SCIENTIFIC AMERICAN

OCTOBER 1987
$2.50

The next revolution in computers, the subject of this issue, will see power increase tenfold in 10 years while networks and advanced interfaces transform computing into a universal intellectual utility.

Wired Glove *gives a computer user the sensation of handling objects on the screen; the image of the hand mimics the user's movements.*

Knowledge

To reach new heights of productivity, take a

Knowledge is the capital of the future. Texas Instruments can help you manage it as a strategic resource for greater competitiveness – on the factory floor and throughout your operations.

Throughout American business, industry, education and government, the drive for greater competitiveness has become a strategic management concern.

Never has the need for increased productivity been so pervasive and so

SCIENTIFIC AMERICAN

October 1987 Volume 257 Number 4

DEPARTMENTS

THE COVER shows a wired glove that enables computer users to manipulate the images on a graphics monitor with their own hands (see "Interfaces for Advanced Computing," by James D. Foley, page 126). Through the glove interface a computer recognizes finger movements and generates appropriate tactile sensations, creating the impression of touch. A separate sensor (not visible) detects hand position and orientation. The computer is made by Symbolics Inc.

THE ILLUSTRATIONS

Cover photograph by Jon Brenneis

Page	Source
56	Armando Garcia, IBM Corporation, Thomas J. Watson Research Center
58	Johnny Johnson
60	IBM Corporation, Thomas J. Watson Research Center
63-64	James Kilkelly
66	Jon Brenneis
68-73	Jerome Kuhl
79	Jon Brenneis
80	Steve Allen, courtesy of The Fairchild Semiconductor Corporation
81-88	Hank Iken
90	Animals, Animals/Holt Studios, Ltd.
92-98	Joan Starwood
116	Control Data Corporation
118-119	Ian Worpole
120	W. J. Harley, E. I. du Pont de Nemours & Co., Inc. (*left and middle*); A. Corradi, Magnax, Inc. (*right*)
121-123	Ian Worpole
124	IBM Corporation
126	NASA/Ames Research Center
128	Scott S. Fisher, Douglas E. Kerr, Amy F. Wu and Warren J. Robinett, NASA/Ames Research Center
129	Steven G. Reznick, Jolen Flores and Jason G. Williams, NASA/Ames Research Center
130	Carol Donner
132	James Kilkelly
134	Sunil K. Mukerjee and Ali Balali, George Washington University
137	BBN Communication Corporation
138-142	George V. Kelvin, Science Graphics
144	James Kilkelly
146-147	Kitt Peak National Observatory (*left*), Gabor Kiss (*right*)
148	Piet Hut (*top left*), Gabor Kiss (*top right and bottom*)
150	Kitt Peak National Observatory (*top*), Gabor Kiss (*bottom*)
151	Gerald Jay Sussman
152	Gabor Kiss
155	Larry Fagan and Mark Friss
156	Edward Bell
157	Doris Wells-Papanek and Michael Kahn (*top*), Andrew Christie (*bottom*)
158	Cliff Wulfman
159	Randy Miller, University of Pittsburgh School of Medicine
160	Edward Bell
161	Andrew Christie
162	Boeing Computer Services Co.
164-166	Gabor Kiss
167	Boeing Computer Services Co.
168	Boeing Commercial Airplane Co.
170	Michael Goodman
171	James Kilkelly
172-173	Michael Goodman
174-176	Andrew Christie

SCIENTIFIC AMERICAN
Established 1845

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Copyright © 1987 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, P.O. Box 953, Farmingdale, N.Y. 11737-0000.

TWA invites business travelers to spend less time with us. TWA's Airport Express.

Advance Boarding Passes Speed You to the Gate.

With TWA, you can arrange to have your roundtrip boarding passes even before you get to the airport, so those of you without luggage to check can go directly to the gate.

This special Airport Express service will save you much time.

Express Check-In for First Class and Business Class Passengers.

TWA realizes the business world moves fast, so we've made sure business travelers can too. Those First Class and Business Class passengers traveling on widebody flights will benefit from our special express check-in. This is a counter reserved specifically for these passengers to quickly assist them with baggage, reservations, or whatever other services they might need.

TWA's Personal Service Commitment.

TWA is determined to bring you the best personal service in the airline industry. So we've assigned an In-Flight Service Manager to every flight to make sure your trip is as hassle-free and comfortable as possible.

In addition, our Chairman of the Board has put together a Quality Control Team. A group of 30 very picky people who fly TWA, assessing the service they receive, on a checklist of over 100 service items. They report directly back to him with their comments, good and bad. And if anything needs improvement, it gets improved. Fast.

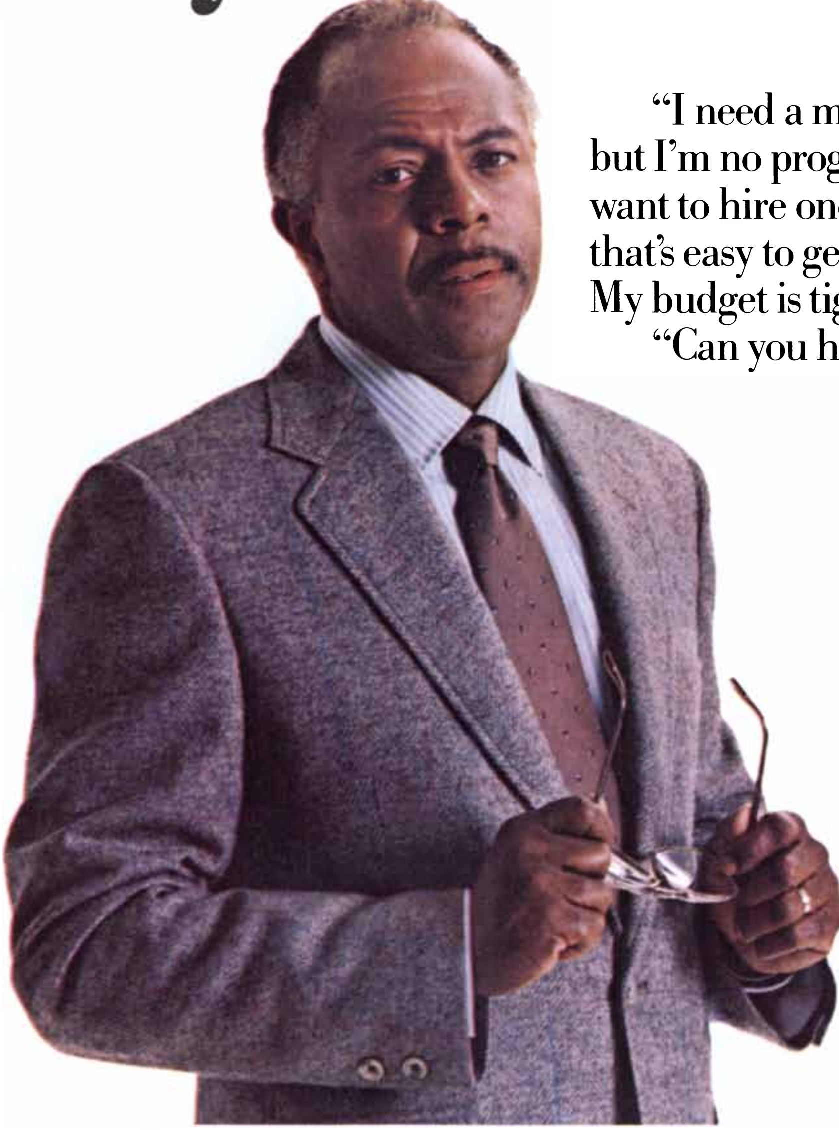
Because at TWA, great service is a top priority.

TODAY'S TWA.
FIND OUT HOW GOOD WE REALLY ARE.

TWA

Why IBM has a full

"I need a mid-size computer, but I'm no programmer and I don't want to hire one. I want software that's easy to get, install and learn. My budget is tight.
"Can you help me?"



Like the north and south poles, people who buy mid-range computers can be alike, yet opposite.

Their needs may be worlds apart, but each asks, "Can you help me?" And IBM answers, "Yes."

The fact is, IBM can help more kinds of customers, more ways, than anybody. We offer an extraordinary range of computing power, and all of it is available to any of our mid-range customers.

But there's one thing we don't have, and neither does anybody else: a "one-size-fits-all" design that can satisfy everyone without compromise. The needs of a small law firm, a retail chain and a multinational corporation, for example, are too diverse for one architecture to serve each equally well.

So IBM offers a choice of mid-range systems, System/3X and System/370. Neither is "better" than the other, they're just better suited for different kinds of jobs. You may even want both.

And since we give you a choice, we'll help you make the right one. Together, we'll analyze your needs for software, connectivity and training. We'll look ahead to future growth, and we'll work to protect your previous investments.

We'll fit our system to your needs, not vice versa.

Power to start with, and grow with.

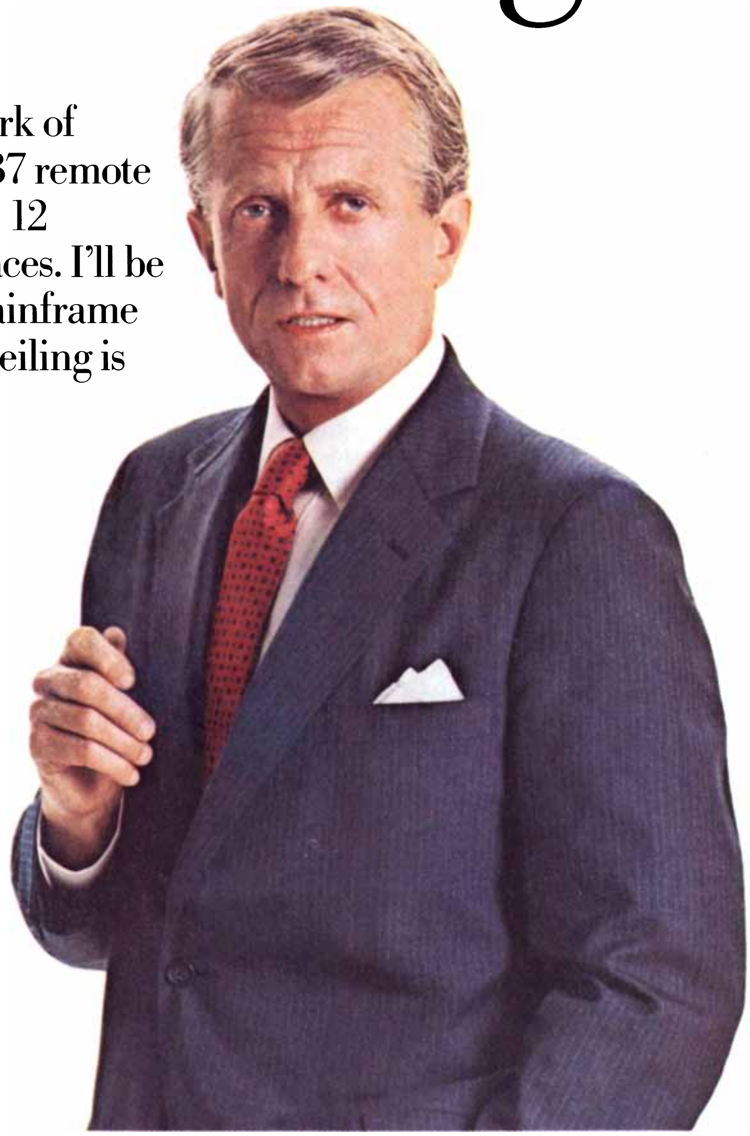
The System/3X family includes the System/36 and System/38.

The IBM System/36 is our most affordable and easiest-to-learn system. With over 4,000 programs available, it's ideal as a small company's first computer, with plenty of room for growth. It's also

-range mid-range.

"I'm building a network of mid-range computers in 37 remote locations connecting with 12 mainframes in 5 other places. I'll be sending both data and mainframe applications. My growth ceiling is pretty much unlimited.

"Can you help me?"



extremely connectible. So a larger business can build a network of System/36s, or include them in existing networks with PCs and mainframes.

The IBM System/38 makes it easy to develop applications, is renowned for its versatile data base, and connects with PCs, hosts and System/36.

Mainframe architecture for smaller neighborhoods.

Our System/370 architecture brings large system computing to mid-size environments with the IBM 9370 Information System and the newly updated IBM 4381 Processor family.

The 9370s are surprisingly compact, yet they not only connect with IBM's biggest mainframes and other systems (through a variety of networks), they can run thousands of mainframe programs.

For sending mainframe power out to departments, or for communicating between departments, 9370s are a terrific choice.

IBM's new 4381s bring similar benefits, plus additional ones: added performance, added growth, and the ability to run MVS/XA, the most powerful IBM mainframe operating system.

The IBM system of choice.

But regardless of architecture, the most important considerations for any system are the people who'll be using it and the challenges they face.

And that's why IBM's full-range mid-range is the right idea. No matter where you fit into it, the fit feels good.

IBM®

LETTERS

To the Editors:

I find it incredible that "'Drivin' My Life Away'" ["Science and the Citizen," SCIENTIFIC AMERICAN, August] did not mention as one possible explanation of the higher automobile-accident death rates per capita in the West the fact that people there simply drive more miles. The reasons mentioned (poor roads, longer time to reach medical care and so on) undoubtedly do contribute to the higher death rates. The lower population density, however, means that everything is farther apart. Since people have more miles and hours on the road, it is hardly surprising that they also have more accidents.

BRUCE WALKER

San Pedro, Calif.

To the Editors:

The discovery of geographic hot spots in the occurrence of automobile accidents, reported in "Science and the Citizen" in the August issue, is a superb parody of epidemiological research. It affords priceless insight into the problem of finding a cause-and-effect relation by means of statistical inference.

You note that "106 of Manhattan's 1,428,285 residents died in [a] three-year period, an average annual death rate of 2.5 per 100,000; in Esmeralda County, Nev., where 13 of 777 residents died, the rate was 558 per 100,000."

Did 13 *residents* of Esmeralda County die, or did 13 people, some from other counties or other states, die in a county whose population was 777? If the dead were all Esmeralda County residents, one might conclude that it is absolutely safe for nonresidents to travel through the county. This is a very important point because most traffic does move *through* the county on a heavily used, undivided highway joining Nevada's two population centers.

Dividing the number of deaths by population introduces a correlation into the statistics where one may not have existed before. A county with no permanent residents, newly created by the Nevada legislature, illustrates the extreme case. There is considerable traffic through the new county because it includes the proposed site for a national nuclear-waste repository. If there is ever a fatal accident, even just one, that county will forevermore have the highest per capita death rate in the nation, and it will be infinite! Does that mean that every driver in that county is doomed to die?

STANLEY CLOUD

Department of Physics
University of Nevada
Las Vegas

[EDITOR'S NOTE: Actually the study reported in "Science and the Citizen" did mention that highway death rates are higher in rural areas even after adjustment for the greater distances driven there. The study also found that the high rates are not explained by the large number of nonresidents traveling through sparsely populated counties.]

To the Editors:

In "Drought in Africa" [SCIENTIFIC AMERICAN, June] Michael H. Glantz rightly points out the importance of nonmeteorological factors, but his prescription of shifting from cash crops to subsistence crops, although popularly accepted, is not necessarily the best policy. In a dissertation recently defended at the University of Wisconsin at Milwaukee, the geographer David Iyegha demonstrated for Ondo State in Nigeria that food production can be easily maintained in the presence of cash crops through intercropping. Indeed, farmers can enhance production of both kinds of crops through judicious management of their plots.

The real problem lies, as Glantz rightly recognizes, in governmental policies that discriminate against farmers to the benefit of city dwellers or punish residents of politically troublesome regions. Indeed, as Djilali Sari showed in *Le désastre démographique de 1867–1868 en Algérie,* government taxation policies can so impoverish farmers that the slightest deficit in rainfall can cause famine. In order to avert such disasters in the future, African governments must be encouraged to avoid both subsidies to city dwellers and large-scale "show projects," which have seldom succeeded.

BRUCE FETTER

Department of History
University of Wisconsin
Milwaukee

To the Editors:

There are conflicting views of the role of cash crops (as opposed to food crops) in the development process. From my perspective a few cases in which cash crops have neither been in conflict with nor taken precedence over the production of food crops do not negate the general case, in which the growing of cash crops (to generate sorely needed foreign exchange) interferes with or takes precedence over food-crop production. There are scores of reports, books and articles that document the continued production of cash crops during drought-catalyzed famines in the West African Sahel and in Ethiopia in the early 1970's and again in Ethiopia in the early 1980's. While food production declined sharply, cash-crop exports were maintained and even increased.

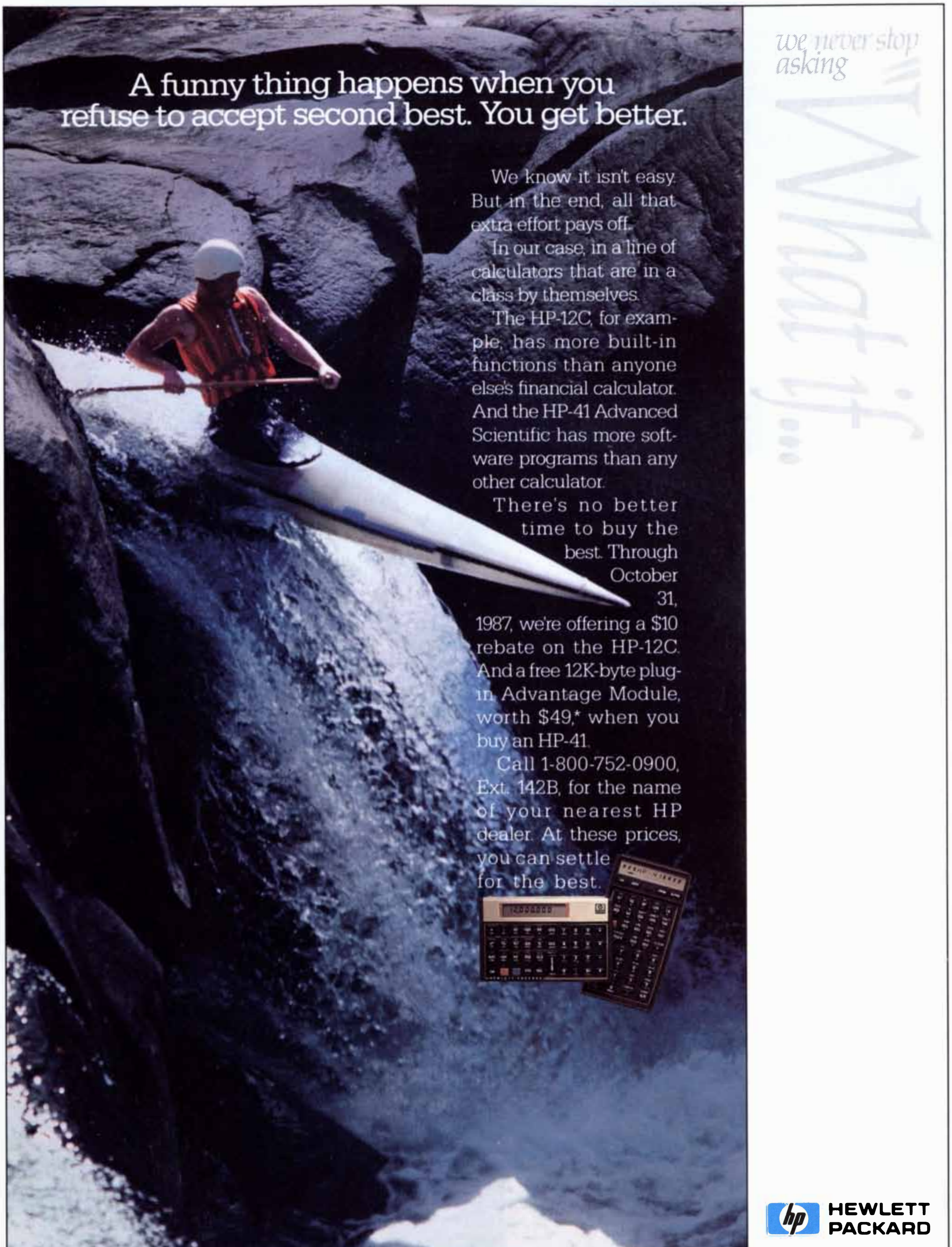
To be sure, there are instances in which cash crops and food have been grown side by side, but when a crisis arises, cash crops have taken precedence. Professor Fetter cites an example from Ondo State in Nigeria. In Borno State and elsewhere in Nigeria, however, cash crops still receive inordinate attention from the government even during periods of drought-induced food stress.

MICHAEL H. GLANTZ

To the Editors:

In an effort to provide a more balanced portrayal of the uncertainties connected with Scheiner's halo ["Science and the Citizen," SCIENTIFIC AMERICAN, May], we wish to raise two points. First, the laboratory evidence cited as being supportive of the hypothesis involving "diamond-shaped" ice is not nearly as strong as suggested in your article. Second, there is no need to resort to this exotic form of ice to produce the halo in the first place: polycrystals of ordinary hexagonal ice can do the job.

Internal crystal structure must be distinguished from external crystal shape. "Cubic ice"—ice *Ic*—has cubic symmetry in its stacking of water molecules, and "hexagonal ice"—ice *Ih*—has hexagonal internal symmetry. Crystal shape is often related to the internal structure, but the relation may not be completely straightforward. Edward Whalley's hypothesis for Scheiner's halo involves octahedrally shaped crystals of ice *Ic,* and some other halos result from pyramidal forms of ice *Ih,* although the commonest halos are produced by simple hexagonal prisms of ice *Ih.* Struc-



*U.S. list price. Offers valid U.S. only.

PG12705


© 1987 BASF Corporation
Create a new world where lightweight plastics can outfly metals.
Aerospace designers are limited by their materials, not their dreams.
At BASF, we looked at the design limitations of metals and saw the need for a radically new generation of materials. The result: strong, lightweight, carbon fiber reinforced plastics. These Advanced Composite Materials will enable future designs to carry more, faster, farther.
In one industry after another, from aerospace to automotive, our broad-based technologies help us create new worlds by seeing in new ways.
The Spirit of Innovation
BASF


WILSON
McLEAN

ture and shape are not necessarily related at all, however. A common form of ice *Ih* in the atmosphere is spherical: frozen water droplets.

The experiments of Erwin Mayer and Andreas Hallbrucker, which you cite as supporting Whalley's hypothesis, involve the rapid quenching of small droplets on a cryoplate. The formation of ice *Ic* was observed; however, the relevance of these results to the freezing of atmospheric droplets is doubtful for two reasons. Both the extremely rapid rate of cooling and the presence of a substrate (the cryoplate) may influence the crystalline structure of the resulting ice, and both are uncharacteristic of atmospheric processes. In addition these experiments demonstrated only that the ice so formed had a cubic crystalline structure. No claim was made about the shape of the submicroscopic crystals. They might be octahedral as required by Whalley's hypothesis (and as stated in your article), but then again they might grow into cubes or some other form inappropriate to Scheiner's halo.

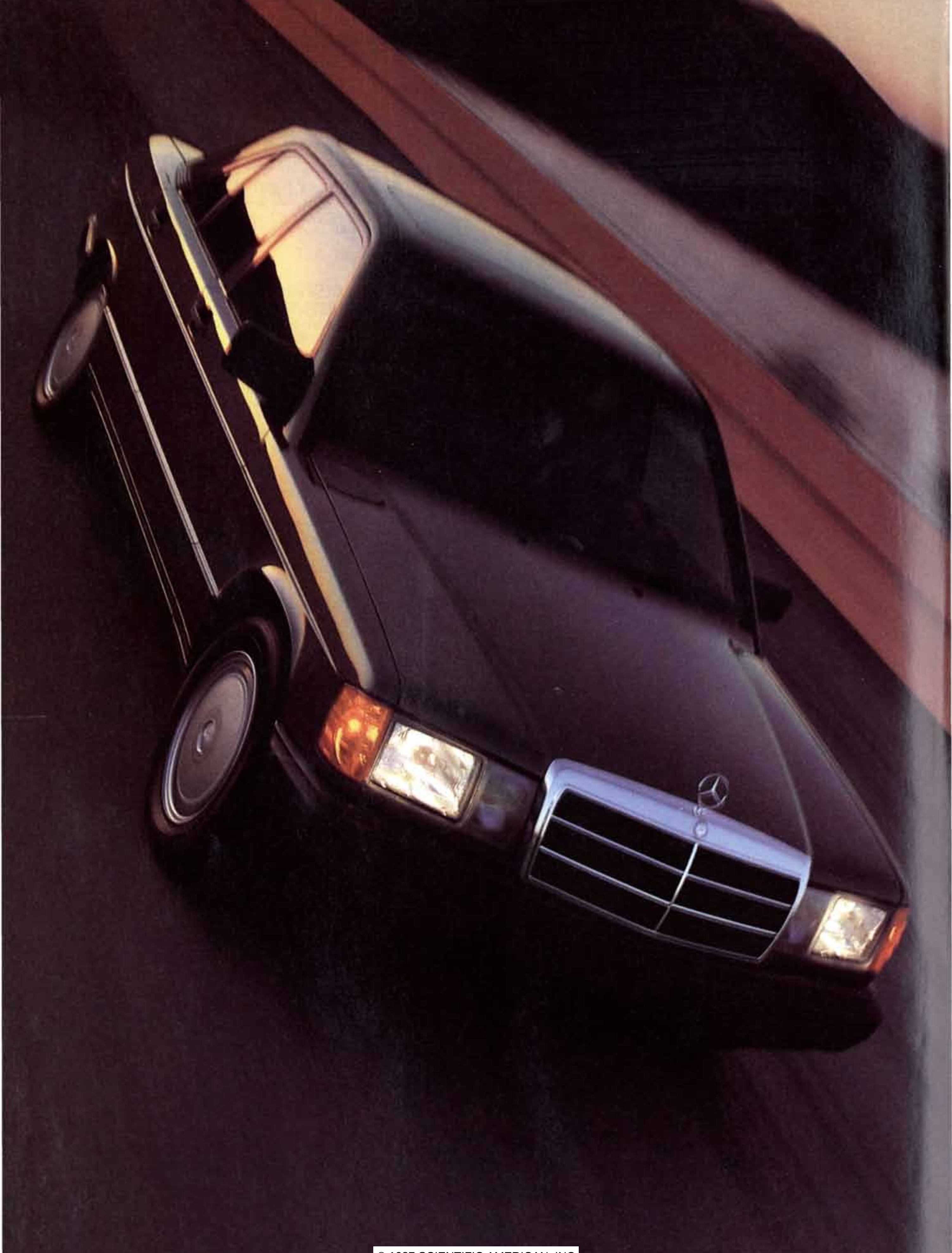
We have suggested an alternative explanation for Scheiner's halo (*Journal of the Atmospheric Sciences,* in press). As noted in your article, single hexagonally shaped crystals of ice *Ih* are incapable of producing a halo that has the right width. The same is not true of twinned crystals, however. Such a twin consists of a pair of crystals in contact at a specific orientation. For the commonest (hexagonal) ice-crystal twins, we have shown that a particular pair of crystal faces (one face from each member of the twin) forms a 70.5-degree prism, just as the octahedral crystals in Whalley's proposal do. Such twins might give rise to Scheiner's halo.

What is so attractive about the explanation involving twins of ice *Ih,* we think, is that this kind of ice is not uncommon in the atmosphere. This is in marked contrast to octahedral crystals of ice *Ic*. Ice *Ic* has never been observed in the atmosphere, and octahedrally shaped ice crystals have never been observed—anywhere. We acknowledge that there is really very little positive evidence for either our proposal or Whalley's. After all, the last sighting of Scheiner's halo was in 1920.

ANDREW J. WEINHEIMER

CHARLES A. KNIGHT

National Center for Atmospheric Research



50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

OCTOBER, 1937: "Despite great projects for harnessing water power for the generation of electric current and the millions of kilowatts being made available by the simple process of making water work as it runs down hill, steam remains the principal generative force for the production of mass power. In the year ending on March 31, steam turbines produced about 70 percent of the 109 billion kilowatt-hours consumed in this country, and if the struggle of turbine manufacturers to fill orders for central station equipment is any criterion, this percentage is going to endure or become even larger."

"Arsenic, though highly poisonous to most forms of life, is food for some microbes. Other microbes eat selenium, an element with which the soil occasionally plays Borgia. Here are two elements that contribute to the vitality of some live beings and to the early death of others. It is one of the great mysteries of bio-science why a living mechanism can use only certain elements, and finds either useless or toxic the rest of the 90-odd known to chemists. We may wonder if man really gets along, with a mere 15 elements, as famously as he might with a number of additional ones. The fascinating history of the vitamins may be repeated with the chemical elements."

"Engineers at the University of Illinois, Urbana, working in a 'research residence,' have determined that a seven-room home, under conditions existing in that locality, may be artificially cooled and air conditioned throughout a four-month summer season at an operating cost as little as 50 cents per day. During the period of tests, outdoor temperatures ranged from 84.5 to 102.8 degrees."

"The discovery was recently made that, contrary to general knowledge, Matthew Brady, famous photographer of the Civil War, was not the first man to make war pictures. The credit really belongs to an unsung and unknown hero (or heroes) who took pictures of the Mexican War in 1847. The pictures were taken by the daguerreotype process."

"A new building code has been adopted by New York City. Among other provisions in keeping with recent progress in the building industry, it permits the use of welding in place of riveting in the construction of steel buildings. This action removes the last serious obstacle to the general use of welding in building construction."

SCIENTIFIC AMERICAN

OCTOBER, 1887: "A reporter of the *Evening Post* lately interviewed Mr. Edison, who said that the commercial phonograph is now the most interesting thing to him. The stories he tells of what his perfected phonograph will do are so extraordinary that he scarcely expects people to believe him, and yet he says that the apparatus is so simple, so effective and so immediately useful that he is certain of its rapid introduction into business. He said: 'The merchant or clerk who wishes to send a letter has only to set the machine in motion, and to talk in his natural voice and at the usual rate of speed into the receiver. When he has finished, the sheet, or "phonogram," as I call it, is ready for putting into a little box made on purpose for the mails. The receiver of a phonogram will put it into his apparatus, and the message will be given out more clearly, more distinctly, than the best telephone message ever sent. For musicians the phonograph is going to do wonders, owing to the extreme cheapness with which I can duplicate phonograms and the delicacy with which the apparatus gives out all musical sounds.'"

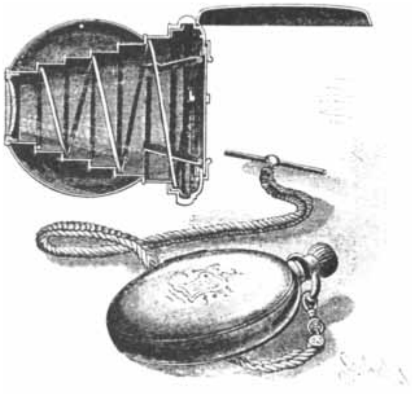
"The type writer is creating a revolution in methods of correspondence, and filling the country with active, competent young ladies who are establishing a distinct profession and bringing into our business offices, lawyers' offices, editorial sanctums, etc., an element of decency, purity and method which is working a perceptible change. The art of dictation is almost a new art, but it is spreading rapidly, and business men are beginning to understand that much of their lives has been wasted in the mere mechanical drudgery of letter writing."

"Photography has long been considered the trusty ally of counterfeiters and forgers, but it cannot be looked upon in that light any longer, as M. Gobert, of the Banque de France, has succeeded in converting this art into a most efficient detective agent. His process consists of taking a greatly magnified photograph of the suspected coin or document, on which any erasure or defect can then easily be detected."

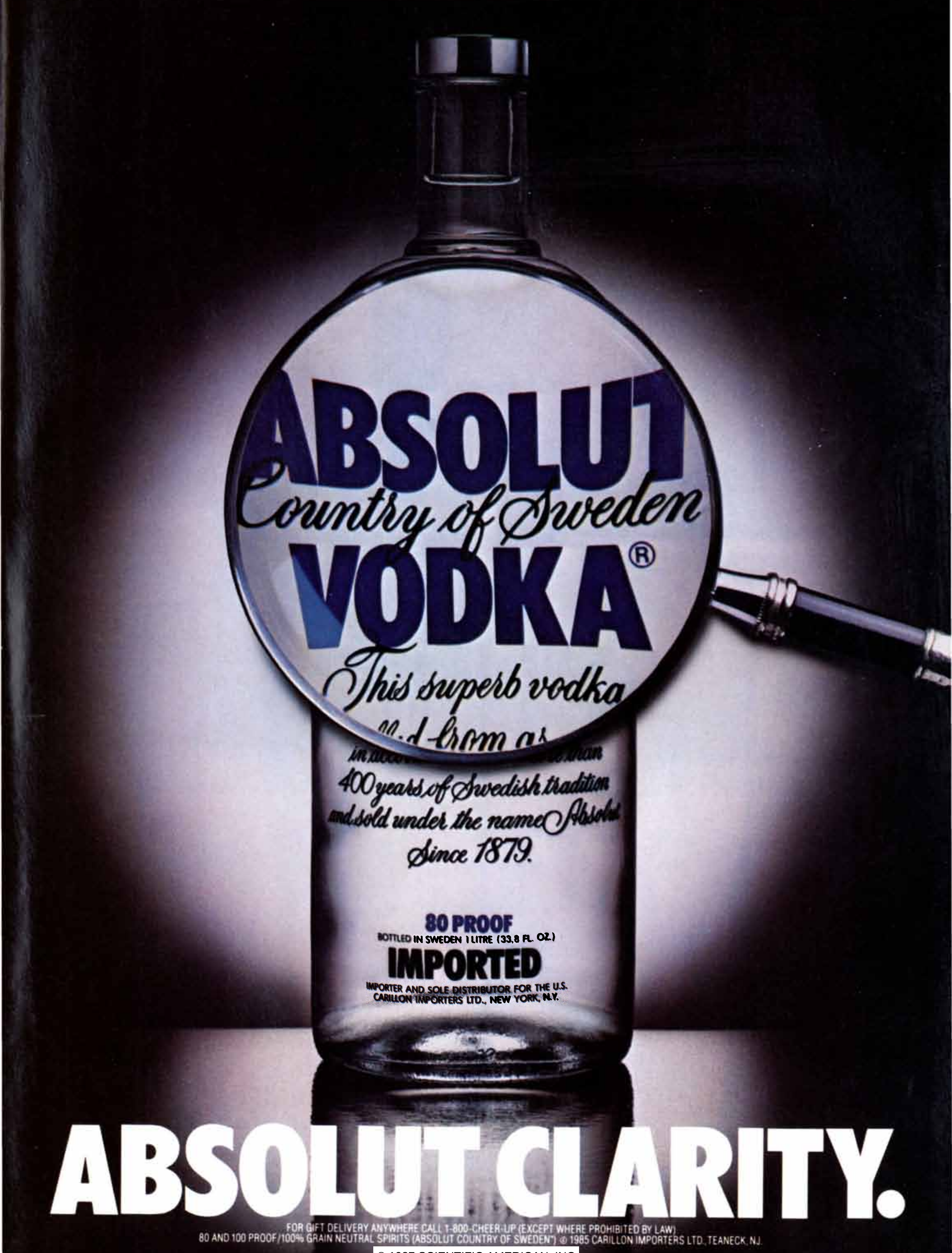
"With regard to the general application of electricity to street car propulsion, there is a very great future in store for us, and the time is very near when horses on street cars will be entirely abandoned. A great advantage (apart from economy) is that we shall be able to travel at a greater speed."

"The first of this season's races for the America's cup, Sept. 27, was won by the American yacht *Volunteer,* which came in 20 minutes ahead of the British challenger *Thistle.* The second race, Sept. 30, was won by *Volunteer* by the superiority of her windward work. The cup therefore remains here."

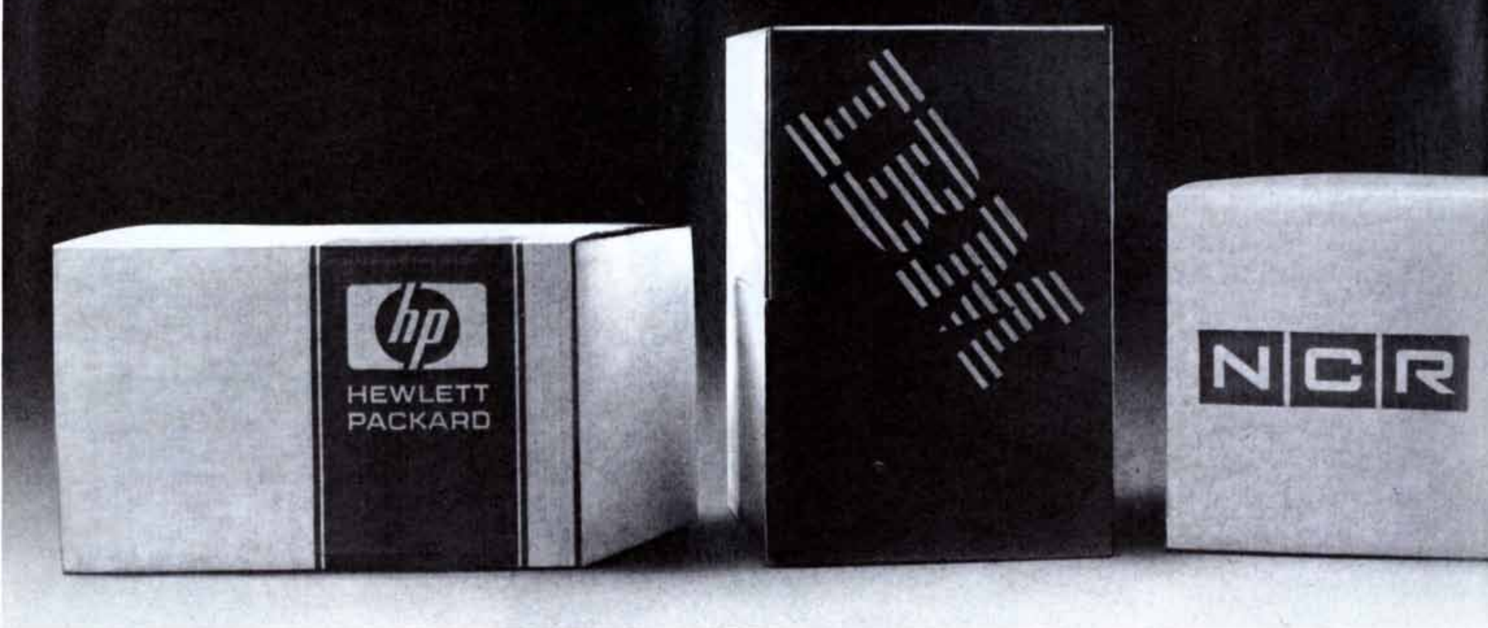
"Numerous ways have been invented to compress the essentials of a photographic apparatus into a compact space, but we call to mind none more effective than the watch camera illustrated herewith. By means of other special attachments, we see no reason why a genuine timepiece may not be combined with the photographic watch in such a way that a race horse can be instantly photographed at the same moment the stop mechanism of the watch is manipulated."



