodland period correlated with the arrival of the cultivation of squash and maize? What factors explain the subsequent disappearance of base camps from Meadowcroft and from the Cross Creek area in general? If our project had accomplished nothing else, it could stand as an unprecedentedly successful effort to create a detailed interdisciplinary data base, one that will make it possible to conduct an intensive study of the ecology of a single region over a very long span of prehistoric time. There is good reason to hope this wealth of information will prove useful not only to prehistorians but also to scholars in other disciplines.

SCIENCE/SCOPE

Batteries that may live seven times longer than the spacecraft themselves will be employed on Intelsat VI communications satellites to ensure performance. The nickel-hydrogen batteries provide power when the satellite slips into Earth's shadow and the solar cells no longer generate electricity. They have demonstrated 6,000 discharge cycles in tests, the equivalent of 70 years in geosynchronous orbit. Hughes Aircraft Company heads an international team building Intelsat VI for the International Telecommunications Satellite Organization.

A new video graphics projector that's brighter and sharper than conventional projection TV may be the next addition to office computer systems. The Hughes projector displays monochromatic computer-generated alphanumerics, symbols, and graphics. It could be used for displaying dynamic computer data and facsimile video pictures in board rooms and other areas, and for teleconferencing. The projector uses a device called a liquid-crystal light valve, a cousin of displays in digital watches. This device intensifies the image from a cathode-ray tube and projects it onto a screen up to 12 feet wide. The picture is so bright and has such high resolution that the viewing room needn't be darkened.

A novel engineering tool for producing the AMRAAM missile is expected to save the U.S. government and Hughes millions of dollars and months of work. A full-scale prototype of the Advanced Medium-Range Air-to-Air Missile has been completed using actual engineering drawings, materials, and processes. The purpose of this "precision physical model" is to refine AMRAAM's design and detect potential manufacturing problems, especially those stemming from late improvements. Among other things, the model has been used to determine routes and lengths for wire harnesses so that mating connectors will line up. It also was used in designing handling and test fixtures, and to show how its components react to vibration. AMRAAM is in full-scale development for the U.S. Air Force and Navy.

NATO early-warning aircraft are being equipped with a communications system that uses four primary encoding techniques to hamper enemy eavesdropping or jamming. The Joint Tactical Information Distribution System (JTIDS) provides E3A AWACS aircraft and NATO ground command centers with secure voice and digital communications. One JTIDS encryption technique is spread spectrum, in which a signal is expanded over a large bandwidth. With frequency hopping, a second method, frequencies are changed many times a second. Another technique, time division multiple access, assigns certain users to specific time slots no longer than a fraction of a second. Finally, to verify messages, JTIDS repeats messages automatically. Hughes is supplying JTIDS to NATO and the U.S. Air Force.

The Hughes Tucson facility, located in picturesque Southern Arizona, is a large, modern manufacturing complex with capabilities for producing advanced missile systems developed by Hughes. We have openings for experienced and graduating engineers to work on such advanced systems as the electro-optical Maverick, radar-guided Phoenix, TOW, and AMRAAM, the Advanced Medium-Range Air-to-Air Missile. Please send your resume to Professional Employment, Dept. S2, Hughes Aircraft Company, P.O. Box 11337, Tucson, AZ 85734. Equal opportunity employer.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

For additional information please write to:
P.O. Box 11205, Marina del Rey, CA 90295

Island Epidemics

Epidemics are patterns in time and space that can best be perceived when they are studied in a small, isolated population. An example is provided by a study of a century of measles epidemics in Iceland

by Andrew Cliff and Peter Haggett

Ever since Charles Darwin studied the finches of the Galápagos Islands in 1835 it has been recognized that oceanic islands can serve as large-scale laboratories for the investigation of biological processes. Not long after Darwin did his work it became clear that remote islands, which offer a simpler site for study than the crowded continental mainland, can also make a valuable contribution to the understanding of the spread of disease. In the late 1840's Peter Panum, a Danish physician, made a study of the 1846 measles epidemic on the Faeroe Islands that has become one of the classic works of epidemiology, the discipline concerned with the patterns of disease in human populations. Today most advances in the study of diseases caused by viruses are made in laboratories constructed by man; for instance, the virus that causes measles was isolated in such a laboratory by John F. Enders and Thomas C. Peebles of the Harvard Medical School in 1954. Nevertheless, islands continue to play a significant role in the elucidation of how diseases spread under epidemic conditions.

An example of an island laboratory is the North Atlantic island of Iceland. The island republic has an area of about 40,000 square miles, roughly the same as that of Kentucky. It lies just south of the Arctic Circle, between 64 and 66 degrees north. The island is sparsely settled; its population has reached 200,000 only in recent years. Until air travel became common it was in an epidemiological sense isolated from many population centers in Europe. The most frequent contacts were with the Scandinavian countries, Denmark in particular. As a result of the severe winters residents of the island were isolated even from one another.

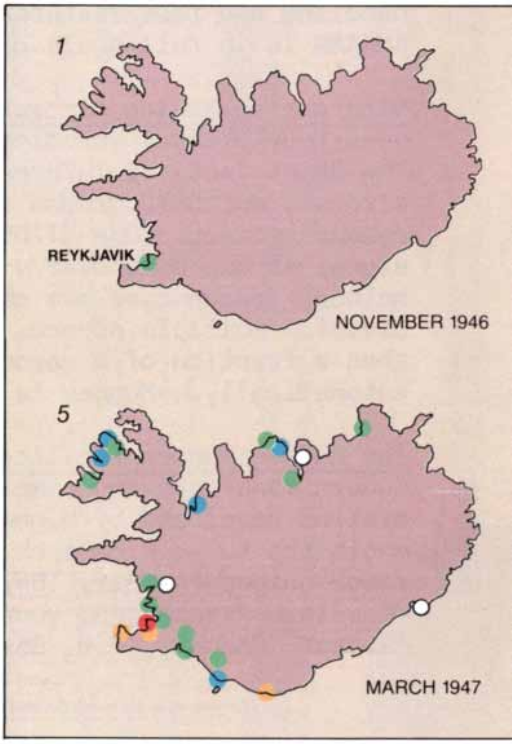
Since 1896 Iceland has had an unusually complete and trustworthy system of medical records that cover some 50 medical districts on a monthly basis. The combination of isolation, a small population and good medical records makes it possible to trace the pathways by which disease is transmitted in Iceland with more precision than such pathways can be traced in any other country.

The detail available in Iceland's public-health records could be employed to study many communicable diseases. Among the possible subjects measles is of particular interest for two reasons. The first reason is that measles is an excellent example of a disease that spreads rapidly in epidemic waves. It is highly infectious and readily recognized, and the epidemics tend to come at regular intervals. The second is that although in the U.S. measles has been all but eradicated by vaccination programs, it remains a serious threat to public health in the poorer countries of the world. In such countries the death rate in measles epidemics can be as high as 30 percent, particularly among children. Recent United Nations statistics show that in the world as a whole measles (and its complications) is one of the 10 most frequent causes of death.

Working with the unique public-health records of Iceland, we have shown that measles tends to spread in a hierarchical pattern according to the size of the community: from the capital city of Reykjavik to regional centers to villages and then to isolated farms. The records also show that when social and technological development changed the relations between the capital and the hinterland, the epidemiology of measles was profoundly altered. Such work could help in gaining an understanding of how the pathways along which measles spreads are determined. If the underlying factors are found, it might be possible to interrupt the pathways by selective programs of vaccination.

The virus that causes measles is a member of the family of paramyxoviruses, which includes the viruses that cause mumps and parainfluenza. When smallpox was eliminated after a long campaign and was dropped from the World Health Organization's list of infectious diseases in 1979, the attention of epidemiologists turned to other viruses. The measles virus became a prime candidate for elimination. One reason for thinking it could be eliminated is that if a community is small enough, the virus periodically disappears without any human intervention.

The periodic disappearance is due to the fact that in a small community the chain of measles infection is readily broken. Having measles confers a lifelong immunity to the virus, so that in such a community after a major epidemic the susceptible population is very small. Therefore even if the virus is present in the community, those who are infectious will not have contact with enough susceptible people for the chain of infec-



MEASLES EPIDEMICS in Iceland generally have the three stages that are shown in the sequence of maps. The circles are the medical districts reporting cases of measles in the epidemic lasting from November, 1946, through

tion to remain unbroken. In a large community, on the other hand, the pool of those susceptible to the disease is large enough for some chains of infection to remain unbroken. As a result the virus is endemic, or continuously present.

How large must a human population be to maintain the measles virus on an endemic basis? Maurice S. Bartlett of the University of Oxford concluded from his study of U.S. and British cities that between 4,000 and 5,000 reported cases of measles per year would suffice to sustain the disease continuously. Bartlett's work was done in 1957. The level of measles attacks and the rate of reporting of the disease to public-health authorities that then prevailed in the U.S. and the United Kingdom imply that 4,000 to 5,000 measles cases per year corresponds to a population of about 250,000. In a community with a population of less than 250,000 the virus apparently dies out, only to be reintroduced later by infected individuals coming from an area where infection is present. Large cities, where the disease is endemic, can serve as permanent reservoirs of infection. A person who introduces the virus into a small community where it was not present is called an index case. As we shall see, index cases are of much significance for tracing the pattern of measles epidemics.

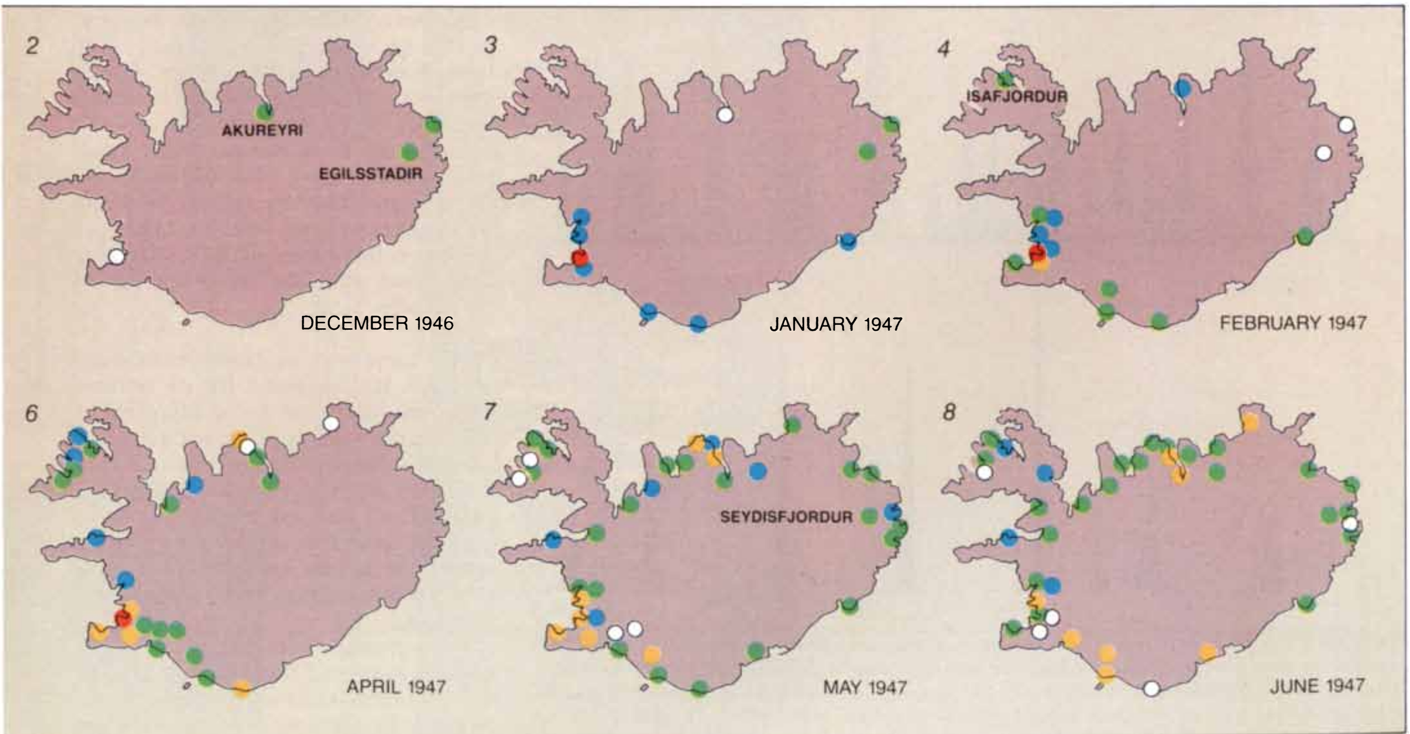
It is clear that the pattern of epidemics in a small community depends on the frequency of contact between the small community and the larger centers of population where the virus is endemic. The frequency of contact in turn depends on the size of the small community, its distance from the big cities and the efficiency of the transportation networks that connect the small community and the large ones. If a community is small or remote, it will escape some of the epidemic waves that emanate at regular intervals from the cities.

Such demographic and geographic factors have enabled us to divide human communities into three groups. The division is done according to the form of the curve that traces the number of cases of measles in the community. The peaks in the curve correspond to epidemics. In large communities the curve has the form of a series of Type I waves. Such waves are regular and continuous: between epidemics the number of measles cases never reaches zero. Medium-size communities have Type II waves, which are regular but not continuous: between epidemics no measles cases are present. The smallest communities have Type III waves, which are discontinuous and irregular: because of their size and remoteness such communities miss some epidemics.

A decade after Bartlett's study Francis L. Black of the Yale University School of Medicine reexamined the relation between population size and epidemic waves but instead of working with cities he measured rates of infection in populations that are spatially even more isolated: the populations of 18 oceanic islands. Of the 18 islands only Hawaii, which then had a population of 650,000, showed the continuous Type I pattern. The Type II pattern was found on islands with a population as small as 10,000. Islands with fewer than 10,000 residents showed the Type III pattern.

After the work of Bartlett and Black a logical next step was to examine how measles spreads both between Type II islands and on a Type II island. For this purpose the 18 islands in Black's sample provided a convenient starting point. In the group of islands Iceland appeared to be the best laboratory for the work we planned to do (with J. K. Ord of Pennsylvania State University and the Icelandic health authorities).