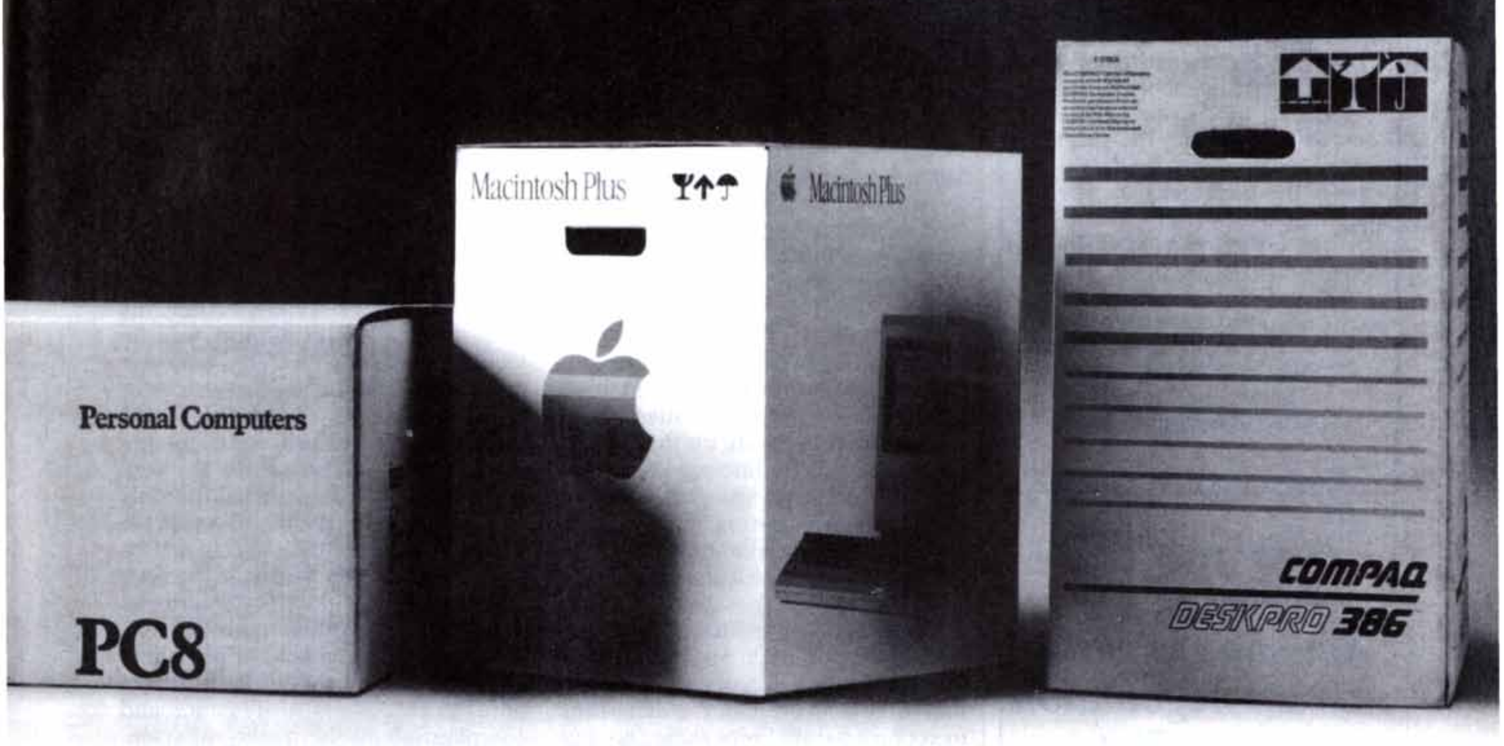
A watch camera


ABSOLUT
Country of Sweden
VODKA®
This superb vodka
400 years of Swedish tradition
and sold under the name Absolut
Since 1879.
80 PROOF
BOTTLED IN SWEDEN 1 LITRE (33,8 FL. OZ.)
IMPORTED
IMPORTER AND SOLE DISTRIBUTOR FOR THE U.S.
CARILLON IMPORTERS LTD., NEW YORK, N.Y.
ABSOLUT CLARITY.
FOR GIFT DELIVERY ANYWHERE CALL 1-800-CHEER-UP (EXCEPT WHERE PROHIBITED BY LAW)
80 AND 100 PROOF/100% GRAIN NEUTRAL SPIRITS (ABSOLUT COUNTRY OF SWEDEN™) © 1985 CARILLON IMPORTERS LTD., TEANECK, N.J.

3M data cartridge available in five



tape backup is now popular models.



The makers of even the most popular PCs know it's insane to go to market without a reliable backup.

So the hunt began.

They each sat through hundreds of action-packed hours of blank, scrambled and frozen screens to see which backup system backed up best.

In the end, they each chose 3M data cartridge tape technology.

Why?

We've been covering computer and human errors almost longer than computers and humans have erred together.

Not only did we invent and patent data cartridge tape technology, we've had 16 years to make it better. Through every technological breakthrough, we've consistently proven to be the best way to back up data.

Just ask Apple.® And NCR.® And HP.® And IBM.® And COMPAQ.®

Trademarks/owner: Apple/Apple Computer Inc.; NCR/National Cash Register; HP/Hewlett-Packard; IBM/International Business Machines Corp.; COMPAQ/COMPAQ Computer Corp.

THE AUTHORS

ABRAHAM PELED ("The Next Computer Revolution") is vice-president of systems and director of computer sciences at the Research Division of the IBM Corporation. His B.S. and M.S. are from the Technion–Israel Institute of Technology and his M.A. and Ph.D. are from Princeton University. He has been at IBM since 1974, when he joined the corporation's Thomas J. Watson Research Center. After working at other IBM research centers, he returned to the Watson center in 1983 as director of technical planning and controls for the corporation's research division. Peled has been in his current position since 1985, managing the Watson center's department of computer science as well as overseeing computer-science activities at IBM research centers and laboratories around the world.

GEOFFREY C. FOX and PAUL C. MESSINA ("Advanced Computer Architectures") are both involved in establishing computer facilities based on parallel processing for the California Institute of Technology. Fox, who is professor of physics and associate provost for computing, has been the principal investigator in a joint Caltech/Jet Propulsion Laboratory program that encourages the application of parallel computing in science. Messina is leader of a related project, the Concurrent Supercomputing Initiative at Caltech (CSIC), which aims to set up a parallel-processing facility equaling the performance of conventional supercomputers. Fox studied mathematics at the University of Cambridge as an undergraduate and received a Ph.D. in theoretical physics from the university in 1967. After holding appointments at various academic and scientific institutions in Britain and the U.S., Fox went to Caltech in 1970 and became professor in 1979. Messina has a B.A. (1965) from the College of Wooster and an M.S. (1967) and a Ph.D. (1972) from the University of Cincinnati, where he also managed the university's computer services. After completing his studies he joined the Argonne National Laboratory, where he became director of the mathematics and computer-science division in 1982. Messina was appointed project leader of the CSIC this spring.

JAMES D. MEINDL ("Chips for Advanced Computing") is vice-president for academic affairs and provost of the Rensselaer Polytechnic Institute. He was educated at Carnegie-Mellon University, obtaining his Ph.D. in electrical engineering in 1958. After working with the Westinghouse Electric Corporation and the Army Electronics Laboratory, he went to Stanford University in 1967 as associate professor of electrical engineering. He was made full professor in 1970, served as director of the Stanford Electronics Laboratories from 1972 to 1986 and moved to Rensselaer last year. Meindl was the author of "Microelectronic Circuit Elements" in the September 1977 issue of SCIENTIFIC AMERICAN.

DAVID GELERNTER ("Programming for Advanced Computing") is associate professor of computer science at Yale University. After his undergraduate education at Yale he went on to earn a Ph.D. in computer science from the State University of New York at Stony Brook. He writes that his interest in programming languages and parallelism "was motivated originally by the need for tools to support new kinds of program structures, particularly in artificial intelligence, but then I became interested in them in their own right. I'm still interested in artificial intelligence, particularly in medical applications and in memory models; in my spare time I grow flowers, write fiction and complain about the quality of scientific prose." Gelernter's introductory text *Programming Linguistics* is to be published by W. W. Norton & Co.

MARK H. KRYDER ("Data-Storage Technologies for Advanced Computing") is professor of electrical and computer engineering at Carnegie-Mellon University. He studied at Stanford University, where he got a B.S. (1965), and at the California Institute of Technology, where he got an M.S. (1966) and a Ph.D. (1970). After spending two years as a visiting scientist at the University of Regensburg in West Germany, he joined the staff of the IBM Corporation's Thomas J. Watson Research Center. In 1978 he went to Carnegie-Mellon, where he was made full professor in 1980. Since 1982 Kryder has been director of the university's Magnetics Technology Center.

JAMES D. FOLEY ("Interfaces for

WITHOUT DATA GENERAL, INTEGRATING YOUR COMPUTERS IS LIKE PITTING MAN AGAINST MACHINE.

DATA GENERAL GIVES YOU THE BEST SOLUTIONS FOR COMPUTER INTEGRATED MANUFACTURING SYSTEMS.

Are the levels of your manufacturing operation locked in hand to hand combat? Our total integration solutions can make them all work together. Hand in hand.

Our computers and solutions span key areas. Linking Engineering with Manufacturing. Planning with Sales.

We give you advanced productivity solutions. Like TEO,™ our technical automation system. And CEO,® our business automation system. To combine with major CIM applications.

Data General is firmly committed to industry communications standards. Like MAP, SNA, X.25 and Ethernet.® They help you forge different systems into a single information mainstream.

What's more, our MV/Family computers are price/performance leaders. Which make these solutions more affordable.

If your company is wrestling with today's complex manufacturing needs, call 1-800-DATAGEN (Canada 1-800-268-5454).

Data General
a Generation ahead.

Advanced Computing"), professor of computer science at George Washington University, is studying the application of expert-systems techniques to user-computer interfaces. He received his B.S. at Lehigh University in 1964 and an M.S. (1965) and a Ph.D. (1969) from the University of Michigan. A year later he became assistant professor of computer science at the University of North Carolina, where Frederick P. Brooks, Jr., introduced him to the promise of artificial realities. In 1976 he moved to the U.S. Bureau of the Census to design graphics for presenting statistics. He went on to G.W.U. a year later. Foley is a coauthor of *Fundamentals of Interactive Computer Graphics* and a fellow of the Institute of Electrical and Electronics Engineers.

ROBERT E. KAHN ("Networks for Advanced Computing") is president of the not-for-profit Corporation for National Research Initiatives, which he founded last year to foster research in information processing. He got his bachelor's degree at the City College of the City University of New York in 1960 and his M.A. (1962) and Ph.D. (1964) in electrical engineering from Princeton University. After two years as assistant professor at the Massachusetts Institute of Technology he went to Bolt, Beranek and Newman, Inc., where he helped to develop the national computer network ARPANET. In 1972 Kahn moved to the Defense Advanced Research Projects Agency, where he subsequently became director of its information-processing-techniques office.

PIET HUT and GERALD JAY SUSSMAN ("Advanced Computing for Science"), respectively professor of astrophysics at the Institute for Advanced Study and professor of electrical engineering at the Massachusetts Institute of Technology, are working on a project to combine astrophysics with computer engineering in a "computational observatory." Hut earned an M.S. at the University of Utrecht and a Ph.D. from the University of Amsterdam. He has been a member of the institute since 1980, except when he taught astrophysics in 1985 at the University of California at Berkeley. An avid dilettante of the Japanese language and culture, Hut has collaborated with colleagues in fields ranging from particle physics, geophysics and paleontology to computer science. Sussman received his undergraduate degree (1968) and his Ph.D. in mathematics

MACSYMA

automates symbolic mathematics. And yields enormous improvements in speed, accuracy and modeling power.

If you work with quantitative models in scientific or engineering disciplines, MACSYMA can increase your modeling power. MACSYMA combines symbolic and numeric computation. And enables you to accurately manipulate symbolic expressions in a fraction of the time required manually.

Wide range of capabilities

MACSYMA offers the widest range of capabilities for symbolic computations in applied mathematics of any commercially available program. For example:

Algebra: MACSYMA can manipulate large expressions, expand, simplify and factor expressions, handle matrices and arrays, and solve systems of equations.

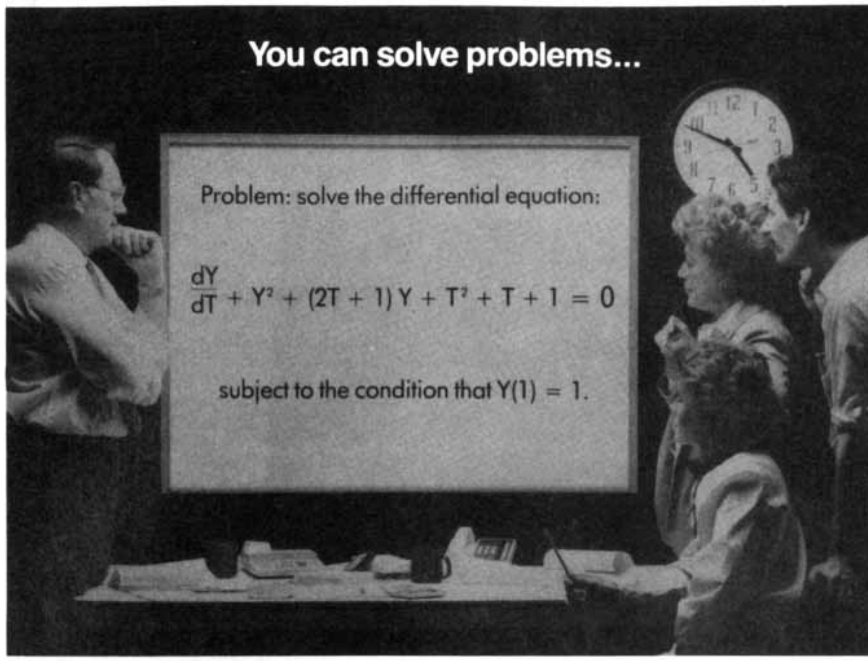
Calculus: MACSYMA can differentiate, perform definite and indefinite integration, take limits, expand functions in Taylor or Laurent series, solve differential equations, and compute Laplace transforms.

Numerical analysis: You can perform numerical analysis using MACSYMA® language, compute with arbitrary precision arithmetic, use a library of numerical analysis routines, and generate FORTRAN code.

Graphics and interfaces: MACSYMA can generate report-quality graphics in 2D or 3D, with perspective, hidden line removal, and captions. MACSYMA also interfaces with the mathematical text processors 'TeX' and 'troff'.

Broad base of applications

Throughout the world thousands of scientists, engineers and mathematicians are using MACSYMA in such diverse applications as aeronautical design, structural engineering, fluid mechanics, acoustics, CAD, electronic and VLSI circuit design, electromagnetic field problems, plasma physics, atomic scattering cross sections, control theory, maximum likelihood estimation, genetic studies, and more.



Symbolically...

```
(C1) DEPENDS(Y,T)$
(C2) DIFF (Y,T)+Y^2+(2*T+1)*Y+T^2+T+1;
```

(D2) $\frac{dY}{dT} + Y^2 + (2T+1)Y + T^2 + T + 1$

```
(C3) SOLN:ODE(D2,Y,T);
```

(D3) $Y = -\frac{\%C\,T\,\%E^T - T - 1}{\%C\,\%E^T - 1}$

```
(C4) SOLVE(SUBST([Y = 1, T = 1],D3),%C),NUMER;
(D4)    [%C = 0.5518192]
(C5) SPECIFIC SOLN:SUBST(D4,SOLN);
```

(D5) $Y = -\frac{0.5518192\,T\,\%E^T - T - 1}{0.5518192\,\%E^T - 1}$

and Numerically.

```
(C6)  FORTRAN(D5)$
           Y = -(0.5518192*T*EXP(T) - T - 1)
     1         /(0.5518192*EXP(T) - 1)
```

Available on many computer systems

Current systems include:

- Symbolics 3600™ Series
- VAX & MicroVAX II
- SUN-2 & SUN-3
- Apollo
- Masscomp

Other versions will be following soon.

For an information kit about all the ways MACSYMA can work for you, just call

1-800-MACSYMA.

In Mass., Alaska or Hawaii only, call (617) 621-7770.

Or please write to us at

Computer-Aided Mathematics Group
Dept. M-SA3
Symbolics, Inc.
Eleven Cambridge Center
Cambridge, MA 02142

MACSYMA

The most comprehensive software for mathematical computing.

symbolics™

(1973) from M.I.T., where he has been a member of the Artificial Intelligence Laboratory since his freshman year. He is coauthor of an introductory textbook on computer science.

GLENN D. RENNELS and EDWARD H. SHORTLIFFE ("Advanced Computing for Medicine") work together in the Medical Computer Science Group at the Stanford University School of Medicine. Both of them are also practicing physicians at Stanford, Rennels as a resident in anesthesiology and Shortliffe as associate professor of medicine. Rennels got his B.A. at Dartmouth College in 1977, his M.D. at the Dartmouth Medical School in 1980 and a Ph.D. in medical information sciences from Stanford last year (with Shortliffe as his thesis adviser). Shortliffe got his A.B. at Harvard College in 1970. He then joined the medical-scientist training program at Stanford, where he earned his Ph.D. in medical information sciences and his M.D. in 1975 and 1976 respectively. His doctoral dissertation was on a rule-based computer program that advises physicians on antimicrobial therapies. Shortliffe was assistant professor of both medicine and computer science at Stanford from 1979 to 1985, when he became associate professor.

ALBERT M. ERISMAN and KENNETH W. NEVES ("Advanced Computing for Manufacturing") are both with the Boeing Computer Services Company, where Erisman is director of the Engineering Technology Applications Division and Neves is manager of research and development programs for the Engineering Scientific Services Division. Erisman received his doctorate in applied mathematics at Iowa State University in 1969 and then joined Boeing. His research has focused on high-performance scientific computing and sparse-matrix algorithms, which are applied to analyze large-scale circuits such as electric-power systems. Erisman has also taught courses at the University of Washington and at Seattle University. Neves, who holds a Ph.D. in mathematics from Arizona State University, worked first at the IBM Corporation and at the Nuclear Power Generation Division of the Babcock & Wilcox Company, where he was a senior mathematician. In 1975 he joined Boeing. Among his projects is a high-speed computing program, currently equipped with advanced workstations and parallel processors.

SCIENCE AND THE CITIZEN

Back with a Vengeance

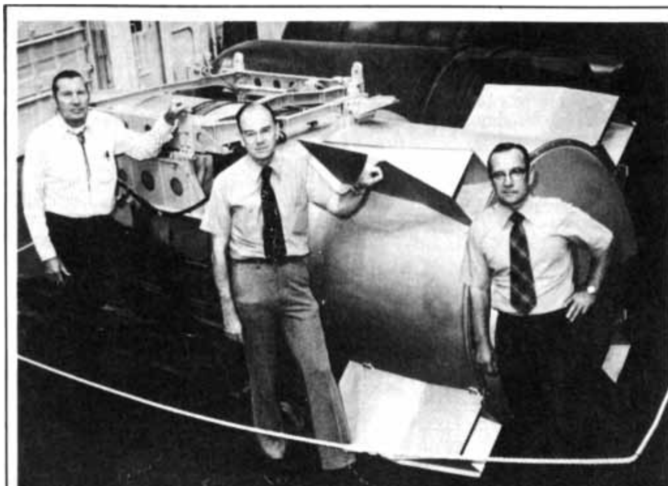
The Pentagon quietly rolls big H-bombs out of retirement

While the Reagan Administration has recently professed a desire to trim its nuclear arsenal (if, of course, the Soviet Union does likewise), the Pentagon has quietly taken a giant step away from that goal. This step, which boosts the gross destructive power of the U.S. arsenal by some 20 percent according to one estimate, was revealed earlier this year in a single sentence of testimony by Robert B. Barker, Assistant to the Secretary of Defense for Atomic Energy. "The B53 bomb," Barker told a House Appropriations subcommittee, "which was in inactive reserve, is being returned to operational status."

The B53 is not just another nuclear bomb: it is the most destructive weapon the U.S. has ever possessed. Its explosive yield, according to *Nuclear Weapons Databook*, is equal to the yield of nine million tons of TNT or 750 Little Boys (one of which devastated Hiroshima) or 30 MX-missile warheads. Built in the early 1960's, the B53's were deployed on B-52 stratofortresses of the Strategic Air Command; by 1983 the command had mothballed the bombs and replaced them with smaller bombs and cruise missiles.

The Department of Defense declines to explain why it is bringing the aging bombs back in service, but Barker's testimony suggests a possible motive. Just before mentioning the B53 decision, he described efforts to develop so-called earth-penetrating warheads, intended, in his words, to "hold hardened, underground, deeply buried targets at risk." Robert S. Norris of the National Resources Defense Council suggests that U.S. war planners have recently become convinced that modern, relatively low-yield warheads, even when delivered by highly accurate missiles, might not be powerful enough to crush buried Soviet command centers unless they burrowed deep into the earth first. Developing strategic earth-penetrating warheads has proved difficult. Weapons designers have built needle-nosed shells that can plunge hundreds of feet into the earth; now they are trying to design nuclear devices that can fit in these shells. "The problem," Milo D. Nordyke of the Lawrence Livermore National Laboratory says, "is to build a warhead that will survive that very severe impact and still detonate."

Have doubts about the effectiveness of new weapons led the Pentagon to resurrect the "big old bomb"?



B53 BOMB

WEIGHT:
8,890 POUNDS

LENGTH:
148 INCHES

DIAMETER:
50 INCHES

YIELD:
9 MEGATONS

B53 THERMONUCLEAR BOMB sits on a dolly behind three employees of the U.S. Department of Energy. The bombs were built at a DOE plant in Burlington, Iowa, in the early 1960's.

Air Force officials may also have fallen back on the "big old bomb on a B-52" because they doubt the reliability of new strategic delivery systems, James P. Rubin of the Arms Control Association says. Technical glitches, he points out, have plagued the B-1 and Stealth bombers and the MX missile. Rubin suggests that the officials' concerns may be heightened by President Reagan's proposal to deploy a Strategic Defense Initiative system jointly with the Soviet Union. Gen. John T. Chain, Jr., head of the Strategic Air Command, recently stated that if the U.S.S.R. deploys an SDI-type shield against ballistic missiles, the Air Force will need to build up its bomber force to maintain deterrence.

Whatever its motives, the Pentagon's resurrection of the B53 belies its own stated policies, according to Josephine A. Stein, a congressional science fellow. Defense officials, she observes, claim they must explode warheads at the Nevada Test Site to ensure that new warheads are "safe" (against accidental detonation) and that aging ones are "reliable" (reliably destructive). Stein contends that by the military's own standards the B53 bombs must be suspect on both counts. Indeed, in 1980 the U.S. Arms Control and Disarmament Agency told Congress that new bombs entering the U.S. arsenal were safer than the B53. To comply with the 150-kiloton limit imposed on nuclear tests by the Threshold Test Ban Treaty, however, the military can only test the reliability or safety of a drastically stripped-down version of the B53. Such a test, Stein says, would be virtually meaningless.

Defense officials have also maintained that they can wage a "limited nuclear war" against the U.S.S.R., de-

For those analyzing developments in the PC industry, Compaq offers an executive summary.

Compaq still

In the midst of the clamor surrounding the new IBM® PS/2 personal computers, one thing is perfectly clear to people who really know PC's. COMPAQ® personal computers still work better. They enhance your productivity *within* the industry standard, and give you maximum performance from the world's largest library of business software.

Demand for the 12-MHz COMPAQ DESKPRO 286 has nearly doubled since the PS/2 introduction.

The 12-MHz COMPAQ PORTABLE III is the smallest, most powerful full-function portable there is.

Still the performance leader

COMPAQ personal computers prove superior in overall performance.

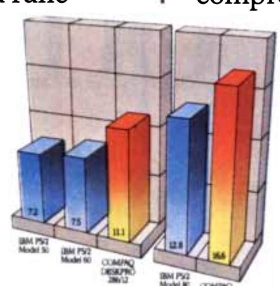
Take speed. The COMPAQ DESKPRO 286® runs your software up to 20% faster than its PS/2™ counterpart. And, the COMPAQ DESKPRO 386™ sets all records for speed in advanced-technology, industry-standard personal computers.

Consider flexibility. Compaq offers 5¼" diskette drives, and allows you to add 3½" drives. In fact, you can add up to four different storage devices on all COMPAQ desktops.

Examine compatibility. We let you use all the industry-standard software and expansion boards that you already own.

Look at expandability. Our industry-standard slots enable you to add many extra functions. So you can configure your system exactly the way you want it.

Finally, compare portability. You can't. The 12-MHz 80286-based COMPAQ PORTABLE III™ is the undisputed leader. It offers the performance of a desktop without any of the compromises you'll find in other portables.



COMPAQ PC's outperform PS/2 PC's in tests for processor speed & disk performance, using the Norton Performance Index Version 4.0.*

Compaq applies innovative technology within the industry standard, *without* sacrificing compatibility.

Earn higher returns on your investment

American business has $80 billion invested in

It still simply works better.

*Index based on a scale that rates the original IBM PC as 1.00.

works better.

Two-way compatibility

Compaq has become famous for its legendary compatibility and connectivity. Our personal computers will run thousands of programs, without modification, far faster than other computers. And they work with all the other compatible computers in your office, without time-consuming diskette conversions.

The COMPAQ DESKPRO 386 sets the standard for high-performance, advanced-technology desktop computing.

the current PC standard, including 72 million software and hardware products, and hundreds of millions of hours in training.

Compaq designs its computers to protect your investment. And because they do more, they also maximize it.

As for the future, all 80286- and 80386-powered COMPAQ personal computers will run the new MS OS/2 operating system, allowing you to directly access up to 16 megabytes of memory. And they'll run all the applications developed for OS/2™. Again, much faster.

We don't burn bridges, we build them

At Compaq, advances are measured by our ability to push technology forward, without leaving you behind.

The new MS OS/2 operating system will run on all 80286- and 80386-based COMPAQ personal computers.

COMPAQ computers let you incorporate developing technology, and take advantage of the latest technology in a way that's fully compatible with the hardware, software and add-ons you already own. So Compaq protects your investment.

These are all reasons why recent surveys show COMPAQ owners are the most satisfied personal computer users.

Call 1-800-231-0900, operator 39, for information and the location of your nearest Authorized COMPAQ Computer Dealer. In Canada, call 416-449-8741.

COMPAQ®

stroying military targets while minimizing "collateral damage" to civilians. But collateral damage, rather than military advantage, would be the only significant result of an attack on Soviet command centers by stratofortresses loaded with nine-megaton bombs, Stein maintains. By the time the subsonic bombers arrived at their targets, she points out, the Soviets would have had hours in which to launch their own missiles, and questions of military advantage would be moot. *—John Horgan*

Animal Passions

A study of laboratory-animal use founders in acrimony

What controls should society impose on scientists who use laboratory animals? A report on the topic gestating at the National Research Council, the research agency of the National Academies of Sciences and Engineering, is stirring strong feelings. Animal-welfare champions on the panel that produced the report complain that the benefits to animals of such controls have been slighted. But one panel member, Arthur C. Guyton of the University of Mississippi School of Medicine, felt so strongly that investigators' interests had been underrepresented that he submitted a "minority report on the plight of the scientific investigator using animals, especially large animals."

Some panelists wanted the report to acknowledge that research animals have at times been mistreated and that some animal experiments have been misleading. Christine Stevens, president of the Animal Welfare Institute, said in late August that she would refuse to sign the existing report because it was "completely unbalanced, with severe omissions of essential information." W. Jean Dodds, chief of the New York Department of Health's Laboratory of Hematology, who has avoided breeding animals with genetic defects, said that the report gave "no perspective on the variety of views within the scientific community." She also would not sign without changes.

A major issue in the dispute is the 1985 amendments to the Animal Welfare Act that require institutions using some warm-blooded animals to have animal-care committees that include at least one outside member to oversee research, and to maintain certain standards in animal facilities. The NRC panel initially voted to recommend that these regulations be extended to cover research on rats, mice and birds but subsequently reversed itself. Guyton's submission depicts the changes that have been made as "a succession of compromises with the animal rights–welfare movements"—which he charges is the antivivisectionist movement under a new name. Guyton argues that "it is very doubtful that any but the very large animal research programs can survive unless many of the new rules are rescinded, which doubtless will not occur without the very strong support of the National Research Council."

The panel was also divided over the use of animals from pounds for long-term or repeated experiments. Bills have been proposed in both houses of Congress that would forbid any use of pound animals by researchers supported by the National Institutes of Health. Michael E. DeBakey, who pioneered many successful techniques for heart surgery and who is a member of the NRC panel, has recently published articles pointing out that purpose-bred dogs for animal research cost between $275 and $600 each, compared with $5 to $55 for a pound animal. DeBakey says most pound animals are not lost pets, as they are often portrayed to be, but abandoned or stray animals that would be killed anyway. Indeed, fewer than 2 percent of the 10 to 15 million pound dogs and cats killed each year go to research laboratories, and most are used in acute studies in which they are anesthetized and never regain consciousness.

Although the report is under independent review, it may never present a consensus that could guide legislators. Guyton's minority statement argues that "strong measures should be taken to insure that pound animals are made available to medical research." If the NRC report is ever issued, it seems likely that it will have several appended disclaimers and additions by individual members of the panel. *—Tim Beardsley*

PHYSICAL SCIENCES

Hearts of Darkness

Evidence grows that black holes lurk at the center of galaxies

Black holes, objects so dense that not even light can escape from their gravitational field, were once considered curiosities of relativistic theory not likely to be found in the real universe. Then in the early 1970's investigators spotted a handful of stars—the most notable is Cygnus X1—wheeling around invisible partners. After years of seeking other explanations, astronomers have concluded that the unseen partners are almost certainly black holes, the collapsed remnants of stars several times as massive as the sun that exploded into supernovas.

ANDROMEDA GALAXY and its companion M32 (circular object above Andromeda) *may have a black hole at their core.*

Recently four workers have reported that black holes millions of times as massive as the sun may serve as the hubs of two of the Milky Way's closest neighbors: the great spiral galaxy in Andromeda and its small elliptical companion, M32, which reside two million light-years from the earth. These reports, together with somewhat more circumstantial evidence that our galaxy itself contains a central black hole, suggest these fantastically compact objects may play a crucial role in the evolution of many—and perhaps all—galaxies. "That's three out of three with no losers," says Douglas O. Richstone of the University of Michigan.

Richstone and Alan Dressler of the Mount Wilson and Las Campanas Observatories report they have found signs of a 70-million-solar-mass black hole in the Andromeda galaxy and a


To Grand Expectations.
PRODUCE OF FRANCE
BOTTLED IN FRANCE
Grand Marnier
Liqueur
PARIS — FRANCE
750 ml (25.4 fl.oz) - 80° Proof
Product of France. Made with fine cognac brandy. 80 proof.
For gift delivery anywhere call 1-800-CHEER-UP (except where prohibited by law).
©1985 Carillon Importers, Ltd., Teaneck, N.J.

10-million-solar-mass black hole in M32. Their observations corroborate earlier findings on M32, by John L. Tonry of the Massachusetts Institute of Technology, and on Andromeda, by John Kormendy of the Dominion Astrophysical Observatory in Victoria, British Columbia. These objects, Richstone points out, should dwarf the black holes in binary star systems in size as well as mass: whereas Cygnus X1's dark companion probably measures only a few hundred meters across, Andromeda's black hole might be as wide as the earth's orbit around the sun.

The recent findings all derive from analyses of the Doppler shifts of light near each galaxy's center. Because the light is shifted in the direction of both longer and shorter wavelengths, indicating that the sources of the light are moving both away from and toward the earth, the workers conclude that stars are orbiting the center at high velocities. They then estimate how much invisible mass would be sufficient to keep the stars locked into these tight, rapid orbits. The technique cannot be applied to our galaxy, since dust obscures its nucleus. Still, various workers, notably Charles H. Townes of the University of California at Berkeley and Reinhard L. Genzel of the Max Planck Institute for Physics and Astrophysics in Munich, have proposed that the intense infrared and radio radiation observed emanating from our galaxy's center indicates the presence of a black hole.

Other galaxies have been suspected of harboring large black holes. About a decade ago investigators proposed that the giant elliptical galaxy M87 contains a black hole in its nucleus, but subsequent studies revealed flaws in the conclusions. Kormendy has analyzed light from galaxies beyond Andromeda and M32 and found tentative evidence of black holes in at least three of them. He and his colleagues acknowledge, however, that doing a conclusive study of galaxies beyond Andromeda and M32 with only ground-based observations is difficult. "It may not be possible until the space telescope flies," Dressler says.

Tonry thinks binary systems such as Cygnus X1 still represent the best evidence that black holes exist, since theorists seem unable to devise even farfetched alternative explanations of the data. He suggests it is conceivable—although highly unlikely—that the invisible cores of Andromeda and M32 consist of clusters of white dwarfs or neutron stars. Kormendy points out that such a cluster would probably collapse into a black hole anyway. Alternative explanations of the data from Andromeda and M32 "are getting crazier and crazier," he says. "Nature can be pathological, but it's usually not."

The findings, if confirmed, could bolster theories linking black holes to both galaxies and their oddly luminous cousins, quasars, according to Ramesh Narayan of the University of Arizona. "The picture many people like to believe," he explains, "is that a galaxy forms first and then its center collapses into something more compact"—a black hole. If the collapse of the nucleus continues, the young galaxy may pass through a stage in which the energy released by stars and other matter plunging toward the black hole outshines the rest of the galaxy; to a distant observer it would then appear not as an extended stellar cloud but as an extraordinarily intense point of light—a quasar. After the black hole consumed most of the matter within its gravitational grasp, Narayan says, the quasar might finally evolve into a less luminous, more normal-looking galaxy. —*J.H.*

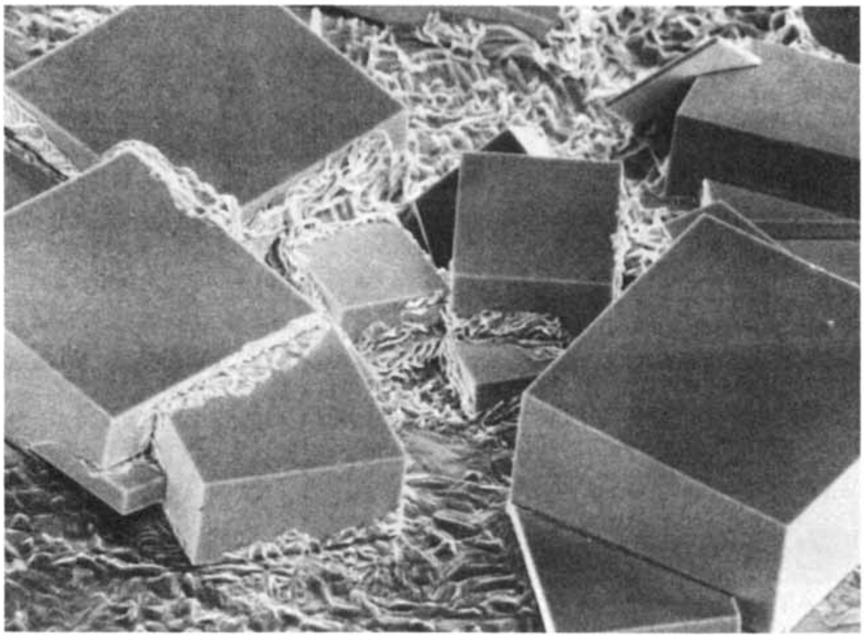
Getting Warmer

Research in superconductivity posts more remarkable advances

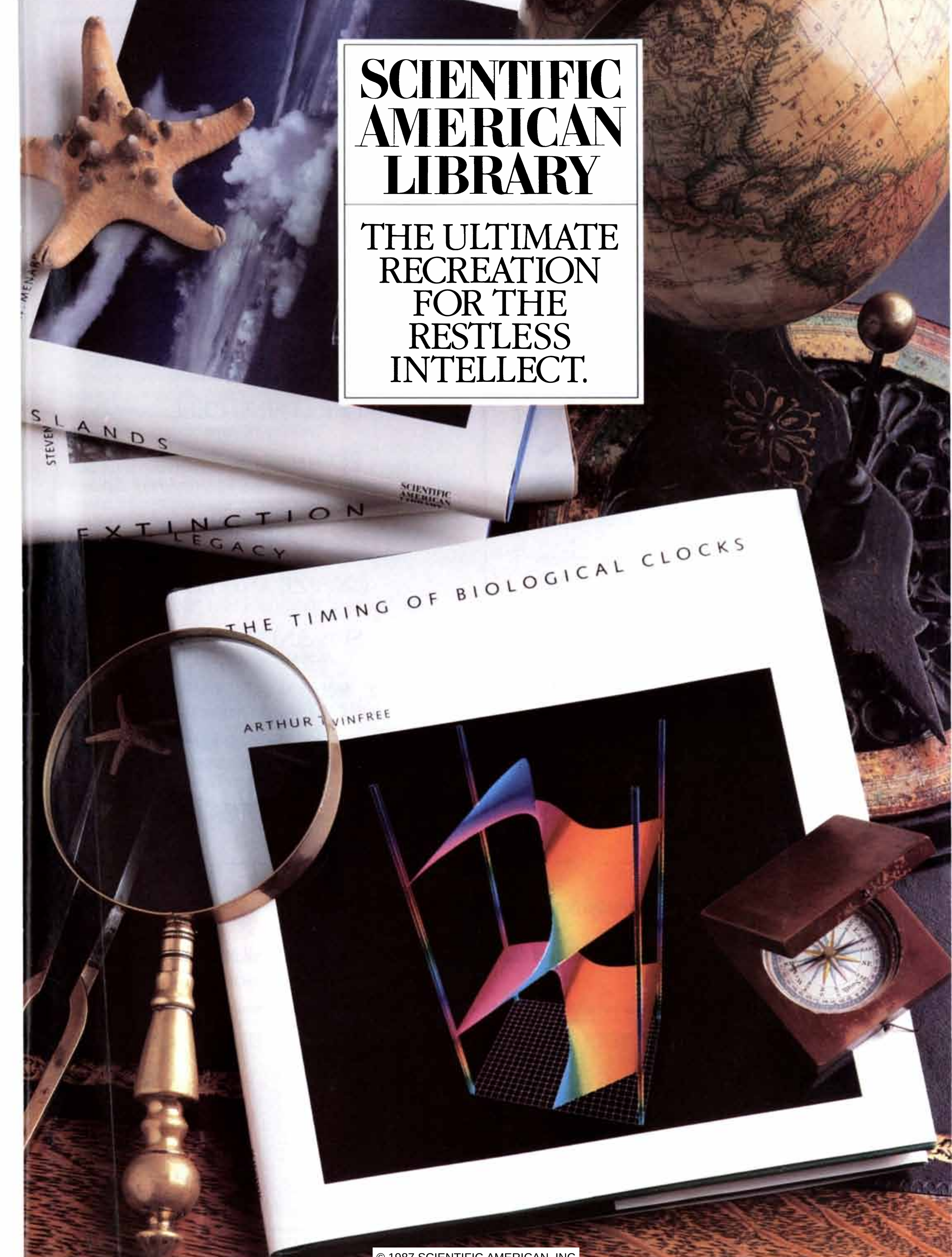
The search for practical room-temperature superconductors is heating up. At North Carolina State University, Jagdish Narayan has determined the atomic structure of a new crystalline phase of yttrium barium copper oxide, the material that has been the focus of most of the recent excitement. Narayan believes this phase becomes superconducting below 290 degrees Kelvin, a spectacular increase in temperature over the previously known phase, which superconducts only below 95 degrees. At the Lockheed Palo Alto Research Laboratory, Chao-Yuan N. Huang maintains he has detected reproducible signs of superconductivity in one compound—he would not say which—at 52 degrees Celsius, or about 151 degrees Fahrenheit.

In spite of such progress, John K. Hulm of the Westinghouse Research and Development Center, who headed a National Academy of Sciences panel on the new materials, cautions that "a whole new set of inventions" will be necessary before they can be widely used. Narayan's

By probing the new superconductors' structure workers seek to understand their strange properties

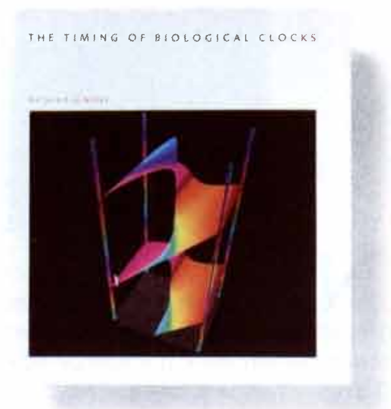


SINGLE CRYSTALS of $YBa_2Cu_3O_x$, which superconducts at up to 90 degrees Kelvin, were grown at the IBM Corporation by Debra L. Kaiser and Frederic Holtzberg.


SCIENTIFIC AMERICAN LIBRARY
THE ULTIMATE RECREATION FOR THE RESTLESS INTELLECT.
EXTINCTION
THE TIMING OF BIOLOGICAL CLOCKS

THE TIMING OF BIOLOGICAL CLOCKS

Arthur T. Winfree

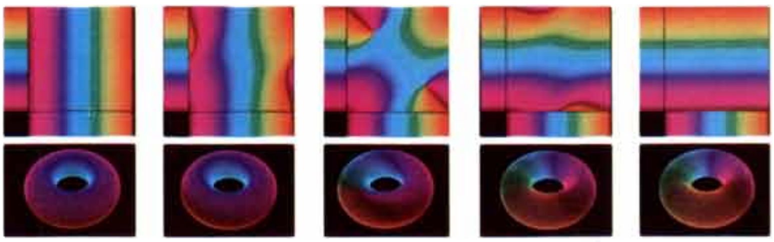


In 1931, Wiley Post flew around the world in eight days and became the first human being to experience jet lag. The fatigue and disorientation he felt happens to all of us when rapid long-distance travel knocks our "internal clocks"—our

In the early 1930s, Wiley Post used the Winnie Mae to study jet lag and its affect on pilot performance. Courtesy of the National Air & Space Museum, Smithsonian Institution.

circadian rhythms—out of kilter with local time. Arthur Winfree describes jet lag as "that disconcerting sensation of time travelers that their organs are strewn across a dozen time zones while their empty skins forge boldly into the future."

Jet lag, biorhythms, mosquito insomnia, temporal isolation experiments, the sleep movement of plants, "forbidden phases" of sleep when one cannot awake spontaneously—these are some of the many fascinating aspects of circadian rhythms that Dr. Winfree explores in *The Timing of Biological Clocks.*



Winfree's innovative use of gradient color to express the passage of time helps us visualize the biological cycles that govern the processes of life. Courtesy of Arthur Winfree.

Winfree shows that the most critical property of biological clocks is their ability to be reset on cue, enabling them to regain synchrony with a changing environment (as when we travel across

Flowering in the morning glory, as in many other plants, is timed by a circadian clock. Courtesy of Travis Amos.

time zones) or to adjust the body's 25-hour rhythm to the 24-hour solar day.

Reporting experiments on animals, plants, and single cells, he not only illustrates the principles that guide the resetting of biological clocks but reveals that each of these clocks has a vulnerable phase when a suitably intense cueing stimulus can produce a thoroughly unpredictable resetting—perhaps even annihilating the clock's rhythm entirely.

The graphics that Winfree uses are as innovative as his insights. By using gradient color rather than the conventional clock dial to express the passage of time, Winfree helps us visualize the true continuities—and discontinuities—of the internal cycles that govern the processes of life.

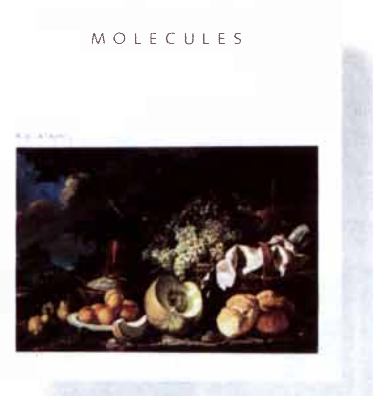
Has this cave salamander, living in temporal isolation, lost its circadian rhythmicity? Courtesy of Chip Clark.

Arthur T. Winfree is one of the world's foremost theoreticians of circadian rhythms. Trained as a biophysicist, Winfree received a MacArthur Grant for his work on biological clocks. Formerly at the Institute for Nonlinear Science at the University of California at San Diego, he is now with the Department of Ecology and Evolution at the University of Arizona at Tucson. Professor Winfree is also the author of the classic work *The Geometry of Biological Time* and *When Time Breaks Down,* a technical monograph on circadian rhythms.

199 PAGES, 233 ILLUSTRATIONS

MOLECULES

P. W. Atkins



Everything we breathe, touch, eat, drink, and smell is made of molecules. They determine the taste of a barbecue and the fragrance of a rose. Yet, to many of us, molecules are an unfamiliar, abstract notion, hidden far below the surface of visible reality.

In this highly readable work, P. W. Atkins now makes the molecular world understandable, accessible, and familiar. You will meet the molecules that make sugar sweet, that make soap "soapy," and that give rise to the brilliant palette of colors unveiled in autum foliage. You will see molecules that can quell pain and alleviate depression as well as those that can induce addiction and wreak havoc on our environment.

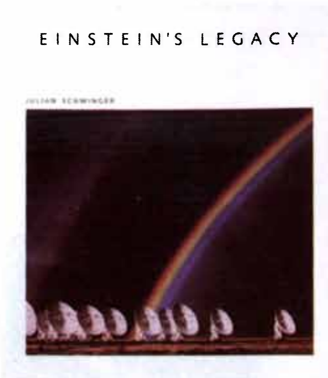
Lavishly illustrated and filled with original insights, MOLECULES portrays the structures of the substances that make up our everyday world and our daily experiences, showing us their shapes, their interrelations, why they behave as they do, and how they carry out their tasks. It gives us a more imaginative appreciation of the richness, economy, and variety of nature.

P. W. Atkins lectures in physical chemistry at the University of Oxford. He is the author of *The Second Law* (for the Scientific American Library) and the widely used textbook *Physical Chemistry,* now in its third edition.

256 PAGES, 290 ILLUSTRATIONS

EINSTEIN'S LEGACY

Julian Schwinger



Nobel laureate Julian Schwinger tells the captivating story behind one of the twentieth century's greatest achievements, the theory of relativity.

The groundwork was laid first by Isaac Newton and then by James Clerk Maxwell. But it was left to the sixteen-year-old Einstein to glimpse, for the first time, that the unifying theories of these two giants were incompatible.

In a lively and compelling narrative, Schwinger recounts Einstein's quest for the reconciliation of the conflict—a quest that ultimately led to the unification of matter and energy and of space and time in Einstein's special and general theories of relativity.

Schwinger carries this story to the space-age experiments that have been designed to confirm Einstein's legacy—and possibly lead to a still more comprehensive theory.

Julian Schwinger was awarded the Einstein Prize in 1951, the National Medal of Science in 1964, and the Nobel Prize for Physics in 1965. He is currently University Professor of the University of California.

"Delightful and instructive.... An ideal gift for the curious non-expert."—NATURE

250 PAGES, 200 ILLUSTRATIONS

"The difference [between this and other book series] comes in the quality of the

"Distinguished researchers take the trouble to talk simply and directly about the work they a

The largest forms of life reproduce at a slower rate and live much longer than smaller organisms. On Size and Life *examines why.*

A molecular view of the plasma membrane. Electro
axopod of the heliozoon Echinosphaerium nucleofi

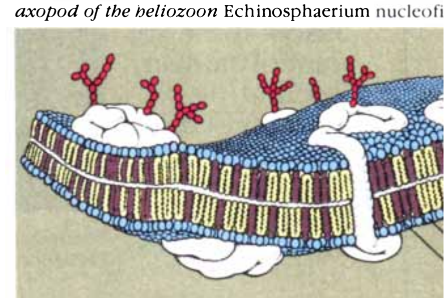
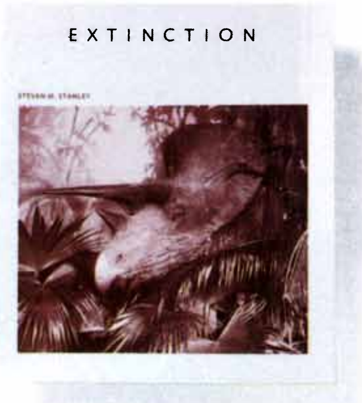


PHOTO CREDITS—Daphnia: James Bell, Photo Researchers, Inc./Hummingbird: Crawford H. Greenewalt, VIREO/Robert Wadlow, 8′ 9 1/2″ tall, and his father: Wide World Photos/Elephant: Peter Joh
Black Star/Important anatomical divisions of the brain: Painting by Bill Andrews/Coca shrub, source of cocaine: W. H. Hodge, Peter Arnold, Inc.

EXTINCTION

Steven M. Stanley



Since the rise of multicellular life, a handful of mysterious cataclysms has swept our planet. These geologically brief outbreaks of mass extinction have decimated tens of thousands of thriving species, from huge dinosaurs to microscopic algae.

In EXTINCTION, Steven M. Stanley gives us a fascinating account of these dramatic global disasters, investigating what the rich storehouse of rock and fossil clues—and the revolutionary findings of plate tectonics—can tell us about their origins, progress, and implications.

In light of recent speculation that meteor impacts may have ended the reign of the dinosaurs through an early version of "nuclear winter"—about which Stanley has much to say—EXTINCTION is a timely as well as intriguing geological detective story.

Steven M. Stanley, professor of paleobiology and director of graduate studies at Johns Hopkins University, is a Guggenheim Fellow and winner of the Schuchert Award of the Paleontological Society.

242 PAGES, 183 ILLUSTRATIONS

OTHER VOLUMES IN THE LIBRARY:

MATHEMATICS AND OPTIMAL FORM
Stefan Hildebrandt and Anthony J. Tromba

ISLANDS
H. W. Menard

PERCEPTION
Irvin Rock

DRUGS AND THE BRAIN
Solomon H. Snyder, M.D.

THE SECOND LAW
P. W. Atkins

CONSTRUCTING THE UNIVERSE
David Layzer

THE LIVING CELL
Christian de Duve

POWERS OF TEN
Philip and Phylis Morrison and the Office of Charles and Ray Eames

FIRE
John W. Lyons

ON SIZE AND LIFE
Thomas A. McMahon and John Tyler Bonner

SUN AND EARTH
Herbert Friedman

EYE, BRAIN, AND VISION
David H. Hubel (in preparation)

THE SCIENCE OF STRENGTH
J. E. Gordon (in preparation)

SAND
Raymond Siever (in preparation)

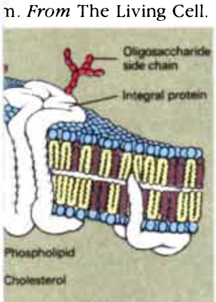
...thors, who combine eminence in their fields with proven gifts of exposition."

—Nigel Calder, NATURE

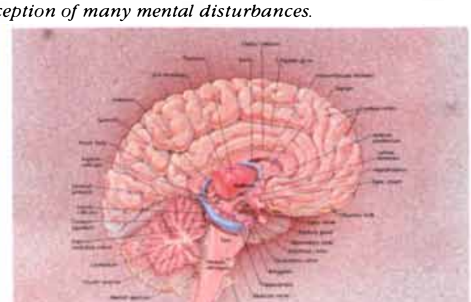
...doing and how it is solving yet another part of the mystery of the human and natural worlds."

—Steven Rose, NEW SCIENTIST

...micrograph of a section through an ...n. From The Living Cell.



Drugs and the Brain *explores the drugs that have led to a new understanding of the brain and that have changed our conception of many mental disturbances.*



...on, Natural History Photographic Agency/Plasma membrane illustration by Neil O. Hardy/Electron micrograph courtesy of Dr. L. E. Roth and Dr. Y Shigenaka/Timothy Leary: Gene Anthony,

new phase of yttrium barium copper oxide, for example, tends to be unstable in its present, impure form. Moreover, the unusual atomic structure of the new superconductors, which are ceramics, presents difficulties for those trying to use them. They consist of crystalline grains that are anisotropic: their electrical conductivity and other properties vary depending on the axis along which measurements are made. In a bulk specimen the grains all point in different directions and are not in complete contact. To achieve high current densities it may be necessary to point them all in the same direction and ensure close contact. Douglas K. Finnemore of the Department of Energy's Ames Laboratory says workers there have achieved some success using magnetic fields to line up the grains during manufacture.

Bulk specimens of the materials suffer from brittleness, and identically prepared samples vary in their superconducting behavior. Improved manufacturing techniques have increased both the maximum current and the maximum magnetic field the materials can tolerate while retaining their superconducting characteristics, but the levels fall far short of those needed in a bulk conductor.

Other limitations could pose problems in constructing hybrid semiconductor/superconductor chips, which could lead to computers of unprecedented speed. Superconducting ceramics must be heated during manufacture to about 900 degrees C., a temperature that semiconducting devices cannot tolerate. Furthermore, ceramic superconductors are susceptible to poisoning: ions from other materials diffuse into them, disrupting their superconducting behavior.

The Reagan Administration has quickly begun to seed the new field. The budget for superconductivity at the Argonne National Laboratory will increase from $1.5 million last year to about $10 million next year. The Department of Defense will spend $150 million over the next three years, and other agency budgets are slated for increases. Argonne, the Ames Laboratory and the Lawrence Berkeley Laboratory have been named as superconductivity research centers.

The Administration has also implemented policy measures intended to reinforce this country's competitive advantage. The president renewed a proposal to amend the Freedom of Information Act to limit the access of foreign companies to Federal research, and science attachés from Embassy Row were excluded from a two-day Government-industry colloquium that was held in Washington this past summer. Can such moves be effective? After all, notes Deborah C. Runkle of the American Association for the Advancement of Science, the first high-temperature superconducting ceramic was developed in Switzerland. *—T.M.B.*

BIOLOGICAL SCIENCES

Threat to the Spotted Owl

A petition to list this bird as endangered has been filed

The timber industry is a significant source of revenue in the Pacific Northwest, where annual sales of cut lumber total in the millions of dollars and from 3 to 6 percent of manufacturing jobs are associated with forestry. No one knows the exact rate of logging, but more than 900,000 acres have been cut in the past 25 years.

Much of the logging is in national forests: pristine old-growth conifer stands hundreds of years old. There independent loggers have run into a problem: the northern spotted owl, *Strix occidentalis caurina*, a raptorial bird one and a half feet tall with enormous territorial requirements. In order to hunt, mate and raise their young successfully, each breeding pair (there are approximately 2,000 in existence) needs between 1,000 and 4,000 acres of undisturbed old-growth forest, a requirement that adds up to millions of acres of prime logging country.

SPOTTED OWL lives exclusively in old-growth forests of the Pacific Northwest.

To preserve the bird and its habitat, a petition calling for placement of the owl on the Endangered Species List was filed with the U.S. Government in late July by the Sierra Club Legal Defense Fund on behalf of several leading environmental groups. The petition is a countermove to an environmental impact statement drafted by the U.S. Forest Service and scheduled for publication in 1988, which outlines management options for the spotted owl. The option favored by the Forest Service calls for 550 protected habitat areas (one per breeding pair of birds) ranging in size from 1,000 to 2,000 acres. Because 75 percent of these areas are outside prime timber regions and the plan provides for only 25 percent of the owl population, many biologists think it caters more to industry than to the spotted owl.

Some state officials believe that is the correct emphasis. According to the Oregon State Department of Forestry, "If too many acres of old growth are set aside for the owl, there will be irreversible economic damage to...Oregon as a whole." Ann N. Hanus, the state's economist, says restrictions on logging would lead to the loss of 1,740 to 4,200 jobs in that state and increase the need for social services and welfare.

Destruction of a nonrenewable resource for economic reasons is unpopular with many biologists. Rocky Gutiérrez of Humboldt State University, an expert on the spotted owl, expresses it this way: "The U.S. is one of the few countries in the world that says animals have an inherent right to exist. To ignore that commitment [in order] to allow the timber industry to make a lot of money is unfortunate." He thinks government should "focus on the negative impact logging has had on renewable resources—such as salmon fishing—that are far more profitable and less costly to harvest than timber."

"What we really need," says Whitney Tilt, the director of the endan-

S C O R P I O

SO ADVANCED, IT EVEN PLANNED FOR THE FUTURE.

There's nothing quite like it. The performance of its 2.9-liter V-6. Its luxurious interior with electrically reclining rear seats. The confident ride and handling of its fully independent suspension. And of course, the security of Anti-lock Braking.

But the new Scorpio from Germany offers advanced thinking that will never show up in a road test. It's thinking aimed at making even the day-to-day happenings of automobile ownership a rewarding experience.

SPECIAL CONSIDERATION
As an owner of Scorpio you are given the special consideration of a free loaner car at participating dealers when you come in for service.*

A HELPING HAND
You deserve something more lasting than a wave and a smile after delivery. Scorpio RoadCare,* a Ford Auto Club Program, provides everything from emergency road & tow reimbursement to custom trip routing and up to $500 emergency travel expense reimbursement. And much more.

LASTING PROTECTION
Scorpio comes with the longest warranty of any new European car in its class sold in America. Its 6-year/60,000-mile limited major component warranty lasts a full two years longer than those offered by Audi, Mercedes-Benz, Saab, Volvo and BMW. Certain restrictions and deductibles apply; ask your dealer for a copy.

A HANDSOME RETURN
Plans have even been made to help ensure a handsome return on your Scorpio at trade-in time. It's Scorpio's unique Guaranteed Resale Program.*

Scorpio. More than a driving machine. It's a total driving experience. From $24,600 to $27,100.** For more information, call 1-800-822-9292, or visit any one of 700 Lincoln-Mercury dealers specially selected to offer this advanced German sedan.

*Certain restrictions apply. See your dealer for copies of these programs.

**Price as shown—$25,653.00. Manufacturer's Suggested Retail Price excluding title, taxes and destination charges.

IMPORTED FROM GERMANY FOR SELECT LINCOLN-MERCURY DEALERS.

SCORPIO

Advancing the art of driving.

gered-species program for the National Audubon Society, "is a balance between economic interests and environmental concerns. If we reduce the issue to owls versus jobs, we lose everything." —*Laurie Burnham*

Sonic Punch

Dolphins and whales generate "bangs" that may stun prey

Stun grenades occupy a key place in the weaponry of commandos. The disorientation the weapons create can provide a decisive advantage in a surprise assault; by the same token sonic booms have been used to confuse hijackers. As so often happens, nature appears to offer a model for such innovations: recordings of dolphins and whales in the wild suggest that some of them generate intense pulses of sound that may stun fish, rendering the prey helpless.

It is well known that several species of dolphins and other toothed whales find prey by echolocation, emitting ultrasonic "clicks" in rapid succession and listening for echoes returning from objects in the water. According to Kenneth Marten of the Long Marine Laboratory at the University of California at Santa Cruz, the clicks may be "jet-engine loud" if the object is far away. Marten and his fellow worker Kenneth S. Norris wondered whether the clicks might not also disturb the sensitive lateral lines of the prey: organs in fish that detect minute movements in the water. Furthermore, several anecdotal reports describe fish as appearing to be stunned immediately before being eaten, and whale stomachs often contain fish that seem undamaged.

Yet Marten and Norris were not able to show that even very loud echolocation clicks affect prey. Recently the fish-stunning hypothesis has regained ground. Several investigators, starting with Virginia L. Cass, formerly at the La Jolla Southwest Fisheries Center of the National Marine Fisheries Service, found that wild bottle-nosed dolphins and killer whales produce banging noises while feeding. Tape recordings of the bangs show that they are much lower in frequency than clicks and so coincide with the hearing range of the prey; they are also much louder and last about 1,000 times longer.

The recordings feature ascending trills of clicks followed by what sounds exactly like a gun firing—or a stun grenade exploding. Sometimes the bangs sound like machine-gun fire. Similar noises are made by dolphins in threatening social interactions, suggesting that for a dolphin a bang might be the equivalent of bared fangs for a dog. Analysis of the sounds rules out a nonvocal source, Marten believes. Sperm whales have also been recorded making banging noises, although it is not known whether they were feeding.

Marten points out that the recordings do not prove the fish-stunning hypothesis. Bangs are not always produced when toothed whales and dolphins feed, and captive dolphins have not been heard to make the sounds (Marten speculates that they might be unbearably loud in a confined tank). He plans to investigate the effect of recorded bangs on captive prey fish. —*T.M.B.*

Eyeing Myopia

Research suggests how reading could lead to nearsightedness

Can myopia, or nearsightedness, be caused by too much reading? A long history of observations suggests that it can—including a reported increase in myopia among Eskimos after the advent of compulsory education. Research has also suggested that myopia is associated with elongation of the eye: light from distant images tends to focus in front of the retina instead of on it. But what is the link between reading and ocular elongation? A hypothesis based on recent experiments with chicks has been put forward by a team led by Josh Wallman of the City College of the City University of New York. If the peripheral regions of the retina are understimulated, the workers report in *Science*, the resulting reduced neuronal activity there may cause parts of the eye to elongate. The entire eye becomes myopic as a result.

To test the hypothesis the team obscured the vision of one eye of each recently hatched chick (the unobscured eye served as the control). In one group the entire eye was covered with a translucent plastic shield (imagine seeing through a Ping-Pong ball—perceiving light but not seeing shapes). In another group only part of the eye, and thus only part of the retina, was shielded.

Measurements made at two and six weeks of age showed that all the obscured eyes had enlarged and become myopic. Yet there were major differences between the groups. In chicks whose eyes were partially occluded, elongation occurred only in the regions that had been visually deprived. In contrast, the totally occluded eyes enlarged uniformly. Myopia, then, may be caused by the retina's localized control of eye growth rather than by processes involving the entire eye, such as focusing, Wallman suggests. Similar results were achieved when chicks' optic nerves were cut, which shows that "little patches of retina, without input from the brain," can influence eye growth.

Why is local control significant? Both human infants and chicks are hyperopic, or farsighted, at birth: the eye is too short for the optical power of the lens, and so light rays converge behind the retina. During normal development the length of the eye tends to increase, thereby moving the image plane closer to the retina. According to Wallman's theory, different retinal regions independently control this increase in length. If there is a lot of neuronal activity in a region of the retina (that is, if clear vision has been attained), growth inhibitors may slow the change. Conversely, low activity may stimulate the secretion of growth promoters.

Wallman suggests that neurons in peripheral regions of the retina are unstimulated by reading, just as the chicks' occluded eyes were unstimulated by seeing through white plastic. The neurons near the center of the retina, or fovea, Wallman says, look at small parts of the visual field, such as the dot of an *i*. The variation in letters and in the position of white space along a page provides enough stimulation to fire many nerve impulses in this region. The neurons in the retina's periphery, on the other hand, have receptive fields that are at least 10 times larger; hence they average the black type with the white background into a uniform and unchanging gray. The responses of these unstimulated neurons dwindle, that section of the eye elongates and the force pulls the rest of the eye into myopia. —*Elizabeth Collins*

Skin of Frog...

A newly revealed defense has a potent effect against microbes

Did potion-concocting witches already know something that a worker at the National Institutes of

Health has just discovered? Michael Zasloff of the NIH has identified a new class of antimicrobial compounds in the skin of the African clawed frog, *Xenopus*. The peptides, which Zasloff calls magainins (from the Hebrew word for shield), constitute a vertebrate chemical-defense system that seems totally distinct from the immune system.

Xenopus is a standard research animal, and Zasloff, chief of the human-genetics branch at the National Institute of Child Health and Human Development, has for several years routinely made incisions in female *Xenopus* to extract eggs. After surgery the wounds are sutured and the animals are returned to a tank of murky, bacteria-laden water.

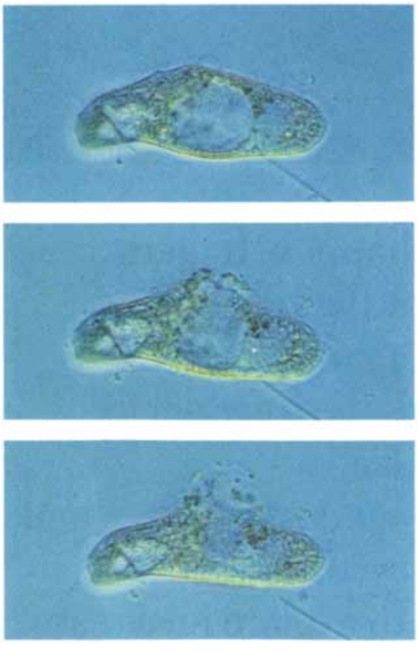
One day last year Zasloff suddenly wondered how it could be that the wounds almost always healed cleanly, showing no signs of immune-system activity such as inflammation or puss in spite of the "continual assault" from bacteria in the water. He says that as soon as the question formed in his mind he knew there had to be some undiscovered defense at work; insects, for example, also have a simple chemical defense against infection.

Identifying the defensive compound was not simple: Zasloff found no antimicrobial activity in the mucus secreted by *Xenopus* skin, and crude extracts of skin also showed little activity. But when he added substances that inhibited the action of natural peptide-destroying enzymes present in skin cells, he found potent antibacterial activity in skin extract. Zasloff eventually traced the activity to two peptides, each consisting of 23 amino acids. The peptides were similar at all but two of their amino acid sites, and they differed in their target organisms. That would have been cause enough for celebration; when Zasloff found that fungi and even hardy protozoa—including some forms of the malaria parasite—succumbed, he "couldn't believe it." His findings have been published in *Proceedings of the National Academy of Sciences.*

To clinch the case, Zasloff had the molecule synthesized and confirmed its effects. He has since characterized two different molecules from *Xenopus* that have similar properties. Zasloff thinks magaininlike compounds will also be found in other animals, including human beings.

Computer modeling by Richard J. Feldmann of the NIH suggests that magainin molecules might form helixes that are the right length to span a cell membrane. Zasloff believes a small group of magainin molecules could organize to form a pore in the membrane of an invading organism. The pore would allow ions to leak in or out of the cell, fatally disrupting its osmotic or electrochemical balance. When the common single-celled organism *Paramecium* is exposed to magainin II, its contractile vacuole, which normally empties wastes to the exterior, stops working and the protozoan bursts. Why magainins do not harm *Xenopus*' own cells is, according to Zasloff, "the number-one question."

Years of research lie ahead before it will be known whether therapeutic applications are possible, but the facts so far are not discouraging. Magainins do not harm red blood cells, and they are not very toxic to mice. NIH workers are already investigating their effects on infection of cells by the AIDS virus and also on cancer cells. —*T.M.B.*



PARAMECIUM, a protozoan, swells and bursts in the presence of magainin II.

MEDICINE

Torn Genes

Evidence mounts that defective or absent genes promote tumors

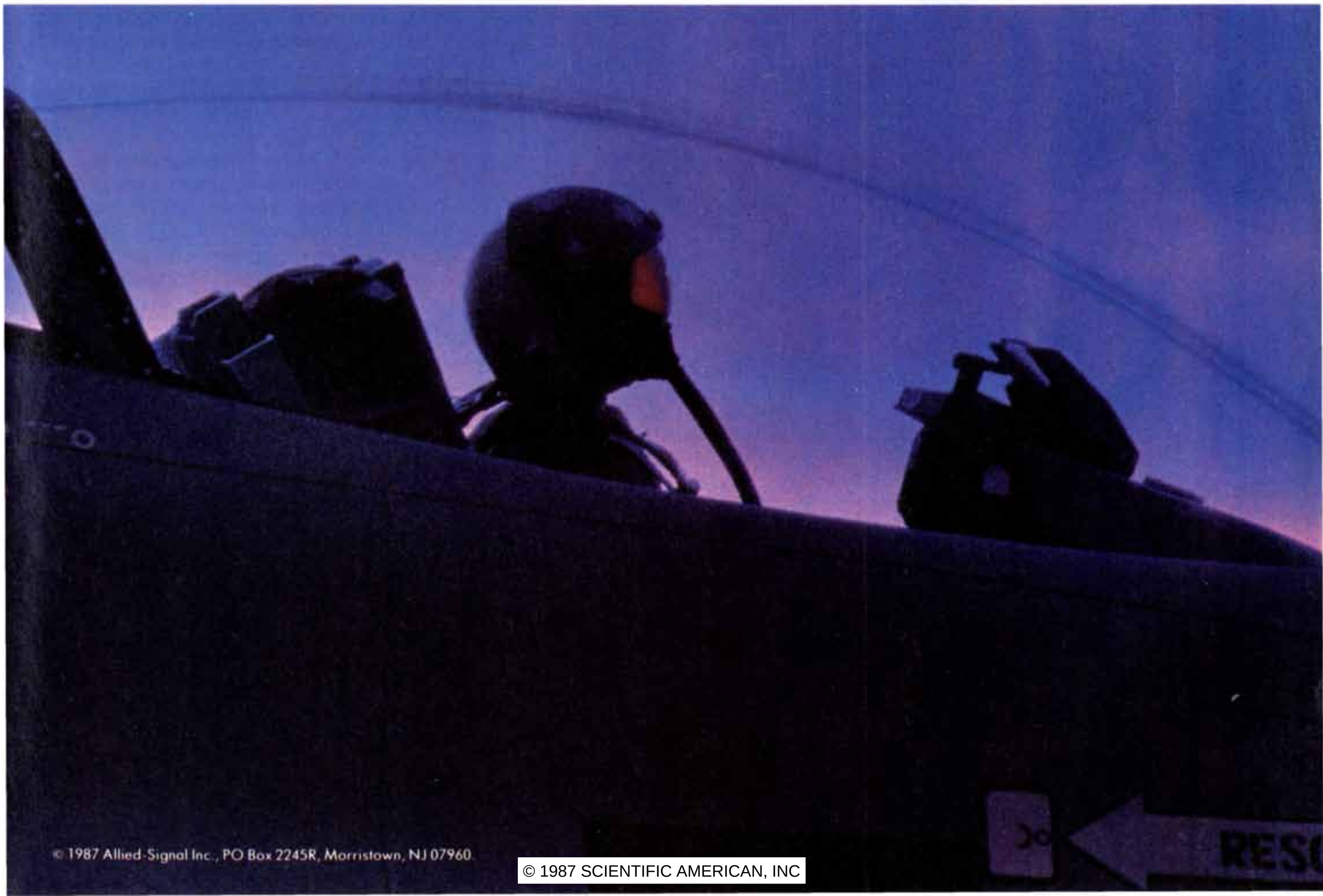
Colorectal cancer, the second-commonest cancer among citizens of developed nations, has begun to yield some of its key secrets to detailed genetic analysis.

Part of the work was carried out by a team led by Walter F. Bodmer, director of the Imperial Cancer Research Fund in London. Writing in *Nature*, the investigators describe how they tracked the gene for a rare inherited syndrome called familial adenomatous polyposis (FAP) to chromosome 5 by using a DNA "marker"—a short stretch of DNA—that binds to human DNA near the gene. The syndrome typically causes hundreds of polyps to develop in the bowel during adolescence, some of which become cancerous unless the bowel is removed. The probe can in principle be used to test DNA from people in families affected with FAP in order to detect the condition before it shows itself.

Meanwhile Ellen Solomon, also at the Imperial Cancer Research Fund, was analyzing cancer cells from ordinary colorectal tumors—that is, tumors from "sporadic" cases, which account for the great majority of colorectal cancers. Her research team found that in a substantial proportion of the tumors—at least 20 percent—DNA from the same general region of chromosome 5 as the FAP gene (and presumably including it) had been lost from one of the two copies of chromosome 5.

"It seems a very plausible explanation," Bodmer says, that this stretch of DNA on chromosome 5 might tell cells how to make some chemical that inhibits cell growth, preventing the formation of polyps. If one of the two copies of the gene is not functioning, either because it was lost during an error in cell division or because a defective gene was inherited (as in FAP), cells cannot produce enough inhibitory substance. The bowel lining grows excessively and produces a polyp, perhaps setting the stage for cancer if the other copy is also lost.

Knowing the approximate location of the FAP gene should make it possible to devise other, more precise probes that might be better predictors of FAP. They might improve the diagnosis and classification of colorectal cancer. "There is nothing at all," Bodmer says, "to stop us from finding the gene" that causes FAP and seems to be implicated in other colorectal cancers. Bodmer notes that if the hypothetical polyp inhibitor



could be identified, it might be possible to treat the disease by replacing the inhibitor or somehow compensating for its lack.

Similar results, also published in *Nature*, have been obtained with a rare inherited cancer syndrome of middle age, multiple endocrine neoplasia (MEN) type 2A. Two international teams have independently found DNA markers that bind close to a gene on chromosome 10 that apparently causes the syndrome. The probes do not yet bind close enough to the gene itself to make them particularly useful for detecting the disease in affected families. Separately, investigators in England have found that tumors from patients with MEN, like colorectal tumors, also often lack genetic material at a specific site. In this case the loss is from chromosome 1, not from the site of the gene causing the syndrome. Nancy E. Simpson of Queen's University in Kingston, Ontario, who led one of the two groups that traced the MEN2A gene to chromosome 10, says, "It's very hard to understand how that result fits in."

Both new results support the idea that loss of one copy of critical genetic information might contribute to other cancers. Jorge J. Yunis of the University of Minnesota Medical School has provided related circumstantial evidence: a wide variety of mutagens and carcinogens cause visible breaks in human chromosomes at critical "fragile sites." Chromosome breaks at these sites are seen in many tumors, although many cancer researchers are not yet convinced that such breaks are causes—as opposed to consequences—of cancer development. Perhaps coincidentally, one of the breaks appears close to the FAP gene. —*T.M.B.*

Dysaphrodisiacs

Many prescribed drugs have a side effect: sexual dysfunction

Many commonly used drugs interfere with sexual function in both men and women. *The Medical Letter on Drugs and Therapeutics*, a newsletter published in New Rochelle, N.Y., recently listed 91 drugs that have been reported to cause side effects such as loss of libido, impotence or abnormal ejaculation in men and the prevention of orgasm in women. Antihypertensives, antacids and antipsychotics are cited most often.

Tandy Computers: Because there is no better value.™

The New Tandy® 1000 TX

The most affordable 80286-powered PC compatible made in America.

Our new 1000 TX features an 8 MHz 80286 microprocessor, for far greater processing power than ordinary PCs. This brings true 16-bit technology, previously found only in "AT®" class machines to an affordable PC.

The Tandy 1000 TX is outfitted with a new 720K 3½" disk drive, and there's room to add a second internal 3½" or 5¼" disk drive.

The 1000 TX includes features you'd expect to pay extra for, like monochrome and color graphics adapters, a printer adapter, joystick adapter and an RS-232C serial port. We also include MS-DOS 3.2®, GW-BASIC—even our new integrated software: Personal DeskMate™ 2.

The Tandy 1000 TX comes with 640K RAM and five card slots for expansion. Add more memory, an internal modem, a hard disk card—or an adapter for connecting the 1000 TX to your workgroup environment.

Come to Radio Shack and see the new Tandy 1000 TX today— only $1199. (25-1600)

Send me a new 1988 computer catalog.

Mail To: Radio Shack
Dept. 88-A-81
300 One Tandy Center
Fort Worth, TX 76102

Name ____________
Address ____________
City ____________
State ____________
ZIP ____________
Phone ____________

Radio Shack®
The Technology Store™
A DIVISION OF TANDY CORPORATION

Price applies at Radio Shack Computer Centers and participating stores and dealers. Monitor sold separately. AT/Reg. TM IBM. MSDOS/Reg. TM Microsoft Corp.

How widespread is the problem? Although recognized by certain clinicians, the prevalence of drug-induced dysfunction is of "unknown magnitude," according to Jerome L. Avorn of the Harvard Medical School. Few large-scale studies have been done—in part, Avorn suggests, because much of the research on pharmaceutical side effects is sponsored by the drug industry, which is not eager to uncover this kind of adverse effect. In addition the delicacy of the topic has limited both data collection and the search for solutions, such as a switch to another drug. Physicians are often not adept at asking about sexual problems, or they assume older patients are asexual; patients are often too embarrassed to mention a problem, or they assume it has nothing to do with their drug.

The drugs at issue act in many places in the central and peripheral nervous systems. The pathways by which they affect sexual function are still unclear—at least partly because sexual function is itself poorly understood. Still, some progress has been made. Sexual dysfunction was first recognized as a significant side effect of antihypertensives about 10 years ago, according to James E. C. Walker of the University of Connecticut School of Medicine. At that time Aldomet (methyldopa), a nerve blocker used to decrease blood pressure, was found to cause impotence. This connection between antihypertensives and sexual dysfunction is not surprising. Drugs that lessen hypertension—high arterial blood pressure—are also likely to affect sexual functions that depend on engorgement with blood.

Drugs may also impair sexual function by interfering with emotional response. For example, dopamine is an important neurotransmitter. Many of the implicated drugs block dopamine receptors in the central nervous system, and dopamine has been thought to be directly involved in sexual response.

Dopamine blockage is also thought to affect sexual function indirectly by raising levels of prolactin, a hormone that stimulates milk production in women and has an unknown role in men. Dopamine normally controls prolactin levels, according to Richard Spark of Beth Israel Hospital in Boston, but in the wake of dopamine blockage the rise in prolactin depresses the so-called hypothalamic-pituitary-gonadal axis. The hypothalamus is prevented from releasing a hormone called GnRh in the pulsed fashion necessary to stimulate the release of other hormones from the pituitary. Without these hormones normal levels of testosterone are not produced in the testis. The same axis, Spark notes, controls female sexual and reproductive function—but with a different target (the ovary rather than the testis) and a different type of hormonal pulse.

What can be done about sexual dysfunction as a side effect? Because of recent advances in neuroscience, nerve transmission can be analyzed at the molecular level, making it possible to target the effects of new drugs more specifically. Physicians might be well advised to take sexual histories before giving drugs that could cause side effects. Such baseline data are necessary to separate preexisting problems from those caused by the drugs. *—E.C.*

The Bad Seed

Toxins in food may be partially responsible for brain disease

A wartime food shortage on the Pacific island of Guam has led neuroscientists, 45 years later, to establish one of the few known links between environmental factors and brain disorders. There is evidence that a rare toxin present in cycad seeds, the main staple of the Guamanian diet during World War II, produces a neurological syndrome that is known as ALS-PD. Because the syndrome combines the clinical features of amyotrophic lateral sclerosis (ALS, or Lou Gehrig's disease), Parkinsonism and a dementia associated with Alzheimer's disease, the findings lend credence to the suggestion that these diseases too may have environmental causes.

Medical investigators have long been puzzled by the disease the Chamorro people of Guam call "rayput," or lazy. ALS-PD among the Chamorro was first noted in neurological literature at the turn of the century. In the late 1940's ALS-PD death rates were as much as 100 times the rates observed for ALS in the U.S., and the mean age of onset was almost 20 years less. By 1955, however, the incidence of ALS-PD on Guam had begun to decline. As genetic and viral agents for the syndrome were ruled out, environmental influences became suspect.