Iceland is the largest island of the Type II group and has the second-largest population of the 18 islands. Its population is relatively small compared with the population of mainland regions, however, and the stern terrain and harsh climate of the interior have largely limited human settlement to the coastal lowlands. The coastline is deeply indented by fjords and the Icelandic communities tend to be remote from one another. Indeed, from an epidemiological point of view it is often appropriate to think of Iceland more as an archipelago than as a single island.



June, 1947. Blue stands for districts where one or two cases were reported; green, three to 20 cases; orange, 21 to 200 cases; red, 201 or more cases. Open circles stand for districts where no cases were reported but where cases had been reported the previous month. The epidemic was the eighth of 16 waves of measles that swept the island between 1896 and 1975. In the first stage cases appear in the capital city, Reykjavik (*1*). In the second stage measles is carried to the cities of Akureyri, Isafjordur and Egilsstadir and to the area surrounding Reykjavik (*2–4*). In the third stage the disease spreads around each of the regional urban centers and reaches the remote communities (*5–8*).

The Icelandic demographic records show that the population increased from fewer than 50,000 in the late 18th century to about 200,000 in 1970. About half of the population lives in Reykjavik and its suburbs, a fourth in the other 11 towns and the rest in the rural areas. The fourfold growth in population over the past two centuries has not been steady. The increase has been repeatedly interrupted by epidemics of infectious disease. The most potent of them was the smallpox epidemic that swept the island in the 1780's, but there were also, in addition to the epidemics of measles, significant outbreaks of influenza, typhoid fever and whooping cough.

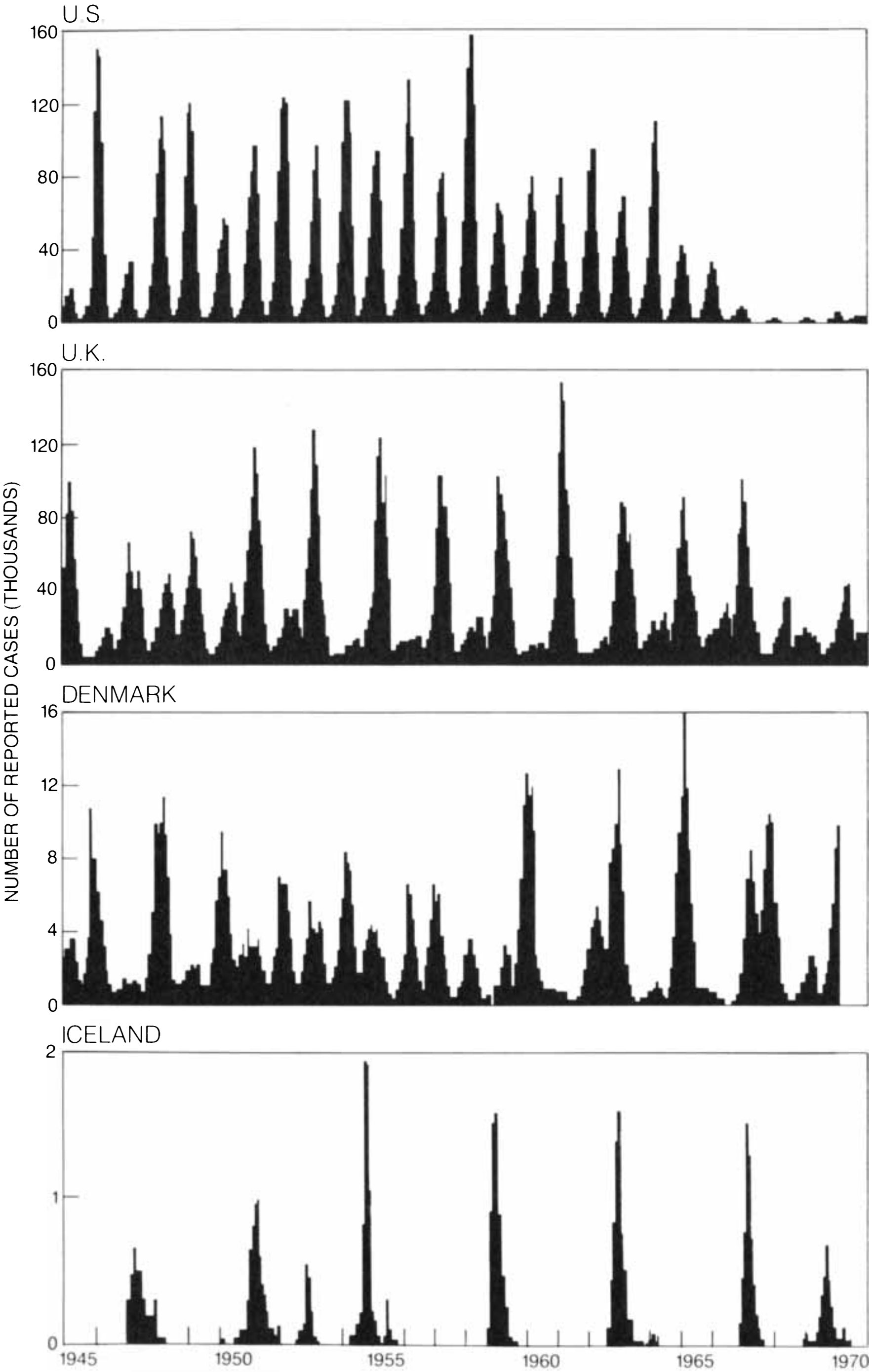
Since 1896 the chief physician of Iceland has published a detailed annual account of public health called *Heilbrigdisskýrslur* (literally "Health Reports"). The account includes monthly reports of the number of cases of major diseases in each of the 50 medical districts of the island. In addition it includes narratives by local medical officers describing the course of the epidemics in their district. The narratives give details of the severity and pattern of spread of each epidemic, the external source of the disease (if the source was known) and how the virus was diffused from village to village or from farm to farm. Even more detailed accounts are available in manuscript in the national archives.

Thus *Heilbrigdisskýrslur* and the manuscripts in the national archives provide two significant types of information. The district statistics yield a picture of the intensity and duration of an epidemic wave. The medical officers' narratives make it possible to reconstruct in detail the spread of the disease between geographic areas.

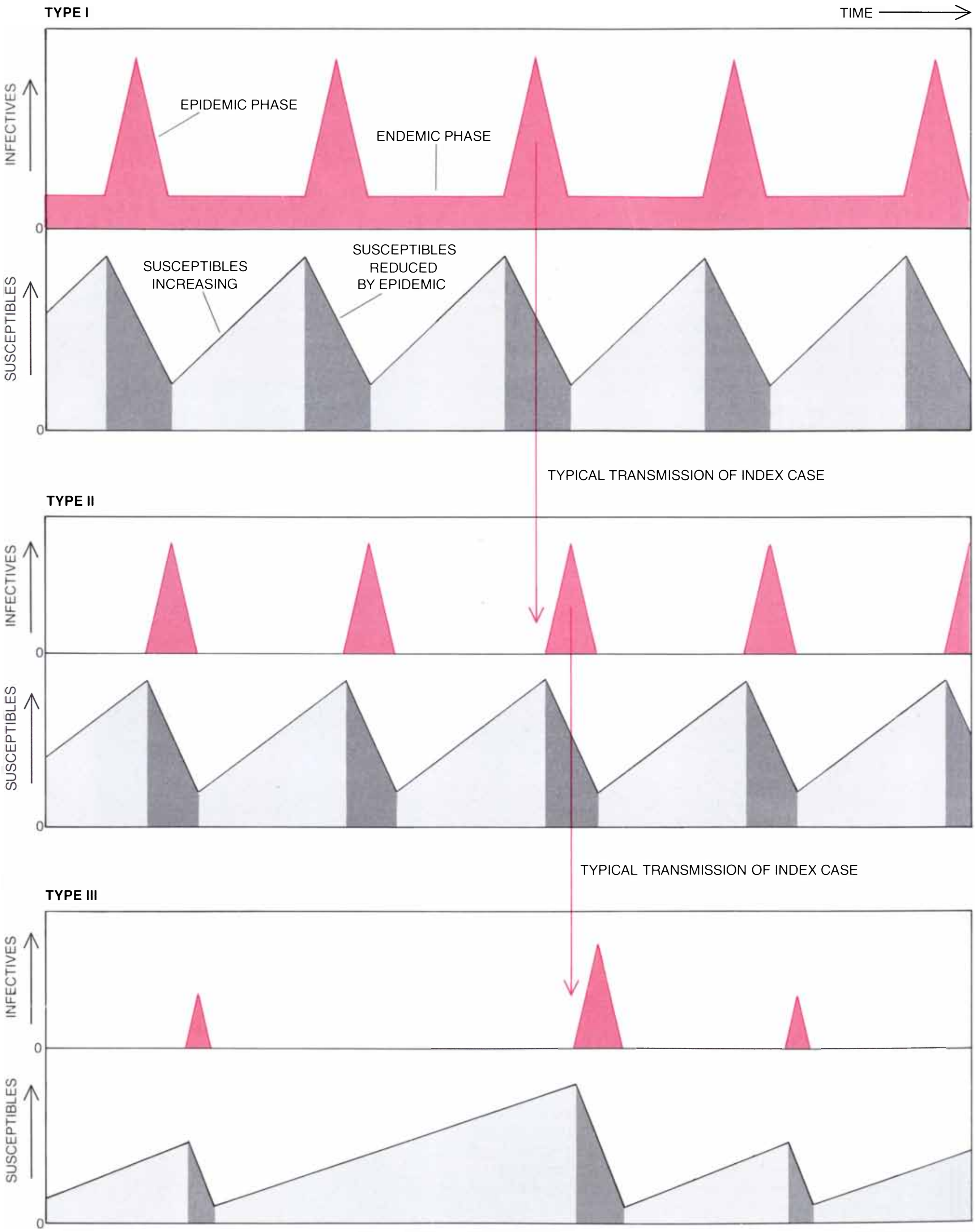
The records show that from 1896 through 1975 there were 16 distinct waves of measles. (Since our work was done, there have been two more waves.) The epidemics varied in size from wave 1 in 1904, which included 822 reported cases, to wave 6 in 1936–37, which included 8,408 reported cases. In most countries medical reporting has improved greatly over the decades, but in Iceland even the early measles records appear to be fairly accurate. The waves lasted on the average for 19 months, and (leaving aside the unusually small and localized wave of 1904) in a typical epidemic only five of the 50 medical districts escaped infection. Between the waves there were periods averaging three years when the island was free of the measles virus.

The narratives in *Heilbrigdisskýrslur* show that in order for an isolated Icelandic farmstead to be attacked by measles the virus must be carried a considerable distance by one infected individual or, what is more probable, by a chain of infected individuals. The virus must be carried several hundred miles across the sea to an entry port, then to a town or village and finally to the farmstead.

The general mechanism whereby measles is spread through human contact has been known for more than a century. In most epidemics, however, and in particular those that come in densely populated urban areas, the pathways of transmission are obscured by the large numbers of infected people mixing with the susceptible population.



POPULATION SIZE strongly affects the pattern of measles epidemics, as is indicated by these four curves showing the number of reported measles cases in the U.S., the United Kingdom, Denmark and Iceland in each month from 1945 through 1970. In the U.S., which had a population of 200 million in 1970, measles epidemics recurred annually until the mid-1960's, when a vaccination program began to eliminate the measles waves. Between epidemics measles was always present at a low level in the U.S. population. In the U.K., with a population of 50 million in 1970, and Denmark, with a population of five million, there was a similar pattern except that the measles epidemics tended to come respectively at intervals of two and three years. In Iceland, with a population of 200,000 in 1970, the epidemics were less regular. Between epidemics there was a period averaging three years when the virus was not reported on the island.



EPIDEMIC WAVES can be divided into three types according to the form taken by the curve that corresponds to the number of measles cases. In each panel the upper curve stands for the number of people in the community who are infected. The lower curve stands for the number who are susceptible to measles. Large communities show Type I waves (*top*). The virus is endemic: between epidemics there are always a few cases of measles. Medium-size communities show Type II waves (*middle*). These waves are regular but discontinuous; the population is too small to sustain the disease on an endemic basis. Between epidemics the virus dies out and is reintroduced from the large community during an epidemic there (*arrows*). Small communities show Type III waves (*bottom*). These waves are discontinuous and irregular; the size and the remoteness of the settlements protect them from some epidemics coming from the larger communities.

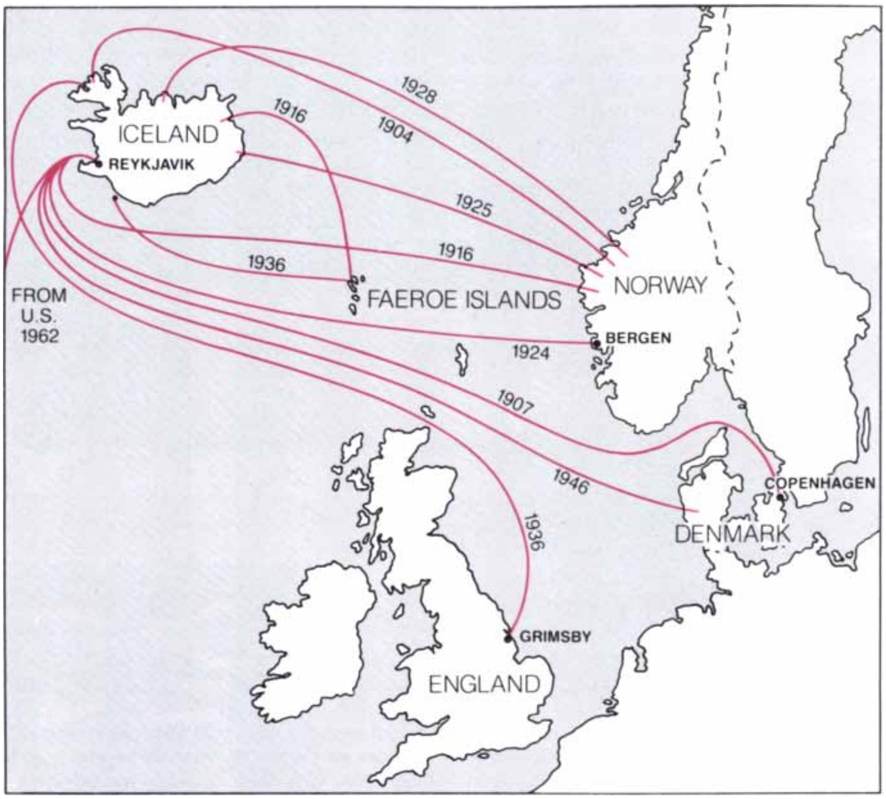
In formulating hypotheses on the spread of infection in such areas it has generally been necessary to adopt the simplifying assumption that contact in a human community is a kind of churning process where the motions of an individual are random. The particular value of *Heilbrigdisskýrslur* is that the reports enable the epidemiologist to discard that simplifying assumption and replace it with the actual movements of people carrying the virus between settlements.