The seeds of the cycad plant drew the most suspicion. Flour made from these lime-size seeds was often the only food available during the war, and the Chamorro also used the seeds in medicinal poultices. Furthermore, cycad seeds are known to contain BMAA, an unusual amino acid that resembles a proved neurotoxin in the grass pea (a relative of the sweet pea) called BOAA. BMAA became the focus of intensive research in the 1960's, but the results were seen as inconclusive. By the mid-1970's only one proponent of the cycad hypothesis remained: Leonard T. Kurland of the Mayo Clinic.

In 1980 Kurland persuaded Peter S. Spencer of the Albert Einstein College of Medicine to revive studies on BMAA. Spencer had developed a way to test in primates the relation between BOAA and lathyrism, a motor-neuron disease prevalent in India, Bangladesh and Ethiopia, where large quantities of the grass pea are consumed. He hoped to use a similar approach to test the cycad hypothesis. Beginning in 1985 Spencer and his colleagues, in collaboration with the Third World Medical Research Foundation, examined the effects of BMAA in macaque monkeys. Their results, published in *Science*, show that macaques fed BMAA exhibit ALS-PD-like symptoms within weeks of initial exposure.

It is not clear exactly how BMAA works, although it appears to block indirectly the glutamate receptor sites involved in nerve-cell regulation. It is known that the substance is a "slow" toxin: it may take 30 years or more before BMAA damage manifests itself. Chamorro who left Guam as teen-agers are still at high risk of developing ALS-PD in middle age. Fortunately both BMAA and BOAA are rare; a person who has never eaten either cycad seeds or grass peas is unlikely to come in contact with the chemicals.

Indeed, the list of the substances whose role in causing brain disease has been clearly demonstrated is quite short. Besides BMAA and BOAA, the only other celebrated member is MPTP, a deadly contaminant of designer street drugs. The list may soon get longer. Spencer expects his study to prompt a broad search for neurotoxic agents. "We are now justified in looking for environmental triggers for ALS, Parkinsonism and even Alzheimer's disease," he says, "and we need to throw a very wide net." He hopes that the benefits of such research will extend beyond the U.S., but in Guam cycad is still on the menu. *—Karen Wright*

TECHNOLOGY

Back to the Future

A new concept may restore propellers to large jet airliners

If a $2-billion gamble by the General Electric Company pays off, the engines powering some large airliners will look very different after 1992. GE is leading a race to develop a jet engine that gains extra thrust from external propellers. GE believes the engine will consume 25 percent less fuel (its main selling point) while maintaining existing levels of speed and passenger comfort.

Actually most airliner jet engines already have a "fan," or propeller, which is contained within the engine shroud. The reasons are economic. Theory says that jet propulsion is most efficient when the speed of the exhaust outflow from an engine is the same as the speed of flight. Conventional civilian airliners are designed to travel slowly compared with the maximum exhaust speeds that jet engines can achieve, and so it is advantageous to harness the jet exhaust by means of a turbine connected to a large fan instead of letting the high-speed flow escape without doing further work. As the fan bites into the airstream, it sweeps most of the air from the intake around the engine core rather than through it; this is the "turbofan" principle. To be sure, the "bypass" air moves more slowly than the jet exhaust, but it actually creates more thrust because there is more of it. Manufacturers have tended to use ever bigger fans to provide increasing bypass. Exposing the fan allows it to be bigger still, producing an "ultrahigh-bypass" engine (*see illustrations on pages 166 and 168*).

The Unducted Fan Engine, as GE calls its version, has two propellers that spin in opposite directions on the same axis; the propellers push from the back instead of pulling from the front. The fans are connected directly to turbines in the engine. As a result the blade tips travel faster than the speed of sound, producing a distinctive and annoying growl. Safety was also a concern, but the blades—which are made of a composite material—have passed the "rubber-chicken test," in which artificial birds and pieces of tire tread are shot at them in order to simulate a collision.

The noise problem has been countered by giving the propellers a twisted, scimitarlike shape. Early tests used two eight-bladed fans. The McDonnell Douglas Corporation and GE have now started flight tests of an engine equipped with a ring carrying eight blades and one carrying 10 blades. This design ensures that blade tips on different rings do not all pass each other simultaneously, which should mean less noise. The tests are being done at Edwards Air Force Base with an MD-80 aircraft on which one engine has been replaced with an Unducted Fan.

Other companies are not far behind. Pratt & Whitney, in partnership with the Allison Gas Turbine Division, has an ultrahigh-bypass engine that it calls a Propfan on the test stand; flight tests are planned for the end of the year. The Pratt & Whitney/Allison engine uses a gearbox to lower the speed of the fan blades.

The Boeing Commercial Airplane Company has announced plans to use the Unducted Fan in its 150-seat, twin-aisle 7J7 aircraft, which should be available by 1992; at that time McDonnell Douglas hopes to offer both the Unducted Fan and the Propfan in its MD-91 and MD-92. —*T.M.B.*

It's a Dirty Job

Microbes could do it: process coal and ore and control pollution

Most people would not associate modern biotechnology with grimy materials-processing or pollution-control industries. Yet those are the industries that stand to benefit by exploiting the remarkable appetite certain bacteria and fungi show for rocks and metals. Just as more familiar microorganisms are harnessed to produce antibiotics and other pharmaceuticals, so these microorganisms could be exploited to process coal, extract metals from low-grade ores and detoxify contaminated terrain. The microbes require little energy input (they generate their own energy through metabolism), and they could conceivably do the job cheaply and with a minimum of damage to the environment.

Such triumphs of bioprocessing already happen naturally. While carving out an ecological niche for themselves underground or underwater in the course of millions of years, various microorganisms have evolved the capability to ingest metals and other inorganic elements in minerals as part of their metabolic processes or detoxification mechanisms. In fact, some microorganisms thrive on nothing more than rock, water and air. Others are able to degrade minerals indirectly, through reactions caused by the by-products of their metabolism. "These tiny microbes are found all over the earth, from the tops of mountains to the bottom of seas...and they have been processing materials for about 3.5 billion years," says Gregory J. Olson of the National Bureau of Standards. The trick is to find them, characterize the process by which they transform raw materials and then find ways to promote the process.

The Electric Power Research Institute (EPRI), in conjunction with the Battelle Memorial Institute and the University of Hartford, has already found several fungi that can degrade certain types of coal. They have noted that when a common fungus responsible for the rotting of wood, *Polyporus versicolor*, feeds on lignite (brown coal), it releases an enzyme that turns the material into a water-soluble liquid having nearly the heating value of the solid lignite. The Houston Lighting and Power Company has set out on a similar quest to find a bacterium that can convert lignite directly into methane, the main constituent of natural gas.

In spite of these efforts, "the biologic conversion of coal should be considered a long-term application of the 'bugs,'" says Linda Atherton of EPRI. The likeliest near-term use is the removal of sulfur, nitrogen and metals from coal before it is burned. As Atherton points out, "there are bugs that have the ability to use each of the impurities found in coal." Cleaning coal before burning it would decrease atmospheric pollution and reduce the amount of ash.

Microorganisms can also be employed to extract copper and uranium from ores that are not worth mining by conventional techniques. The process often amounts to no more than sprinkling water on piles of ore; bacteria, which are ubiquitous in the rock, do the rest. They oxidize sulfur and iron for energy, converting the copper or uranium into a water-soluble form that is carried in the runoff. Other metals, such as zinc, lead, gallium and even gold and silver, could be processed similarly.

In addition to extracting metals from their natural ores, microorganisms could also be used to treat waste in order to reclaim valuable metals such as copper and strategic metals such as cobalt, chromium and nickel. The National Bureau of Standards is investigating the possibility in collaborative research projects with the American Iron and Steel Institute and the Office of Naval Research.

It is not always necessary to find an existing microbe with a developed "taste" for a particular element. Indeed, it has proved surprisingly easy to create new cultures of bacteria with a taste even for toxic metals by carefully selecting and cultivating strains. Whereas some bacteria readily accumulate the metals internally, others attach them to compounds that precipitate out of solution or incorporate them into volatile substances that escape into the atmosphere. Such behavior could be fruitfully applied to detoxify contaminated water or soil. —*Gregory Greenwell*

OVERVIEW

But Is It Art?

Science can both deepen and resolve issues of authenticity

John Keats wrote in *Ode on a Grecian Urn:* "'Beauty is truth, truth beauty,'—that is all / Ye know on earth, and all ye need to know."

Is it really that simple? Consider the crouching deer sold in the 1960's to the Freer Gallery of Art in Washington. The ceramic sculpture was thought to be Chinese Anyang whiteware some 3,000 years old, but a dating method called thermoluminescence analysis recently showed it was fired less than 200 years ago. "It's a fake, probably 20th century," W. Thomas Chase, the Freer's conservator, said as he drew the graceful figurine from a dusty cabinet in a back room of the museum. "Too bad," he added, gazing at it wistfully. "It's a charming piece."

Forgery is not new: Roman sculptors copied classical Greek statues and sold them as originals. Neither is the use of technology to analyze art: one of the first objects Wilhelm Roentgen examined with the X-ray machine he invented was a painting. But in recent years techniques that can probe an art object at levels inaccessible to the eye have grown more sensitive, while the demand for art has led to a rise in the number and sophistication of forgeries.

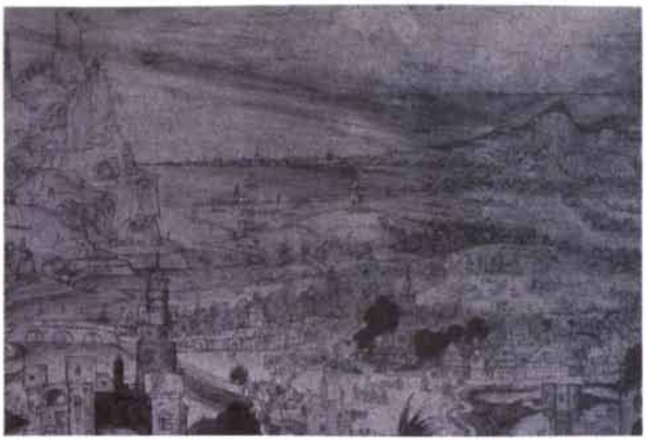
The result is that the aesthete's eye—no matter how educated and refined—is no longer the sole arbiter of art. Increasingly museums rely on scientific instruments to help them gauge the worth of objects, both those they are considering acquiring and those already in their collections. Yet even with all this combined scholarly and scientific acumen, judging the authenticity of art remains a profoundly difficult task. Earlier this year, for example, the *New York Times* reported that X-ray and chemical tests had cast doubt on a famous bronze cat in the Metropolitan Museum of Art's collection of ancient Egyptian art. Months later officials at the museum say they still cannot determine whether the cat is ancient or modern.

Analytic techniques are generally developed not to ferret out fakes but to help unravel the genesis and history of genuine works. Pieter Meyers of the Los Angeles County Museum of Art notes that "authentication is only a fringe benefit that comes from scholarly analysis of materials." For example, Maryan W. Ainsworth of the Metropolitan uses a technique called infrared reflectography in order to study the charcoal or leadpoint "underdrawing" that lies under many paintings. The painting is irradiated with heat lamps and viewed with a digital video camera equipped with an infrared filter.

Displaying an infrared image of a 16th-century Flemish landscape, Ainsworth pointed out the delicate, deftly sketched underdrawing of figures, trees and castles, some of them not visible in the painting itself. "It's like stumbling on an archive no one has used before," she said. "This expands our connoisseurship." Almost reluctantly, she added that the knowledge gained of a particular painter's style can also help a scholar to detect the anomalous underdrawing of a different painter, perhaps a forger.

The sheer number of forgeries on the market means that questions of authenticity arise quite often. Oscar W. Muscarella of the Metropolitan maintains that more than 50 percent of the Near Eastern artifacts that dealers show him are fakes, or at least wrongly attributed; many of the works eventually end up in museums or private collections.

What instruments do museums use to protect themselves from fakes? Al-



16TH-CENTURY FLEMISH LANDSCAPE, when irradiated with heat lamps, reveals a dark "underdrawing" (right) *to a video camera equipped with an infrared filter. The infrared reflectogram was made by Maryan W. Ainsworth at the Metropolitan Museum of Art.*

most all have the old standbys: microscopes and X-ray machines. Major museums have considerably more than those. The National Gallery in Washington, for example, has an X-ray fluorescence spectrometer and a gas chromatograph for analyzing respectively the elemental and chemical components of materials. Among other things, such tests can determine whether a painting thought to be of 19th-century origin contains titanium white, a pigment first manufactured in the 20th century.

Museums seek out research laboratories to gain access to still more powerful machines. The Metropolitan, for example, has analyzed the distribution of different elements—and the pigments they make up—in paintings with the help of a reactor at the Brookhaven National Laboratory. To carry out the analysis, called neutron-activation radiography, technicians place the painting in front of a nuclear reactor and allow it to absorb the thermal neutrons. Elements in the painting reemit the radiation at characteristic intervals in the form of gamma rays, which are recorded on photographic plates.

New techniques are constantly being developed. The Getty Conser-

SCIENCE/SCOPE®

Galaxy II and INTELSAT satellites are now transmitting U.S. television programming to Japan using a hookup provided by Hughes Aircraft Company. NHK, Japan's leading television broadcaster, transmits from its New York studio to Hughes' Brooklyn, New York ground station. The signal is sent to Hughes' Filmore, California ground station via Galaxy II, and then beamed to an INTELSAT satellite over the Pacific Ocean. The signal is then relayed to a Japanese ground station north of Tokyo and fed into the local NHK studio. The daily broadcasts include segments of major U.S. and European news and entertainment programs, plus live on-the-scene reports from NHK bureaus in North America and Europe. Galaxy II is one of three domestic satellites owned and operated by Hughes Communications, Inc., a subsidiary of Hughes.

Hughes is reducing software costs to customers by up to 40 percent via the comprehensive Hughes Software Engineering Environment (SE^2®). SE^2 is an integrated system of computers, workstations, communications, specialized software tools, and comprehensive training and user support. The system lets engineers and managers focus their efforts on software development, quality, and productivity, without having to supervise a complex computer facility. Operating successfully for several years on many diverse domestic and international real-time, computer-controlled projects, SE^2 provides a complete life cycle of software development and support. Hughes is continually updating SE^2, including a 1986 upgrade to comprise a total Ada® capability.

Gate array technology improves the airborne radar performance and increased operational modes for the U.S. Air Force F-15. Pioneered by Hughes, gate arrays are used extensively in the new Hughes APG-70 radar system. They provide the capability to increase the speed of the system's programmable signal processor (PSP) to more than 34 million complex operations per second, five times faster than its predecessor. Gate array technology enables the APG-70 to incorporate operating modes that were unavailable in earlier tactical airborne radars, greatly enhancing the entire F-15/APG-70 weapon system's operational capabilities. The system was developed under contract from McDonnell Douglas, which builds the F-15 for the U.S. Air Force.

Advanced composites of graphite and Kevlar have been evolving as replacements for metals. These composites have a unique advantage over metal because they do not expand or contract when exposed to heat or cold, and therefore are unaffected by the temperature extremes of space. Hughes has been using graphite composites in the construction of spacecraft since the early 1970's, including the Leasat satellites built to provide communications for the U.S. Navy. Until now, however, graphite has been difficult to mold into complex shapes, but Hughes has made it possible to produce complicated new forms. The company is developing intricate spacecraft coupling rings seven feet in diameter in which graphite composites will replace aluminum.

Hughes Ground Systems Group is applying its airspace management experience to the exciting challenges of worldwide air traffic control. These systems will be designed to ensure service 24 hours a day, 7 days a week. They will support distribution of processing among multiple computers linked via local area networks. The many challenges include design and development of hardware and software to support advanced display and man-machine interface technology, and using satellite technologies for future ATC applications. To help design the next generation of air traffic control systems, send your resume to Hughes Ground Systems Group, Employment Dept. S2, P.O. Box 4275, Fullerton, CA 92634. Equal opportunity employer. U.S. citizenship required.

Ada is a registered trademark of the U.S. Government (Ada Joint Program Office).

For more information write to: P.O. Box 45068, Los Angeles, CA 90045-0068

HUGHES
AIRCRAFT COMPANY

Subsidiary of GM Hughes Electronics

vation Institute in Marina Del Rey, Calif., for example, is studying how laser fluorescence spectroscopy can be employed in the analysis of oils and resins in paint. When excited with a laser, these materials fluoresce at wavelengths that indicate their composition. Workers are also developing a way to digitize micrographs of the surface of a sculpture and analyze them with a computer, which looks for the surface irregularities that distinguish a genuine sculpture from a fake.

Occasionally a technical examination will provide an unequivocal insight. Recently conservators at the Isabella Stewart Gardner Museum in Boston examined under an infrared camera a painting attributed to Rembrandt and spotted a signature: it belonged not to Rembrandt but to one of his students. Far more often the data are subject to widely varying interpretations. Thermoluminescence analysis illustrates this point. The technique exploits the fact that clay absorbs radiation with time, releasing it only when heated. By heating the clay and measuring the radiation it releases, workers can determine roughly when the clay was last heated. If a sculpture has been repaired and reglazed or has survived a fire, however, testing will give the date of these events rather than the original firing. In addition, knowing that testers commonly take samples from the bottom of the leg of a pot or a statue, forgers may glue an ancient leg onto a modern piece; they have also beamed X rays at ceramic objects to boost their radiation content and hence their apparent age. Only with additional analysis can technicians determine if X-ray irradiation has occurred.

Gary W. Carriveau of the National Gallery notes that most information is useful only when compared with data from tests of works known to be genuine. The data base is often deficient, he says: "In most areas we've really only scratched the surface."

The need for a technical data base is critical for authenticating modern art. Modern art is particularly easy to fake because forgers need not mimic the effects of aging. Moreover, as Eugena Ordonez of the Museum of Modern Art in New York says, "modern art is not always technically challenging." Forgers may have a more difficult time, however, reproducing a Klee or a Miró at the microscopic level. With this in mind, Ordonez is systematically studying paintings in the museum's collection with a scanning electron microscope, an X-ray-diffraction machine and a cross-polarizing microscope. Through the polarizing microscope, for example, Ordonez says she can instantly see that a rather drab-looking fleck of paint from a Monet is actually composed of many different pigments.

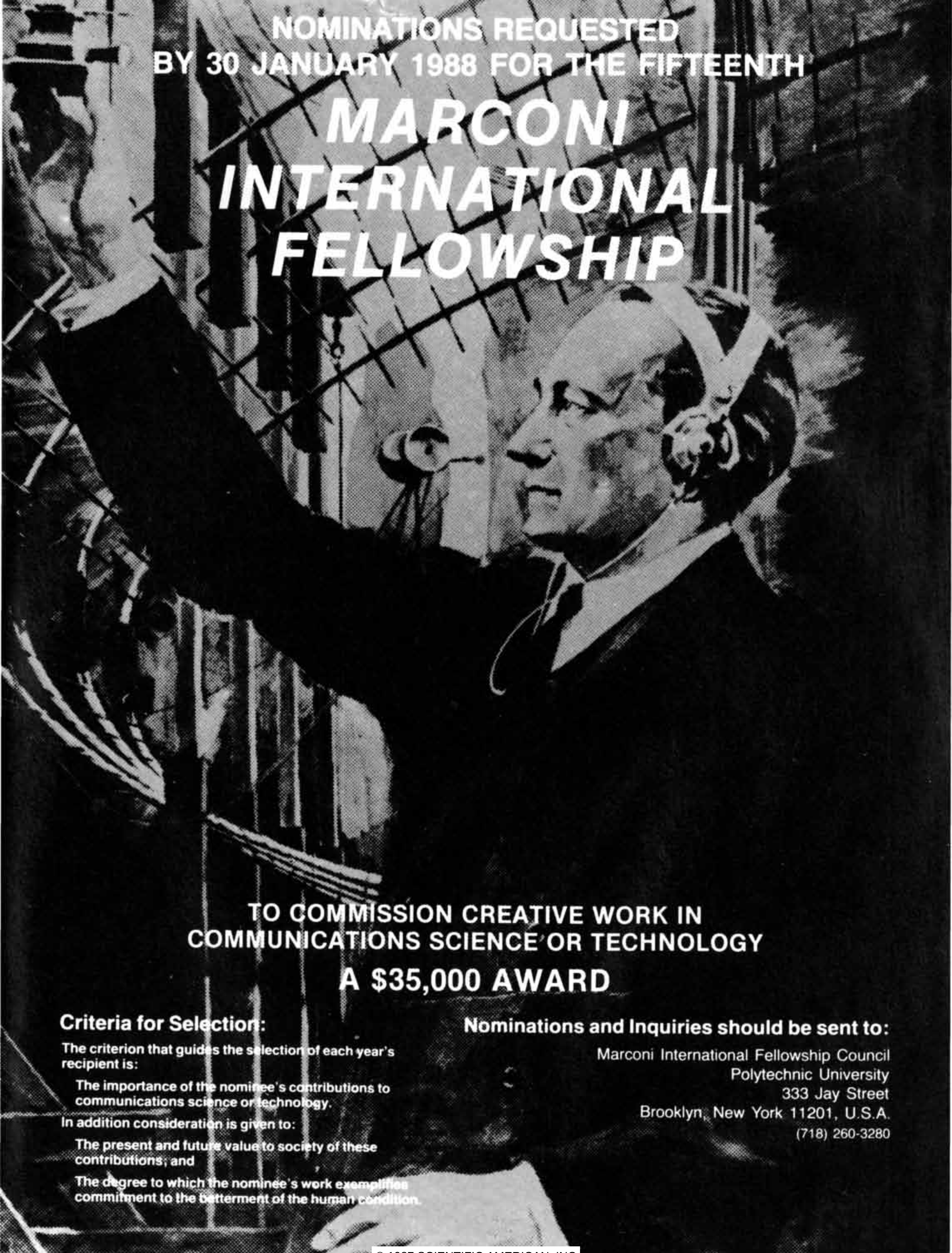
Of course, a determined forger can also learn what materials and methods an artist has used. Indeed, as the art of technical analysis advances, so does the art of forgery. "They read the same journals we do," Stuart J. Fleming of the University of Pennsylvania Museum observes. Forgers dip sculptures in acid or bury them in dung heaps to lend them an "ancient" patina. They employ lasers to map the surface of ancient coins and then micromachine this topology onto fake coins. They are also taking advantage of high-quality printing techniques to make bogus prints.

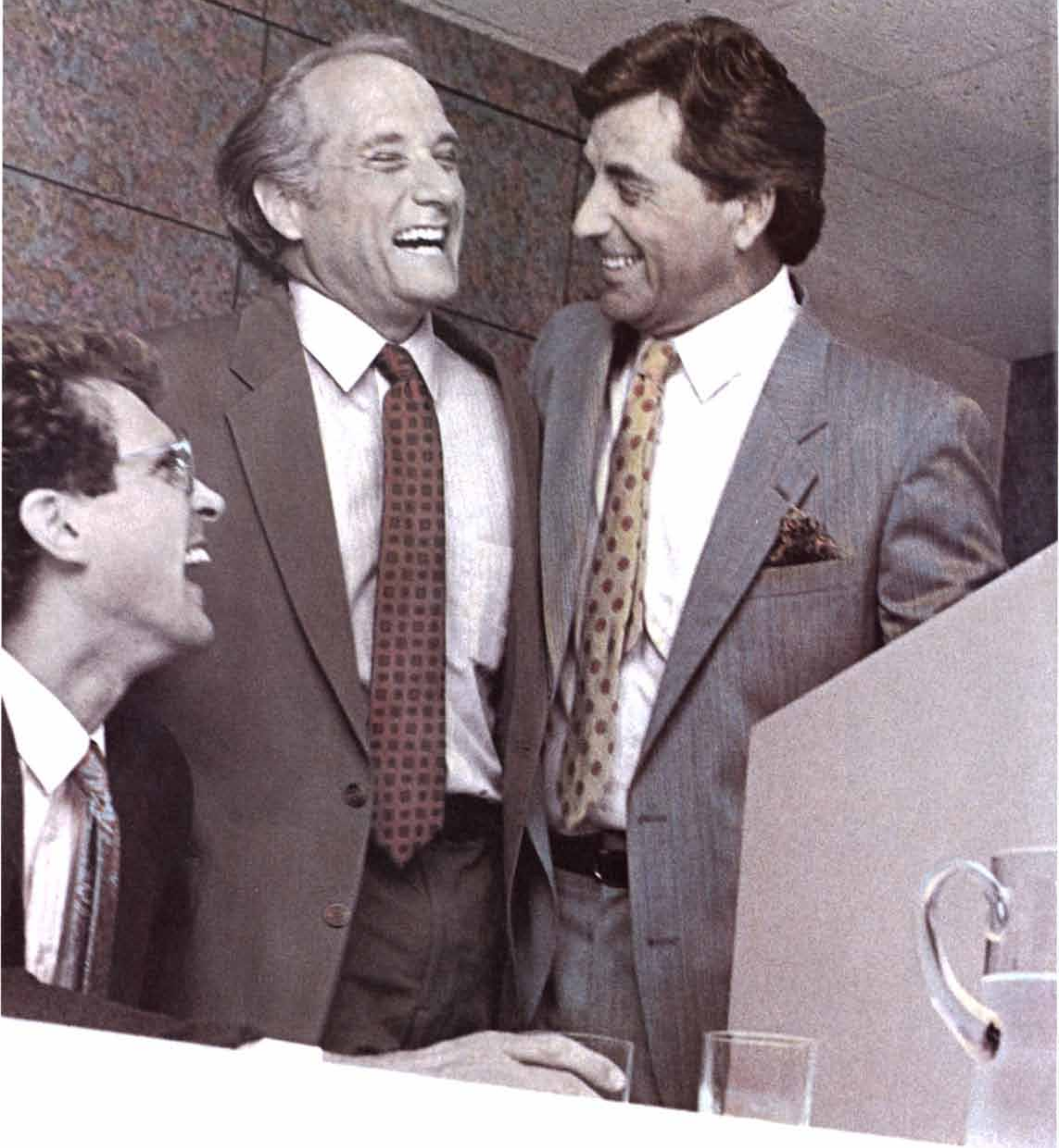
The scholar's job is further complicated, Fleming says, by the fact that there are "degrees of authenticity." Rubens, for example, is known to have started paintings and then turned them over to students to complete. "Some pieces are out-and-out fakes," Chase notes. "Others are a pastiche of genuine fragments, or have been altered later for popular tastes."

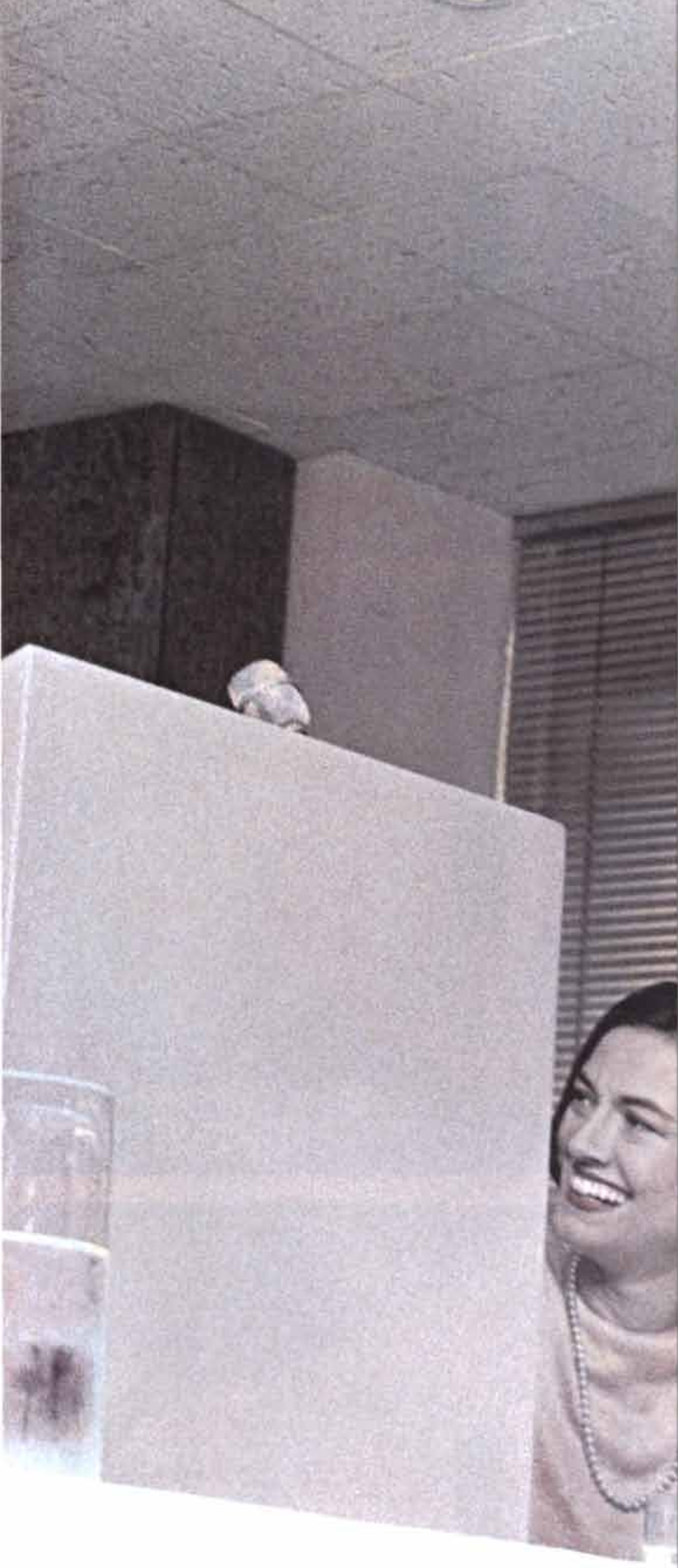
This range of possibilities is important for investigators to keep in mind, particularly when testing small samples. Robert L. Feller of Carnegie-Mellon University recalls a colleague who found a sample of zinc white, a relatively modern pigment, in an early-18th-century American painting. "He concluded the painting was inauthentic," Feller says, "but it turned out that it had been retouched." Feller adds somewhat bitterly: "He was paid nothing for his work, and he was discredited."

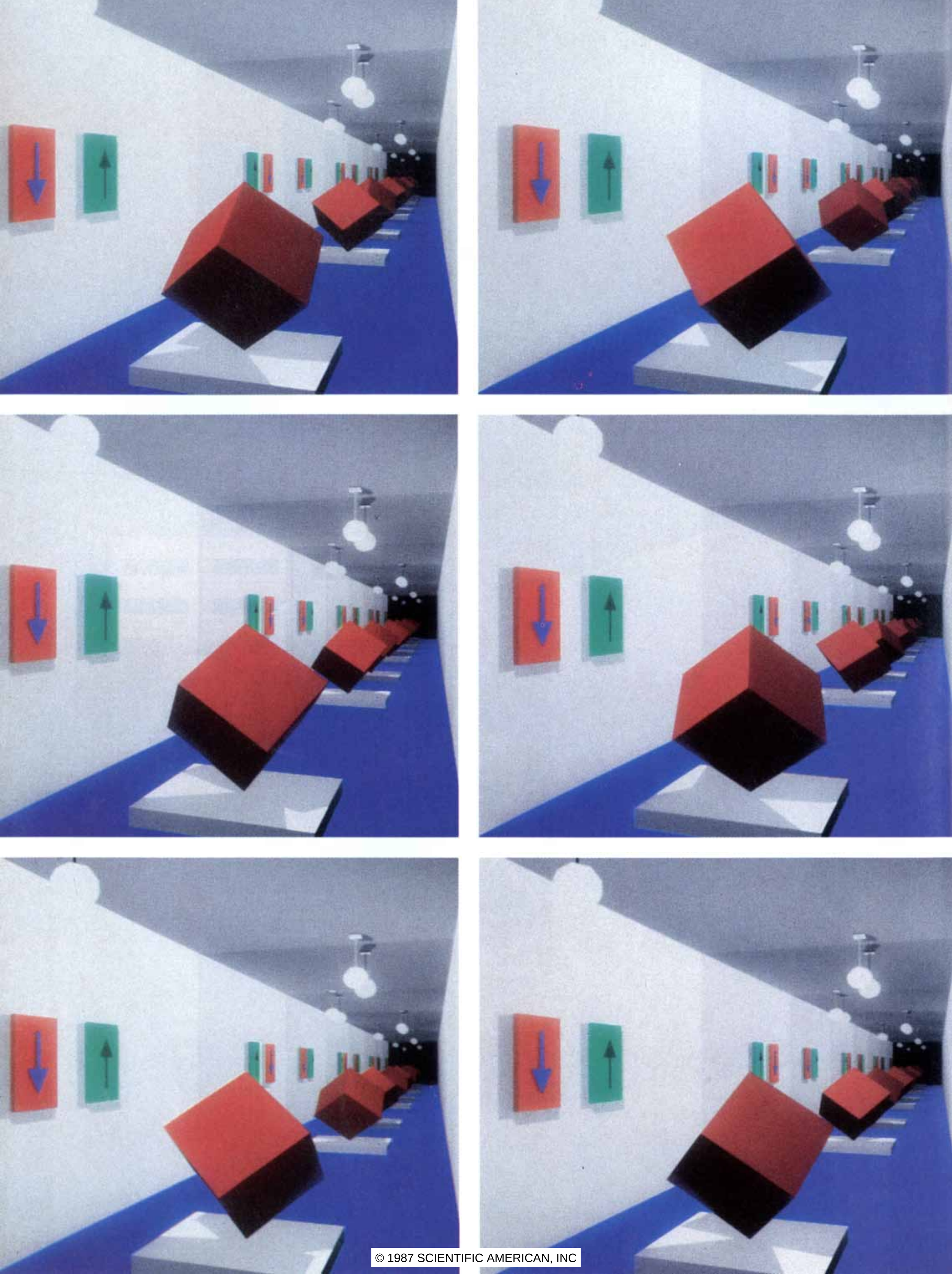
Some scholars admit their inability to rid art of its ambiguity occasionally frustrates them. James R. Druzik of the Getty Institute complains that it is almost impossible to prove the authenticity—as opposed to the inauthenticity—of any particular object: "You spend huge amounts of money proving something is authentic, and someone says, 'This hand looks funky.' It's still a stylistic issue." Others are more sanguine. "Sometimes one can only deal with probabilities," Meyers says. "We have the opinion of one individual against another individual." Carriveau adds: "There are no scientific truth machines."

Perhaps Keats was right in spite of it all. —*John Horgan*


NOMINATIONS REQUESTED
BY 30 JANUARY 1988 FOR THE FIFTEENTH
MARCONI
INTERNATIONAL
FELLOWSHIP
TO COMMISSION CREATIVE WORK IN
COMMUNICATIONS SCIENCE OR TECHNOLOGY
A $35,000 AWARD
Criteria for Selection:
The criterion that guides the selection of each year's recipient is:
The importance of the nominee's contributions to communications science or technology.
In addition consideration is given to:
The present and future value to society of these contributions, and
The degree to which the nominee's work exemplifies commitment to the betterment of the human condition.
Nominations and Inquiries should be sent to:
Marconi International Fellowship Council
Polytechnic University
333 Jay Street
Brooklyn, New York 11201, U.S.A.
(718) 260-3280







SCIENTIFIC AMERICAN October 1987 Volume 257 Number 4

The Next Computer Revolution

In less than 50 years computers have become essential to industrial society; in the next phase they will grow more powerful by at least an order of magnitude and become a ubiquitous intellectual utility

by Abraham Peled

When it was first invented, the computer was an interesting laboratory curiosity. Today it is inconceivable that contemporary industrial society could exist without it. The domestic and international financial industry, manufacturing and transportation all rely on electronic flows of information. Technologists who design materials or biologically derived pharmaceuticals depend on the computer, as do physicists who are exploring the nature of energy and matter. The way computing has permeated the fabric of purposeful intellectual and economic activity has no parallel. Computing is perhaps the most exciting technological enterprise in history. It certainly is to those of us taking part in it: imagine being engaged in a field that changes exponentially!

And now computing appears to be entering a new passage. In this phase, by means of developments in hardware and software, computing will grow more powerful, sophisticated and flexible by an order of magnitude in the next decade. At the same time the technology will become an intellectual utility, widely available, ultimately as ubiquitous as the telephone. Visual and other natural interfaces will make the machines easier to use, and a flexible high-capacity network will be capable of linking any combination of individuals who need computing, whether they are physicians trying to reach a difficult diagnosis, investment bankers structuring a deal, aeronautical engineers creating a new airframe, astrophysicists modeling the evolution of the universe or students studying for an examination.

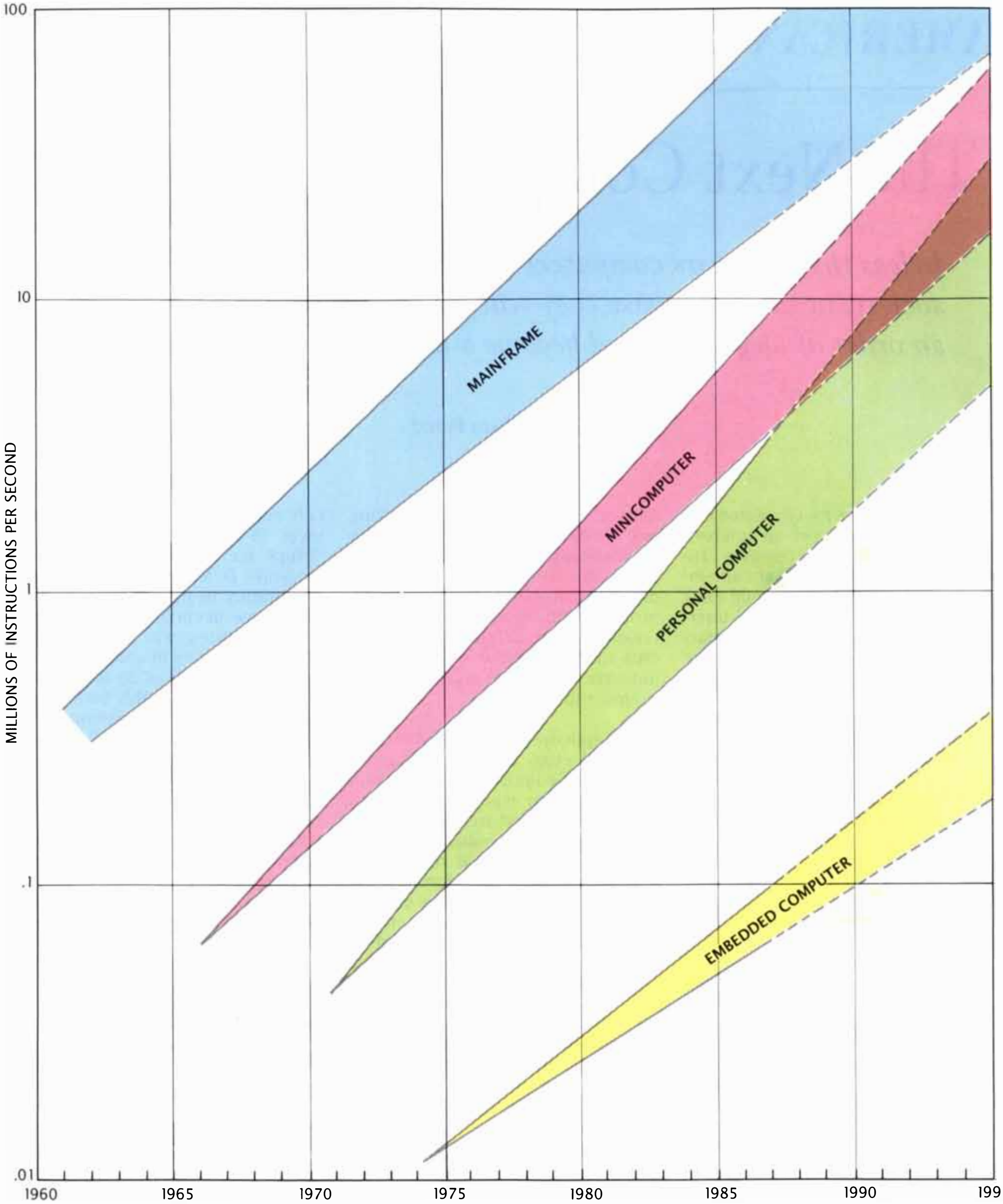
Although the emergence of such an intellectual utility represents a profound change in society's relation to the computer, it is the direct result of well-established trends that have been carrying the industry forward since the end of World War II. The most important of these forces is the miniaturization of electronic components. Miniaturization has been primarily responsible for a sustained reduction in the cost of computing at a rate of from 20 to 30 percent per year over a period of three decades. As the size of a transistor, gate or other element etched on a chip declines, the speed of operation increases proportionately and the density of the elements per area of chip rises geometrically.