Examination of the official records shows that the main reservoir from which measles was brought to Iceland is northern Europe. When the virus reaches the island, it can spread along several routes. Wave 1 of 1904, which was confined to the fishing villages and isolated farms of northwestern Iceland, provides an interesting example of the process of virus introduction and transmission.

The statistical data for wave 1, the first wave to be recorded in *Heilbrigdisskýrslur,* are not complete. There are 822 cases in the published records, but 1,993 cases are noted in the manuscript reports in the national archives. Some 40 percent of the 1,993 cases came in August, the peak month of the epidemic. Most of the cases were in Isafjordur, the main town of the region, where the physician was so overwhelmed that he was not able to keep detailed records; hence the discrepancy between the official total of 822 and the probable total of about 2,000. Twenty-three deaths attributable to measles were reported.

The path of movement of the disease was carefully noted in the narrative: "The disease came from Norway in April to the whaling stations in Hesteyri and Isafjordur. In the first place it stopped and there were no further infections, but in the latter it spread because the doctor was not informed at an early date. It spread most rapidly after a confirmation ceremony at the village of Eyri on May 22, 1904." From Eyri the epidemic was transmitted throughout the western part of the island.

This short excerpt from the official record throws light on three factors that are of much significance in understanding the epidemiological history of measles in Iceland. The first factor is the crucial role of the index case, whose spatial mobility is necessary to start a measles outbreak. In wave 1 the index cases were the crew of a Norwegian whaling ship; in other waves the index cases were fishermen. The second factor is the contribution of communal activity to the spread of the disease. Communal activities served to bring the susceptible population together in the critical period when the virus had recently been brought into the area. The confirmation held at Eyri on May 22, 1904, is an example of a gathering of the local populace after which the virus was rapidly transmitted. The outbreak in Isafjordur began with "an adolescent girl who had been confirmed at Eyri and came from there to the town for domestic service."



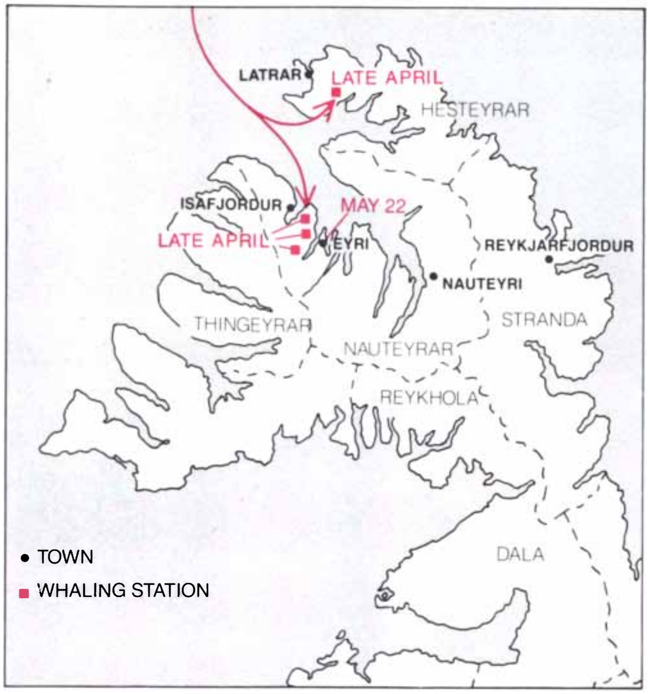
INDEX CASE is an infected individual who introduces measles into a population where the virus has died out. The map shows the routes that were taken to Iceland by the index cases who began measles epidemics on the island (where the route is known) from 1896 to 1975.

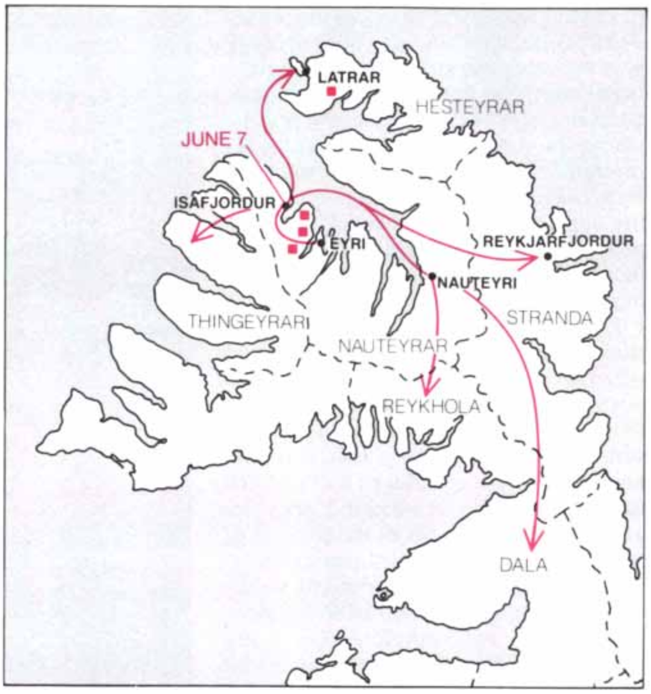
The third factor that influenced how fast and how far the epidemic spread was the countermeasures implemented against it. The countermeasures generally took the form of interrupting contact between those who were infected and the rest of the people in the community: putting ships' crews in quarantine, putting measles patients in fever hospitals or isolating an infected farm household.

When the movement of the index cases and the resulting spread of the disease are traced in all 16 measles waves, it becomes clear that measles epidemics in Iceland tend to have three stages. In the first stage the virus is introduced to Reykjavik from outside Iceland, or is introduced from elsewhere in Iceland to Reykjavik soon after it arrives on the island. In the second stage the infection spreads from the capital to the provincial urban centers: Isafjordur in the northwest, Akureyri in the north and Egilsstadir and Seydisfjordur in the east. At the same time the virus diffuses out from Reykjavik into the region in southwestern Iceland surrounding the capital. In the third stage the disease spreads from the regional centers into their hinterland and begins to reach the outlying farms and fishing villages.

The account in *Heilbrigdisskýrslur* of the progress of the 16 waves of measles through the medical districts can also be utilized in a quantitative way to measure the time taken by the virus to reach a particular area. Such measurements confirm the qualitative three-stage model. The official record shows that Reykjavik is reached on the average within 1.5 months of the time the virus arrives on the island. The figures for the regional centers are as follows: Isafjordur 6.5 months, Akureyri from five to eight months, Egilsstadir five months and Seydisfjordur 9.3 months. The only small towns and villages reached before the provincial centers are those near the capital. The isolated parts of northwestern Iceland are generally not reached until from 12 to 18 months after the start of the outbreak in Reykjavik.

Another chronological pattern that is of interest is the seasonal trend in the diffusion of the disease. If the movements of the index cases in the 16 waves are classified according to the month in which the disease was transmitted, it becomes evident that the spread of the disease has seasonal peaks and dips. In northern temperate climates measles, like many other infectious diseases of childhood, most frequently spreads in the winter. In Iceland, on the other hand, over much of the period since 1896 the transmission of measles has shown two peaks: until 1946 there was a peak in





MEASLES EPIDEMIC OF 1904 was largely confined to the sparsely populated northwest peninsula of Iceland. The lines on the map show the routes of the index cases who spread the disease. The virus was brought to Iceland in late April by the crew of a Norwegian whaling ship (*left*). Measles appeared first in a group of whaling stations. At a confirmation service at the nearby village of Eyri on May 22 many residents of the area were exposed to the measles virus. From there the infection was carried to Isafjordur, the largest city in the area, in early June (*right*). From Isafjordur the measles virus was dispersed throughout the entire peninsula during the summer months.

May and June in addition to the winter maximum. The local medical officers offer two reasons for the island's having a summer peak. One reason is that human mobility, which was severely limited by the subarctic winters, increased after the spring thaw. The other is that the major communal activity of haymaking began in June.

Tracing the movements of the index cases around the island can yield much information about the geographic and chronological patterns of measles waves. In order to apply the categories derived from Bartlett and Black, however, a count must be made of measles cases in each community over a period of time. When the number of cases and their timing are noted, it is found that since 1945 Iceland as a whole has shown the Type II pattern: the waves were fairly regular but between epidemics the measles virus apparently disappeared from the island. From 1896 through 1945, on the other hand, the pattern on the island seems to have been that of Type III.

If the number of cases in the individual communities is examined, a more complex pattern emerges. The largest communities have shown a consistent Type II pattern throughout the 20th century. This includes Reykjavik and its hinterland and the four outlying cities of Isafjordur, Akureyri, Egilsstadir and Seydisfjordur. The small towns and villages other than those in the neighborhood of the capital, however, showed waves of Type III.

The small communities vary in the number of epidemics they escaped. Half of the 16 waves failed to reach the offshore island of Flatey. Substantial parts of the northwestern peninsula and of the north coast, which are sparsely populated and remote, missed two or more of the 16. As a rule medical districts with a population of 2,000 or more were reached by all the waves. If a district had a population of fewer than 2,000, its chance of missing a particular epidemic depended on how remote from the urban centers it was.

One of the fundamental goals of epidemiology has long been to formulate testable theories that could serve to project the course of future epidemics. Most of the attempts to devise such theories have had a simple basic structure. The number of people who will be infected in an epidemic wave is derived from the mixing of those who are susceptible with those who are already infected. These relations are expressed in a set of equations where the rate of mixing of the groups is controlled by factors termed diffusion constants. When such equations are calibrated to take account of the clinical history of the disease, the local vaccination rate and the recovery rate, they can generate a fairly good approximation of an epidemic wave train.

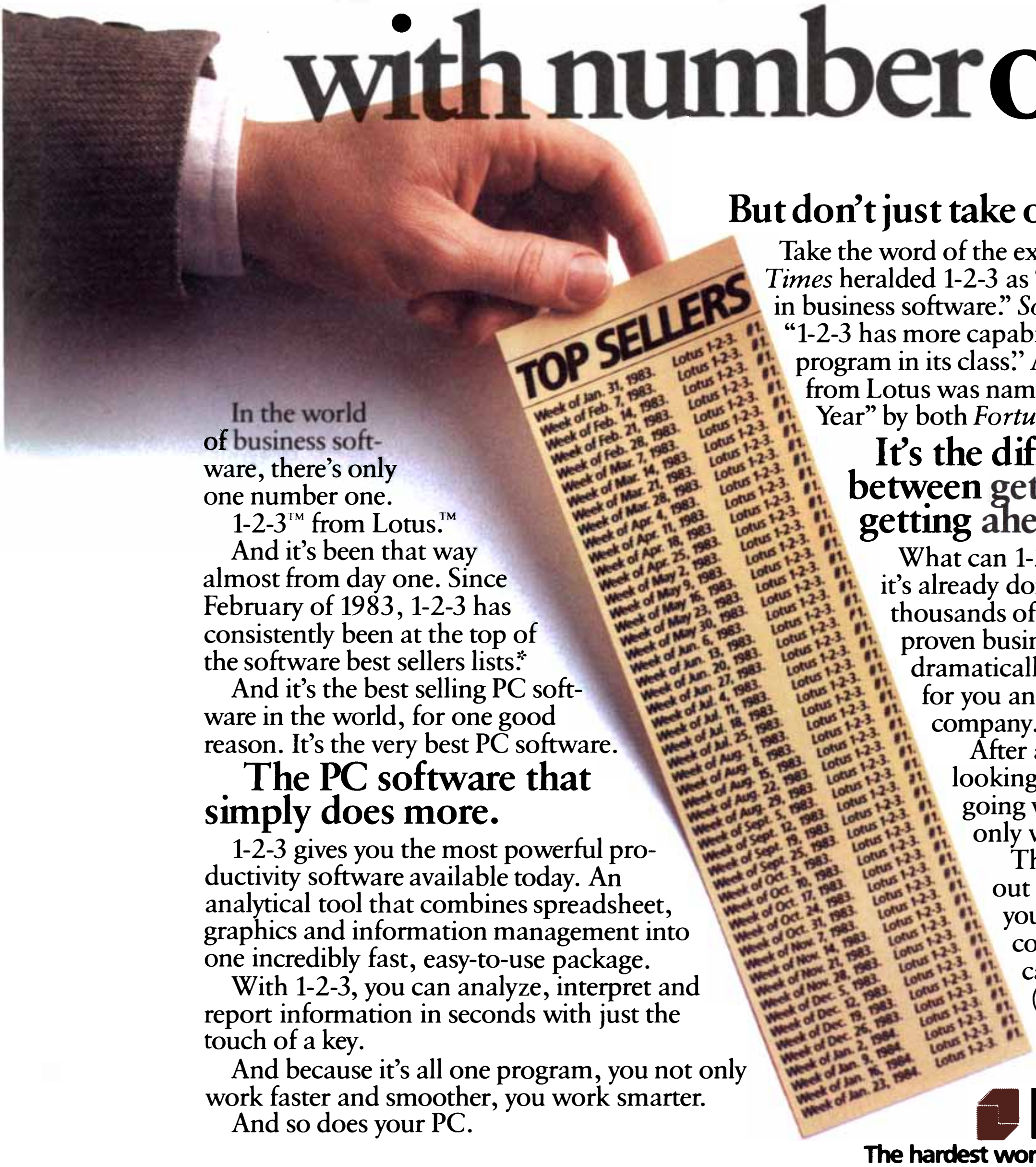
Knowledge of the geographic pattern of contact can be utilized to calibrate such equations on geographic principles as well as medical ones and thereby make the general theory more specific. A single set of equations for the country as a whole can be replaced by a separate set for each medical district. The equations for a district include a diffusion constant that expresses the intensity of contact in the district rather than a generalized constant for the larger area.

Each set of equations is in essence a mathematical portrait of the factors that influence the course of the epidemic in a particular district. Having constructed such a portrait for each medical district, the sets of equations can be linked. The linkage depicts the exchange of infected and susceptible people between adjacent districts. What is meant by adjacent, of course, depends on maps of the movements of the index cases rather than on a map of linear distances. For example, in epidemiological terms Reykjavik can be regarded as a near neighbor of every other Icelandic community no matter how distant the other community is on a straight line.