The process of miniaturization can be expected to continue at its current pace for at least the next 10 or 15 years, pushed forward by difficult yet feasible engineering refinement of current technologies [see "Chips for Advanced Computing," by James D. Meindl, page 78]. X-ray lithography using synchrotron radiation, new materials and better device structures will probably improve the density of components on a chip by a factor of 20 to 40. Such processors will probably be from six to 12 times faster than existing ones.

These improvements will be compounded by the steady increase of parallelism in computing systems. Virtually all computers perform calculations in sequence—one step at a time. Parallel processing is much more powerful, as it enables a computer to operate many times faster by carrying out all or many of the steps in a problem or task simultaneously [see "Advanced Computer Architectures," by Geoffrey C. Fox and Paul C. Messina, page 66].

If miniaturization sets the pace for technological progress, the rate at which software can be developed and deployed will ultimately determine the speed with which computing systems penetrate and transform the industrial, service and scientific enterprises. The reason is that software transforms a computer from a tool that can solve a particular problem in principle to one that can solve it in practice. There is no single or even predominant software technology; progress will come from better structure, more powerful languages and more efficient programming environments [see "Programming for Advanced Computing," by David Gelernter, page 90].

ADVANCED COMPUTER DISPLAY shows an infinite recession of rotating cubes (six frames are presented to indicate motion). The images were made by software designed for SAGE (Systolic Array Graphics Engine), a specialized device that will consist of 16 million transistors. SAGE will manipulate several thousand polygons at once; it will be the prototype for single-chip devices generating many graphic displays in 10 years.



EVOLUTION OF GENERAL-PURPOSE COMPUTING during a 40-year period is charted by colored bands. Four kinds of machines are tracked: mainframes (*blue*), minicomputers (*red*), personal computers (*green*) and embedded computers (*yellow*). Each of the bands defines the range of computing power, in millions of instructions per second (MIPS), that is available from a specific kind of machine at a particular time. The dotted lines represent projections beyond 1987. In any year the computing power of mainframes is greater than that of minicomputers; the latter are more powerful than personal computers, which outperform embedded devices. Furthermore, computing power is cheaper on less powerful machines. For instance, in 1987 the approximate relative cost of executing one million instructions per second on a mainframe computer is 100 units; on a minicomputer it is 40 units, on a personal computer it is three units and on an embedded computer (if one were powerful enough) it would be .15 unit.

In the context of software the term "structure" refers to the efficiency or even elegance with which the subroutines that make up a complete program are put together to enable the machine to carry out its task. The language allows the user to tell the machine how to implement the program or some part of it to carry out a task. The power of a computer language is gauged by the degree of detail that needs to be specified. The most powerful languages enable the user simply to state the mathematical or logical formulation of the problem, in the expectation that the computer can fill in the details.

The programming environment, the array of physical and logical means by which a programmer transmits instructions, has evolved from toggle switches to the keyboard and the mouse, from flow charts and coding sheets to interactive text and graphic representations. Advances in technology promise still more natural means of communication [see "Interfaces for Advanced Computing," by James D. Foley, page 126]. A major part of the software effort is clearly directed at removing the mechanical difficulties impeding the use of a computer so that the only difficulty left will be the conceptual one of solving the problem at hand.

The rapidly falling cost per calculation, precipitated by advances in hardware and software, makes computing available in ever widening price and performance categories: the mainframe, the minicomputer, the personal computer and the embedded computer that resides in the device it controls.

Such powerful systems come into existence through continued miniaturization. The trend rides subtle improvements in lithography (the process by which element patterns are etched on a chip), process technology and fabrication techniques. Today production lithography and process technology can achieve a resolution of from 1.0 to 1.5 micrometers. Optical lithography can probably yield devices as small as .4 micrometer; X-ray lithography can take us to .1 micrometer (the human hair is 100 micrometers thick).

As the size of the devices declines, a new set of problems, called parasitic effects, begins to emerge. On a scale measured in micrometers the electrical and magnetic fields that devices and minute wires generate interfere with and even make impossible the desired operation. Yet field-effect transistor technology (the transistor technology that increasingly underlies all but the highest-performance machines) will probably continue to offer improvements in density and speed. The time that elapses between the receipt of a signal and a change in output, known as a gate delay, will be on the order of 200 picoseconds (200 trillionths of a second) by 1997; density will be on the order of from 16 to 20 million devices per chip. Memory chips offering capacities of 16 million, 64 million and even 256 million bits would be the natural consequence of such progress.

Microprocessors designed for such technology could operate at speeds of from 30 million to 60 million instructions per second (MIPS). The technology could yield a single-chip computer, including memory and input/output adaptors, capable of running at 1 or 2 MIPS; it would be the equivalent of today's personal computer. Any other design point in between is also clearly feasible.

Bipolar technology, the original semiconductor technology, which underlies the critical components of a high-performance machine, can also evolve in this direction. Gate delays of 40 picoseconds are foreseeable. Because bipolar devices consume relatively large amounts of power, heat dissipation limits density. The problem may be solved with new materials, perhaps even high-temperature superconductors.

As the speed with which a computer executes a command increases, the machine is able to utilize and produce a growing amount of information in a given period of time. This is true whether the machine is running an econometric model or generating images of a vibrating protein molecule. Therefore the ability of storage devices from which the computer draws information or into which it loads information must increase [see "Data-Storage Technologies for Advanced Computing," by Mark H. Kryder, page 116]. Indeed, magnetic-recording devices such as disks have seen great improvements in their storage capacity during the past 20 years. In 1967 their capacity was 200,000 bits per square inch; in 1987 it is about 20 million bits per square inch. Continuation of this trend could yield storage densities on the order of 300 million bytes per cubic inch within a decade (the equivalent of 300 novels). Optical storage mediums such as compact discs can provide from five to seven times the density of information that magnetic devices can achieve.

Growth in the speed with which information is generated and in its volume creates a demand for communication capacity [see "Networks for Advanced Computing," by Robert E. Kahn, page 136]. Adding force to the trend is the need to interconnect computers and their users at different locations. Fiber-optic technology is being deployed to meet the expected demand. More than 250,000 miles of fiber-optic cables are already installed in the U.S. At present only about 25 percent of their bandwidth, or capacity, is being used, primarily for voice communication. Devices for getting information into and out of fiber-optic cables at 1.7 billion bits per second are commercially available; experimental versions operate at 10 billion bits per second.

Integrated optoelectronic circuits based on gallium arsenide, a semiconducting material that can efficiently transform light into current or current into light, could offer improvements in speed and cost. Within the next decade high-speed computer networks with a capacity of 45 million bits per second will be widely available. The effect of the availability of such a system, greater by three orders of magnitude than its 1970's predecessor, will be much more than a simple scaling up. Applications involving the rapid movement of visual images will be feasible. New computer-system configurations in which distant machines will provide responsive support as if they were in the same building will be possible. Utilization of these networks will require considerable advances in communications software and new and more efficient protocols, as well as specialized computing elements that embody the protocols in hardware.

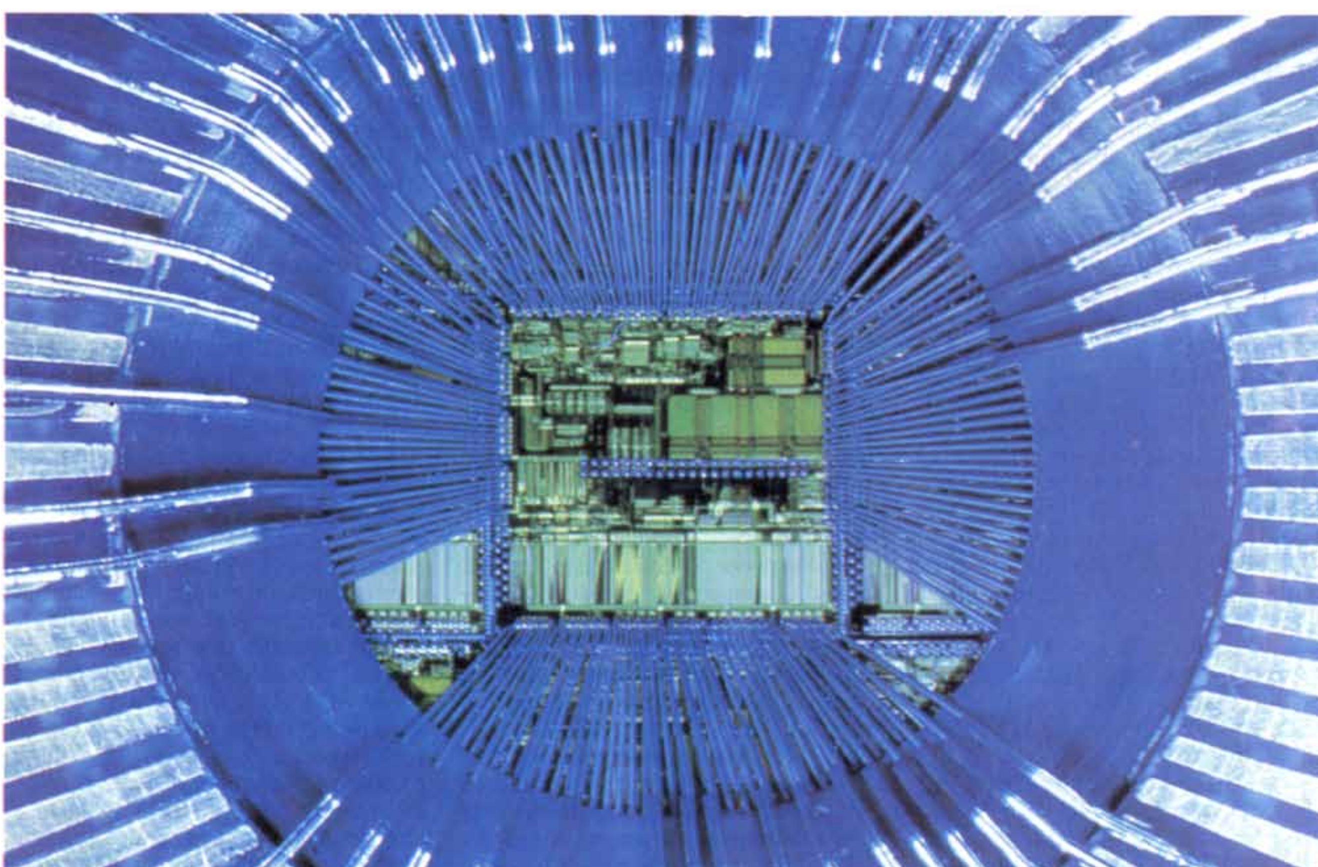
Because of improvements in semiconductor technology and in screen resolution, users of such systems and other kinds of computing facilities will be able to work with high-quality, fast displays that are reasonably priced. The displays will be capable of showing vivid, highly detailed images depicting the results of complex simulations and modeling. It will be possible to interact with the display in order to explore the impact of a change, such as the introduction of a new element into a molecular structure or a change in the camber of an airfoil. High resolution offers another advantage: a screen will be able to display text comparable in quality to that on a fine printed page. Studies at the IBM Research Division have shown that the resulting 20 to 30 percent improvement would lead to reading speeds and retention rates comparable to those associated with printed mediums.

Such devices will require the development of special graphics chips. At IBM Research, for example, we are developing the prototype of this kind of system. It is called SAGE (Systolic Array Graphics Engine). SAGE manipulates a set of several thousand polygons, producing changes in shape and animating them. Through parallel computation it generates a new image in real time as it is needed. SAGE consists of more than 16 million transistors. Someday they will all be etched on a single chip.

"I wish we could make progress on getting the right software as fast as we are building better computers." The answer to this plaint is that substantial progress has indeed been made in the ability to produce software; it is simply dwarfed by the unprecedented rate of progress in the underlying hardware and overwhelmed by the number of new possibilities resulting from that progress. Writing software is inherently difficult because it involves specifying the solution to a problem in sufficient detail to enable a computer to execute it. Choosing among many possible solutions while keeping in mind such constraints as the form in which data are received is a conceptually complex task. The complexity, coupled with the fact that successful software changes continually as it is updated and adapted, compounds the problem.

The creation and evolution of high-level languages are part of the answer to the problem of making it possible for software to realize the potential of hardware. A high-level language expresses instructions succinctly, providing relatively little detail. For example, instead of being specifically instructed to review each individual entity, a computer might be told in a language called SETL to perform a mathematical computation on a group of entities: "Calculate contributions as a percentage of gross income for all tax returns reporting gross incomes of $1,000,000 or more." Such languages are popular in the design and early prototyping of systems. Early prototyping helps to identify and solve many of the conceptual difficulties. It is a very promising way of gradually building software, and it could substantially improve productivity.

A further dimension of improvement will result from the inclusion of languages aimed at facilitating large programming projects. In such projects the challenge lies in the fact that many parts of the system must, for reasons of time and economy, be designed in parallel. The constraint focuses attention on careful interface design. Because these programs are long-lived, the language must accommodate modification and evolution. The main emphasis in this area is on provisions in the language for encapsulation of data and for procedures that allow the data to be used only in predetermined ways. Such considerations often must be traded off against the immediate need for adequate performance.



SINGLE-CHIP COMPUTER measuring 10 millimeters by 10 millimeters can operate at 350,000 instructions per second; the device is one of 87 etched on a wafer. Connectors and probes enable technicians at IBM Research to simultaneously test all the device's circuits. If any are defective, the device can be discarded when the wafer is cut up to harvest the individual computers.

Past decisions favored efficiency. As the cost of computing continues to decrease, decisions will favor structure and modularity. Finally, there will be development and exploitation of programming methods that reduce the time needed to instruct the computer. An example is PROLOG, a language that frees the programmer from the details of sequencing so that he or she can focus on the simple logical relations between program elements.

As the programmer grapples with these challenges the job will be made easier through the power of personal computers combined with easily accessible repositories containing information about programs. These highly interactive systems will provide rapid access to many related views of the program. Such interactivity can be seen in the Cornell Program Synthesizer, where early work showed the effectiveness of a set of language-based tools for learning program skills. More recently the Garden Project at Brown University has mobilized the graphics capabilities of workstations to present the user with multiple views of the program.

Expert systems are both a tool and a product of software technology. Creating software is perhaps the area in which the systems are most extensively used. A typical expert system consists of application logic, which is a set of rules (if *A* and *B*, do *C*), and a general-purpose program—the inference engine—which invokes the appropriate rules for achieving a specified goal and then executes them. This style of programming frees the programmer from the details of data structures and flow of control (detailed specification of the order in which instructions are to be executed). It is also possible to add rules at any time and in any order. This makes the systems particularly suitable for rapid prototyping and the gradual building of the application. Medicine is an area in which intensive work is pushing the technology closer to usefulness [see "Advanced Computing for Medicine," by Glenn D. Rennels and Edward H. Shortliffe, page 154].

Although software appears to lag behind hardware, a retrospective look at the evolution of the computer shows that the two technologies have produced an expanding array of computing capability and that the trend shows every sign of vigorous growth [*see illustration on page 58*]. There has been a steady increase in computing power and a widening range of it available within major price brackets. Technology has made possible the production of considerably smaller but "useful" computers at a substantially lower cost per millions of instructions per second, as well as at a lower total system price. Consequently new price categories have emerged: the mainframe has been followed by minicomputers, personal computers and embedded computers. The shaping of these

THE ECONOMICS OF SYSTEMS DESIGN

In order to buy a system capable of performing a task at the lowest cost, the bank information systems manager described in the text could employ the following equations.

(1) $P_1 = M \times C(M)$ where

P_1 = Price of executing the task on one processor
M = Millions of instructions per second (MIPS) required to execute the task
C = Unit price in \$/MIPS for the specified peak MIPS

(2) $P_N = (e \times M) \times C\left(e \times \frac{M}{N}\right)$ where

P_N = Price of executing the task on some number of slower processors
N = The number of slower processors
e = Expansion factor, the increase in required work produced by dividing the task among several processors

The multiprocessor option is preferable if:

(3) $e < \frac{C(M)}{C(e \times M / N)}$

With $N = 10$ and $e = 2$, the multiprocessor option is better if:

(4) $\frac{C(M)}{C(M/5)} > 2$

A more complete breakdown of the bank's costs is:

(5) $P_N = (k \times M + m \times I_m) \times C\left(\frac{k \times M + m \times I_m}{N}\right) + m \times B$ where

k = MIPS expansion within the given application
m = Number of messages to be exchanged between processors
I_m = Millions of instructions required per average message
B = Price of communication per average message

With $N = 10$, $m = 100$, $k = 2$, $I_m / M = 0.005$ and $m \times B / C(M) = 0.1$, the multiprocessor option is better if:

(6) $\frac{C(M)}{C(M/4)} > 2.8$

If B increases five times, the requirements change to:

(7) $\frac{C(M)}{C(M/4)} > 5$

If I_m is reduced 10 times, the requirements change to:

(8) $\frac{C(M)}{C(M/4.87)} > 2.27$

If $m \times B/C(M) \geq 1$, it is *never* preferable to divide the task between several processors.

kinds of computing machines is a natural result of technology and economics. Denser circuitry consuming less power, associated with less expensive packaging and cooling, makes new design points possible. The increased demand for computing at the lower price generates the additional economies of scale that are the result of volume production.

It is important to note that the smaller systems generally have proportionately less memory, disk and input/output capacity; therefore they are simply not able to execute many large programs and applications. On the other hand, users who have tasks that can run on any of these systems clearly have an increasing number of choices; new applications—all other considerations being equal—will tend to show up in their lowest feasible category. As the peak MIPS for each category increases, the total number of applications that are economically justified increases much faster, since it is the sum of all applications justified at each level.

Experience to date has shown that single processors (or a small number of tightly coupled processors) are most versatile; furthermore, their capacity, throughput (actual processing capacity) and response time are predictable and well behaved for a wide class of applications. The performance, throughput and response time of any multiprocessor configuration depend heavily on the application assigned to it.

For example, a bank-information-systems manager may need to buy a system that has the capacity to operate at 100 transactions per second in order to track withdrawals from an array of automatic teller machines. He or she might buy a single 50-MIPS machine to attain a response time below one second. It is perfectly conceivable that 10 5-MIPS machines might be cheaper, but the calculation is not yet complete. The manager must reckon with the expansion factor: the cost of the hardware and software needed to coordinate the activity of the 10 computers. When the expansion factor is taken into account, the single 50-MIPS machine could be more economic [*see box on preceding page*].

It is also possible that the cost of communication between individual computers will be so great that it equals or exceeds the cost of a single computer having a capacity equal to that of all the individual machines. The point is that as computing systems become available in a growing number of forms the choice will depend largely on data-access and data-sharing patterns instead of on the raw cost of computing.

Such considerations operate in decisions about systems as diverse as those that undergird the operations of a major corporation and those that amplify the creativity of an individual worker. In a large corporation more than 100,000 terminal nodes may be generating 1,000 simultaneous requests. Such numbers increase by an order of magnitude if interenterprise transactions are being handled.

The choice of system will depend heavily on the characteristics of the application. A system for executing commercial transactions may require a one-second response time. Therefore at a given volume of transactions and price the most economic system might consist of four processors, each capable of executing 18.5 MIPS. Geographic distribution of such a system always imposes a penalty. The response time rises in proportion to the number of messages associated with each transaction. Even if very high bandwidth eliminates communication delay, computing overhead (instructions that need to be carried out in each computer to send and receive a message) incurred by the exchange of messages will slow response time. Massive worldwide systems will generate a number of formidable challenges: software distribution, maintenance and updating, as well as the problems that arise from continuous operation and those associated with interenterprise networking. The complexity that is encountered in scaling up will require new algorithms, data structures and concepts.

In contrast to the enterprise-wide systems in which the essence is shared data, the essence of individual computing systems is power. Such systems are designed to help an individual or small group of professional workers by applying computer power to a problem. Their roots are in time-sharing and interactive computing on mainframes, but they have grown explosively since the advent of the personal computer. They promise to produce the most fundamental changes in the nature of computing, making it the universal extender of human intelligence.

A typical personal computer in the next decade may have an order of magnitude more computing power than today's typical machine, and from 10 to 100 times more storage. Its high-quality, high-speed screen will, in less than one second, reproduce a new image that consists of from one million to four million pixels in from 64 to 256 shades of color.

A considerable fraction of the computer's power will go to making the machine easier to use by accommodating a natural human-machine interaction based, for example, on voice or handwriting. Workers at IBM Research have developed a system capable of recognizing 20,000 words if they were spoken with brief pauses between words. The system comprises 60 million instructions, which activate four specialized microprocessors and an IBM PC/AT. Four years ago such a computer would have occupied a room; in five years it will probably take up less space than a card. Continuous speech recognition would require 30 times more computing power, and additional storage and computing would be needed to improve accuracy and reduce sensitivity to ambient noise.

A paperlike terminal allows the user to "write" on a flat liquid-crystal display; the computer recognizes the characters and translates them into commands, text or drawings. The prototype system requires from 2 to 4 MIPS; even the improved algorithms necessary will consume only a small fraction of the processing capacity of future personal computers.

In addition to accommodating recognition of speech and writing, the power of the personal computer will increasingly be exploited to display the results of computation in a manner that is visually intuitive. Pie charts and graphs will be augmented by 3D images presented in many shades of color. These powerful interactive systems will mean that the personal computer can serve as the natural front end of any computing system—the human window opening on a large network or on a supercomputer capable of depicting in a matter of seconds the results of hours of calculations.

The availability of individual computing capacity of great power and flexibility will enable engineers to model such complex phenomena as the behavior and efficiency of an airframe under various flight conditions [see "Advanced Computing for Manufacturing," by Albert M. Erisman and Kenneth W. Neves, page 162]. Often the insights are more penetrating (and are gained at a much lower cost) than those of actual simulation or experiment. The ability to model performance and appearance electronically can also benefit architects, interior decorators and theatrical-set designers. Such computer modeling will accelerate evolution from idea to product. The new insights gained could lead to totally new products or design concepts. In the same way electronic simulation can enable a research scientist to do experiments that would be impossible in nature, such as determining what happens when galaxies collide [see "Advanced Computing for Science," by Piet Hut and Gerald Jay Sussman, page 144].

Where the computation will be carried out depends on its size and the required response time. An engineer modifying the shape of an automobile body may need to invoke 60 million instructions to compute the result and an additional four million instructions to display the vehicle from a particular angle. On a personal computer running at 1 MIPS the job could be done in 64 seconds. On a 20-MIPS computer, to which the personal computer is joined by a high-speed link, the display takes four seconds; the 60 seconds needed for the computation is cut to three. The tradeoff is between the individual's time and the cost of computing. As computing costs decline, the devices will be dedicated to improving the individual's productivity.

Another implicit tradeoff is between modeling the car body or airframe on the computer and constructing an experimental version for laboratory testing. As the massive computing power needed to simulate reality becomes available it will be used for all but the final stages of the design-to-production cycle. Experience shows that computer simulation is much more flexible and economic than prototyping. The result is invariably a better design for anything from the keel of an America's Cup competitor to a faster computer.

A remarkable outcome of the exponentially rising density of components and the deepening sophistication of software is the parallel processor. Certainly the uniprocessor, or conventional computer, will be the principal engine of computation for some time to come. Uniprocessors are versatile and well behaved. Economic motivation coupled with advances in computer science, however, will create many useful applications for parallel processors (and for distributed machines as well).

The quest is definitely for the highest absolute performance. Virtually every pattern of interconnecting computational elements is being explored, from thousands of relatively slow, inexpensive elements to tens of the highest-speed uniprocessors that can be built. Very-high-bandwidth interconnections are essential to allow data and instructions to move among processors rapidly so that idle time is minimized. Small numbers of very-high-speed processors can be joined by optoelectronic circuits driving optical fibers. A large number of processors must be connected by a switching fabric (such as a dense array of field-effect transistors) that can establish connections between any two processors the way a telephone network connects two telephones. A

RP3, an IBM parallel computer test-bed, simulates various kinds of parallel architectures and challenges them with specific kinds of tasks. The effectiveness of an architecture for a particular task can be determined and the cost of a prototype avoided.

network of direct connections between all the processors would be impossible to build. Switching fabrics, as one might imagine, are objects of intense inquiry. An intriguing model posits interconnections patterned after the mammalian brain.

Performance-analysis tools, well developed for uniprocessors, are still rudimentary for parallel processors. At IBM Research investigators are experimenting by interconnecting small numbers of high-speed processors and building highly parallel prototypes. One such device is RP3, a parallel processor whose flexible organization allows it to serve as a test-bed for various kinds of parallel architectures. The RP3 is developed in collaboration with New York University and is partially supported by the Defense Advanced Research Projects Agency. The computer will be instrumented to enable investigators to gather information in real time about how various kinds of parallel machines are able to handle a wide range of applications.

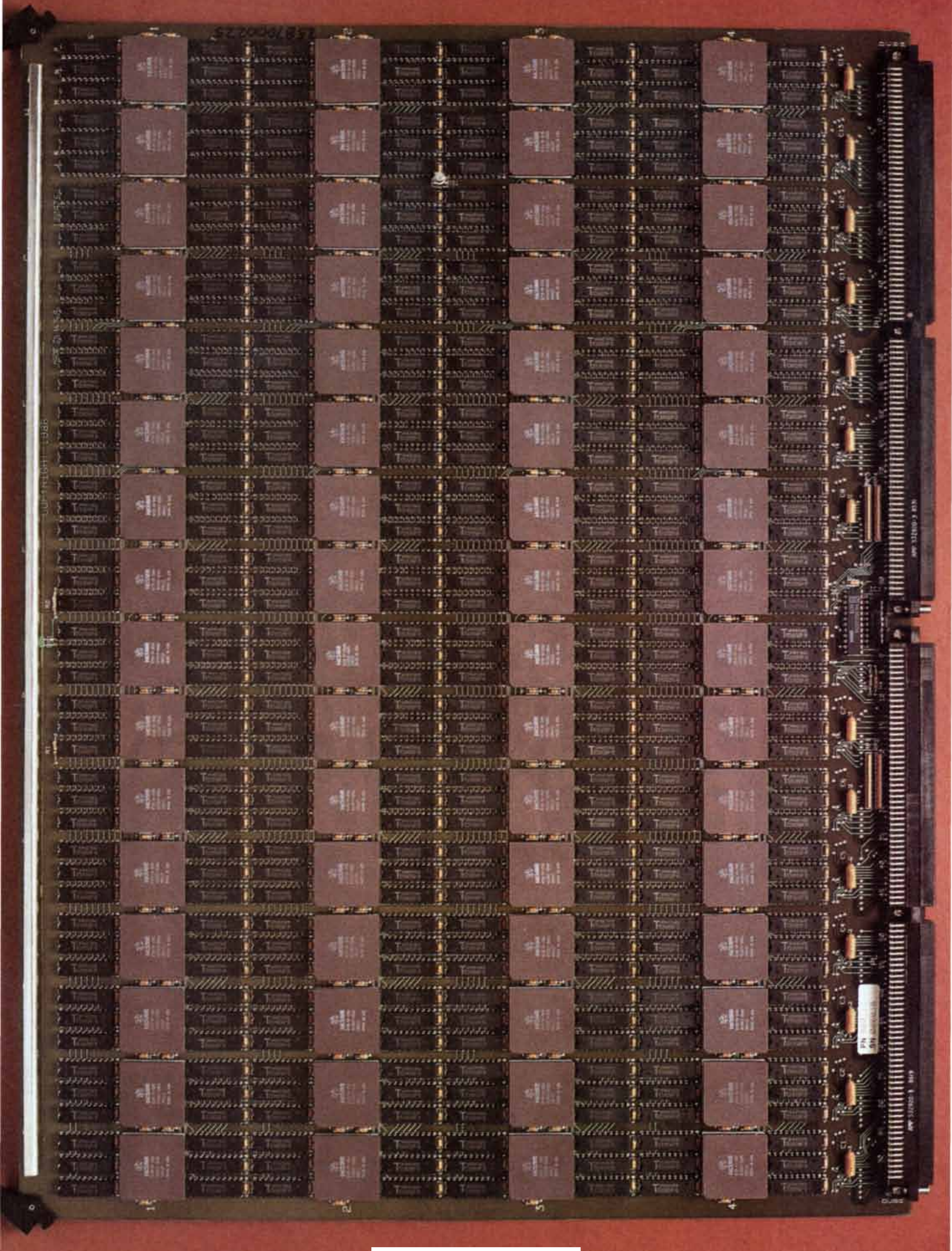
Another type of machine, also at IBM Research, is YSE, a parallel processor that simulates circuitry at the component, or switch-and-gate, level rather than at the level of architecture. It is a highly parallel processor able to execute a simulation several hundred times faster than any uniprocessor. YSE makes it possible to test and debug a number of designs quickly without incurring the expense of a prototype.

The trend toward building highly specialized computers will accelerate as increasingly powerful microprocessors become available. For example, at IBM Research scientists are building Gigaflop 11, a machine capable of executing 11 gigaflops (floating-point operations) per second. A switching fabric called a permutation network moves data among the 576 processors executing a common instruction according to a predetermined plan. GF11's first challenge will be to calculate the mass of the proton based on quantum chromodynamics, the dominant theory that attempts to describe the ultimate structure of matter. The calculation, which requires about 10^{17} floating-point operations for an accuracy of 10 percent, would occupy 15 years of a typical supercomputer's time. GF11 should be able to do the job in four months.

The progress in computing systems will continue—perhaps exponentially and certainly unabated—for at least the next 10 or 15 years. The widespread availability of computers to a growing community of users will amplify creativity and fuel the continued progress.

Currently computing can amplify only simple, relatively routine mental capacities, but steady progress is being made toward an ability to enhance the more analytical and inferential skills. Just as machines capable of extending and amplifying human physical abilities created the Industrial Revolution, so computing—through its ability to extend man's mental abilities—is the engine propelling the current and as yet inadequately named revolution. The journey has only begun.

GF11, a scientific parallel processor capable of 11 billion floating-point operations per second, is being built by IBM Research. Its first assignment will be to derive the proton's mass in four months. On a supercomputer the task would require 15 years.



Advanced Computer Architectures

Conventional computers attack problems one step at a time. Advanced computers are able to do several parts of the problem at once, just as a team of workmen might divide a task to complete it more quickly

by Geoffrey C. Fox and Paul C. Messina

Suppose you were overseeing the construction of a house and you chose to have the entire job done by a single worker. The worker would approach the project sequentially: he or she would complete each of the tasks involved (bricklaying, plumbing, installing wiring) one step at a time, doing the various parts of each job in a certain order.

This seems an unnecessarily slow way to build a house. Many of the tasks, such as the bricklaying, could be completed much faster if they were split up among several laborers all working at the same time; some tasks, such as wiring and plumbing, are independent of each other and could be done concurrently by separate teams of workers. Yet the slow, sequential method—one worker doing the job one piece at a time—is a good analogy to the way most modern computers work.

Most computers have a single processor, which is the unit that does computational work such as addition, multiplication or the comparison of two numbers. Human programmers divide each computational task into a sequence of steps—a program—and the computer's processor executes the sequence one step at a time. This approach is inherently slow for two reasons. First, during each phase of a computation much of the processor is idle; the procedure for multiplying two numbers, for example, requires several steps, and the circuitry that can perform one step may be idle while another step is being executed. Second, like the tasks involved in building a house, many computational tasks can be divided into subtasks that are independent of one another and so could be done concurrently by separate processors.

Designers of advanced computer architectures are developing approaches that overcome both sources of slowness. On the one hand, they are devising ways to increase the speed of single processors by keeping more of a processor's components active. On the other hand, they are designing systems in which many processors are linked together to form what are called parallel processors or parallel computers. In both approaches the aim is to have many computational steps taking place at any given time.

How might one improve the speed of a single processor? A major source of slowness in conventional processors involves access to memory. If data or instructions need to be fetched from a memory bank before a certain computational operation can take place, all the processor's functional units (the logic circuits that perform the individual steps of computations) must remain idle while the memory cycles. At the end of the operation the result of a computation may itself have to be stored in memory, leading to another period during which no functional unit is active. The solution to this problem has been to design machines in which, while one operation is being executed, the next set of instructions to be followed is fetched from memory and decoded, or divided into operations that can be done by the various functional units.

Access to memory can also cause bottlenecks when more than one operand needs to be fetched from memory before an operation can be executed—for example, when two numbers that are stored in memory are to be multiplied; the second operand cannot be fetched while the channel to memory is occupied in fetching the first one. How can one design a memory so that more than one datum can be called up at a time?

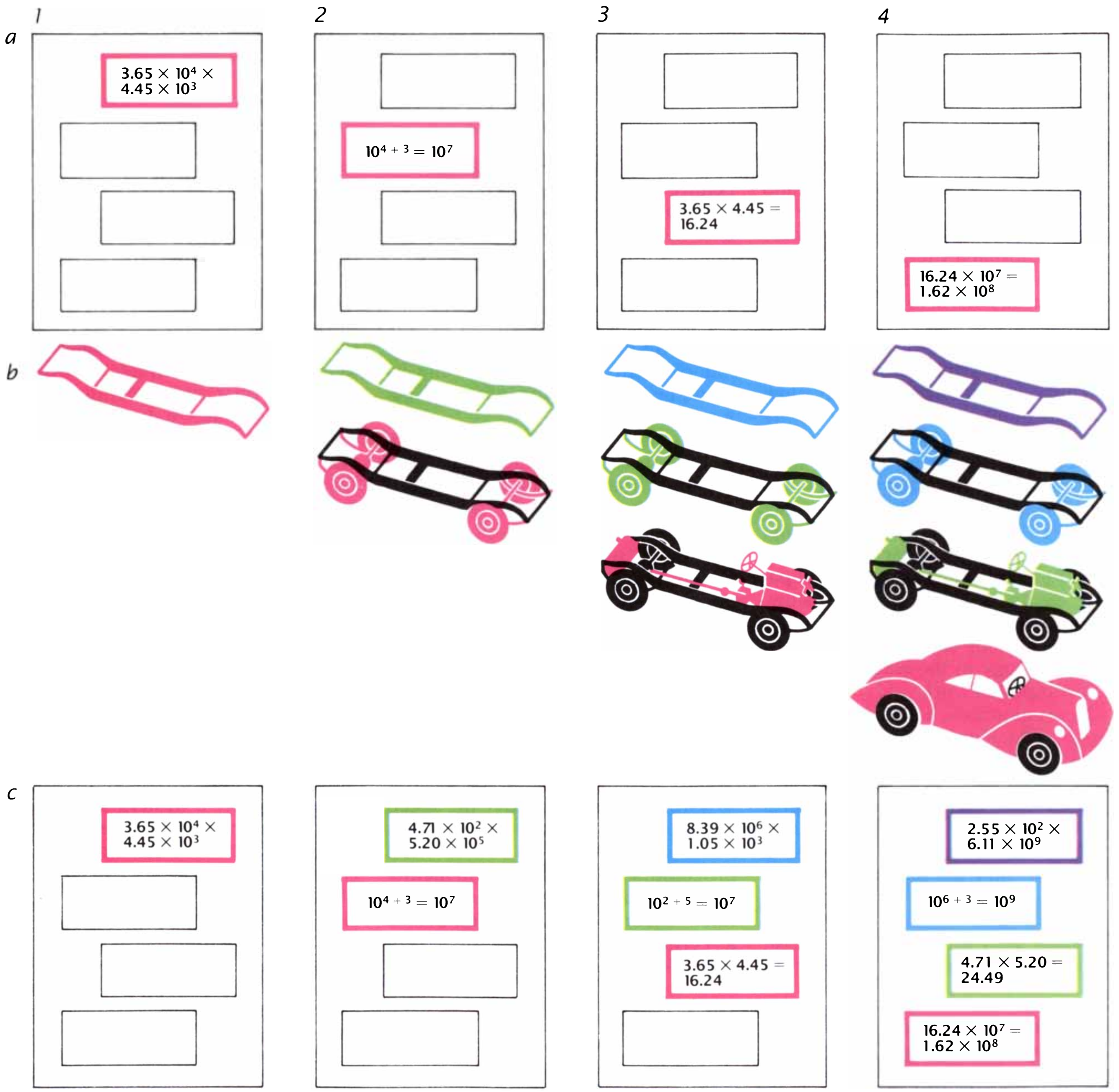
The answer, a system known as interleaved memory, is particularly effective for certain kinds of numerical calculations in which large arrays of numbers are acted on in succession. An interleaved memory typically consists of a small number of separately accessible memory units, perhaps eight or 16 of them. The first memory address (that is, the location in which the first number to be operated on is stored) is in the first memory unit, the second address is in the second unit and so on. If there are eight units, the ninth address is in the first unit, the 10th address is in the second unit and so forth. In this system several memory units can be referred to at the same time through separate channels, and data that are to be operated on simultaneously

PARALLEL PROCESSING, in which a single computing task is divided among several processors (units that do computational work), is exemplified by this board from an NCUBE computer. Each of the 64 brown squares is a processor that has roughly the power of a VAX 11/750 minicomputer. The black rectangles are memory units; six of them (totaling half a megabyte of memory) are dedicated to each processor. The computer into which this board fits may have from one to 16 such boards (or as many as 1,024 processors) in addition to a host board, which contains a master processor that distributes work and data to the other processors. This computer represents a middle ground in parallel processing. Some parallel systems include fewer but more powerful processors (the Cray X/MP, in a certain configuration, consists of four supercomputers linked in parallel), whereas others include a greater number of less sophisticated processors (the Connection Machine can include 65,536 very simple processors).

can be fetched simultaneously, without waiting for channels to clear.

Another source of slowness lies in the actual process of computation. As anyone who has multiplied two seven-digit numbers by hand knows, arithmetic operations can require many small steps. When a processor multiplies two floating-point numbers (numbers that are not necessarily integers), it must first split each number into an exponent and a mantissa (as in scientific notation), then add the exponents, then multiply the mantissas (an operation that in itself can require a large number of steps) and finally express the new product in scientific notation. The functional unit that does floating-point multiplication can be split into segments, each of which executes one of these tasks. Then, in a simple processor, all but one of the segments must be idle at any given time, representing a great waste of computing power.