The results obtained so far suggest that forecasting epidemics is likely to prove as difficult as forecasting snowfall. In both cases there are dangers in forecasting what does not happen as well as in failing to forecast what does. What is clear from the experiments in Iceland is that if more reliable models are to be constructed, it is necessary to understand the geographic pattern of the disease. Estimates incorporating ge-

1-2-3 from Lotus.

Maybe the best way to look out for number one is to go with number one.



In the world of business software, there's only one number one. 1-2-3™ from Lotus.™

And it's been that way almost from day one. Since February of 1983, 1-2-3 has consistently been at the top of the software best sellers lists.*

And it's the best selling PC software in the world, for one good reason. It's the very best PC software.

The PC software that simply does more.

1-2-3 gives you the most powerful productivity software available today. An analytical tool that combines spreadsheet, graphics and information management into one incredibly fast, easy-to-use package.

With 1-2-3, you can analyze, interpret and report information in seconds with just the touch of a key.

And because it's all one program, you not only work faster and smoother, you work smarter. And so does your PC.

But don't just take our word for it.

Take the word of the experts. *The New York Times* heralded 1-2-3 as "the wave of the future in business software." *Software News* said, "1-2-3 has more capabilities than any other program in its class." And in 1983, 1-2-3 from Lotus was named a "Product of the Year" by both *Fortune* and *Info-World*.

It's the difference between getting by and getting ahead.

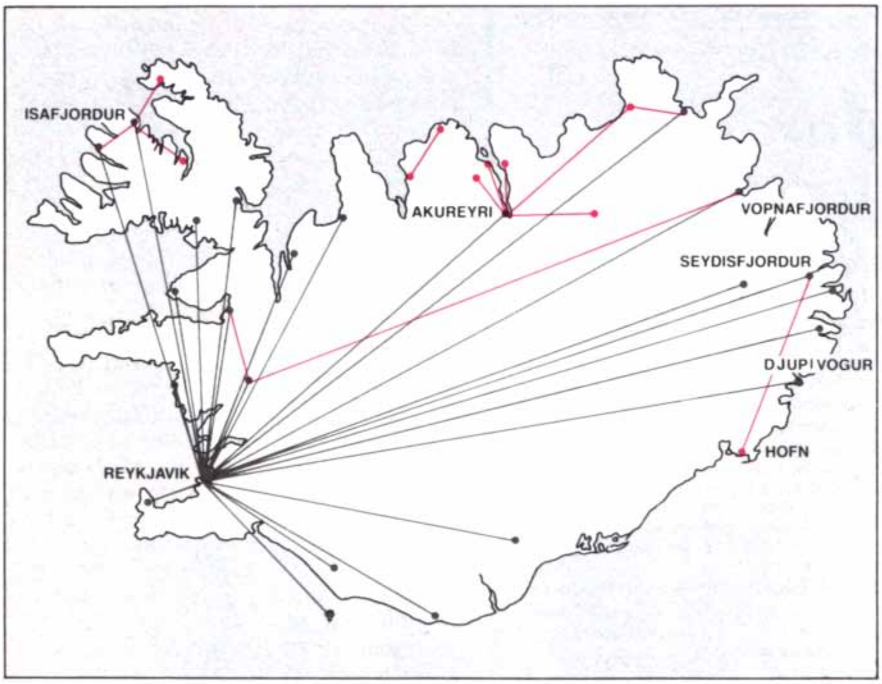
What can 1-2-3 do for you? What it's already done for hundreds of thousands of PC users. Give you a proven business software that can dramatically increase productivity for you and everyone in your company.

After all, when it comes to looking out for number one, going with number one is the only way to go.

The best way to find out what 1-2-3 can do for you is to visit your local computer store. Or call 1-800-343-5414 (In Massachusetts call 617-492-7870).

Lotus
The hardest working software in the world.™

1-2-3 and Lotus are trademarks of Lotus Development Corporation. *Based on Softalk, Softsel Best Selling Software Lists, since month of February 1983 and week of April 25, 1983, respectively.



PATHWAYS OF TRANSMISSION for the measles virus in Iceland center on Reykjavik. The lines depict the routes followed by two or more index cases in the 16 measles epidemics recorded from 1896 to 1975. Most of the frequently traveled pathways radiate from the capital (*black*). There are also small "circulation cells" around the provincial urban centers (*color*).

ographic elements are consistently more accurate than those in which spatial elements are ignored.

Although the geographic pattern of the diffusion of measles in Iceland has been quite stable since 1900, the chronological pattern of the measles waves has changed considerably since World War II. The changes in the chronological form of the waves appear to be the result of the process of modernization, which has linked Iceland to the rest of the world and brought Icelandic settlements closer to one another. As a result the island's unusual seasonal pattern, with its winter and summer peaks, has been eliminated. The summer peak has gradually disappeared, and Iceland currently has the single winter peak found in the northern temperate countries.

The second development has been a sharp change in the shape and timing of the epidemic waves. Before 1945 the measles waves varied greatly in amplitude. In addition the epidemics were separated by intervals that were often irregular and could be quite long. Since World War II the waves have tended to be more regular in amplitude and to be separated by shorter and more uniform intervals.

The reason for the change is that the intensity of contact with the virus reservoirs of Europe and the U.S. has greatly increased. When most travel was by sea, the diffusion of the virus from mainland reservoirs to the island was occasional and irregular. The conversion of Keflavik Airport near the capital into a major staging post for international air travel, however, has led to the island's being at continuous risk of infection carried by air travelers. In the 30 years from 1945 to 1975 air-passenger traffic to Iceland increased by a factor of 15. As a result of the increase in exposure, whenever the susceptible population in Iceland reaches the minimum necessary for an epidemic, there is the risk of a measles wave diffusing out from Reykjavik.

As Iceland has become more closely bound to northern Europe and the U.S. the outlying areas of the island have acquired closer bonds to the capital. The epidemiology of the hinterland now closely resembles that of Reykjavik. Before World War II there was a notable difference between the epidemic waves in the capital and those in other places. The waves in Reykjavik were faster and more intense than the waves elsewhere: they had a shorter time to the infection peak, a shorter duration and a greater amplitude.

Since 1945, however, the two patterns have converged. The speed, duration and intensity of the epidemic waves in the hinterlands are now about equal to those in the capital. The convergence has resulted both from a decrease in the velocity of the waves in Reykjavik and an increase in velocity in the rest of the country. The slowing of the waves in Reykjavik results partly from improvements in health care there. From 1916 through 1972 the number of physicians per 1,000 people in the capital rose from 1.0 to 2.6; the increase in the rest of Iceland was only from .7 to .8.

The velocity and the intensity of the waves outside Reykjavik have increased in part because of advances in transportation and communication. Since World War II a network of roads has been constructed around Iceland, which has reduced dependence on sea travel. Moreover, the island has the most frequently utilized domestic airline system in Europe, as measured by the number of passenger miles per year per 1,000 residents.

Sociological changes have accompanied the decreasing isolation of the villages and farms, and some of these changes have contributed to the changing character of the measles waves. For example, before 1945 the rural areas had an educational system in which a teacher visited outlying farms for a few weeks each term. Since the war this ambulatory system has largely been replaced by boarding schools where farm children spend weeks at a time. The vulnerable school-age population is therefore more concentrated and more accessible to the measles virus than it was in earlier decades.

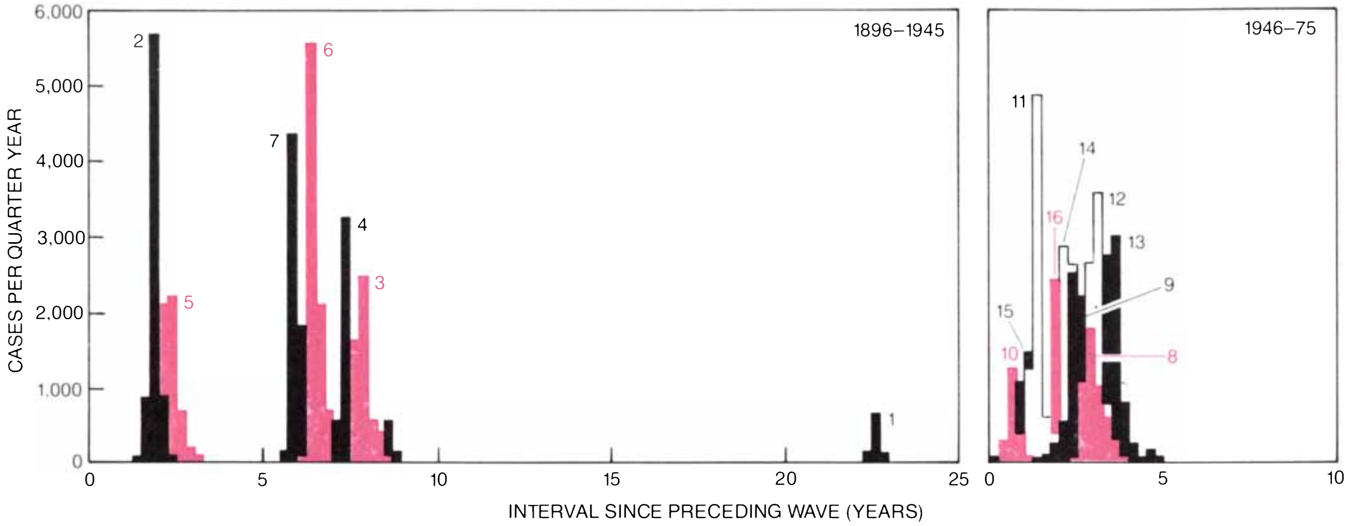
The diffusion of measles in Iceland is not a static phenomenon but a dynamic process that reflects the transformation of the island's social structure. What lessons does the story hold for public-health workers who seek to eradicate the measles virus elsewhere in the world? The advances in transportation and communication in Iceland have made many communities that were once remote from Reykjavik adjacent to the capital in an epidemiological sense. Such communities, which in the past had Type III measles waves, now have Type II waves.

Conquering the measles virus in the world as a whole depends on reversing this trend. Through the administration of measles vaccine large populations will begin to have the properties of smaller ones. In theory a global vaccination program could systematically reduce the size of susceptible populations to the level where the chain of infected people would not be maintained. In terms of the Bartlett model that would mean reducing the waves in all communities from Type I to Type II and from Type II to Type III.

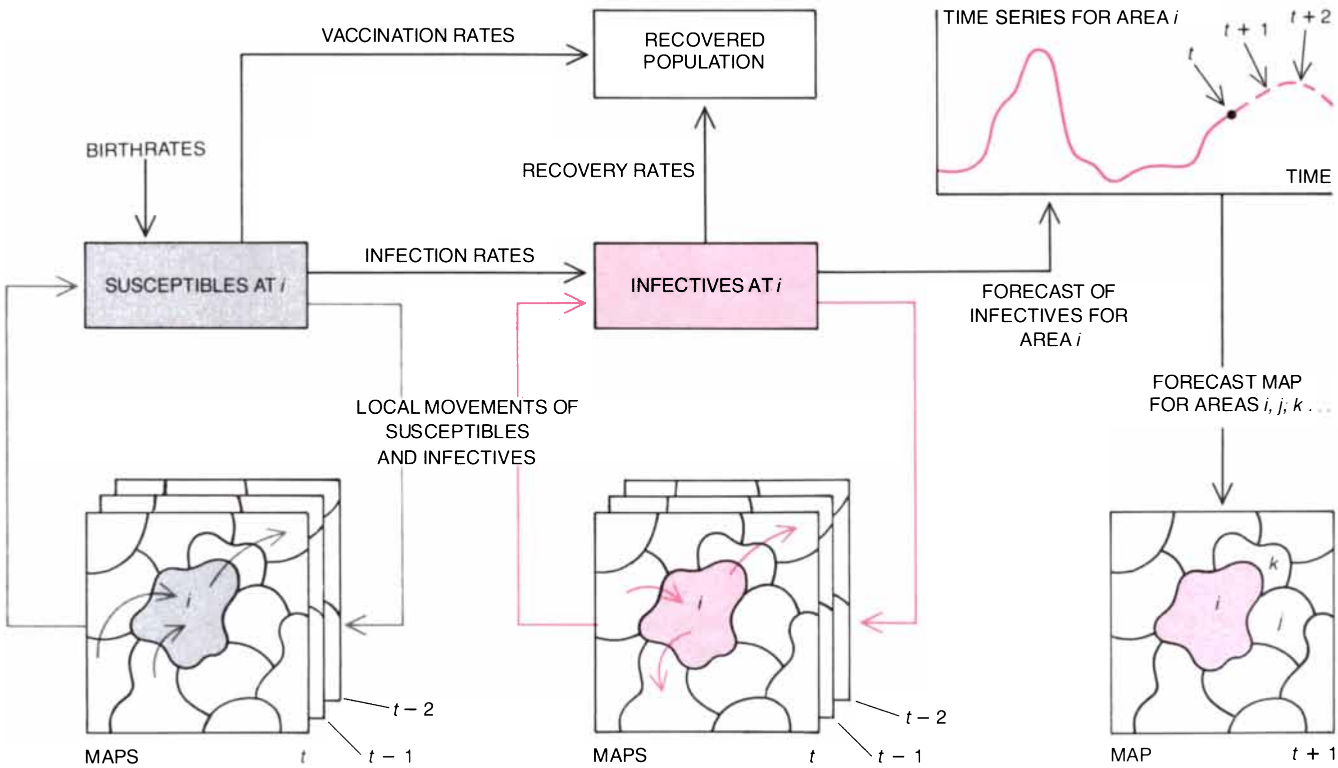
In the U.S. the program coordinated by the Centers for Disease Control (CDC) brought a spectacular reduction in cases of measles. Other countries, notably Australia, are showing an interest in this program. Medical opinion is divided on the feasibility of eliminating measles at the global level: the costs and the logistic problems would certainly be many times greater than those encountered in the elimination of smallpox. Even if the global eradication of measles in this century remains a dream, more parts of the world could undoubtedly become almost free of the disease.

A significant first step would be containing the virus in a decreasing number of reservoir areas of Type I.