A solution to this problem might be analogous in many ways to an automobile assembly line. An assembly



PIPELINING, a technique for speeding up the operation of a single processor, is analogous to the operation of an automobile assembly line. Certain computational tasks require several small steps, each of which is done by a different component of the processor. In a conventional processor (*a*), while one component executes its task the others are idle. In the example shown the task is multiplying two numbers, and the steps are extracting each number's exponent and mantissa (in memory the numbers are coded in scientific notation), adding the exponents, multiplying the mantissas and expressing the product in scientific notation. In an assembly line (*b*) no assembly station is idle: as soon as a step has been completed on one car and that car has been moved down the line, the next car is moved into the station. Thus several operations are performed at once on different cars. A pipelined processor (*c*) operates in much the same way: after an operation has been done on one pair of numbers, another pair is brought in to have the same operation done on it, without waiting for the first pair to undergo every stage of the computation.

line can produce about one car per minute. The feat is not accomplished by having thousands of workers simultaneously working on one car at a time. Instead the process of building a car is divided into many smaller tasks, each done at a particular assembly station. As the partially assembled car goes by a station, the worker at that station executes his or her own task, such as bolting on the lid of the trunk or installing the radio. It is the large-scale repetition of such small tasks that gives the assembly line its inherent speed.

Assembly lines for floating-point arithmetic can also be constructed; they are called pipelines, and they are particularly effective for applications that require many repetitions of the same arithmetic operation. Suppose, for example, a particular computer program involves multiplying many pairs of numbers. Assume, for the sake of simplicity, that each of the small tasks involved in floating-point multiplication takes one tick of the computer's internal clock. At each tick of the clock a new pair of numbers enters the processing segment that performs the first task. If the task requires, say, 10 segments, then at the end of the 10th clock tick the product of the first pair of operands will be ready, the product of the second pair will be nearly ready, the third pair will have two stages left to go, and so on; the 11th pair will have just entered the pipeline, and the 12th and later pairs will be waiting.

The amount by which a pipeline can speed up a computer's operation depends on the pipeline's structural details. Pipelines can be designed for such operations as data transfers and fetches from memory as well as for arithmetic operations. For example, if a computational operation requires two or more items that are in memory, and if a fetch from memory takes several clock ticks, a pipeline can make it possible for the processor to initiate requests for data on successive ticks, without waiting for the first datum to arrive from memory before requesting the next.

When an automobile assembly line is started, some time passes before a completely assembled car emerges. When a car does come off the line, it has been under construction for days, even though the line's rate of production is one car per minute. In a pipeline the time that passes between the beginning of work on the first task and the emergence of the first product is called the latency. The latency of a pipeline depends on the number of segments in it and the amount of time needed to perform each task. A pipeline that has a large latency is efficient only when the number of identical computations to be performed is large, just as an assembly line is efficient only for producing large numbers of cars. (If only a dozen cars were made every time the assembly line was started up, the rate of production would be much lower than one car per minute.)

One application for which pipelines are particularly well suited is called vector processing. In the terminology of computer science a vector is essentially an ordered array of independent numbers. Such arrays are called vectors because in geometry the independent numbers might represent coordinates: an array of three numbers could represent a position or direction in three-dimensional space, whereas an array of four numbers could represent a position or direction in four-dimensional space-time. In the context of computing, vectors can have thousands of elements and need not have any geometric interpretation.

It often happens that similar operations must be performed on every element of a vector. For example, in a "vector multiply" the first element of one vector is multiplied by the first element of another, then the vectors' second elements are multiplied, and so on. Such operations are natural candidates for pipelining, and hardware designed to pipeline vector operations is a mainstay of advanced computer architectures. In many of these systems groups of vector elements are processed simultaneously by separate functional units within a processor. Such machines are not true single-processor machines; they represent an intermediate step between sequential machines and parallel processors.

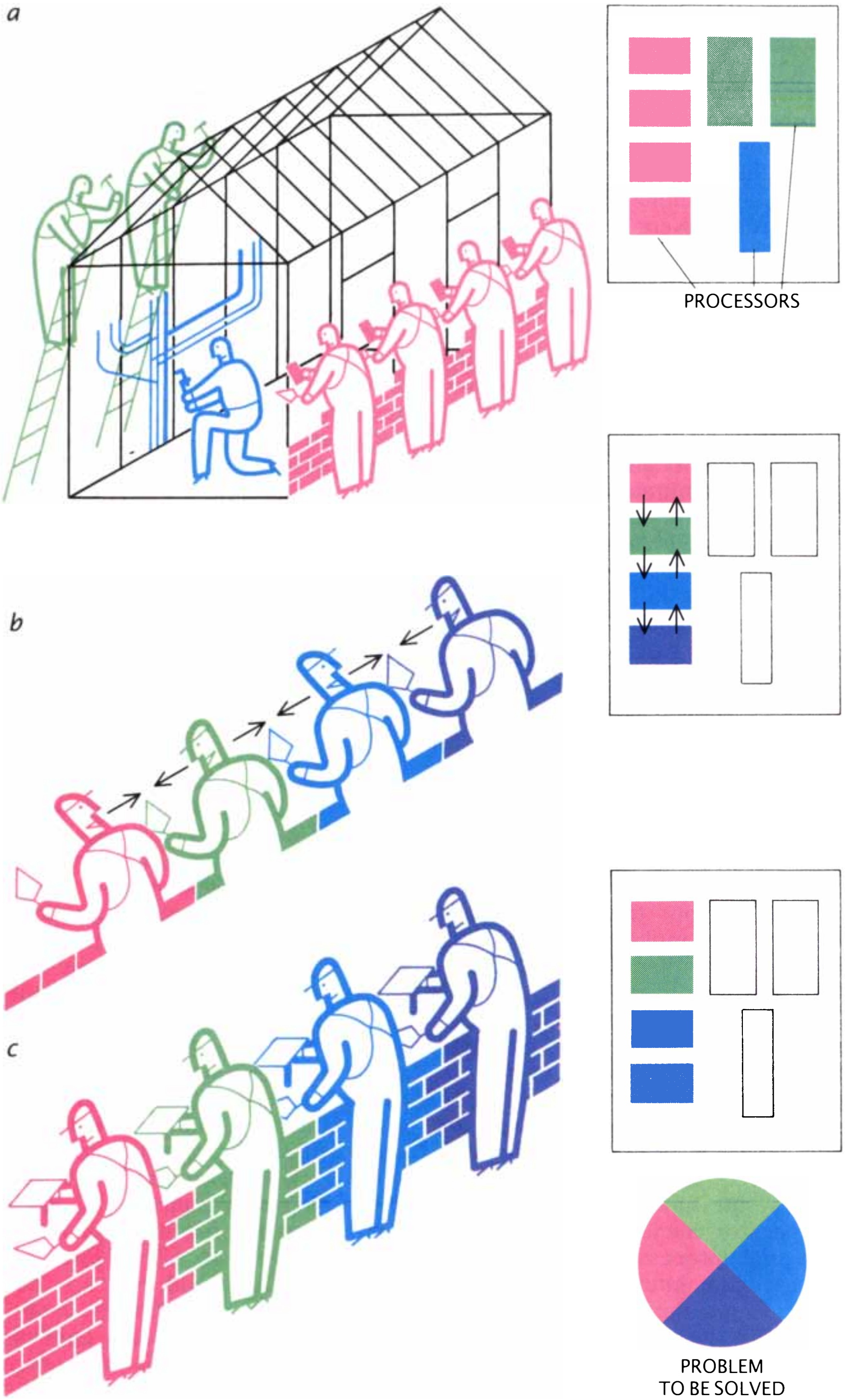
Another class of machines that are virtually parallel computers are the "very long instruction word" (VLIW) machines. The processors in VLIW machines have several copies of each functional unit. During each cycle, at the stage when a conventional processor would fetch a single instruction from memory, VLIW machines fetch several instructions simultaneously. To make this possible the instructions are grouped together into a single "word," which the processor reads from memory. The instructions are then carried out simultaneously by the several functional units in the processor, and so the processor itself acts somewhat like a parallel machine.

In VLIW machines the operations must be scheduled very carefully to avoid certain kinds of conflict, such as those that can arise when one operation requires the results of another. Some VLIW machines that can follow 10 or more instructions simultaneously are just entering the market, and designs have been developed for machines that will be able to do hundreds of operations simultaneously. VLIW architecture is particularly good for executing irregular operations, an application in which pipelined and vector architectures, which require a degree of regularity, do poorly.

All these approaches—interleaved memory, pipelining, vector processing and VLIW machinery—are effective ways to increase the speed and efficiency of single processors. Yet there are some applications in which even very fast single processors are simply not good enough.

For example, a realistic numerical study of quantum chromodynamics (a theory of fundamental particles) would require repeated recalculation of about 100 million values, which represent the strengths of various fields in the four-dimensional space-time region near a proton. The speed with which computations can be carried out depends in part on the rate at which signals can pass from one part of the computer to another. The speed of light in a vacuum, the fastest speed at which such signals could travel, is about nine inches per nanosecond (a nanosecond is a billionth of a second). Hence even if the components of a processor could work infinitely fast, even a physically small single-processor machine could not achieve more than a few billion instructions per second, which is not enough for a quantum-chromodynamics problem. For such applications the answer is to link many processors in a true parallel computer.

In designing a parallel machine the engineer is faced with a large number of choices. For example, how many processors should there be? How should they be connected? How sophisticated should each one be? In order to understand some of the options available to a designer of parallel processors, it is helpful to consider an analogy. Assume once again that you are overseeing the construction of a house, but this time you have chosen to contract with a construction company. In this analogy each worker represents an individ-



CONSTRUCTION OF A HOUSE presents a good analogy for some aspects of parallel processing. Each worker represents a single processor; the construction crew as a whole represents a parallel computer. Just as several kinds of workers—bricklayers, plumbers, carpenters—may be employed in building a house (*a, left*), so it is possible to design a parallel computer that includes several kinds of processor, each specialized for a particular task (*a, right*). Similarly, just as bricklayers must communicate with one another if they are to build walls with even seams (*b, left*), so must the processors in a parallel computer be able to communicate (*b, right*) if they are to work together on a single large problem. The analogy extends to the way work is divided. In building a house each bricklayer may be assigned a particular stretch of wall (*c, left*); each worker executes the same operation but in a different place. Likewise (*c, right*), a computer problem can sometimes be divided in such a way that each processor performs the same operation but on a different set of data (perhaps representing the physical conditions in a particular region of space). Such a division is called a domain decomposition.

ual processor and the construction crew as a whole represents a computer. What are the characteristics of this "computer," and how does it differ from other ways of arranging large numbers of "processors"?

The system is clearly parallel, because many workers undertake different parts of the job at the same time. It is also heterogeneous: workers who have different skills (bricklayers, electricians and so on) do different tasks. One can imagine other situations, however, in which the best approach would be to employ many essentially identical workers, just as many advanced parallel computers are made up of arrays of identical processors.

Another feature of this system is that the individual "processors" are relatively sophisticated: each one of them has several "functional units" (hands, shoulders, legs and so forth), which are capable of operating simultaneously. The system is said to have a coarse grain size, because each worker is relied on to do a substantial amount of work. The effectiveness of the system is not reliant on a sophisticated network of communications among the processors.

This is not to say that communication among workers is unimportant. Bricklayers working next to one another must stay in close communication to be sure they build a uniform wall that has clean seams between sections laid by different workers. In this case the network of connections has a so-called nearest-neighbor topology: information is passed between neighboring workers. Such a system could have a high overhead (that is, it could cause workers to spend more time talking and less time working) if workers needed to pass many messages to others who were far away. For this reason many advanced architectures rely on communication networks based on topologies other than the nearest-neighbor system.

Another important consideration involves the instructions given the workers. In the bricklaying analogy each worker has a task slightly different from that of all the other workers (the wall of a house has irregular features, such as doors, windows, ornaments and so on), and so each worker follows an individualized set of instructions. In a situation where workers have identical tasks—say in a textile factory where each worker operates a loom and the looms produce identical patterns—

identical instructions could be issued to all workers simultaneously. Both of these situations have analogues in advanced computer architectures.

A final characteristic of the construction analogy has to do with the way tasks have been divided among the workers. Where bricklaying is concerned, the job has been divided into many smaller tasks that are similar to one another; each bricklayer "solves" a small piece of the overall problem. This kind of division is called domain decomposition, and it is particularly appropriate when the major difficulty in a problem is the size of the data base for that problem. Here each brick takes very little time to lay, but the overall problem is time-consuming because of the sheer number of bricks to be laid. Each worker is therefore made responsible for a few of the bricks.

Such problems are quite common in computing—the modeling of quantum chromodynamics is a good example—and domain decomposition is the most natural way to approach them. It is important in domain decomposition to ensure that the network in which the processors are connected bears some resemblance to the natural topology of the problem being solved. Here the nearest-neighbor network is a good match for the actual geometry of a house.

Some tasks are not amenable to solution by domain decomposition, however, and are best approached by a method known as functional decomposition. The construction company again provides a good analogy. The company's staff includes managers and bookkeepers as well as workmen, and the workmen include plumbers and electricians as well as bricklayers. The overall task can be divided into subtasks that require the particular skills of each. A parallel computer that relies on functional decomposition can sometimes work significantly faster than a sequential computer, although the degree to which the computing process is accelerated depends on the number of significantly different tasks in the problem. Often functional decomposition can be combined with domain decomposition: the construction task can be divided among bricklayers, plumbers and so forth, and the bricklaying subtask can be divided among several bricklayers.

The analogy to a construction company provides a good general overview of the issues involved in designing parallel computers. Let us now consider in more detail some of the questions faced by design engineers and how they have chosen to answer them.

Given the opportunity to include more than one processor in a system, there is no inherent reason to create a homogeneous system by choosing only identical ones. Why not combine an appropriate mix of special-purpose processors to make a heterogeneous system? After all, many human organizations follow this approach. For example, an airline company that serves some small cities as well as major metropolitan areas would probably choose to operate some small, propeller-driven airplanes, a number of medium-size jets and a few jumbo jets.

A few heterogeneous parallel computers have been built, and many more have been designed. The machines that have actually been built generally contain only two kinds of processors: general-purpose processors and processors designed to do floating-point arithmetic. Other systems have been designed that would combine specialized processors with architectures that are particularly suited to such purposes as artificial intelligence, graphics, data bases and numerical processing, but no large systems have yet been produced commercially. Because homogeneous parallel computers are so dominant among working machines, the rest of our discussion will focus primarily on homogeneous systems.

Once an array of processors has been chosen, how might each processor be told what steps to follow during a computation? In the construction company each worker followed a unique set of instructions. In computer terminology such a system is designated multiple-instruction-stream/multiple-data-stream, or MIMD. With an MIMD system one can split a program into separate pieces, distribute the tasks and the associated data among the processors and let the processors run nearly independently. Another approach is called single-instruction-stream/multiple-data-stream, or SIMD. In an SIMD system a single set of instructions is broadcast to all the processors. At any given step in a computation each processor either executes precisely the same command as all the others or it does nothing.

A good analogy to SIMD systems is a bingo game. In bingo one person calls out numbers in succession. Each number is really a command: "Check your card for the number I just called. If the number appears on your card, cover it. If it does not appear, do nothing." This single instruction causes each player either to do nothing or to take the same action as the other players. The numbers on a player's card are the data available to that player. In some ways the card can be considered a local memory unit: a memory unit available to only one processor. If the caller has a loud enough voice, an arbitrary number of people can play in near unison. All the action takes place in or near the memory units, each of which is associated with a separate instruction processor (that is, a human player).

One natural application for SIMD machines is the quantum-chromodynamics problem we have mentioned above. To solve such a problem each processor might be made responsible for computing the strength of various fields in a particular region of space. Every processor would then have to perform essentially the same operation on its own data set. Other current applications for an SIMD structure include image processing, the analysis of fluid flows and searches through large data bases.

Several systems that have SIMD structure have been built and shown to be effective for many numerical and symbolic tasks. Examples include the Goodyear Massively Parallel Processor, which contains 16,384 individual processors, and the Connection Machine, which contains 65,536 processors. Each processor in these computers is extremely simple, and yet the machines can carry out general computations in addition to the tasks that are more obviously suited to an SIMD structure.

How should the many processors of a parallel computer be connected to one another and to the memory units? The system of interconnections by which processors can share information among themselves and with memory units is one of the most important characteristics of any parallel system, and it is also the area in which most of the variety in parallel systems arises. In one method, called shared memory, all the processors are connected to a common memory system either by a direct connection, by a network of connections or by what is known as a memory bus. A bus is essentially a channel along which several processors pass requests for information and along which the memory unit returns data to the processors.

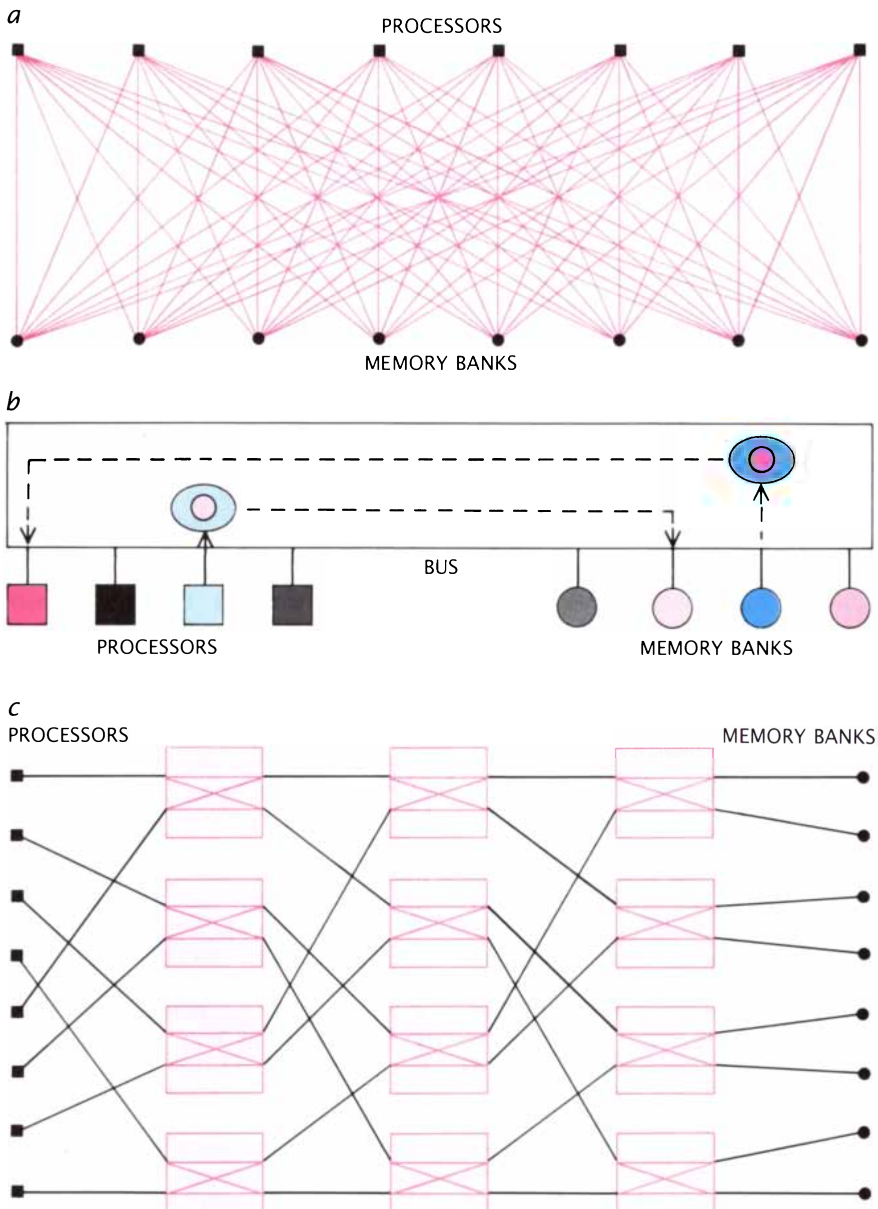
In a shared-memory system every

data location is directly accessible to every processor; if a processor needs data for its task, it simply reads it from memory. Since other processors may be continually modifying the stored data, care must be taken that no processor reads from a memory location before the appropriate value has been stored there. Most of the necessary coordination is done by software, but some assistance from the hardware can be helpful. For example, in some systems certain areas of memory are reserved for elements that can keep track of whether various locations have yet been written into, in order to safeguard against premature attempts to gain access to the data those locations will contain.

One limitation of shared memory is that it may be difficult or expensive to make a memory that can serve a large number of processors simultaneously. If each processor has a direct connection to memory and there are 1,000 processors, and if the memory is split into 1,000 separate banks in order to reduce bottlenecks, one million connections must be made. If, on the other hand, all the processors have access to memory through a common bus, then the bus must be fast enough to service memory requests from all the processors simultaneously. In this respect designing a common memory is similar to managing the traffic on a limited-access expressway. For a given speed limit and number of lanes, each highway has a certain capacity in cars per minute. When more cars attempt to use the highway than its capacity allows, traffic slows and becomes congested. One solution would be to build highways with more lanes (or several parallel highways), but there are practical limits to that approach.

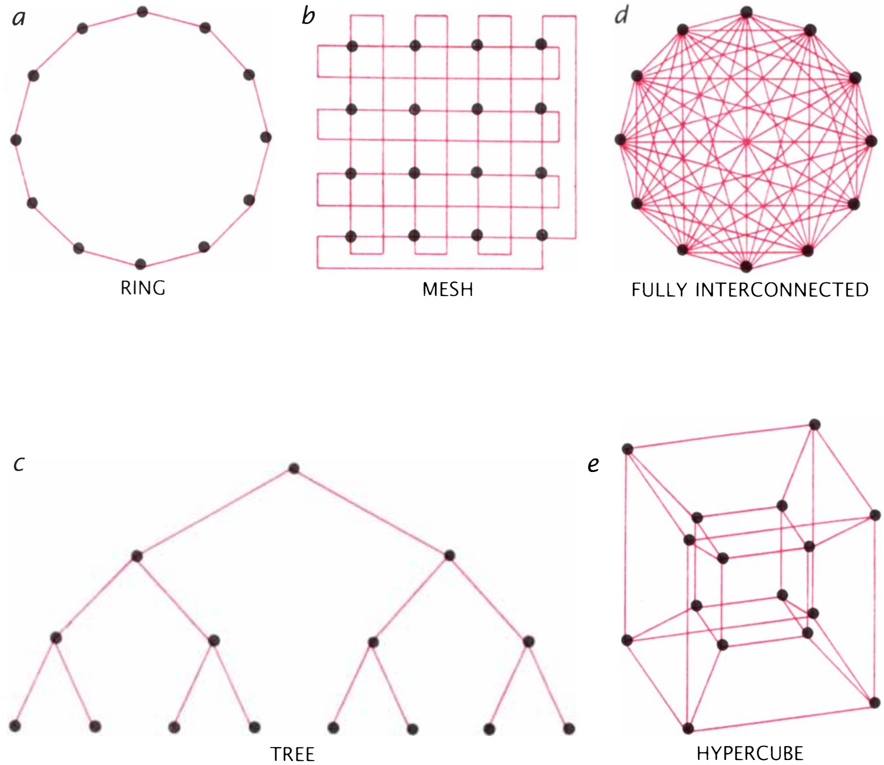
Another answer would be to build a communication network similar to the worldwide telephone network. There are hundreds of millions of telephones in the world, and yet with few exceptions any telephone in the world can be connected to any other. This is certainly not accomplished by stringing wire between all possible pairs of instruments. Instead groups of instruments are connected to separate switching systems. The switching systems are connected among themselves in a hierarchical manner so that any one switching station is responsible for a manageable number of connections. When a connection is requested (by dialing a telephone number), the request propagates by way of several switching stations until that particular instrument is reached.

A disadvantage of network connections is that they increase the latency of access to memory, since each request for information must traverse several stages of the network on its way to the memory unit. In addition, just as in the telephone network, there may be periods of peak activity when the network is saturated with requests for connections, resulting in very long delays—a kind of computational Mother's Day. Various types of



SHARED-MEMORY ARCHITECTURES enable all the processors in a parallel system to have access to global, or common, memory units. In the simplest shared-memory scheme (*a*) every processor is directly connected to every memory bank. One disadvantage of this system is that each processor or memory bank must support a large number of connections. An alternative (*b*) is a "bus": a common communication line along which processors pass requests to memory banks and memory banks return data. Such a bus can become crowded (and therefore slow) when there are a large number of messages to be passed. Another alternative is a so-called omega network (*c*), in which processors are connected to memories by a series of switching boxes, each of which has two input lines and two output lines. In an omega network each processor effectively has a direct line to each memory, but there need not be as many communication lines as in a true direct-connection scheme. The advantage becomes more apparent as the number of processors and memory banks is increased. A disadvantage is that each message must pass through a number of switching stations before it can reach its destination.

network have been designed to minimize congestion while keeping the cost reasonable and the speed of access to memory high.



DISTRIBUTED-MEMORY ARCHITECTURES establish connections among processors that each have sole control over some amount of memory. Two of the simplest schemes are the ring (*a*) and the mesh (*b*). A more specialized architecture is the binary tree (*c*). It is particularly useful for so-called expert systems, which often rely on decision-making processes that can be mapped as trees. One could also connect every processor to every other processor (*d*), but this system requires an unwieldy number of connections. One innovative architecture that is seeing wide use is the hypercube topology (*e*), in which processors are connected as they would be if they were at the corners of a multidimensional cube. (In the example shown they are at the corners of a four-dimensional hypercube.) In this system each processor can pass messages to any other by a relatively short route, but the processors are not required to support very many connections.

If it is difficult or expensive to construct shared-memory machines, why not attach some amount of memory to each processor and connect the processor-memory pairs (or nodes, as they are usually called in this context)? Systems that are built along these lines are said to have local memory (which is also known as distributed memory). In local-memory architectures data must be distributed to the appropriate nodes at the start of the computation. It will generally be necessary in the course of the computation for nodes to get data from the memories of other nodes. This can be done by sending a message to the appropriate node, asking that the information be sent back in a return message. If the two nodes concerned are not directly connected, other nodes in the path between them can forward the message.

Many schemes have been devised for connecting the nodes of distributed-memory systems. The designer of such a system has several aims: communication among nodes should be fast, no node should have to support too many connections and the topology of the connections should somehow match the natural geometry of the problem to be solved. The analysis of quantum chromodynamics provides a convenient example of this last condition. If each node calculates the values of various fields in a certain volume of space, then it should be connected to the nodes responsible for calculating those values in surrounding volumes.

One of the simplest interconnection schemes is a ring, in which each node is connected to two others and the line of connections forms a circle. Another relatively simple scheme is the mesh, in which each node is connected to its four nearest neighbors. Although any number of nodes can be connected in a ring or a mesh, the number of connections each processor must support is small (two or four respectively). The processor within each node can have access to its own memory at an appropriate speed, and the communication channels are required to handle traffic among only a few nodes.

These simple topologies do have some drawbacks. The number of nodes involved in sending a message from one node to another can be quite large (in a ring as many as half of the nodes may be involved) resulting in long delays. If a node cannot find anything to keep itself busy while it is waiting for data to arrive, its resources will be idle; if data from other nodes are needed often, a lot of computing power will be wasted.

A way to alleviate that problem is to resort to a slightly more elaborate topology known as the *n*-cube or hypercube scheme. In this approach nodes are connected as they would be if they lay at the corners of a multidimensional cube. For example, eight nodes would be connected in a 3-cube, whereas 16 nodes would be connected in the topological arrangement that models a four-dimensional cube. (The one- and two-dimensional cubes are respectively a straight line and a square.) In general, a *p*-dimensional cube can connect 2^p nodes. The longest path a message might have to follow is the logarithm to the base 2 of *p*. Commercial systems are in use today that link 2^{12} nodes in a hypercube topology.

Given the differences between the topologies available to shared-memory systems and local-memory systems, and also given the different strengths of the two approaches, it is only natural to try to combine the two. For example, one could imagine a hybrid system in which each node consists of a modest number of processors that share a memory. The nodes, which would be called clusters in this case, could be connected in a mesh or hypercube topology. The converse of this system, in which a number of nodes that each have local memory also share a common memory, is already found in a few systems.

As long as one is experimenting with the number of processors in a computer and the ways in which processors are connected, why not experiment as well with the way the operation of the machine is controlled? Almost all computers built to date, including parallel computers, are controlled by programs. A programmer writes out a set of instructions, which is then translated into machine language: the code that describes the actual computational steps to be undertaken by the computer. This program determines, in a precise and absolute way, the flow of work through the computer.

Parallel architectures make it possible to imagine radically different ways of controlling the computer's operation, most of which are still in the experimental phase. One approach is known as dataflow. In a dataflow architecture a problem is represented as a directed graph: a collection of points, called nodes, connected by arcs. Nodes represent mathematical or logic operations and arcs represent the flow of data from one operation to another. In a physical machine nodes might be represented by memory-processor pairs and arcs by physical connections between the processors.

A node in a dataflow machine does not execute its operation in a predetermined sequence specified by a program. Instead it waits until it has received all the data required to carry out its operation, and then it executes the operation and passes the answer on to the next node. A crude analogy might be a relay race, in which the second runner starts as soon as he or she receives the baton from the first runner. In the case of a dataflow machine, however, a processor might wait for results from many other processors, not just one.

One could imagine a dataflow system in which a very large number of nodes all execute instructions with no concern for what most of the other nodes are doing. Because each performs its appointed task only when the necessary data arrive, there is no danger of acting prematurely, say by fetching data from a memory location before the most up-to-date value has been placed there. Likewise there is no danger of interfering with other processors, say by trying to read the same memory location at the same time. Moreover, there would never be an instance of two messages trying to use the same communication channel at the same time; the dataflow approach thus avoids the commonest bottleneck of parallel architectures.

The opposite approach, known as demand-driven architecture, is also possible. Demand-driven machines are similar to dataflow machines in that one can visualize them as a collection of nodes at which operations take place and arcs along which data travel. The difference is that an operation at one node is carried out only when another node requires the output of the first. Expressions are evaluated only when the result is required, whereas in dataflow architectures every operation is executed as soon as possible even if it turns out not to have been necessary.

Demand-driven architectures do require an inherent time delay: the time it takes for a message to pass from one node to a "higher" node requesting that the higher node execute its operation. The advantage of dataflow and demand-driven architectures is that they may be much easier to program reliably than conventional parallel machines. A user would merely need to specify the operations to be carried out at each node and the flow of data between nodes, without worrying about the exact mechanics or timing of the algorithm to be followed.

Today's computationally inclined scientists have substantial but limited experience with advanced architectures. Several promising parallel architectures, such as SIMD machines, the Connection Machine, the California Institute of Technology's Hypercube and various shared-memory machines, have been explored and found viable in a broad range of applications. On the other hand, the currently available parallel hardware and software are rudimentary. There have been few broad quantitative comparisons of different machines. Hence, although we are confident that parallel machines will be extremely important in the future, and that they promise much greater performance than sequential machines, we cannot predict either the specific architectures or the performance levels that are likely to be found in the coming years.

Parallel processing has come as far as it has largely because farsighted funding by the Government and industry of many university-based projects has led to an impressive number of commercial endeavors spanning a wide variety of architectures. We believe the near-term future for new hardware is most promising. We are not as sanguine about software. We do not see the same vigorous funding going into this crucial but unglamorous area. It has taken more than 30 years to build up today's base of sequential software. Ingenious software techniques may make some of this software base applicable to newer machines, but any major progress will require a significant effort to restructure and rewrite the basic applications in commercial and research areas.

One crucial step will be the establishment of some standards. Such languages as FORTRAN, COBOL and LISP provided the building blocks for the fundamental scientific, business-related and artificial-intelligence work on sequential machines. These are certainly imperfect languages, but they were good enough to allow the development of a foundation of sequential software that could evolve as the hardware developed. No similar understanding and agreement concerning analogous software standards has yet developed in the parallel world.

The future will undoubtedly be affected by new technology; improvements in semiconductor technology, and perhaps in technology based on the new high-temperature superconductors, will certainly lead to new tradeoffs in the design of hardware. There will certainly be new hybrid machines in which optical interconnections link high-speed digital processors. Such advances would retain the essential models of computing we have described here.

More radical is the possibility of computing that is purely optical, which would allow fast analog computations of various algorithms that today can be implemented digitally only in very clumsy and time-consuming ways. It is not known whether such optical computers would necessarily be limited to a few specialized areas of signal processing or whether they could be developed into general-purpose machines.

Finally, we must also mention the recent work that has been done to develop so-called neural-network models. These are intuitively attractive because they provide a computational model comparable to that of the human brain. Neural-network machines are potentially far superior to conventional computers in such areas as pattern recognition, although they are much weaker at such humdrum tasks as floating-point arithmetic. Investigators have already been able to build some simple neural-network hardware, and we expect this new model of computing to undergo rapid evolution both in hardware and in algorithms.

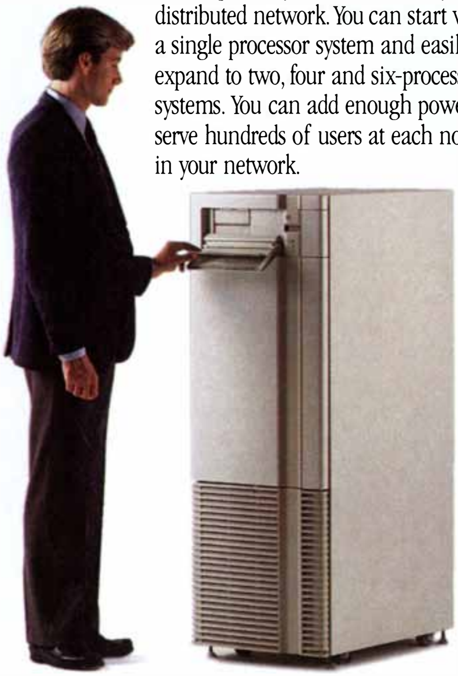

SAFARI VAN
FULL-SIZE SIERRA PICKUP
ONE GREAT TRUCK LEADS TO ANOTHER.
Drive home a 1988 full-size Sierra Pickup from GMC Truck, and it could be the start of something bigger than both of you.
That's because Sierra, like every vehicle from GMC Truck, offers so much that it demands to be part of a collection of similarly gifted vehicles. And Sierra is a great place to start, with its ample room for three adults, standard V-6 or optional V-8 power and the ability to carry 4' X 8' plywood sheets between the wheelhousings on long-beds.
And Sierra's standard equipment list includes a rear-wheel anti-lock brake system (operates only in two-wheel drive), electronic fuel injection on gas engines, power steering, power brakes and more.
TRAVEL FIRST CLASS.
Safari offers a similar degree of refinement in a very versatile van. Select a Safari with the newly available Touring Package and enjoy luxurious upholstery, rich carpeting, ample storage pockets and compartments, and supreme comfort for up to eight passengers with available seating. Here is very grand touring in a very manageable size.
Why not buckle up now and visit your nearest GMC Truck dealer? There's never been a better time to start your GMC truck collection — or a better choice of fine GMC trucks to choose from.
See your dealer for the terms and conditions of this limited powertrain warranty.
GM 6/60 QUALITY COMMITMENT PLAN
For a catalog and the name of your nearest dealer, call 1-800-VALUE88 (1-800-825-8388).
GMC TRUCK
IT'S NOT JUST A TRUCK ANYMORE.

Our new small systems give you the big edge.

NonStop CLX.™

Distributed processing for every business location.

It's the newest member of the Tandem NonStop family. It can run a fully distributed network. You can start with a single processor system and easily expand to two, four and six-processor systems. You can add enough power to serve hundreds of users at each node in your network.



LXN.™

Integrates UNIX™ into the Tandem OLTP network.

Our lowest-cost system can run UNIX applications and can access the Tandem OLTP network—all from any workstation. The LXN can support up to 32 users and take a huge workload off your host computer. We are the first to bring OLTP features to UNIX in this price range.

UNIX and UNIX System V are trademarks of Bell Laboratories.

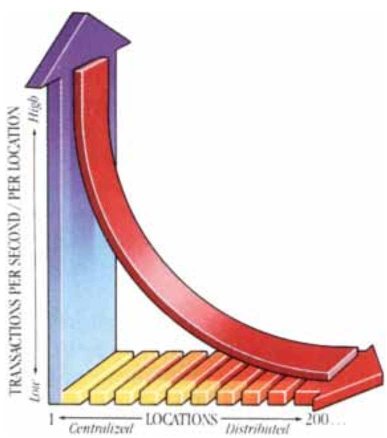
The only high-performance, low-cost systems for on-line transaction processing.

CHOOSE ANY SITE.

Now you can afford to deploy your resources where they're needed. Branch offices, warehouses, the factory shop floor, retail stores, wherever. These systems require no special computer room environment. They can also talk to other major computer systems.

HIGH PERFORMANCE AT ANY LEVEL.

Only Tandem is equally productive whether you're processing transactions in high volume from a central site or from many distributed sites. And we can deliver a lower cost per transaction at a lower cost per location.

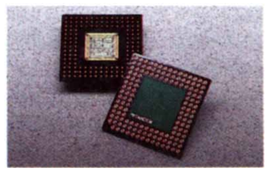


EFFICIENT EXPANDABILITY.

As you add users, add new processor and memory boards. You can expand in any increment you choose. You never buy more than you need. And there are no expensive rewrites of application software.

ADVANCED COMPONENTS, FEWER PARTS.

Our custom chip designs deliver more performance on less power and require less space. And having fewer parts increases each system's reliability.

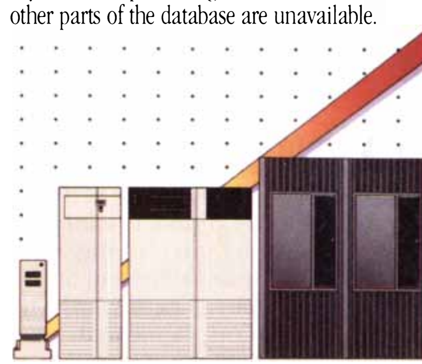


EASY SERVICE.

To reduce maintenance costs, special software guides office workers through routine service procedures. You can also let us do it, or we can do it cooperatively. All critical components can be replaced on the spot, without interrupting the application. So fewer service calls are required, and they take less time. Other software allows us to diagnose system faults from remote Tandem service centers.

SQL DATABASE TECHNOLOGY.

SQL, the industry standard for relational database management systems, is easier and more productive for both the user and the programmer. But only Tandem's NonStop SQL™ combines this productivity with high-performance OLTP. There's no sacrifice in performance, no loss of data integrity. And local processing can continue even when other parts of the database are unavailable.



LET US DEMONSTRATE.

Whenever there's a need for constantly current information, efficient expandability and unbeatable price performance, Tandem technology proves consistently superior. Compare us to any other OLTP system. You'll see why companies in every major industry choose Tandem.

For information, write: Tandem Computers Incorporated, 19191 Vallco Parkway, Loc. 4-31, Cupertino, CA 95014. Or call 800-482-6336

TANDEMCOMPUTERS

The technology leader in on-line transaction processing.