Although that is an appealing vision, there is still a large gap between epidemiological theory and public-health practice. Furthermore, in populations that are partially vaccinated there is a risk of raising the average age of infection; adult measles infections are in some cases more serious than childhood ones. To avoid such hazards the distance between theory and practice must be diminished. Geographic reconstruction of the way diseases move through human populations can help to diminish that distance and help in the design of programs to contain infectious diseases. In this effort the islands of the world could well prove to be a uniquely valuable epidemiological laboratory.



EPIDEMIC FORECAST is a prediction of how measles might spread in the immediate future. The forecast is based on a number of factors that are shown schematically in the illustration for area *i* at time *t*. Each arrow corresponds to an equation. The susceptible population of the district is determined by the birthrate, the vaccination rate and the movement of susceptible people into and out of the district. The infected population is determined by the infection rate and the recovery rate. Combining these factors yields an estimate of the level of measles in area *i* at time $t + 1$. Linking the equations for all the medical districts yields the pattern of the epidemic across Iceland at time $t + 1$.



INTERVALS BETWEEN MEASLES WAVES have become shorter and more regular since 1945. The bars stand for the number of measles cases per quarter year in each of the 16 waves. The panel at the left is for the seven waves through 1945. The panel at the right is for the nine waves after 1945. The position of the bars on the horizontal axis indicates the interval since the preceding wave. Before 1946 the intervals were irregular and could be quite long; the longest interval was more than 20 years. Since 1946 the intervals have been short and regular, averaging about three years. In the same period the epidemic waves have also become somewhat more uniform in amplitude.

THE AMATEUR SCIENTIST

In which heating a wire tells a lot about changes in the crystal structure of steel

by Jearl Walker

Many of the useful properties of iron and steel depend on the crystal structure of the metal. Heating or cooling it may drastically alter its properties because of subtle changes in the crystal structure. Charles F. Walton, a metallurgist and mechanical engineer in Cleveland, has devised a demonstration of how profoundly common steel, a binary (two-element) alloy, changes with heating. His demonstration is simple to do, but it presents several puzzles, not all of which have been solved in detail.

Walton heats and cools a 60-inch length of piano wire (No. 29 steel wire). He suspends the wire horizontally between two post terminals mounted on wood boxes and attaches it to a Variac, which enables him to send electric current through the wire in controlled amounts. The Variac is turned up quickly to about 55 volts in order to supply about 14 amperes to the wire. (Although the current greatly exceeds the Variac's limit, it is kept on for too short a time to damage the device.)

The wire is heated so much by the current that it expands, sags and soon glows red. When the current is turned off, the redness diminishes and the wire contracts. Here the first puzzle arises. For an instant the wire gets red again and sags once more. Thereafter it resumes cooling. For some reason the cooling wire releases energy in a sudden red wink. What is the source of the energy? Why is it not released continuously in the course of cooling?

The second puzzle has to do with the cooling rate of the wire. Walton reheated the wire, turned off the current and wrapped a wet sponge around a section of the red-hot wire for a few seconds. The water rapidly cooled that section, but the rest of the wire took a while longer to cool. The rapidly cooled section was so brittle that it snapped easily between Walton's fingers. The point of the broken piece was hard enough to scratch glass. The slowly cooled part of the wire did not snap easily, and a broken point did not scratch glass. How does rapidly cooling a section alter its ductility and hardness?

The third of Walton's puzzles appears in the magnetic properties of the wire. At room temperature the wire is attracted to a small magnet held nearby. When the wire is red-hot, it shows no discernible response to the magnet. Why should the magnetism of a material depend on its temperature? Indeed, what in the cool wire is responsible for the magnetic attraction?

Walton suggests that anyone setting out to repeat his experiment obtain wire of the same size (.075 inch in diameter). A thinner wire oxidizes so much when it is heated that it will serve for only a few experiments. A thicker wire requires too much current to heat. You must be very careful not to touch the wire while the current is flowing. *It is lethal!* Always turn off or unplug the Variac before you apply the magnet or the wet sponge.

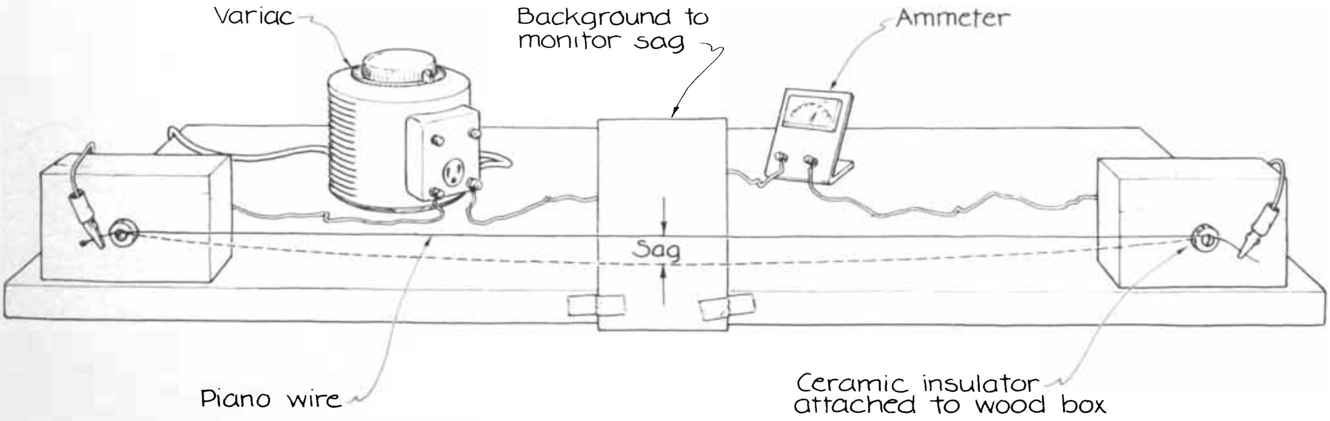
Walton's puzzles can be solved by beginning with iron, the major constituent of steel. Iron is allotropic, meaning that in the solid state it can exist in several different crystalline forms. A crystal is often described in terms of the arrangement of the smallest possible unit of atoms in it. The rest of the crystal is a series of repetitions of the unit cell. The unit cell of iron at room temperature is arrayed with an atom at the center of a cube formed by eight other atoms at each corner. The formation is called a body-centered cubic structure. Iron with this array of atoms in its crystals is termed alpha iron or ferrite.

Iron usually consists of many independent crystal regions called grains. Within each grain the unit cells are uniformly oriented, but the orientation of the grains is random. Grains form when hot iron cools and crystals begin to grow at nucleating sites. The crystals continue growing until eventually they come in contact with one another, forming a matrix of grains.

When iron is heated to a temperature of 910 degrees Celsius, its crystals are transformed from the body-centered cubic structure to the face-centered cubic structure, which is characteristic of gamma iron or austenite. The cubic cell has atoms at the corners and at the centers of the faces. Again there are grains, each grain consisting of a crystal at a certain orientation.

The transformation from alpha iron into gamma iron requires energy to rearrange the atoms into the new structure. Another form of iron, delta iron, appears at a temperature much higher than any in Walton's experiment. If iron is heated even more, it melts.

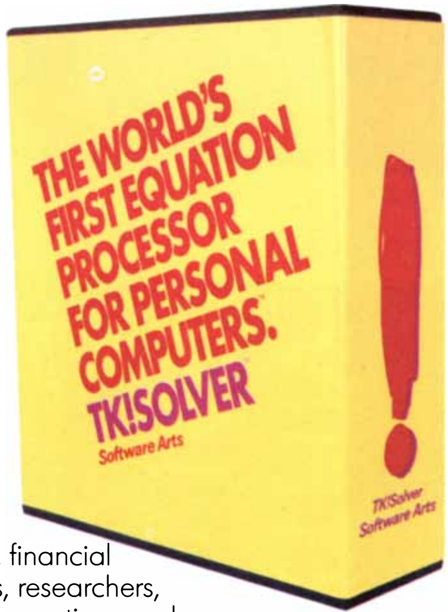
When alpha iron is heated, each addition of energy at first simply raises the



Charles F. Walton's arrangement for heating piano wire

EQUATIONS PROCESSED

NO PENCIL. NO PAPER. NO MANUAL LABOR.



The TK!Solver® program will take on your toughest problems—linear, quadratic, simultaneous equations, whatever. Then stand back. Because TK!Solver turns your personal computer into a simple, yet powerful, desktop equation processor.

Whether your problem is a simple formula or a model consisting of many equations, TK!Solver can help improve your productivity. Once the equations are written, enter the known values, press the ! key, and TK!Solver gives you the answer.

Engineers, scientists, architects, financial analysts and planners, educators, researchers, and other professionals who use equations and mathematical models can work more creatively with TK!Solver.

TK!SOLVER GIVES YOU:

BACKSOLVING

If the programs you use now require you to rewrite the same equation to solve for different unknowns, TK!Solver can dramatically improve your productivity. Enter your problem once and then solve for the unknowns no matter where they are in your equation.

ITERATIVE SOLVING

If TK!Solver can't solve an equation directly, take an educated guess at the answer. Type the ! key and the TK!Solver program starts with your guess and performs repeated approximations to converge on the answer.

LIST SOLVING

Given a list of input values, TK!Solver automatically calculates the equation for every value in your list. For example, if you want to know how different interest rates will affect monthly loan payments, enter a list of interest rates and let TK!Solver calculate the payment amount for each value.

UNIT CONVERSIONS

Any type of unit conversion—Fahrenheit to Celsius, meters to feet, dollars to deutschemarks, newtons to dynes—can be made without altering your equations. Just define the numerical relationship between two units of measurement and the TK!Solver program automatically converts the variable value to the unit you specify.

TABLES AND PLOTS

Quickly generate tables and plots of your results on your screen or printer.

AVAILABLE NOW

You can run the TK!Solver program on the IBM® PC and XT and compatible machines, the Digital™ Professional™ 350, the Digital™ Rainbow™ 100, the Wang Professional Computer, Apple® //e, and on the following personal computers using MS™-DOS: TI Professional Computer, GRiD Compass Computer,™ Canon AS-100, Eagle® 1600, Toshiba T300, and the Zenith Z-100.™

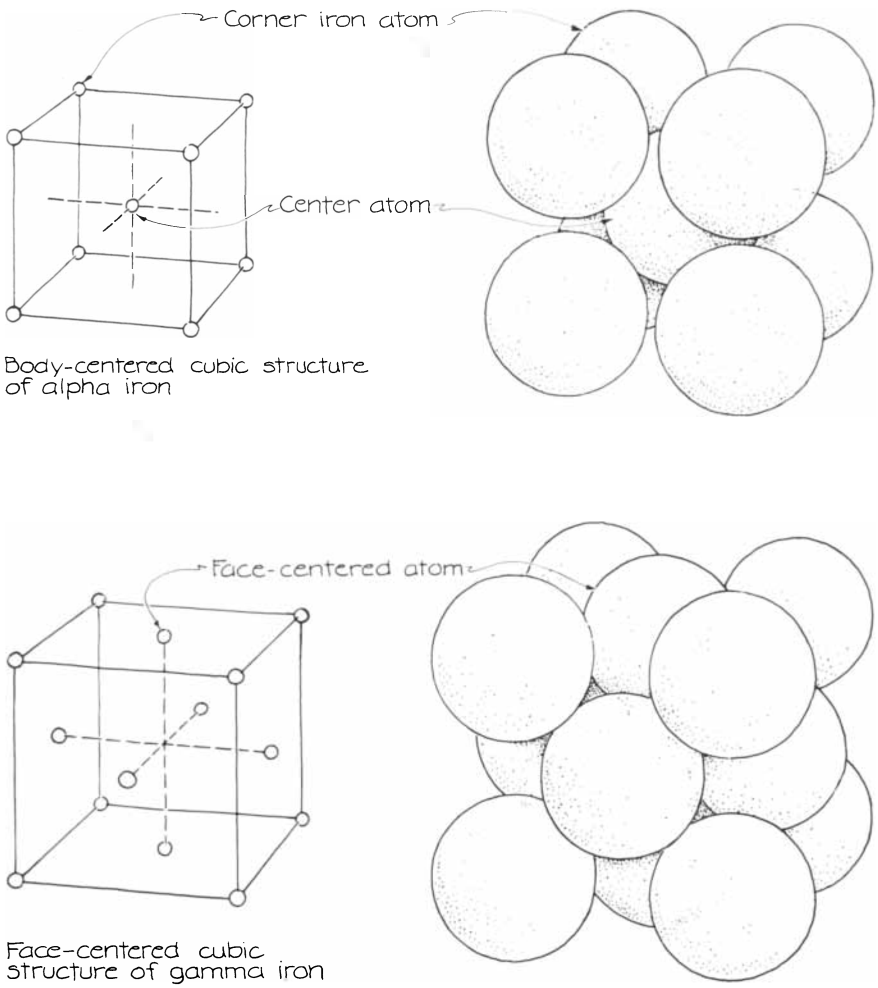
SEE IT TODAY

There's more. Lot's more. But you'll have to see it to believe it. And that's easy. Bring your own equations into your nearest computer retailer and ask to see the TK!Solver program in action.
The world's first equation processor for personal computers.

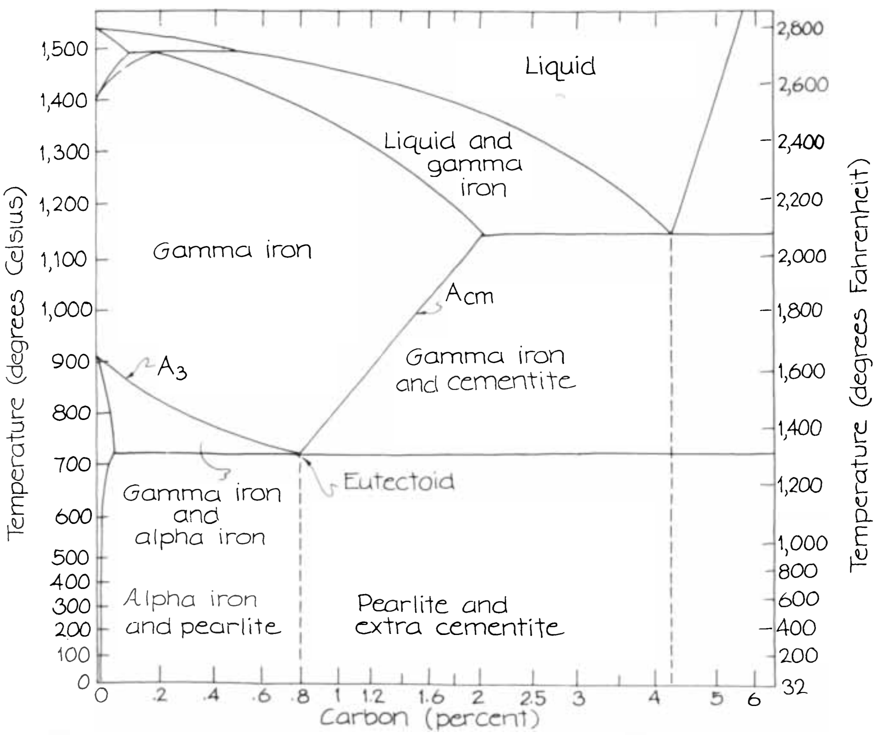
TK!Solver®

By Software Arts,™ creators of VisiCalc®

27 Mica Lane, Wellesley, Massachusetts 02181 617-237-4000



The structure of a unit cell in two types of iron crystal



A phase diagram for an alloy of iron and carbon

temperature. Once the transition point is reached the temperature must remain constant until enough energy is added to transform all the crystals into gamma iron. Only then can heating again raise the temperature.

The converse is also true. When gamma iron cools, its temperature drops until it reaches the transition point. Then heat must be removed until the crystals have reverted to alpha iron. Only then can the temperature begin to drop again.

The transition point between the alpha and the gamma form of iron is similar to the freezing and melting point of water. When ice is heated, its temperature increases until the melting point is reached. The temperature cannot rise further until the ice is fully melted. When water is cooled, it must remain at the freezing point until it fully freezes. Only then can its temperature decrease.

Walton's piano wire consists primarily of alpha iron. When he sends current through it, the collisions of the electrons of the current with the crystal structure of the wire generate heat. Eventually the wire is transformed to gamma iron. Additional heating makes the wire so hot that it soon radiates in the red part of the visible spectrum.

When the current is turned off, the wire cools, dimming the visible emission. At the transition temperature the rearrangement of the face-centered cubic crystals into body-centered cubic crystals releases energy, which momentarily reheats the wire to the point where it again turns red and sags. The glow is brief because the energy is quickly lost through radiation and by convection in the air. The brief wink of red light in Walton's experiment is the energy released by the transition of gamma iron to alpha iron. It is evidence that the atomic arrangement of one crystal form requires more energy than the atomic arrangement of another form.

To solve another of the puzzles in Walton's experiment one must consider the carbon content of steel. There are of course many alloys of steel, but I refer here only to the binary alloy of iron and carbon. The analysis is aided by a phase diagram of the kind shown in the bottom illustration at the left. The ordinate represents the temperature of the alloy, the abscissa the amount of carbon in the iron.

When steel is liquid, carbon can readily dissolve in the iron. Even when the metal is solid, however, carbon may mix with the iron crystals to form what is called a solid solution. The solubility of carbon in such a solution plays an important role in Walton's experiment.

The second phase diagram [*see illustration at left at top of page 151*] identifies a point where the wire is in the gamma-iron state at a temperature of 1,200 degrees C. and the carbon content is .4 percent, that is, there are four parts of

8087 SUPPORT and FORTRAN for your PC

87FORTRAN/RTOS™ - our adaptation of the Intel Fortran-86 Compiler generates in line 8087 code using all 8087 data types including 80-bit reals and 64-bit intergers. The compiler uses the Intel large memory model, allowing code/data structures of a full megabyte, and supports overlays. Includes RTOS and support for one year............. **$1350**

87PASCAL/RTOS™ is Intel's ISO-Standard Pascal with 8087-8088 exceptions. These make it possible to use all the 8087 data types directly, while generating modules in one of the three Intel Memory Models. Includes RTOS and support for one year...... **$1350**

RTOS DEVELOPMENT PACKAGE includes 87FORTRAN, 87PASCAL, PL/M-86, Utilities, TX Screen Editor and RTOS... **$2500**

RTOS—
REAL TIME MULTI-TASKING/ MULTI-USER EXECUTIVE
RTOS is a Micro Ware configured version of iRMX-86. Includes ASM-86, LINK-86, LOC-86, LIB-86, and the ROM Hex Loader..... **$600**

OBJ→ASM™ - a multipass object module translator and disassembler. Produces assembly language listings which include public symbols, external symbols, and labels commented with cross references. Ideal for understanding and patching object modules for which source is unavailable......... **$200**

PC TECH JOURNAL REVIEW:
"The Micro Ware package is preferable ... it executes the basic operations more rapidly and Micro Ware provides a free update service."

87BASIC™ includes patches to the IBM Basic Compiler and both runtime libraries for USER TRANSPARENT and COMPLETE 8087 support. Provides super fast performance for all numeric operations including trigonometrics, transcendentals, addition, subtraction, multiplication, and division.............. **$150**

87MACRO™ - our complete 8087 software development package. It contains a "Pre-processor," source code for a set of 8087 macros, and a library of numeric functions including transcendentals, trigonometrics, hyperbolics, encoding, decoding and conversions... **$150**

87DEBUG™ - a professional debugger with 8087 support, a sophisticated screen-oriented macro command processor, and trace features which include the ability to skip tracing through branches to calls and software and hardware interrupts. Breakpoints can be set in code or on guarded addresses in RAM......... **$150**

FOR→BAS™ - a library of interface routines which allow MS Fortran programs to call the IBM Basic Compiler library and access features such as the RANDOM NUMBER GENERATOR, SOUND, PLAY, DRAW and SCREEN commands.................... **$150**

8087-3 CHIP...... $175
- with 180-day warranty and 8088 exchange.

64K RAM Upgrade............... **$50**

87/88GUIDE - an excellent tutorial on writing 8087 code and interfacing it with compilers. Full of code that runs!........ **$30**

MATRIXPAK™ manages a **MEGABYTE!** Written in assembly language, our runtime package accurately manipulates large matrices at very fast speeds. Includes matrix inversion and the solution of simultaneous linear equations. Callable from MS Fortran 3.2, MS Pascal 3.2, and Micro Ware 87MACRO, 87BASIC, and RTOS compilers................. each **$150**

MICROSOFT FORTRAN 3.2
MICROSOFT PASCAL 3.2 These new IEEE compatible compilers support both double precision and the 8087... each **$259**

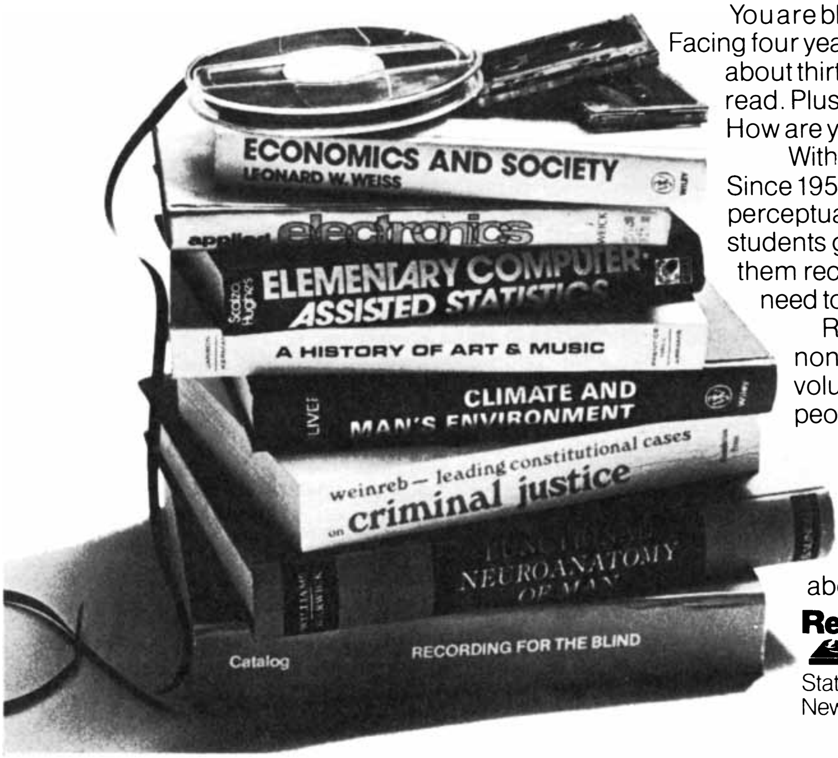
MICROSOFT C COMPILER includes Lattice C and the MS Librarian........ **$350**

XENIX.............................. **CALL**
LOGITECH Modula-2/86................ **$445**
FLOAT87 for MS C Compiler.............. **125**
Multitool Word.......................... **269**
SuperSoft Fortran 66..................... **299**
SSS 8087 Support......................... **50**
Computer Innovations C86 V 2.04........ **345**
STSC APL★PLUS/PC...................... **545**
87BASIC+................................. **75**
HALO Graphics........................... **125**
GRAPHMATIC............................. **125**
ENERGRAPHICS.......................... **295**
Professional BASIC....................... **295**
Kidger Optical Design Program.......... **3000**
COSMOS REVELATION.................. **CALL**
dBASE II............................... **CALL**
SuperCalc III........................... **CALL**
MAYNARD Electronics Boards.......... **CALL**

Micro Ware
P.O. Box 79
Kingston, MA
02364
(617) 746-7341

You Can Talk To Us!

Close your eyes. Now have someone read this to you.



You are blind. A student. Facing four years of college. With about thirty-two textbooks to read. Plus fifty supplemental texts. How are you going to manage?

With Recording for the Blind.

Since 1951, we've helped over 53,000 blind, perceptually and physically handicapped students get through school. By sending them recordings of the books they need to read. Free.

Recording for the Blind is non-profit, and supported by volunteers and contributions from people like you who can imagine what it's like to be blind.

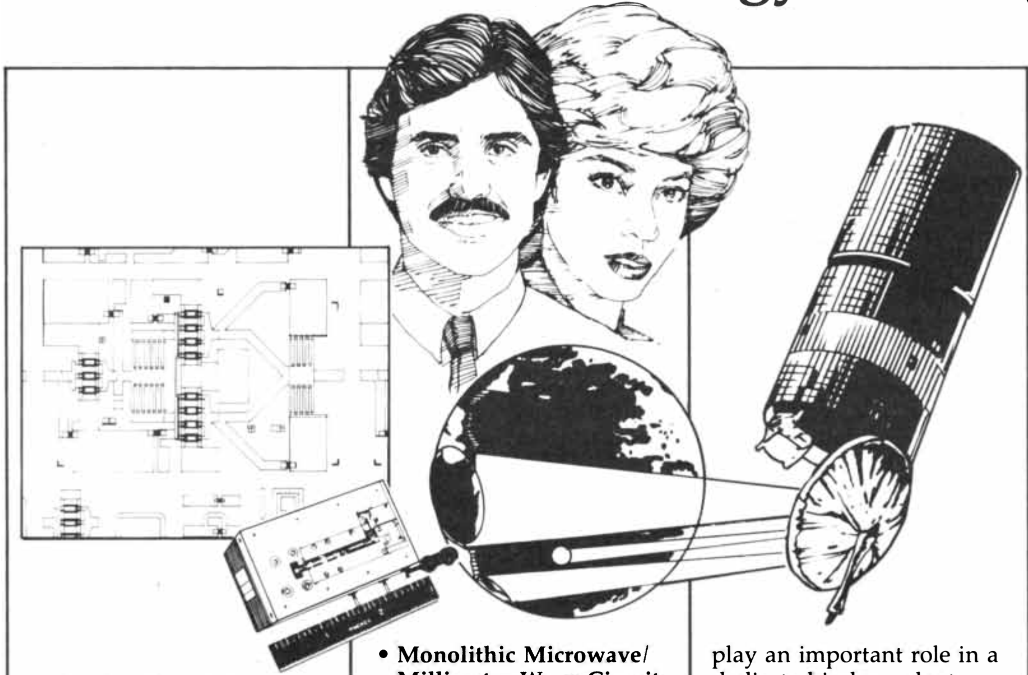
Your tax-deductible donation will help our students meet their educational goals. We'd all be grateful.

If you want to know more about us, write:

Recording for the Blind, Inc.
an educational lifeline.
Station E, 215 East 58th Street
New York, New York 10022, (212) 751-0860.

Opportunity. Far beyond the ordinary.

Top career challenge in solid state millimeter/microwave technology is at Hughes.



Here's an exciting opportunity to build a rewarding career in the expanding fields of solid state microwave and millimeter-waves at Hughes Electron Dynamics Division's, Torrance Research Center, located in Southern California.

We are the leader in millimeter/microwave technology. And our rapid growth coupled with an expanding worldwide market, has opened immediate opportunities in a variety of related areas including:

- **GaAs FET Devices Development and Manufacturing**
- **Monolithic Microwave/ Millimeter-Wave Circuits**
- **Wideband Receivers and Power Amplifiers**
- **MIC Development and Manufacturing**
- **High Speed Digital Circuits**
- **Project Management**
- **Mechanical and Packaging Engineering**

Your related microwave experience and/or BS, MS, or PhD degree (AA for Research Assistants) can be the key to a great future. In addition to excellent salary and fringe benefits, we offer small organization atmosphere backed by big company versatility. Your creativity and initiative will be encouraged and rewarded because you will play an important role in a dedicated independent engineering project team—rather than working on a small portion of the development of large systems. This, coupled with our expanding business, means a much better chance for you to grow, personally and professionally. Also, you'll enjoy a stimulating lifestyle in our modern beach area community, in the heart of Southern California.

If you're ready for a challenge that's far beyond the ordinary, act now. Please send your resume to: Lisa Annett, Hughes Electron Dynamics Division, Dept. AS-5, P.O. Box 2999, Torrance, California 90509.