Chips for Advanced Computing

In 1959 the number of transistors that would fit on a chip was one; now it has surpassed a million. As limits are reached, the pace is slowing, but by 2000 there will be chips with a billion components

by James D. Meindl

Perhaps the most significant event of the past 50 years has been the emergence of a society whose chief product is information. The driving force of this transformation is the development of electronics technology, particularly the integrated circuit fabricated on a single chip of silicon. Since the advent of the integrated circuit in 1959, the number of transistors that can be squeezed onto a chip has increased from one to several million. As a result the performance of integrated circuits has improved by a factor of more than 10,000. Remarkably, while the chip was undergoing dramatic improvements in performance and complexity, its cost remained virtually unchanged.

In considering such phenomenal technological progress the question that naturally arises is: How long can it continue? One way to think about the question is to ask when a chip with one billion transistors, corresponding to the initial stages of gigascale integration (GSI), will become commercially available. To determine the date it is necessary to analyze the factors that serve as limits on GSI. Those factors form a hierarchy with five levels corresponding to fundamental limits, material limits, device limits, circuit limits and system limits. Each level of the hierarchy can be considered from three different points of view, which are respectively theory, practice and historical analogy.

Theoretical limits are based on accepted principles of science. In effect they constitute design limits on GSI. Practical limits, on the other hand, depend not on scientific principles but on fabrication processes and equipment. Practical limits help to dictate what is possible at a particular time within the framework of commercially available technology. Historical analogy can help to illuminate the long-term future of the information revolution by way of comparison with the mature Industrial Revolution. Together these perspectives can answer the question of when GSI will come about. Having answered that question, one can look beyond currently projected limits to devices now in the laboratory that may trigger the revolution's next phase.

In this article I shall take up the theoretical perspective first, then the practical one and finally the analogical contribution of history. From a theoretical perspective the fundamental limits are laws of nature that cannot be changed. Such constraints are by definition independent of the material employed to make a device, the design of the device itself, the circuit in which it is used and the system in which the circuit is found. Fundamental limits form the first (and most general) level of the hierarchy mentioned above. As such, they are extreme limits beyond which lie only forbidden zones of operation.

Fundamental limits come from several areas of physical science. For example, thermodynamics shows that there are random statistical fluctuations in the energy level of atoms and electrons in the crystalline semiconductor of which a transistor is made. Now, the energy level of the electrons bears an intimate relation to the signal that is being processed by the transistor. If that signal is not considerably stronger than the random fluctuations, the fluctuations will be mistaken for the signal itself. Hence the fluctuations impose a minimum switching energy. Because they vary with temperature, the minimum energy required to switch the device rises as the temperature increases. Other fundamental limits come from quantum mechanics and from electromagnetic theory or the speed of light.

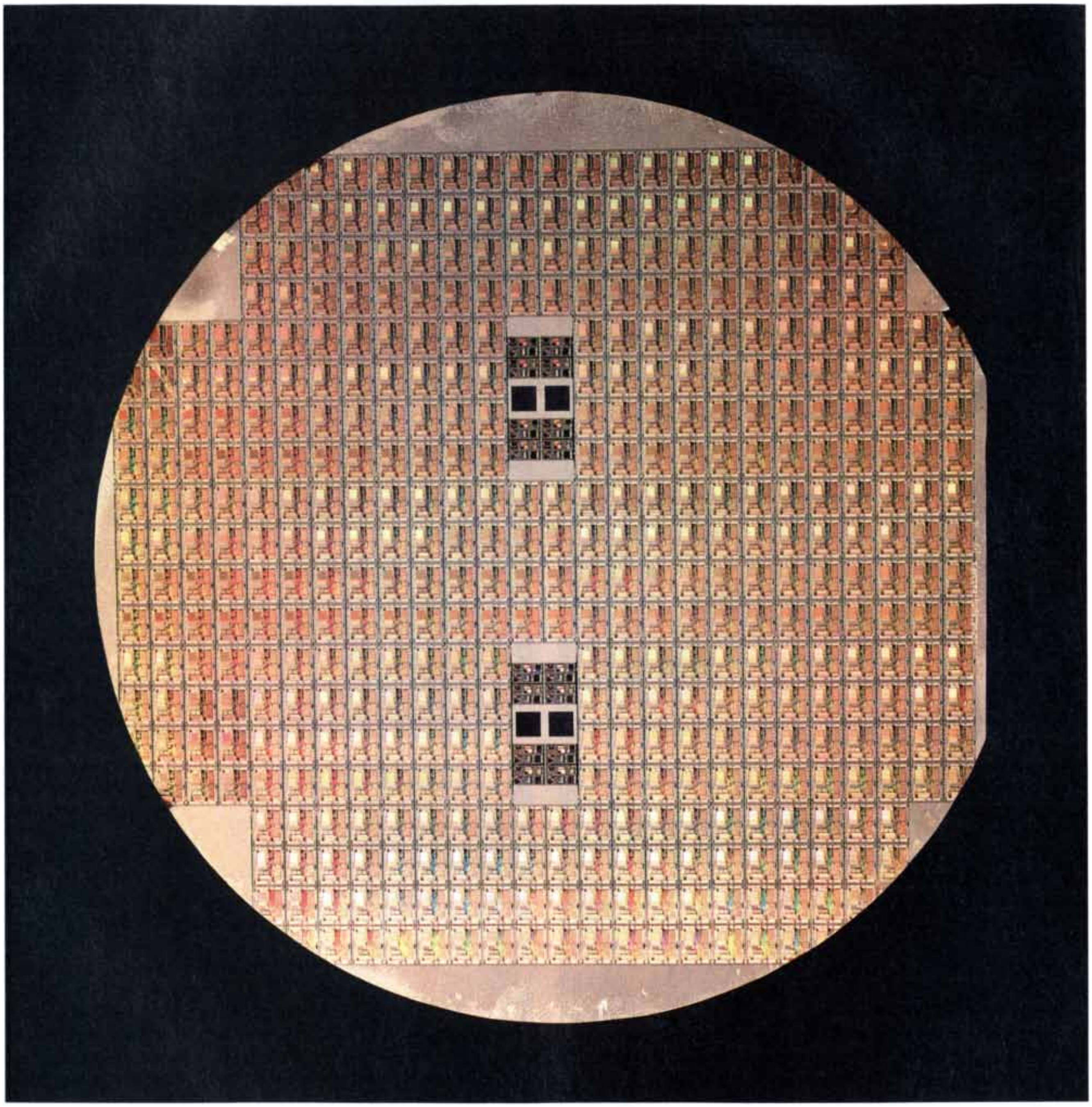
Material limits depend on the chemical composition and structure of a substance but not on the configuration of the device itself. In the case of microelectronic circuitry the materials of greatest interest are semiconductors. The atoms that make up the crystalline lattice of a semiconductor are held together by electrons called valence electrons; each silicon atom has four such electrons. Under ordinary conditions the valence electrons are tightly bound to the atoms of the lattice.

The situation can be altered by "doping," or adding to the lattice impurity atoms that have more or fewer valence electrons than silicon has. Arsenic, a common dopant, has one more valence electron than silicon. The extra electron is normally free to wander through the lattice. If an electric field is applied to the material, the free electrons, which carry a negative charge, conduct an electric current. Dopants such as boron, on the other hand, have one less valence electron than silicon. The "hole" resulting from this deficit acts as a positively charged carrier of current. By choosing the appropriate type and concentration of dopant the electrical properties of the semiconductor can be precisely controlled.

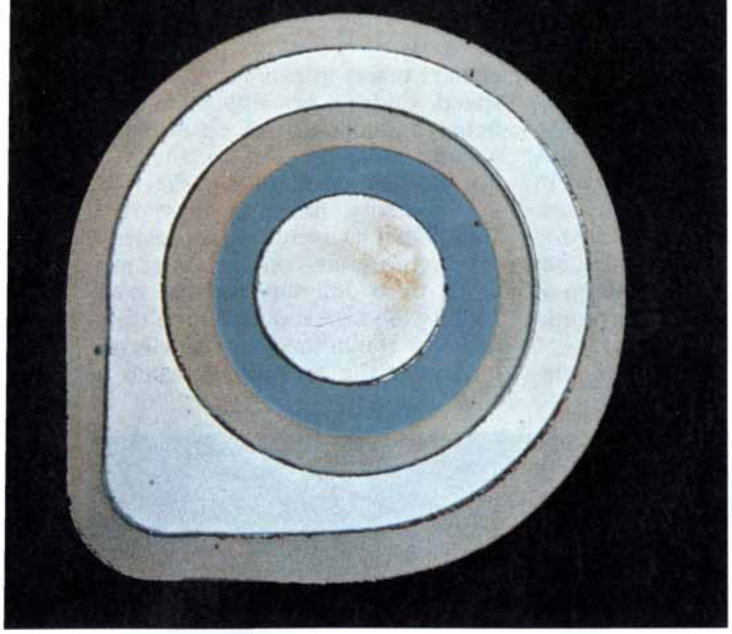
Several properties of silicon render it uniquely well suited to serve in integrated circuits. One such property is the size of its bandgap: the difference in energy between valence electrons and conduction electrons. If the gap is too small, relatively slight increases in temperature will boost many electrons into the conduction band, interfering with the precise control of electrical properties required in the device. Silicon, however, has a bandgap of 1.12 electron volts, sufficiently large for the material to maintain excellent semiconductor properties over a wide temperature range near 300 degrees Kelvin (27 degrees Celsius, near room temperature).

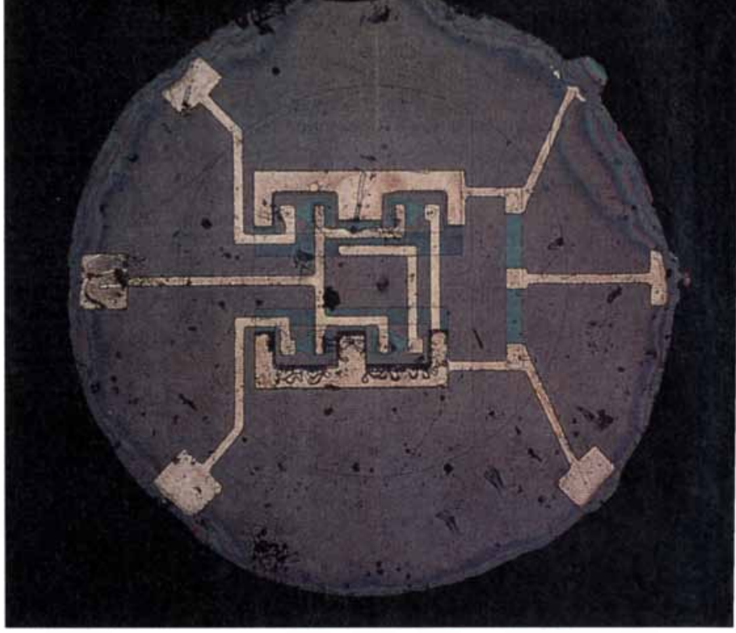
That is not all. In addition, silicon is an abundant elemental semiconductor that can be formed into almost perfect crystals at relatively low cost. What is more, its native oxide, silicon dioxide (SiO_2), is an excellent insulator with some desirable attributes for making integrated circuits. No other material has such a combination of virtues, which helps to explain the dominance of silicon in the fabrication of integrated circuits. Some other material limits, however, have in recent years elevated another semiconductor to a position of great importance: gallium arsenide (GaAs).

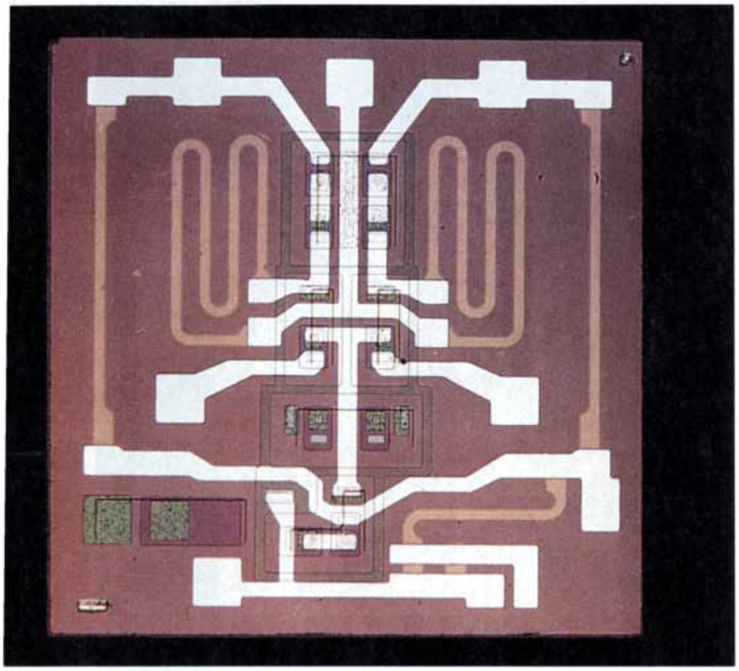
The principal reason that chip designers resort to gallium arsenide is speed. Under a low applied electric field conduction-band electrons in gallium arsenide drift six times as fast as they do in silicon. Electron-drift mobility, however, may overstate gallium arsenide's superiority. A more accurate comparison is provided by a different material limit, namely the time required for a carrier under the influence of an electric field near the breakdown value to

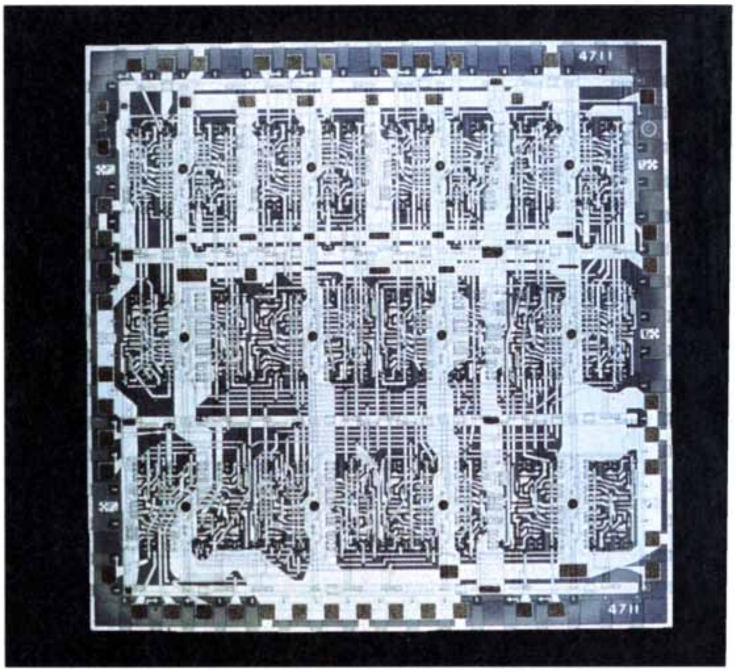


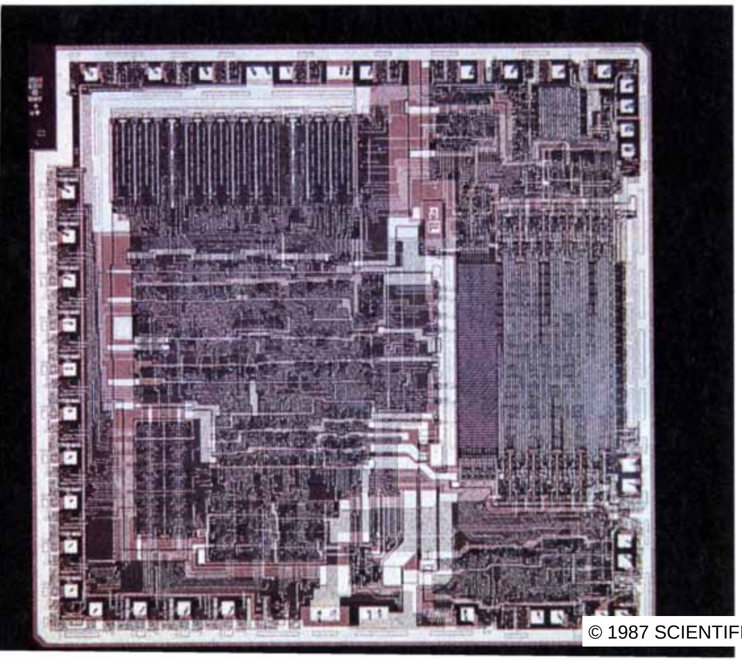
SILICON WAFER contains 470 computer chips in the final stages of manufacture. The wafer, shown at its actual size, is a slice of a silicon "ingot." The chips are fabricated together on the wafer; the rectangles near the center are test circuits for monitoring fabrication processes. After testing to detect defective chips, the chips are cut apart with a diamond-tipped saw and the functional ones are packaged. These chips are of the type called PACE1750 A, manufactured by the Performance Semiconductor Corporation. Each one is the central processing unit of a minicomputer; they are used in military avionics and other systems.

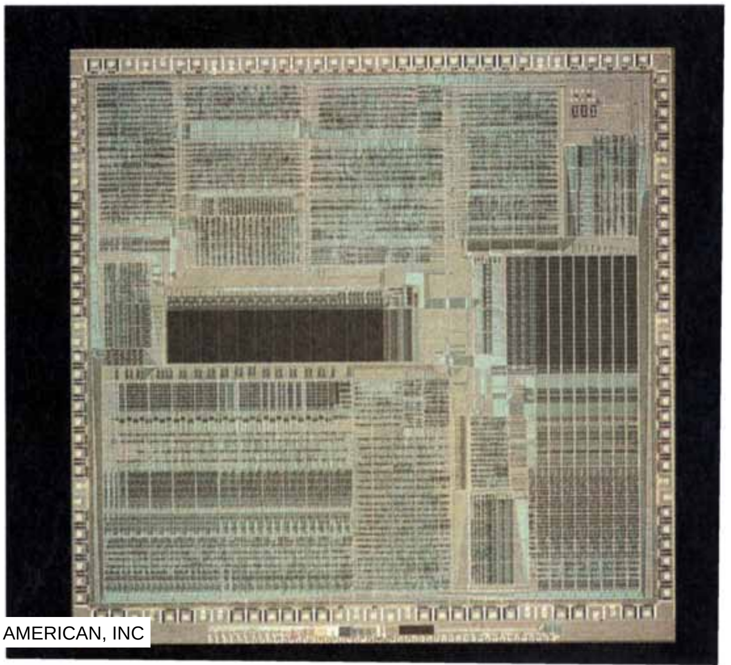







4711
4711



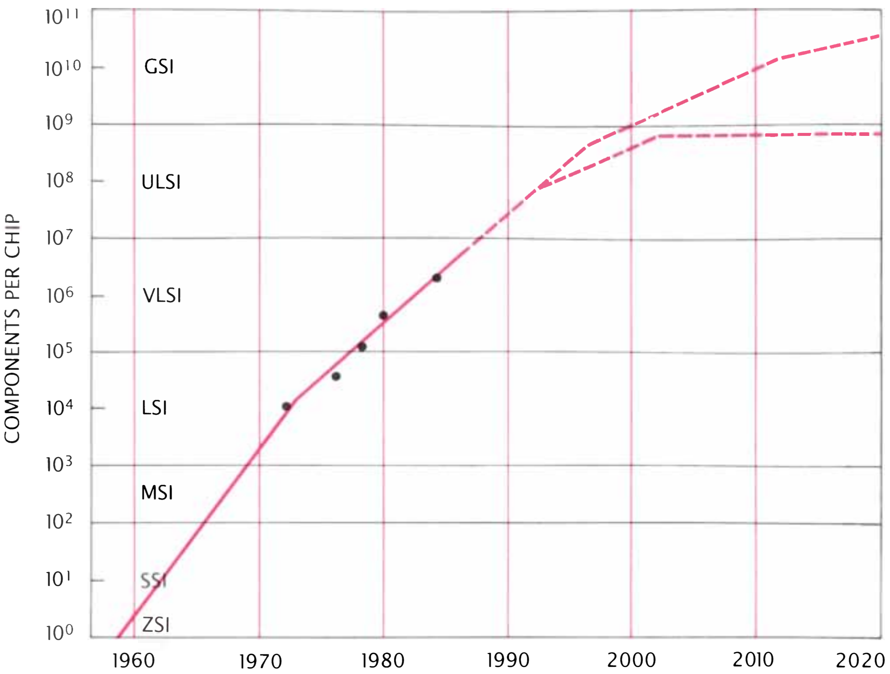


undergo a drop in electric potential of, say, one volt. (The breakdown electric field is one that shakes so many valence electrons loose from the lattice that a self-ionizing avalanche begins.) By this measure gallium arsenide is about 2.5 times faster than silicon.

From the viewpoint of several material limits, then, gallium arsenide offers advantages over silicon in speed. As transistors become smaller, however, that advantage may be offset by other material factors. A transistor can be made to switch faster by applying more power to it, but so doing increases the buildup of heat in the device. For extremely small devices the speed of switching may be limited by the capacity of the substrate to conduct heat away from the device. Because silicon has three times the thermal conductivity of gallium arsenide, very small silicon devices may be able to switch just as fast as those made of the ostensibly "faster" material.

Thus at each level of the hierarchy opposing limits must be balanced against one another. Nowhere is this balancing process clearer than at the device level. Device limits are numerous, because they encompass all the material limits as well as additional limits based on the device's size and geometry. In spite of such profusion, the key problem for GSI at the device level can be identified with reasonable certainty: to determine the smallest possible dimensions for a metal-oxide-silicon field-effect transistor, or MOSFET.

A common type of MOSFET consists of two islands of silicon doped to increase the concentration of negative-charge carriers (*N*-type silicon) in a layer of the same material doped to increase the concentration of positive carriers (*P*-type). Bridging these islands, called the source and the drain, there is a layer of silicon dioxide that serves as an insulator. Deposited on the insulator between the source and the drain is a metal electrode: the gate. A positive "input" voltage on the gate attracts electrons to the interface of the silicon substrate and the insulator. The electrons form an induced channel, allowing "output" current to flow from source to drain and sending a logic signal to the next stage of the circuit. In the absence of an input signal no channel is formed, and no output current results.



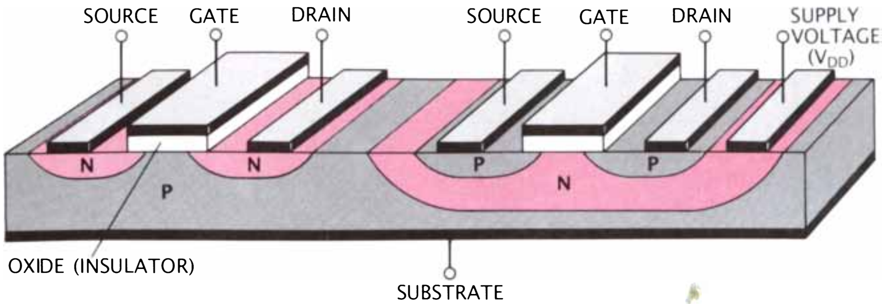
NUMBER OF COMPONENTS PER CHIP doubled annually in the 1960's. In about 1972 designers ran out of unused space on the chip for additional components, and the rate fell somewhat. Nevertheless, according to the more optimistic projection, gigascale integration (GSI)—a one-billion-component chip—will be achieved by the year 2000. (The projections differ in assumptions about limits imposed by chip-fabrication processes.)

Since metal-oxide-silicon technology has become the dominant one throughout digital electronics, achieving GSI rests on the continuing attempt to scale down the MOSFET. The process of scaling down begins with the definition of a scaling factor, often called *S*. All the lateral and vertical dimensions of the MOSFET are then reduced by a factor of *S*. (Thus if the scaling factor were 2, the height and width of the device would be decreased to half their original values.) In addition, the supply voltage is reduced by the same factor, which keeps the strength of the electric field constant and prevents an increase of the stress on the device.

The parallel reduction of size and electric-field strength yields some remarkable advantages. The time required to switch the device, which depends on the length of the channel, decreases by a factor of *S*. The power dissipated per unit of chip area remains constant, and so the problem of heat removal is not made worse. The packing density of the transistors on the chip, which depends on the area of the device, increases by a factor of S^2. Perhaps best of all, the energy consumed in each switching operation, which depends on the power and the switching time, decreases by a factor of S^3. Thus the result of scaling is a chip that has more devices switching faster and using less power to do so.

Given such advantages, the designer of chips would like to know how far scaling can go. The answer lies in the minimum allowable length of the MOSFET's channel. That minimum emerges in part from the interplay of the supply voltage and the doping concentration, which is the concentration of impurity atoms (usually boron) in the *P*-type substrate of the channel. At every junction between *N*-type and *P*-type materials, charge carriers migrate to the opposite side of the interface, where their concen-

HISTORY OF CHIPS is one of increasing density, as six chips made by the Fairchild Semiconductor Corporation from 1959 to 1985 show. At the top left (1959) is the first planar transistor. At the top right (1961) is the first planar integrated circuit on a single chip; it includes four transistors along with other components. At the middle left (1964) is the first consumer-oriented linear integrated circuit; it has five transistors. At the middle right (1968) is a bipolar logic array; it has 180 transistors. At the bottom left (1978) is the first 16-bit chip to include an entire central processing unit; it has 20,000 transistors. At the bottom right (1985) is the CLIPPER CPU; it has 132,000 transistors.



COMPLEMENTARY MOSFET'S provide the basis for the dominant integrated-circuit technology. A MOSFET is a metal-oxide-silicon field-effect transistor. Electric charge in a semiconductor such as silicon can be conveyed by electrons (carriers of negative charge) or "holes" (carriers of positive charge). In *N*-type silicon electrons predominate; in *P*-type silicon, holes. Each type also contains a smaller quantity of carriers with the opposite charge. When the voltage on the gate is "off," no current flows from source to drain. When the voltage is "on," minority carriers are drawn up under the gate, thereby forming an induced channel and enabling a large output current to flow from the device. The illustration shows how a pair of complementary MOSFET's (that is, one with an *N*-type channel and the other with a *P*-type channel) can be formulated on a single chip.

tration is lower: electrons migrate to the *P*-type side and holes to the *N*-type side. As a result there forms what is referred to as an electron-and-hole-free space charge region. To operate correctly, the channel must be at least twice as long as this region.

If the channel length is to be the minimum, it is clearly desirable to reduce the length of the space charge region, which can be done by increasing the doping concentration. Increasing the doping concentration, however, has the unwanted effect of necessitating an increase in the supply voltage. If the supply voltage is raised too high, the oxide insulator under the gate will begin to break down as a result of the self-ionizing process I mentioned above. The minimum channel length thus depends on reconciling supply voltage and doping concentration. For a supply voltage and doping concentration typical of those now in use, it can be shown that the minimum channel length is between .1 and .2 micrometer (millionths of a meter). For the moment such limits are truly theoretical: current MOSFET's have channel lengths of from one micrometer to two micrometers.

Whereas scaling down has great advantages for devices, the opposite is true for the thin films of metal that connect transistors on a chip. First, not all connections between devices become shorter as transistors are scaled down. In the past two decades as transistors have become smaller, chips have become bigger. As a result long-distance interconnects (which may link transistors at opposite corners of a chip) have tended to become longer. Their increase in length is frequently described by a "chip scaling factor" (S_c) corresponding to the linear dimensions of the chip.

In contrast, local connections, which link neighboring transistors, do become shorter in proportion to the scaling factor *S*. Even for local connections, however, scaling is not beneficial. Because of the electrical behavior of conductors, the time required for a signal to be conveyed by an interconnect may not decrease as the connector becomes both smaller in cross-sectional area and shorter. Instead the delay time tends to remain constant. Furthermore, the density of current in the connector increases with *S*. If the current density increases sufficiently, the material of the connector may actually be pulled along, leaving voids. This phenomenon, called electromigration, can eventually cause the interconnect to fail completely.

Stepping beyond devices, one reaches the level of circuit limits. On the circuit level one of the most significant problems is to determine the minimum supply voltage at which a basic logic circuit can operate. Perhaps the most basic logic circuit is one called the inverter, which is found as a component of almost all complex microelectronic circuitry. An inverter circuit converts a low-voltage input (a "0" in the computer's digital repertoire) into a high-voltage output (a "1") and does the converse as well.

Perhaps the most remarkable logic circuit in all of GSI is an inverter that includes one MOSFET with an *N*-type channel and one with a *P*-type channel, a combination that is called complementary MOSFET technology. As long as the circuit is not being switched, it is always the case that one of the devices is on (conducting current) and the other is off (not conducting). Because the devices are wired in series, the result of such an arrangement is that the circuit consumes almost no power when it is not being switched. Indeed, by analysis of the circuit's characteristics it can be shown that the minimum possible supply voltage needed to operate the inverter is about .1 volt at room temperature.

Such remarkably low consumption of power is one reason complementary MOSFET designs have rapidly become the dominant technology for processing information. That dominance has been reinforced by another circuit limit: minimum channel length. I have already described how minimum channel length appears as a limit at the device level, but minimum channel lengths for short-channel MOSFET's also depend on the configuration of the circuits in which they are used. It can be shown that a complementary MOSFET configuration makes possible the shortest channels: between .10 and .15 micrometer. The increased speed that accompanies the narrowing of the channel is one reason for the predominance of the complementary MOSFET technology.

The top rung of the hierarchy is the system level, and there the great challenge is formulating a model that can link lower-level parameters to the system's overall architecture and packaging. The only such model I know of was recently formulated by Brian Bakoglu, who was then one of my doctoral students at Stanford University. The model, which we call SUSPENS, is based partly on a formula known as Rent's rule. Rent's rule, which was worked out in the 1960's and became widely known among computer scientists in the 1970's, relates the number of logic circuits on a chip to the number of connections needed to link those circuits and to the number of pins (leads) needed to link the chip to the rest of the system. A clever extension of Rent's rule can provide an average length for the interconnections on the chip.

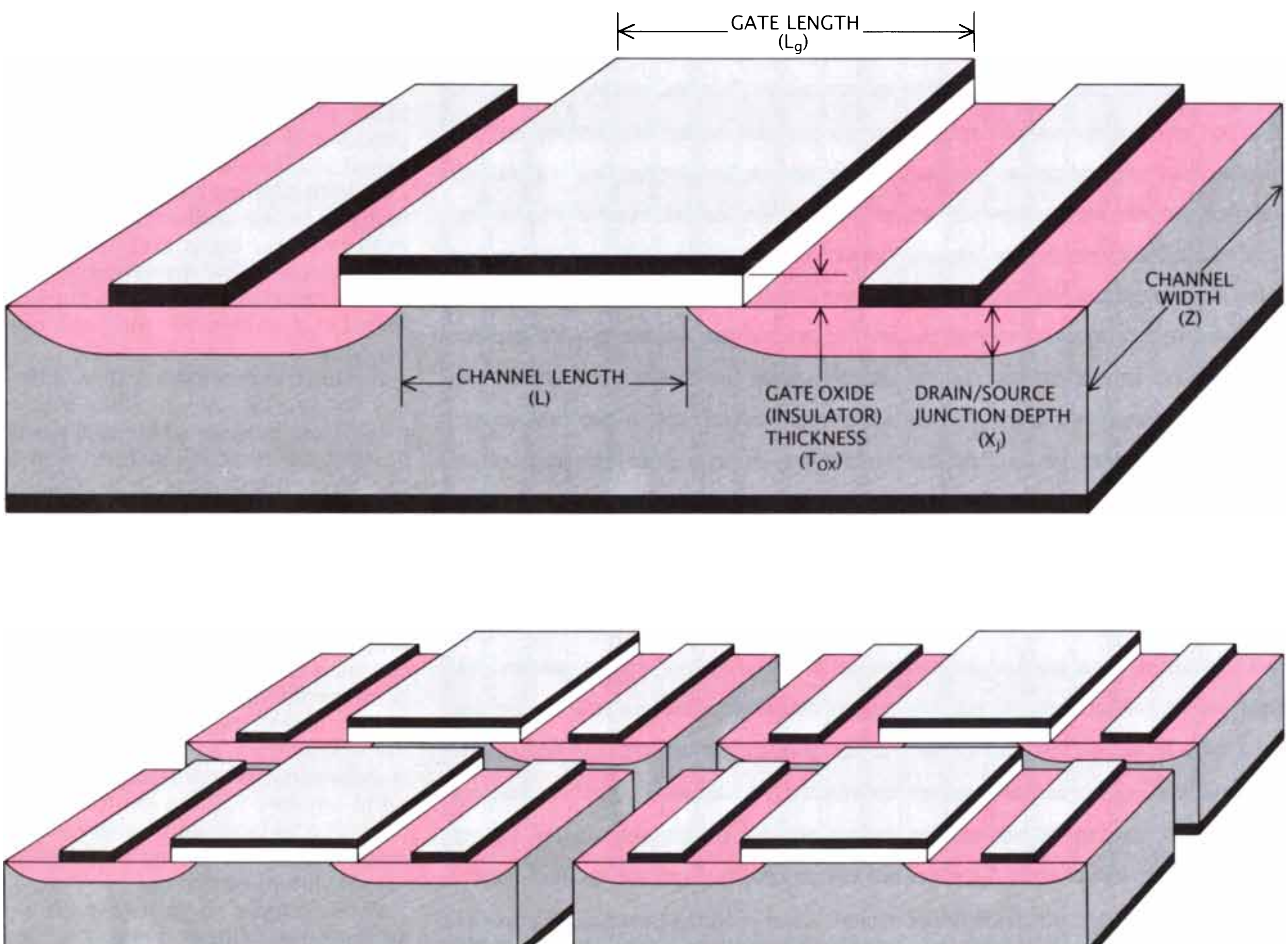
The model based on Rent's rule includes 26 parameters. Of the 26, eight measure chip and system architecture as portrayed by the rule itself. Four correspond to transistor technology, five to chip wiring and nine to connections among modules (packages in the system that contain one or more chips). By inserting into the model values for the 26 parameters a designer can obtain the maximum frequency for the "clock" that paces the system, optimum chip size and total power dissipation. Such information enables a designer to determine the optimum number of transistors per chip and to test the effect that a new technology will have on system performance.

By knitting together limits on many levels, the system model summarizes the hierarchy from the theoretical perspective. The totality of practical limits, on the other hand, can be described by three parameters: minimum feature size, die area and packing efficiency. Minimum feature size is the lateral dimension of the smallest identifiable feature of a MOSFET or a metal interconnection. A prime example is the gate length of a MOSFET, which is somewhat greater than the length of the channel.

In 1960 the average minimum feature size of integrated circuits was about 25 micrometers. By 1980 it had fallen to 2.5 micrometers. If the minimum were to keep falling at the same rate, in 2000 it would have reached .25 micrometer, not much longer than the limits described above at the device and circuit levels. Because of the proximity of those limits, it appears unlikely that the historic rate of decrease—11 percent per year—will be maintained through the 1990's. That conclusion is reinforced by the fact that the minimum feature size has begun to approach the limit of resolution of the optical lithography equipment used to make integrated circuits; the limit of resolution (about .5 micrometer) is determined by the shortest wavelengths of visible radiation.

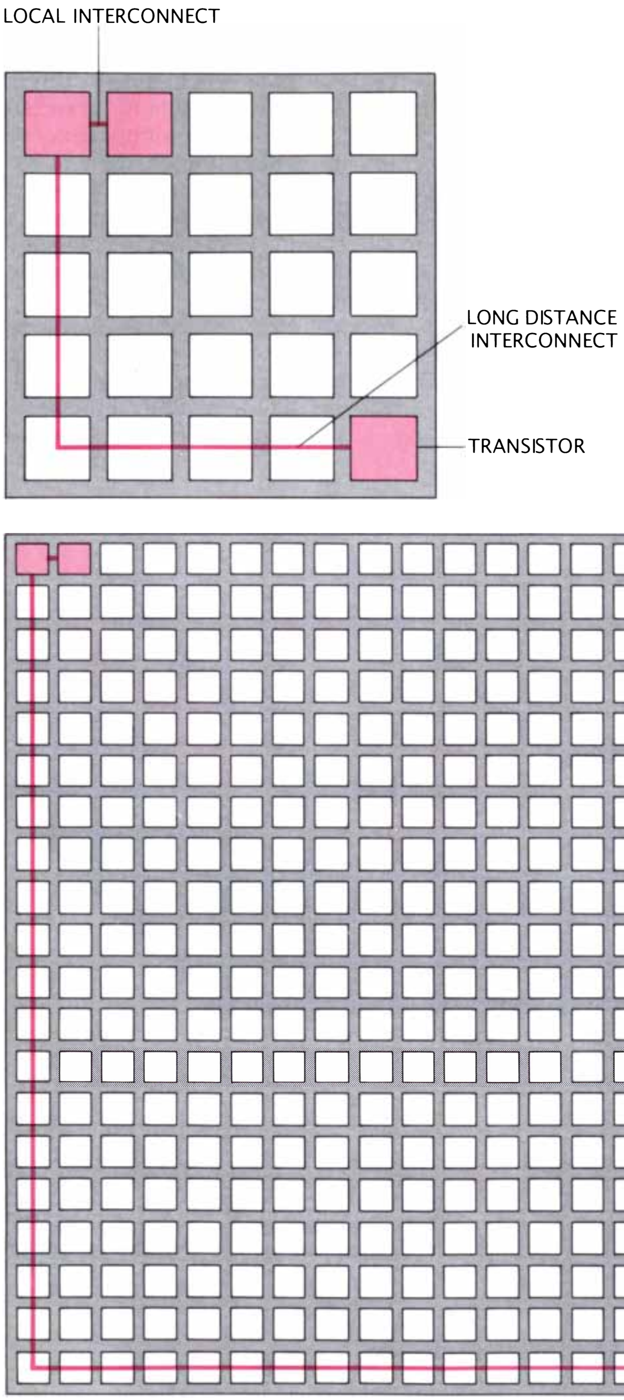
As minimum feature size was decreasing, die area, or the area of an average chip, was steadily increasing. Trends in die area are often analyzed by examining the die edge: the length of one edge of a square chip whose size is typical for its era. In 1960 the die edge was 1.4 millimeters; by 1980 it was eight. The continuation of that rate until 2000 would yield a die edge of 50 millimeters. Yet certain practical limits suggest that the rate will slacken. In optical lithography there is a tradeoff between field of view and resolution. Because chips are generally etched as a unit, the need to resolve smaller features also limits the increase in chip size. A more realistic prediction is a die edge of from 20 to 40 millimeters in the year 2000.

Packing efficiency is, as its name suggests, a measure of how closely transistors can be packed on a chip. Since the 1960's packing efficiency has increased dramatically as the re-



MOSFET SCALING has considerable advantages for transistors. When a MOSFET is scaled down, a scaling factor (S) is chosen. All linear dimensions are reduced by S, as is the supply voltage. Here S is 2. As can be seen in the lower panel, the packing density increases with S^2. Gate delay, the time required for an electron or a hole to traverse the channel, falls with S. Power density remains constant because of the scaling of the supply voltage and current. The energy required for switching decreases with S^3.

sult of improvements in fabrication processes along with innovations in circuits and devices. Until 1972 the increases were rapid. In that year designers ran out of unused real estate on the chip for transistors and connections. Thereafter increases in packing efficiency were slower, depending on ingenious transistor designs, augmenting the number of masking steps in the photolithographic process and even building three-dimensional devices extending above and below the chip surface.



SCALING OF CONNECTIONS between devices is not nearly as beneficial as scaling of devices. As devices have become smaller, chips have become bigger. The chip scaling factor (S_c) measures the proportional increase of an edge of the chip. The illustration shows the scaling of a chip in schematic form. Here S (the proportional decrease in device dimensions) and S_c are both equal to 2. Whereas local interconnects linking neighboring devices decrease with S, long-distance interconnects tend to increase with S_c.