MAKING WAVES IN MILLIMETER/MICROWAVE TECHNOLOGY

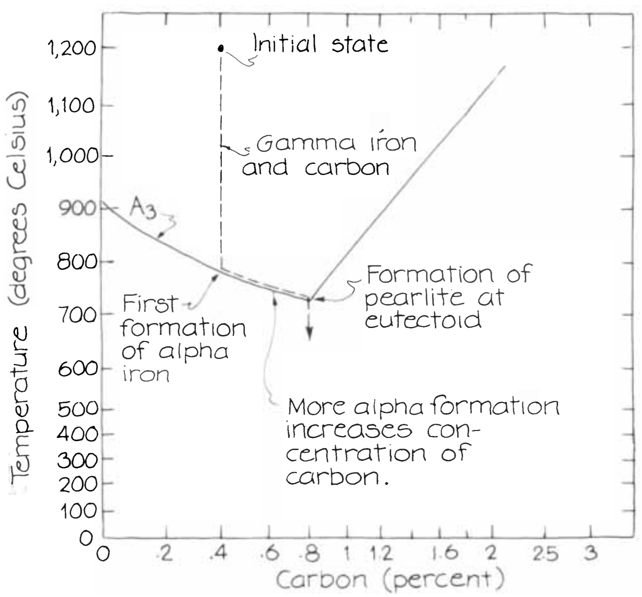
Creating a new world with electronics

HUGHES

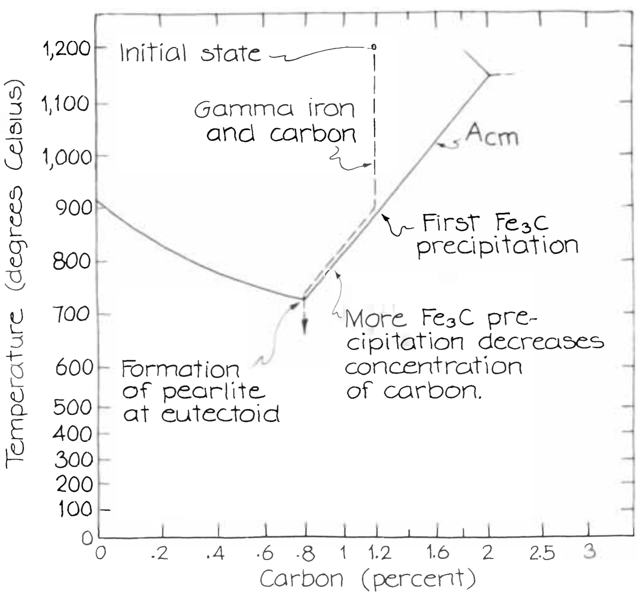
HUGHES AIRCRAFT COMPANY

ELECTRON DYNAMICS

Proof of U.S. Citizenship Required
Equal Opportunity Employer



The cooling of hypoeutectoid steel



The cooling of hypereutectoid steel

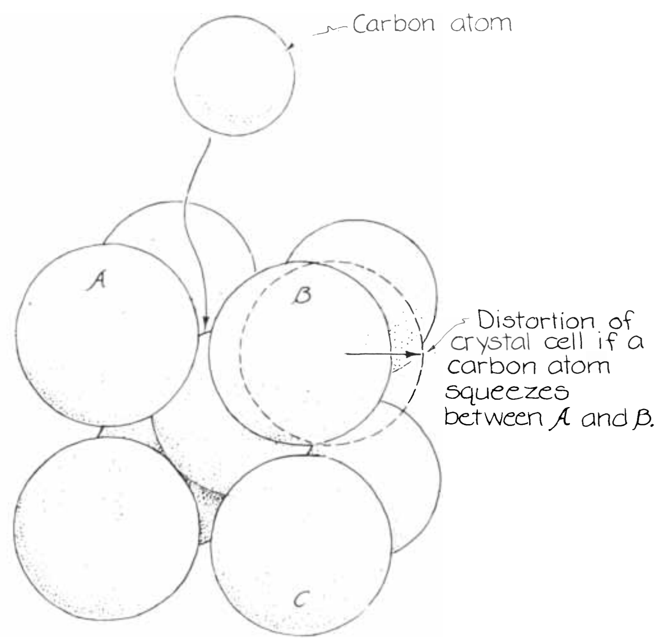
carbon to 1,000 parts of solid solution. The carbon is in solution with the gamma iron in the sense that it is dispersed throughout the crystal structure. The carbon atoms can squeeze into the edges of the unit cells. Since the carbon content is small, only a few of the edges contain carbon.

In the diagram the cooling of this steel is represented by a vertical line extending from the initial point down to a line labeled A_3. That line marks the transition from gamma iron to alpha iron. Until the transition point is reached each removal of energy from the iron reduces the temperature but does not change the crystal structure or the solubility of the carbon in that structure. After A_3 has been reached the next removal of energy forces some of the gamma iron (primarily at the grain boundaries) to switch to alpha iron.

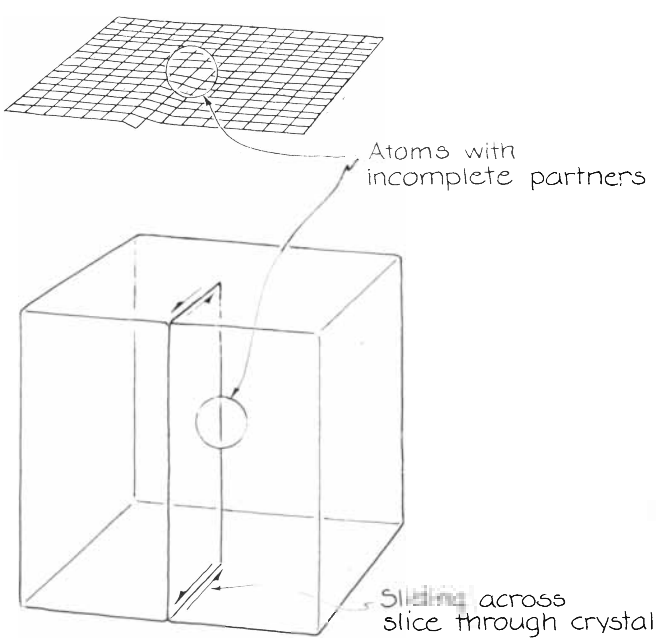
Carbon is almost insoluble in alpha iron (its limit of solubility is only .025 percent), so that the solution of gamma iron and carbon becomes more concentrated. In the phase diagram the cooling path follows A_3 downward to the right as the further formation of alpha iron results in a greater concentration of carbon in the remaining gamma iron. Whereas in pure iron the transition between gamma iron and alpha iron takes place at one temperature, the presence of carbon spreads the transition over a range of temperatures.

Eventually the end of A_3 is reached at a temperature of 723 degrees C. This state, called the eutectoid, marks the highest concentration (.8 percent) possible for carbon in gamma iron. Further removal of energy forces the remaining solution to precipitate alternating layers of alpha iron and clusters of iron carbide (Fe_3C), commonly known as cementite. The combination of alpha iron and cementite is called pearlite. Since further cooling does not alter the mixture, the steel at room temperature is pearlite.

Heating the steel reverses the process. The steel is pearlite until the eutectoid is reached. Further heating begins to transform the cementite into a sol-



How a carbon atom distorts a unit cell



An edge dislocation in a crystal

id solution of gamma iron and carbon. Still more heating begins to change the rest of the alpha iron into gamma iron, decreasing the concentration of carbon in the gamma iron. The path on the phase diagram is upward to the left, along the A_3 line. When no alpha iron remains, the path leaves the line and heads upward parallel to the temperature axis. In this period of heating, the crystal structure and the carbon concentration remain constant.

When steel in the gamma-iron state has a carbon concentration less than the .8 percent characteristic of the eutectoid, the steel is said to be hypoeutectoid. Hypereutectoid steel, which has a carbon concentration greater than .8 percent, cools in much the way I have described it except that initially it precipitates iron carbide instead of alpha iron.

Suppose a steel in the gamma-iron state has 1.2 percent carbon. As the steel cools, the concentration of carbon remains constant until a transition point is reached. In the third phase diagram [*see illustration at right at top of preceding page*] the relevant line is labeled A_{cm}.

If the sample is cooled further, not all the carbon can remain in solution with the gamma iron. The cooling forces some of the carbon (mainly at the gamma-iron grain boundaries) to precipitate as iron carbide. The concentration of carbon decreases. In the phase diagram the cooling path follows A_{cm} toward the eutectoid. There additional cooling results in the formation of pearlite. When the steel reaches room temperature, it has a large amount of iron carbide mixed within the pearlite.

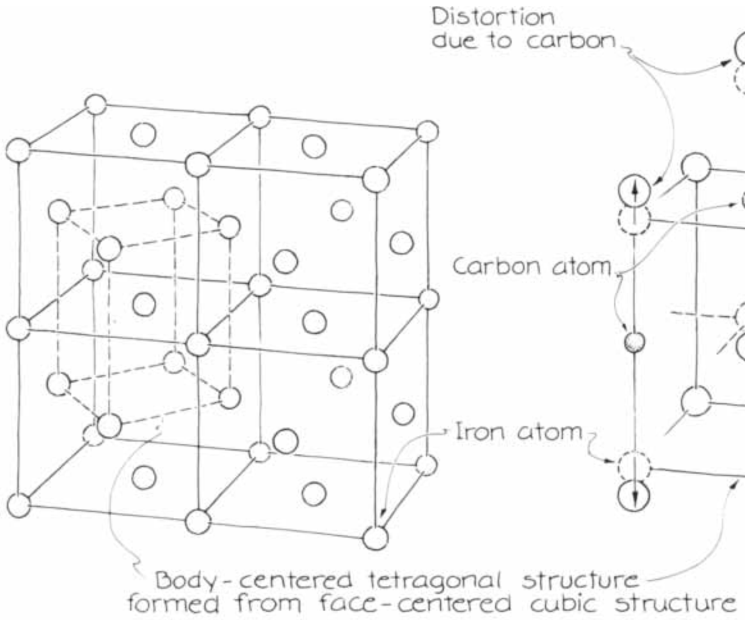
Piano wire in the gamma-iron state contains about .8 percent carbon. After the wire has been heated to red heat it cools fast enough to undercool the temperature of the eutectoid to between 550 and 600 degrees C. Then it suddenly transforms from carbon and gamma iron to pearlite by releasing energy. The wire winks red and sags momentarily. Part of the sag results from thermal expansion as the wire is reheated by the crystal transformation. The rest of the sag derives from the expansion entailed in the rearrangement of the atoms to form pearlite.

The solubility of carbon in iron depends primarily on the space available for carbon among the unit cells of the iron. In the body-centered cubic arrangement of alpha iron a carbon atom can lie on an edge of the cubic cell or at a face center. Since the space available at both locations is less than the size of a carbon atom, the atom must force a corner iron atom out of its proper position. In the illustration at the left at the bottom of the preceding page two possibilities are shown for the position of a carbon atom on an edge. If the atom forces its way between iron atoms *A* and *B* by moving *B*, *B* moves out of its position toward the right. If instead a carbon atom forces its way between atoms *B* and *C*, *B* might be forced upward. In either case the displacement of *B* greatly distorts the crystal structure. Carbon is essentially insoluble in alpha iron because little space is available and its presence distorts the crystals.

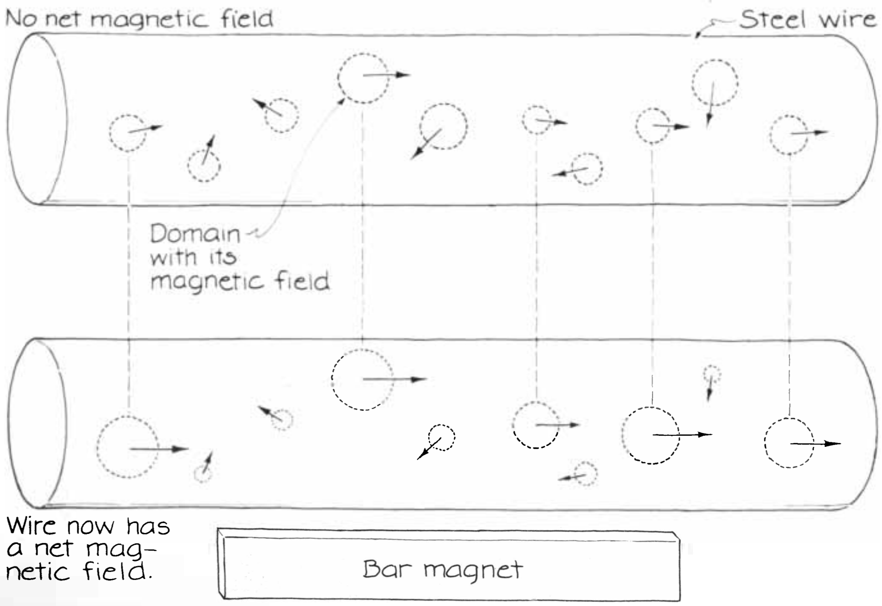
In gamma iron the carbon is limited to the edges of the crystal cell. Although there are fewer nesting places for the carbon than there are in the alpha iron, the spaces on the edges are slightly larger. The presence of the carbon therefore creates less distortion of the crystal than it does in the alpha iron, allowing more carbon to be in solution.

The primary purpose of having carbon in steel is to strengthen the steel. The iron grains almost always contain dislocations that interrupt the regular pattern of ideal crystals and thereby weaken the grains. Carbon strengthens the grains by anchoring the dislocations in place.

A common type of irregularity is an edge dislocation. Consider a uniform cube of alpha iron with each cell inside the cube bonded to adjacent cells. Make



The crystal structure of martensite



How crystal domains align as a magnetic field is applied an imaginary slice halfway through the block and force the opposite sides of the slice to slide in opposite directions by an amount equal to the width of one crystal cell. The cube then contains a line of atoms out of alignment with the surrounding cells.

Such an edge dislocation weakens the grain. If a shearing stress is now applied in such a way as to slide the blocks of crystals even more, the vertical line of dislocation is easily moved through the crystal. Normally the bonding between atoms in the cells is strong. The misaligned atoms, however, are poorly held and can be moved with even small shearing forces.

When carbon is mixed into hot iron and then precipitated as cementite as the iron cools, it tends to collect in the space provided by the edge dislocations. It anchors the dislocations, thereby diminishing the chance that an applied stress can move them through the crystals and rupture the grains. When Walton heats piano wire and lets it cool by radiation and air convection, the carbon atoms have time to diffuse through the iron and form cementite at the dislocation sites. The wire is then ductile enough to be bent without breaking.

When Walton heats the wire and cools it rapidly with water on a sponge, he blocks the formation of cementite. This quenching process proceeds so fast that the carbon cannot diffuse. Moreover, only part of the gamma iron has time to transform into alpha iron. Although the transformation is initially rapid, it soon slows. Hence the section of wire quenched with water consists of a small amount of alpha iron and a good deal of gamma iron. Nearly all the carbon atoms are frozen in place. The mixture is a supersaturated solution of carbon because the concentration of carbon in the iron is greater than is normally possible at that temperature.

The formation of this new structure, which is called martensite, requires neither diffusion nor nucleation. It is a spontaneous change in the existing crystal structure of the gamma iron with carbon atoms on a few edges. The iron atoms change from a face-centered cubic structure to a body-centered tetragonal structure. The carbon atoms in the gamma iron are not given enough time to diffuse and are trapped at their locations in the unit cells. They force iron atoms out of their usual places in the body-centered tetragonal structure.

This distortion of the crystals creates regions of high stress in the martensite. The distortion is also responsible for the hardness of the martensite because it locks dislocations in place within the grains. That is why the martensite Walton creates by quenching the wire while it is hot is hard enough to scratch glass. It is also brittle because of the numerous sites of high stress in the grains. If the wire is bent, it breaks because of the internal stress.

The carbon in martensite is not permanently trapped, but it diffuses so slowly through the iron crystal that it can be regarded as permanently fixed. If the temperature of the martensite is increased, the rate of diffusion increases, allowing the carbon to collect into tiny amounts of cementite. With enough temperature and time the steel regains more of the properties of pearlite.

The third puzzle raised by Walton's work involves the magnetization of the steel wire. Iron is said to be ferromagnetic. One property of such a material is that it has regions called domains, each domain contributing a magnetic field to its surroundings. The material as a whole may seem to be nonmagnetic because normally the magnetic fields from all the domains cancel.

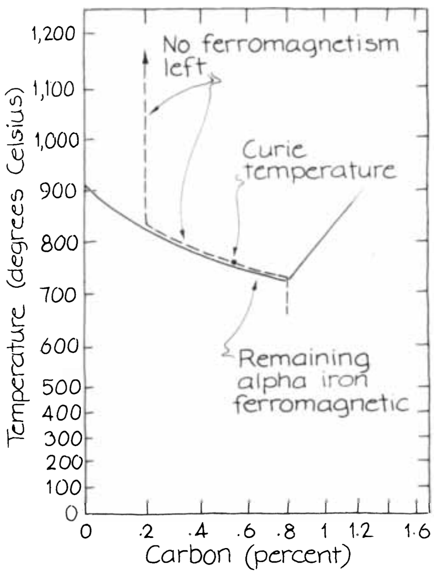
The domains in piano wire cancel in this way until a magnet is brought close to the wire. The field of the magnet realigns the domains so that they give the wire a net magnetic field. The realignment is primarily due to changes in the size of the domains. Any domain with a magnetic field approximately parallel to the magnet's field grows larger at the expense of adjacent domains that are in other orientations.

The result of the growth of domains is an attraction between the wire and the magnet. When piano wire is at room temperature, the attraction is strong enough to move the wire. At a temperature characteristic of gamma iron the attraction is absent.

When the wire is heated, the increase in temperature leads to greater agitation of the atoms and molecules. The agitation begins to disrupt the organization of the magnetic field of a domain. The hotter the substance becomes, the weaker each domain's magnetic field becomes. Eventually the organization of the domains disappears entirely. This event takes place at what is called the Curie temperature after Pierre Curie, who in 1894 found that iron loses its ferromagnetism when it is heated above 768 degrees C.

The ultimate source of the magnetic field in a ferromagnetic substance such as iron is still not understood. Apparently when an iron atom joins a crystal, it partially ionizes (loses one electron or more) as the electrons in its outer orbit become loosely bound because of the proximity of other iron ions. Although the electrons are not entirely free, they are mobile enough to hop between ions.

Every electron has a magnetic field. Although the origin of this field is not understood, the field is as characteristic of an electron as the electric charge is. As the electrons hop between ion sites in an iron crystal they influence one another through what is called an exchange interaction. The individual magnetic fields of the semifree electrons line up in the same direction. An electron with its magnetic field aligned in some other direction would require more energy. Thus the organization of fields from the electrons in a domain that gives the domain a net magnetic field comes from a minimizing of the energy associated with the exchange interaction.



The loss of ferromagnetism

As the temperature increases, the thermal agitation destroys the cooperative behavior of the electrons until, above the Curie temperature, the domain structure is spoiled. At the Curie temperature the semifree electrons still have individual magnetic fields, but there is no overall organization of fields. If a magnet is brought near iron that is hotter than the Curie temperature, the iron is only weakly attracted to the magnet. The magnetic field of the magnet can orient some of the fields of individual electrons, but the alignment is fleeting because of the thermal agitation.

The loss of magnetism can be followed in a phase diagram. Consider the pearlite of a low-carbon steel. When it is heated to the eutectoid and begins to reform into gamma iron and carbon, the transformed iron loses its ferromagnetism. As the heating path continues upward along A_3 the remaining alpha iron maintains its magnetism until the material reaches the Curie temperature. Thereafter none of the iron (including the alpha iron) is ferromagnetic.

Much more can be said about the effect of temperature on the crystal structure of steel. By employing etching processes and photographing specimens through a microscope you can study the precipitation of iron carbide from steels. When the steel contains more carbon than the forms I have discussed, new formations appear. I leave their cause and structure for you to look into.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

COMPUTER RECREATIONS: "DARWIN." Aleph Null in *Software: Practice & Experience*, Vol. 2, No. 1, pages 93–96; January–March, 1972.

68000 ASSEMBLY LANGUAGE PROGRAMMING. Gerry Kane and Doug Hawkins Legenthal. Osborne–McGraw-Hill, 1981.

STEEL MINIMILLS

TECHNOLOGY AND STEEL INDUSTRY COMPETITIVENESS. Office of Technology Assessment, U.S. Government Printing Office, 1980.

THE STEEL INDUSTRY. W. Adams and H. Muller in *The Structure of American Industry*, edited by Walter Adams. Macmillan Publishing Co., 1982.

STEEL—UPHEAVAL IN A BASIC INDUSTRY. D. F. Barnett and L. Schorsch. Ballinger Publishing Co., 1983.

OCEANIC FRACTURE ZONES

A NEW CLASS OF FAULTS AND THEIR BEARING ON CONTINENTAL DRIFT. J. Tuzo Wilson in *Nature*, Vol. 207, No. 4995, pages 343–347; July 24, 1965.

TOPOGRAPHY AND TECTONICS AT THE INTERSECTIONS OF FRACTURE ZONES WITH CENTRAL RIFTS. Norman H. Sleep and Shawn Biehler in *Journal of Geophysical Research*, Vol. 75, No. 14, pages 2748–2752; May 10, 1970.

VERTICAL TECTONISM IN OCEANIC FRACTURE ZONES. Enrico Bonatti in *Earth and Planetary Science Letters*, Vol. 37, No. 3, pages 369–379; January, 1978.

MEMBRANE STRESSES NEAR MID-OCEAN RIDGE–TRANSFORM INTERSECTIONS. Kazuya Fujita and Norman H. Sleep in *Tectonophysics*, Vol. 50, pages 207–221; September 20, 1978.

FORMERLY EMERGING CRUSTAL BLOCKS IN THE EQUATORIAL ATLANTIC. Enrico Bonatti and Andy Chermak in *Tectonophysics*, Vol. 72, Nos. 3–4, pages 165–180; February 10, 1981.

HOW RECEPTORS BRING PROTEINS AND PARTICLES INTO CELLS

INTRACELLULAR SITE OF ASIALOGLYCOPROTEIN RECEPTOR-LIGAND UNCOUPLING: DOUBLE-LABEL IMMUNOELECTRON MICROSCOPY DURING RECEPTOR-MEDIATED ENDOCYTOSIS. Hans J. Geuze, Jan Willem Slot, Ger J. A. M. Strous, Harvey F. Lodish and Alan L. Schwartz in *Cell*, Vol. 32, No. 1, pages 277–287; January, 1983.

pH AND THE RECYCLING OF TRANSFERRIN DURING RECEPTOR-MEDIATED ENDOCYTOSIS. Alice Dautry-Varsat, Aaron Ciechanover and Harvey F. Lodish in *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 80, No. 8, pages 2258–2262; April, 1983.

TURING MACHINES

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM. A. M. Turing in *Proceedings of the London Mathematical Society*, Second Series, Vol. 42, Part 3, pages 230–265; November 12, 1936.

ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM: A CORRECTION. A. M. Turing in *Proceedings of the London Mathematical Society*, Second Series, Vol. 43, Part 7, pages 544–546; May 20, 1937.

ON NONCOMPUTABLE FUNCTIONS. Tibor Rado in *Bell System Technical Journal*, Vol. 41, No. 3, pages 877–884; May, 1962.

THE UNDECIDABLE. Martin Davis. Raven Press, 1965.

INCOMPUTABILITY. C. A. R. Hoare and D. C. S. Allison in *Computing Surveys*, Vol. 4, No. 3, pages 169–178; September, 1972.

ALAN TURING: THE ENIGMA. Andrew Hodges. Simon and Schuster, 1983.

PARASITES THAT CHANGE THE BEHAVIOR OF THEIR HOST

MODIFICATION OF INTERMEDIATE HOST BEHAVIOUR BY PARASITES. J. C. Holmes and W. M. Bethel in *Behavioural Aspects of Parasite Transmission*, edited by Elizabeth U. Canning and C. A. Wright. Academic Press, 1972.

ALTERED COLOR, BEHAVIOR AND PREDATION SUSCEPTIBILITY OF THE ISOPOD *ASELLUS INTERMEDIUS* INFECTED WITH *ACANTHOCEPHALUS DIRUS*. Joseph W. Camp and Harry W. Huizinga in *The Journal of Parasitology*, Vol. 65, No. 4, pages 667–669; August, 1979.

THE LIFE CYCLE OF *OCTOSPINIFEROIDES CHANDLERI* BULLOCK, 1957 (ACANTHOCEPHALA: NEOECHINORHYNCHIDAE), WITH SOME OBSERVATIONS ON PARASITE-INDUCED, PHOTOPHILIC BEHAVIOR IN OSTRACODS. David J. DeMont and Kenneth C. Corkum in *The Journal of Parasitology*, Vol. 68, No. 1, pages 125–130; February, 1982.

ALTERED BEHAVIORAL RESPONSES IN INTERMEDIATE HOSTS—AN ACANTHOCEPHALAN PARASITE STRATEGY. Janice Moore in *American Naturalist*, Vol. 123, No. 4, pages 572–577; April, 1984.

THE INFLATIONARY UNIVERSE

THE FIRST THREE MINUTES: A MODERN VIEW OF THE ORIGIN OF THE UNIVERSE. Steven Weinberg. Basic Books, Inc., 1977.

INFLATIONARY UNIVERSE: A POSSIBLE SOLUTION TO THE HORIZON AND FLATNESS PROBLEMS. Alan H. Guth in *Physical Review D*, Vol. 23, No. 2, pages 347–356; January 15, 1981.

A NEW INFLATIONARY UNIVERSE SCENARIO: A POSSIBLE SOLUTION OF THE HORIZON, FLATNESS, HOMOGENEITY, ISOTROPY AND PRIMORDIAL MONOPOLE PROBLEMS. A. D. Linde in *Physics Letters*, Vol. 108B, No. 6, pages 389–393; February 4, 1982.

COSMOLOGY FOR GRAND UNIFIED THEORIES WITH RADIATIVELY INDUCED SYMMETRY BREAKING. Andreas Albrecht and Paul J. Steinhardt in *Physical Review Letters*, Vol. 48, No. 17, pages 1220–1223; April, 1982.

AN INDIAN HUNTERS' CAMP FOR 20,000 YEARS

MULTIDISCIPLINARY RESEARCH IN THE NORTHEAST: ONE VIEW FROM MEADOWCROFT ROCKSHELTER. J. M. Adovasio in *Pennsylvania Archaeologist*, Vol. 52, Nos. 3–4, pages 57–68; September, 1982.

EVIDENCE FROM MEADOWCROFT ROCKSHELTER. J. M. Adovasio, J. Donahue, K. Cushman, R. C. Carlisle, R. Stuckenrath, J. D. Gunn and W. C. Johnson in *Early Man in the New World*, edited by Richard Shutler, Jr. Sage Publications, Inc., 1983.

ISLAND EPIDEMICS

MEASLES PERIODICITY AND COMMUNITY SIZE. M. S. Bartlett in *Journal of the Royal Statistical Society*, Series A, Vol. 120, Part 1, pages 48–70; 1957.

MEASLES ENDEMICITY IN INSULAR POPULATIONS: CRITICAL COMMUNITY SIZE AND ITS EVOLUTIONARY IMPLICATION. Francis L. Black in *Journal of Theoretical Biology*, Vol. 11, No. 2, pages 207–211; July, 1966.

SPATIAL DIFFUSION: AN HISTORICAL GEOGRAPHY OF EPIDEMICS IN AN ISLAND COMMUNITY. A. D. Cliff, P. Haggett, J. K. Ord and G. R. Versey. Cambridge University Press, 1981.

THE AMATEUR SCIENTIST

STRUCTURE AND PROPERTIES OF ALLOYS. Robert M. Brick, Robert B. Gordon and Arthur Phillips. McGraw-Hill Book Company, 1965.

MAGNETISM IN SOLIDS. D. H. Martin. The MIT Press, 1967.

If it's worth remembering, it's worth 3M data recording products.

3M computer media products remember everything, every time. Their reputation for reliability is backed up by something very rare in the magnetic media industry: over thirty years of manufacturing experience. That experience, plus the fact that 3M controls the entire manufacturing process, assures that the same high quality is built into every computer media product we make. So you can have complete confidence in every 3M data cartridge, diskette, and roll of computer tape you buy.

Insist on the best for all your DP needs. Insist on information processing products from 3M. Inventor of the data cartridge and the world's largest manufacturer of flexible magnetic media for the home, office and computer room.

Look in the Yellow Pages under computer supplies and parts for the 3M distributor nearest you. In Canada, write 3M Canada, Inc., London, Ontario. If it's worth remembering, it's worth 3M data recording products.

3M hears you...

3M

The trophies are few and the purses small, but the reason to run in the Glendevon sheepdog trials is as old as the Scottish hills themselves. Sooner or later, every dog has his day. The good things in life stay that way.
DEWAR'S
"White Label"
never varies.
"White Label"
DEWAR'S
BLENDED SCOTCH WHISKY
100% SCOTCH WHISKIES
PRODUCT OF SCOTLAND
John Dewar & Sons Ld
750 ml
(25.4 FL.OZ.)
Authentic
The Dewar Highlander
Blended Scotch Whisky · 86.8 Proof · © 1984 Schenley Imports Co., N.Y., N.Y.