The combined effect of changes in minimum feature size, die area and packing efficiency has been a rapid growth in the number of components on each chip. Indeed, that growth is one of the most remarkable technological trends ever recorded. Beginning at unity in 1959, the number of devices per chip doubled annually. In the 1970's (as a result of the practical limits described above) the rate declined, but only to the still remarkable level of quadrupling every three years. That level should persist until the early 1990's, when the effect of other limits will probably slow the rate of growth again. Nevertheless, the answer to the question posed at the beginning of this article is that GSI—one billion components per chip—should be achieved by the year 2000.

The full impact of GSI, however, will not be felt immediately. Past experience shows that the volume of production of a new integrated-circuit product does not reach its peak for from three to five years after the product has been introduced commercially. (The delay is necessary in order for demand to grow and full production capability to be established.) Hence the ultimate effect of a chip with a billion components is not likely to be felt until after 2005. Thus both practical limits and theoretical limits suggest that the torrid pace of the information revolution will probably be sustained for the next two decades.

Having traversed the realms of theory and practice, I should like to turn away from science and engineering as such to consider history, which can by analogy serve as a guide to the future of the information revolution. Historians have often observed that a commercially successful technology tends to follow an S-shaped developmental curve. At the beginning development is slow and largely confined to the laboratory. After the commercial significance of the technology is realized as the result of initial commercial introductions, there is substantial investment; a period of rapid advance follows. At the top of the curve, as basic limits are reached, development slows and technological obsolescence sets in.

Various signs suggest that the information revolution is following this pattern, as did the Industrial Revolution that began in the 18th century. By drawing analogies between these two technological upheavals, it may

MULTIFLOW
TRACE 7/200

PROPERTY	VALUE BEFORE SCALING	VALUE AFTER SCALING
LENGTH		
LOCAL	L	L/S
LONG DISTANCE	L	$S_C L$
THICKNESS	W	W/S
WIDTH	H	H/S
RESISTANCE		
LOCAL	R	SR
LONG DISTANCE	R	$S_C S^2 R$
CAPACITANCE		
LOCAL	C	C/S
LONG DISTANCE	C	$S_C C$
DELAY TIME		
LOCAL	RC	RC
LONG DISTANCE	RC	$(SS_C)^2 RC$
CURRENT DENSITY	J	SJ

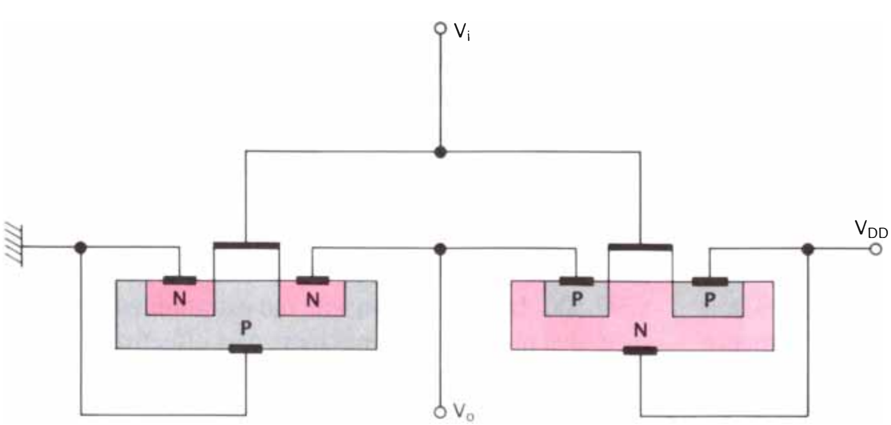
EFFECTS OF SCALING CONNECTIONS include severe disadvantages. As the dimensions of the interconnect are scaled, the delay time of long-distance interconnections (the time required for a signal to traverse the connection) tends to increase exponentially with the product of S and S_c. Current density increases with S, which can ravage the interconnect, leaving empty spaces that cause the connection to fail completely.

be possible to locate the information revolution on the S-shaped curve. I think no one would deny that iron was the most important material of the Industrial Revolution. In that sense it is analogous to silicon, the key material of the information revolution. Iron combined with other elements to yield steel alloys is the analogue of silicon doped with impurity atoms, which forms the starting material for an integrated circuit. The set of analogies can readily be completed at the device level (a transistor and a piston), the circuit level (an integrated circuit and an internal-combustion engine) and the system level (a computer network and a transportation system).

Now, from 1860 to 1900 the annual rate of steel production in the U.S. doubled every four years. At the end of that period the top of the S was reached and the production curve became saturated: from 1900 to 1985 steel production remained roughly constant. Silicon production, on the other hand, is still on the steep part of the S curve. For more than a decade the rate of production has been doubling every two years. Furthermore, as I pointed out above, the achievement of GSI suggests that rapid increases will continue for two more decades. Not long after that, though, the shoulder of the S curve will probably be reached and silicon production may level off, or at the very least its rapid growth may slow somewhat.

It should be noted that even at the top of the S-shaped curve, steel remains the dominant industrial material in terms of tonnage. Yet the production of other materials such as aluminum, titanium, plastics and composites is growing much faster than that of steel. The integrated-circuit industry is likely to follow a similar pattern. Silicon will undoubtedly remain the dominant material. The focus of scientific interest and economic growth, however, will shift to other materials. Among them are cryogenic superconductive integrated circuits, thin films of semiconductors on insulator materials, and substrates formed from several materials by means of molecular-beam deposition. Such possibilities suggest that growth need not end when the top of the S curve is reached.

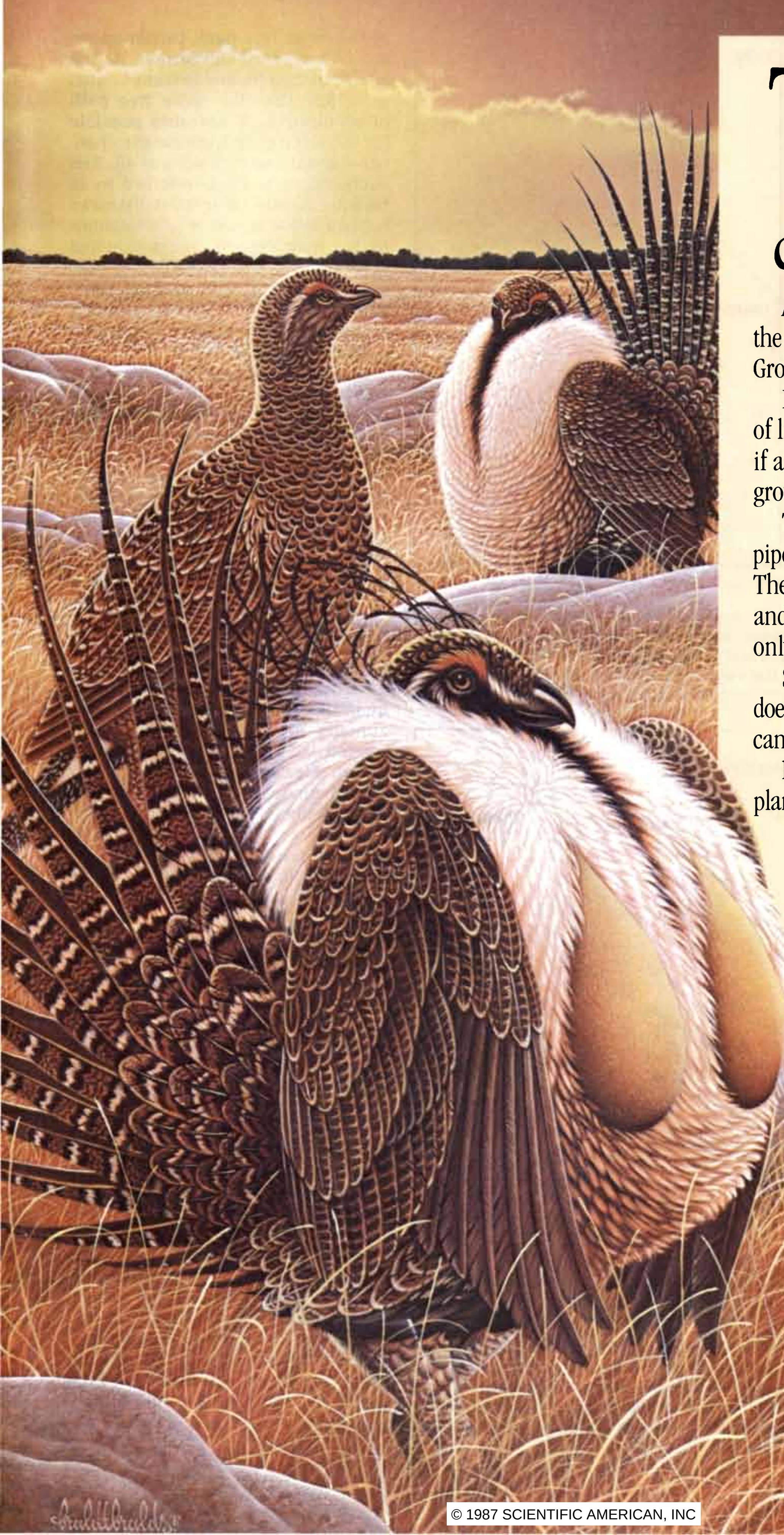
Indeed, an inspiring feature of the history of technology is the appearance of discontinuities: points where the established limits are violated and new vistas suddenly open up. Discontinuities are caused by discoveries or inventions that fundamentally change future prospects. There are several innovations that could have such a dramatic effect on integrated-circuit technology, and they are best understood in terms of the scales of length on which they will have their effects.

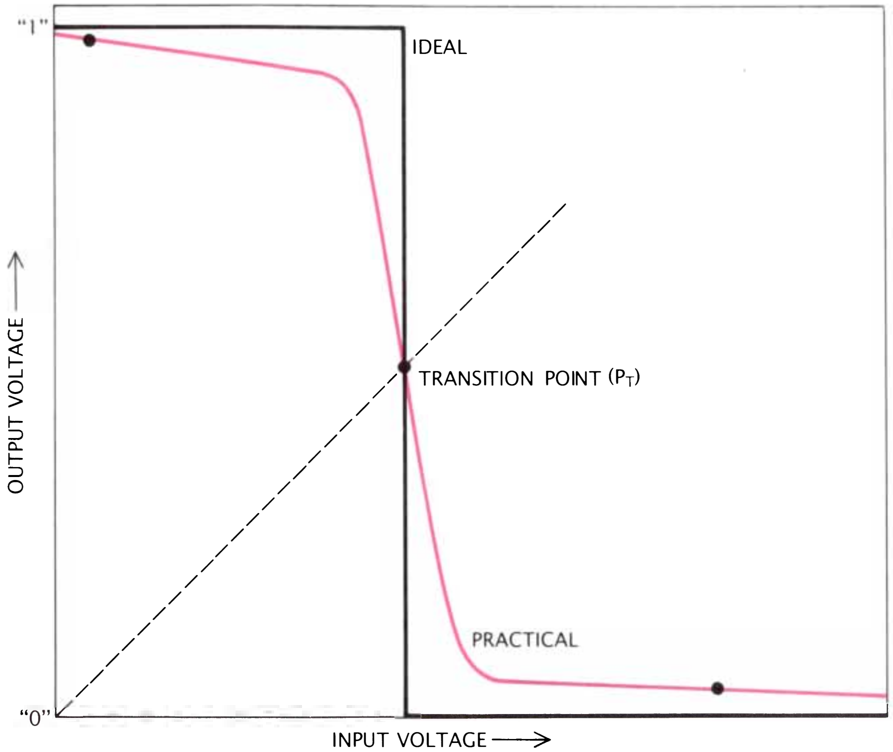


INVERTER CIRCUIT includes complementary MOSFET's. The inverter is a basic logic circuit. It converts a low-voltage input (a "0") to a high-voltage output (a "1") and does the opposite as well. Unless the circuit is switching, one MOSFET is always off (not conducting current) and the other is on (conducting current). Because the devices are wired in series, the circuit consumes very little power when it is not being switched. Low power consumption is a marked advantage of complementary-MOSFET technology.

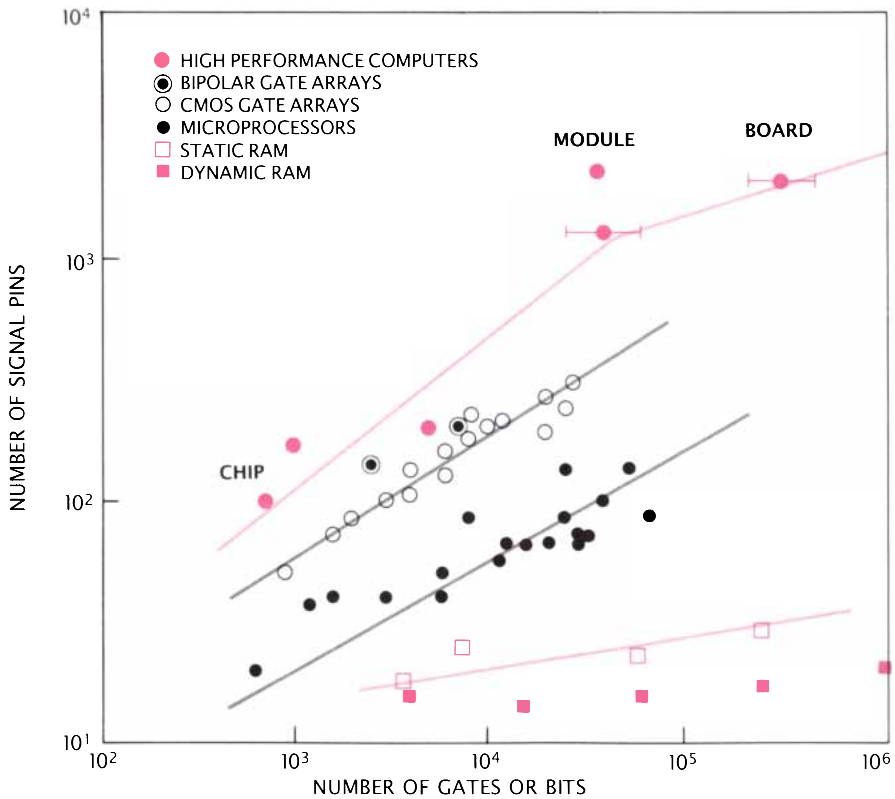
One scale is provided by the distance an electron or a hole can travel in silicon before colliding with a vibrating silicon atom or a dopant atom. That distance is called the mean free path; the longer it is, the faster the carriers travel. The mean free path can be increased merely by substituting gallium arsenide for silicon, as in the MESFET (metal-semiconductor field-effect transistor). A more radical departure is the MODFET (modulation-doped field-effect transistor), in which a thin layer of aluminum-gallium-arsenide is deposited on a gallium arsenide substrate containing no dopant. The absence of impurity atoms increases the mean free path, rendering MODFET's faster than MESFET's, which are in turn faster than MOSFET's.

On a scale just below the length





TRANSFER CURVE for the inverter circuit on page 86 shows output voltage as a function of input voltage. Ideally the curve should be perfectly quantized, or stepped. If that were so, the slope of the curve at the point of transition between a high output and a low one would be infinite. In practice this cannot be achieved, but the curve for the complementary inverter circuit comes closer to it than the curve for any other circuit.



RENT'S RULE describes the relation between the number of logic circuits on a chip and the number of pins (leads) needed to connect the chip to the rest of the system. The rule is an empirical one, worked out in the 1960's as experience with chips was accumulated. Rent's rule forms part of the basis for an overall model linking the properties of materials, devices and circuits with those of the system in which they are employed.

of the mean free path, carrier speed can be increased still further. As the length of the channel becomes equal to or less than the mean free path of an electron, it becomes possible for the electron to traverse the channel without any collisions at all. The collisionless result is referred to as ballistic motion. Over short distances ballistic velocity can be several times higher than the collision-dominated drift velocity that prevails in channels longer than the mean free path. Ballistic transistors constitute another integrated-circuit frontier meriting considerable exploration.

On a still smaller scale is the quantum regime, whose range extends downward from the mean free path, which in silicon is about 10 nanometers (10 one-billionths of a meter). One of the features of the MODFET described above is that it includes two materials joined in a single crystal lattice called a superlattice. The physical properties of one material (gallium arsenide) are such that it acts as a "quantum well," capturing many electrons from the other material (aluminum-gallium-arsenide). By means of superlattice technology it may be possible to construct quantum wells vertically and horizontally in a crystal whose cross section would resemble a checkerboard. Electrons that "tunnel" between squares might be exploited to perform digital operations. Quantum-well devices could be more than 10 times as fast as even the most optimistically imagined MOSFET's of the year 2000.

No three-dimensional quantum-well devices have yet been constructed, but their construction remains an intriguing prospect just over the technological horizon. If the quantum-well devices were made practical, they could send the S-shaped curve moving sharply upward again. Indeed, the current situation could be compared to the one prevailing in the 1950's between the invention of the transistor and that of the integrated circuit. The most rudimentary quantum-well device—the MODFET—has already been formulated. What remains is to extend that concept to three-dimensional structures, provide connections among them and put such devices in the context of new computer architectures. If those things can be done, it may be possible to transcend all the limits currently in view on chip design and fabrication.

How nuclear energy can help defuse the next oil crisis

Nuclear-generated electricity, still the fastest-growing major energy source in America, may be our best defense against another oil crisis.

More and more energy experts are asking the same question: How long before another oil shock torpedoes our economy and threatens our national security?

Oil turmoil

Signs of the next energy crisis:

- U.S. oil imports soared last year, costing the country $27 billion. This year, America's foreign-oil bill is expected to grow even bigger.
- Many oil analysts are saying that in three years or less, as much as 50% of all the oil used in the U.S. will have to be imported. That's a higher percentage than we have ever imported before, even during the oil crises of the 1970s.
- A whopping two-thirds of the world's oil lies under the sands of OPEC nations.

The need for nuclear

Nuclear electricity is a *domestically produced alternative to foreign oil.* Not just at the power plant, where nuclear energy is used instead of oil to generate electricity, but wherever Americans choose electricity (instead of oil) to heat their homes or run their factories.

The 1987 special report on U.S. energy security, ordered by the President and prepared by the U.S. Department of Energy, states that without electricity from nuclear energy, the United States "would be using more oil, paying more for each barrel of it, and feeling much less secure about its energy outlook."

The more we use our own nuclear electricity, the less we'll have to rely on energy from unstable regions of the world.

Nuclear energy for a secure future

With over a hundred operating plants in the U.S., nuclear energy is now our second leading source of electricity. But in spite of all that we have accomplished, the threat of foreign oil dependence remains. Difficult choices still need to be made, but one fact is clear: the more we develop our own energy sources, the more we can control our own destiny.

For a free booklet on energy independence, write to the U.S. Council for Energy Awareness, P.O. Box 1537 (FQ20), Ridgely, MD 21681. Please allow 4-6 weeks for delivery.

Information about energy
America can count on
U.S. COUNCIL FOR ENERGY AWARENESS

Programming for Advanced Computing

Parallel computers pose the major challenge: they demand programs that do many things at once. How do we organize a complex beehive of activities to form a coherent whole?

by David Gelernter

Computers are too slow. Even the fastest conventional machines can't manage more than a couple of hundred million multiplications per second. Where does that leave *hard* problems—simulating the interactions of systems of atoms, for example? Small atomic simulations can run for weeks on the fastest of current-generation machines. Many other important problems are just as hard or harder, from simulations of the gravitational interplay of celestial objects to the fluid-flow problems that are central to weather forecasting to the fast searches of enormous knowledge bases that are becoming important to artificial intelligence.

As a stopgap the traditional sequential computer, which executes its instructions one at a time through a central processing unit, can be made faster by improvements to its circuitry and internal organization. But the long-term answer is parallelism: hooking many computers together and focusing them all on a single problem. Ten identical computers working together on a problem will solve it, in the best case, 10 times faster than a single computer working alone. Accordingly parallel computers are now for sale that incorporate tens or hundreds or thousands of individual processors, or subcomputers. Building a parallel computer poses a number of engineering problems, many of them related to establishing ways for the processors to communicate [see "Advanced Computer Architectures," by Geoffrey C. Fox and Paul C. Messina, page 66]. Unlike strategies for making conventional computers faster, which entail relatively minor adjustments in programming style, it also poses a hard software problem.

To start out, think of a computer program as a kind of machine. A parallel program is a machine that differs from a sequential one as radically as a parallel computer differs from its sequential counterpart. In one sense, of course, a program is merely a document—a set of instructions to be executed by the computer, written in a language both the programmer and the computer can understand. When the instructions are carried out, however, the program becomes an event, a process that transforms data into results. Viewing a program as a machine reconciles its two aspects: the text is the machine before it is turned on and the event is the machine in operation.

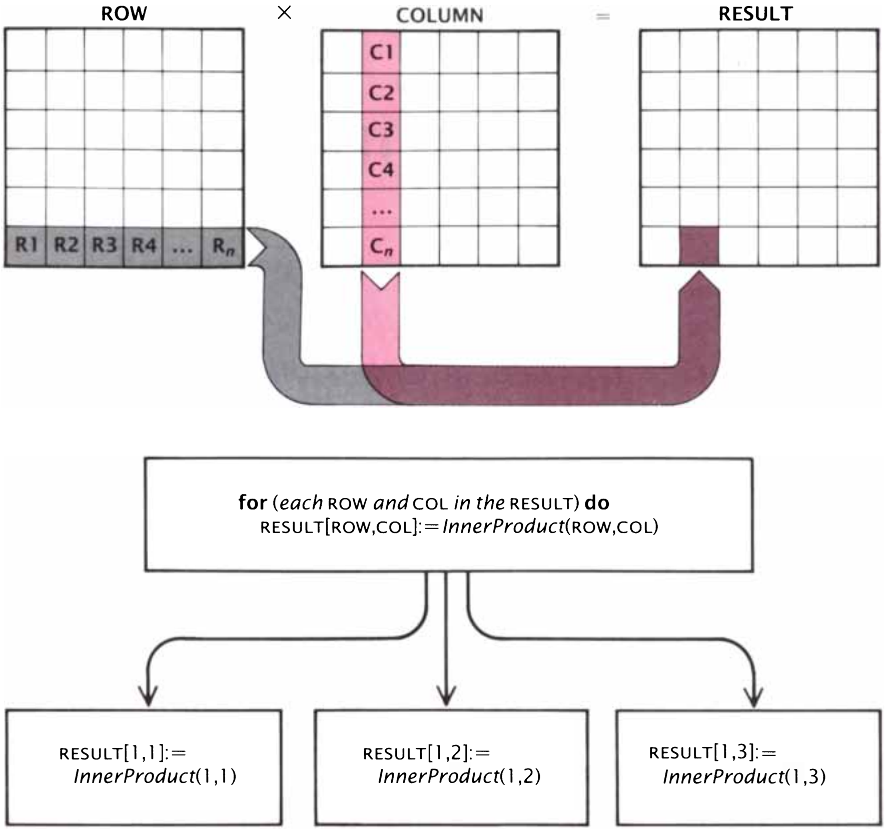
There is nothing a program can do that a mechanical "computing engine" built of gears and sprockets cannot. The software machine is much easier to construct, though; it can be built on a computer terminal using a programming language. If a program itself is a machine, where does that leave the computer—the physical machine on which the program runs? It becomes nothing more than a kind of power source for the software machine. Thus a program can usually be made to run on a wide range of dissimilar computers, and its characteristics can be largely independent of the details of the physical computer it inhabits at any given moment.

We can therefore discuss parallel computing in terms of programs alone, glancing at the physical machinery only in special circumstances where software and hardware are intertwined. A conventional program does one thing at a time, but a parallel program must create and manage many simultaneous threads of activity. It must divide a task into many parts, and it must coordinate them by establishing an orderly flow of instructions, data and results.

To build a new kind of software machine, we need a new kind of construction kit: a new programming language. Conventional languages lack the vocabulary and syntax for specifying parallelism. They enable a programmer to write, roughly speaking, "Do *A*, then *B* and then *C*" but not "Do *A*, *B* and *C* at the same time." But the practical requirement for a new language is only the surface manifestation of a deeper issue. A programming language doesn't leap into existence by itself; it is the expression of a particular *software model,* of a particular approach to constructing programs. So the fundamental question becomes "What should a parallel program look like?" In the many lines of investigation that are now under way I see three broad answers to the question.

The first school sidesteps the issue altogether. Instead of proposing new software models, it relies on the

HONEYBEES EXECUTE A PARALLEL PROGRAM to maintain the hive. Like the bees—individually feeble agents working in concert—a parallel program can bring large amounts of computing power to bear on a problem by establishing multiple processes, or loci of activity. The bees coordinate their activities through visual and chemical signals; similarly, processes in a parallel program must communicate to work together.



MULTIPLICATION OF TWO MATRICES is typical of the problems for which a conventional program can automatically be transformed to a parallel one. Each element in the result represents the "inner product" of a row in one matrix and a column in the other, computed by multiplying each element in the row by the matching element in the column and summing the products (*top*). A conventional program for matrix multiplication centers on an inner-product subroutine that does the pairwise multiplications, nested inside a loop that computes successive inner products (*middle*). Because each inner product can be computed independent of the others, a parallelizing compiler might automatically convert the loop into an array of simultaneous activities (*bottom*).

automatic generation of parallel versions of ordinary programs. One approach within this school extends the strategy by which a single program can be made to run on many different computers. A computer cannot in general understand a programming language without help. Setting a variable equal to the product of two numbers requires just a single line in a "high level" programming language such as Fortran or Pascal, but it entails a number of machine operations, which vary depending on the design of the computer. The command must therefore be translated into the right step-by-step instructions in machine code, the computer's own primitive, digital language. A second program called a compiler usually does the translation.

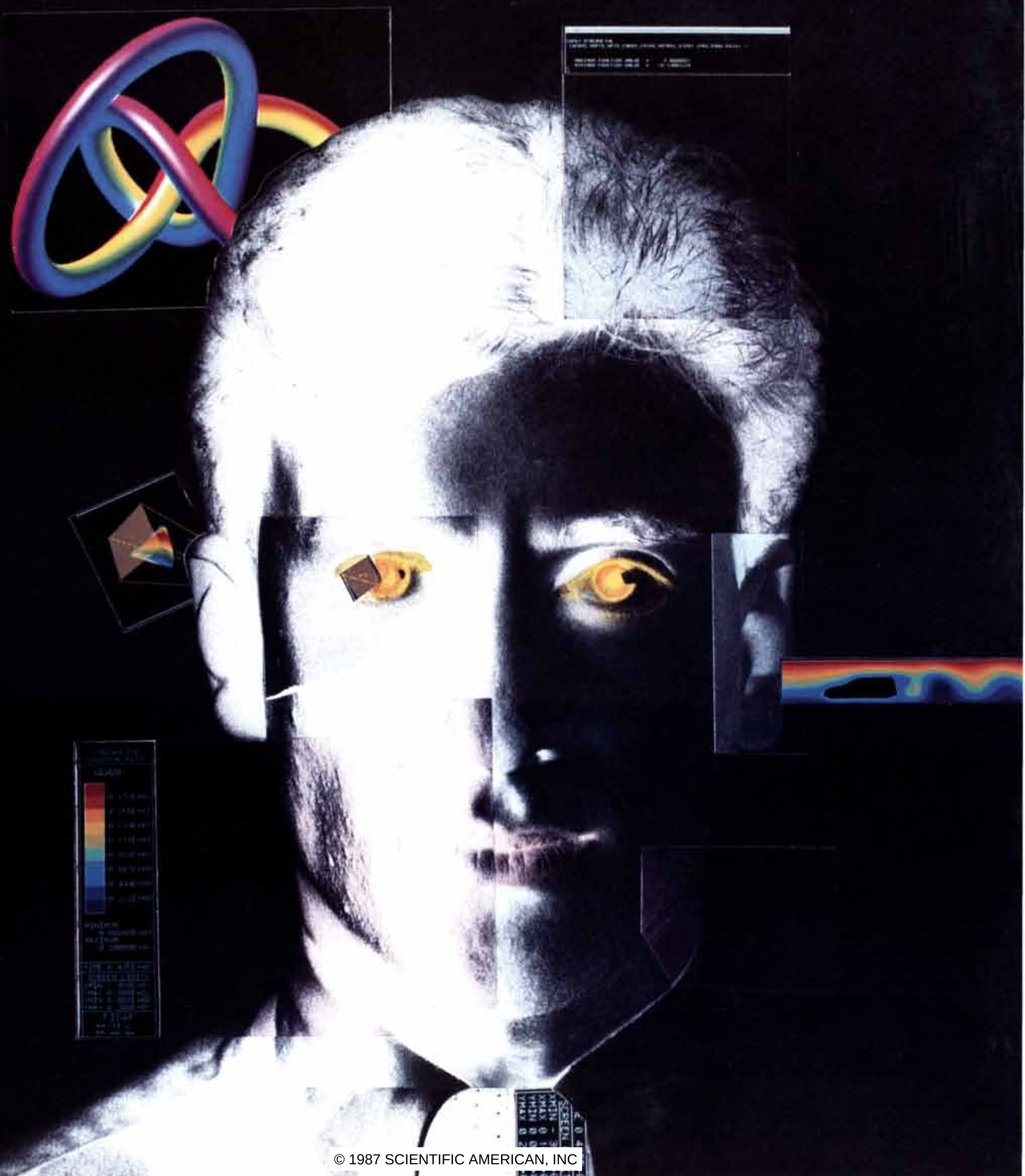
Driven primarily by the famous "dusty deck" problem—the colossal investment of time and money represented by existing programs (a "dusty deck" is an old deck of punched cards)—workers at the University of Illinois, Rice University, Multiflow Computer, Inc., the IBM Corporation and other places have been developing compilers that automatically transform sequential programs into parallel machine code. A parallelizing compiler finds the parallelism hidden in a sequential program by searching its text for operations that in fact can be done simultaneously, even though the program specifies they be done one by one. It then generates code that reflects the implicit parallelism it has found.

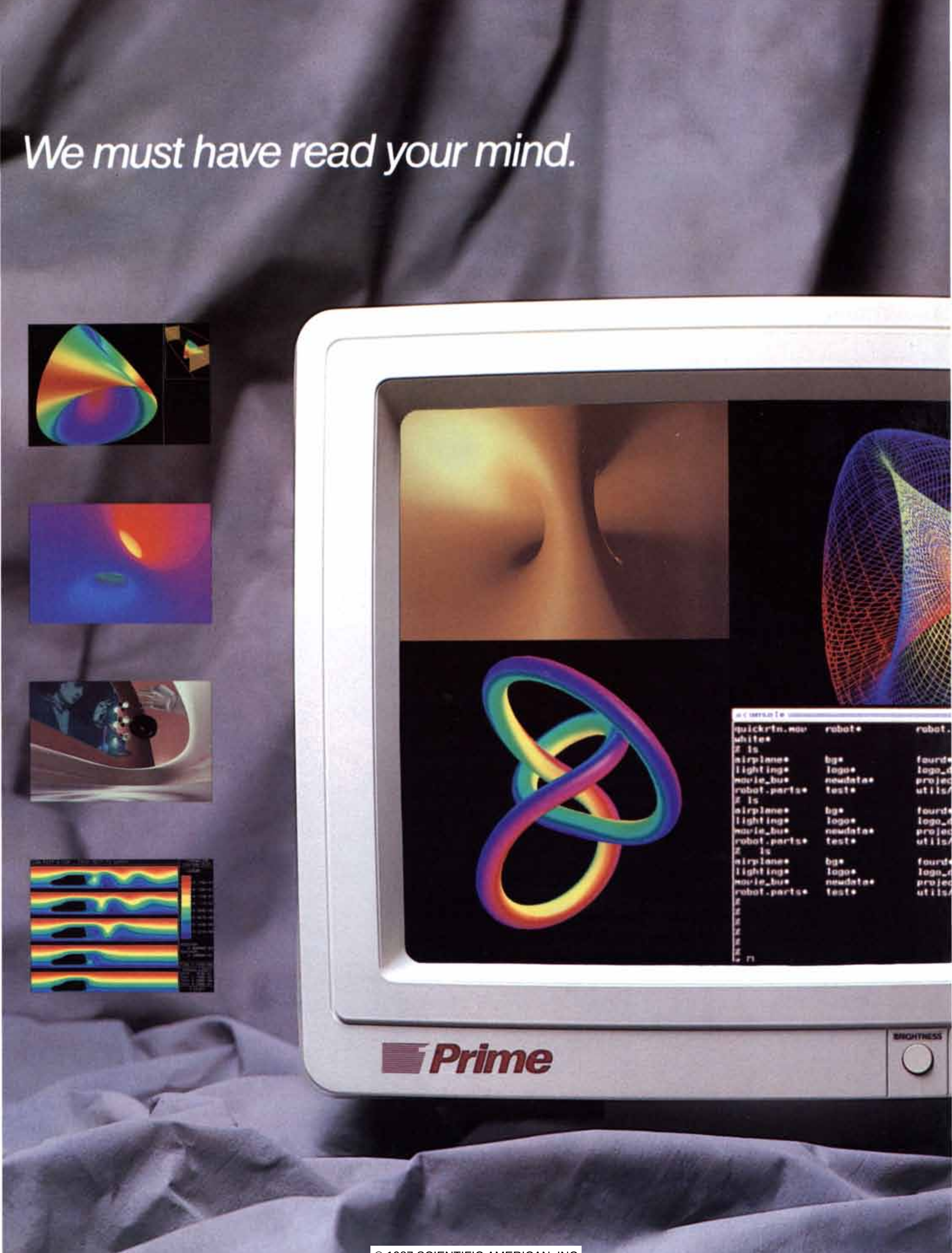
For example, many programs center on a computational loop in which the same operation or series of operations is performed for each of a succession of values. A parallelizing compiler can often tell by inspection whether loop iterations really must be done sequentially, as the program prescribes. If each new calculation depends on values calculated in a previous iteration, sequentiality is required. Otherwise the compiler might well be able to convert the loop into parallel machine code, allowing the computer to do all calculations simultaneously.

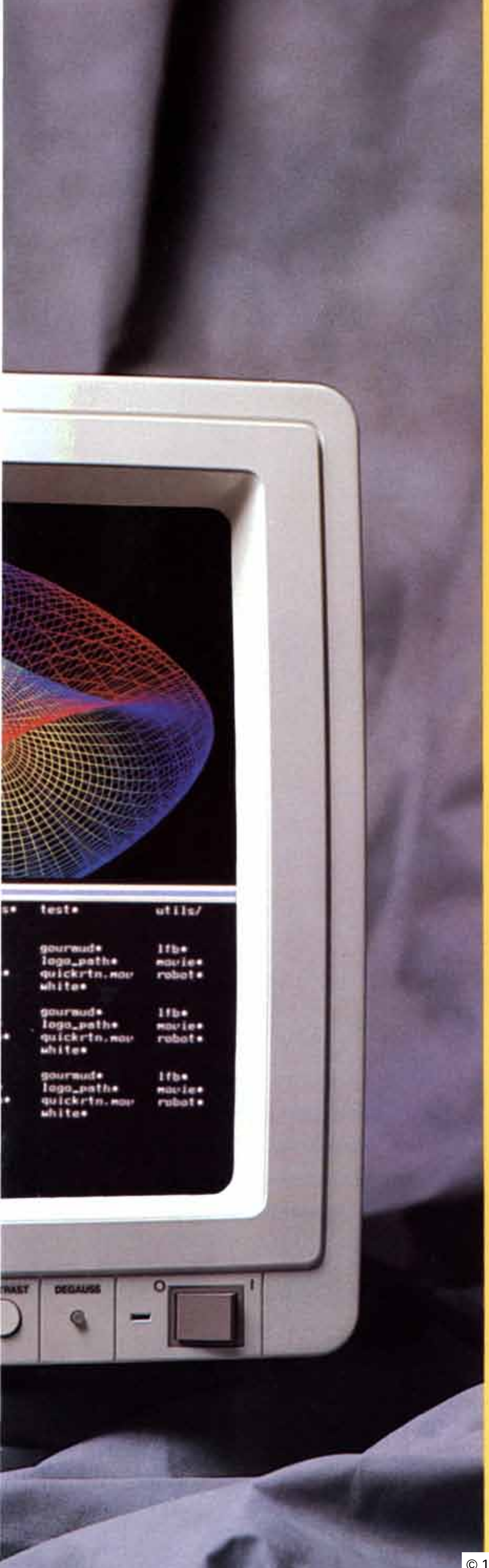
Consider a program for multiplying two matrices, a basic procedure in numerical computing. Each element in the result matrix represents the "inner product" of a row in the first matrix and a column in the second. To get the element in row *i*, column *j* of the result matrix, we multiply each element in row *i* of the first matrix by the corresponding element in column *j* of the second matrix—first element times first element, second times second and so on—and sum the products. A conventional program for matrix multiplication might include two loops, one inside the other—an inner loop for the pairwise multiplications and an outer one for computing successive inner products. On each iteration the outer loop takes a row and a column, does the multiplications by way of the inner loop and sums the products.

Because each iteration is logically autonomous, a parallelizing compiler might convert the loop into parallel code. If we are multiplying, say, two square matrices, each with 500 elements on a side, then by parallelizing the outer loop and doing the inner-product computations simultaneously we can speed up the computation by—in the best case—250,000 times. ("In the best case" here means "on a computer containing 250,000 processors." So far no general-purpose machine this large exists; the smaller number of processors in existing machines sets an upper limit to the speedup by forcing each processor to do several of the potentially parallel steps in sequence.)

Dataflow computers offer another way to take advantage of implicit parallelism. These machines are usually not aimed at the dusty-deck problem, since they tend to be conceived together with a new style of programming. But they address the same basic question: how to avoid explicitly parallel software models. A dataflow-machine program does not specify an execution sequence. The only constraints on its order of execution are those that are forced by the program's logical structure. Any two instructions can be executed by separate processors simultaneously, unless one instruction needs a value the other is directly or indirectly responsible for computing. In that case its execution is postponed until the datum arrives. Data flow from instruction to instruction as the program executes.




We must have read your mind.
Prime



Imagine a workstation with real-time 3D color graphics and all the compute power your applications require.

Prime Computer, a Fortune 500 company and a pioneer in providing engineering and scientific systems, introduces the workstation of your dreams. The PXCL™ 5500 workstation. For over 15 years Prime has struck the perfect balance between engineering experience and innovation. It's only natural that we now bring you a workstation that exemplifies this symmetry and raises it to a higher dimension. A system that perfectly balances high computational processing power with three-dimensional color graphics.

Some workstations give you extraordinary 3D graphics. Others may offer exceptional computing capability. But the PXCL 5500 workstation has the unsurpassed power and sophistication to give you the 3D color graphics you need, when you need them.

Your ideas never looked so good so fast.

The PXCL 5500 system has the graphic capability to make your ideas look even better than you'd imagined. With its customized geometry circuits, it is capable of transforming and rendering the most complex images, turning your ideas into pictures that can represent potential products. At the heart of the system is a variety of custom VLSI dedicated graphics processors, true color capability (a palette of 16.7 million colors), and full RGB and look-up table color selectable on a per-window basis.

Power beyond pictures.

And to all this incredible color, dimension and detail, the PXCL 5500 workstation applies remarkable speed. Over 145,000 3D floating point transformations per second. No more waiting for your workstation to create 3D shaded images or complex simulations. Because the PXCL 5500 system is so interactive, you may find your workstation waiting for you.

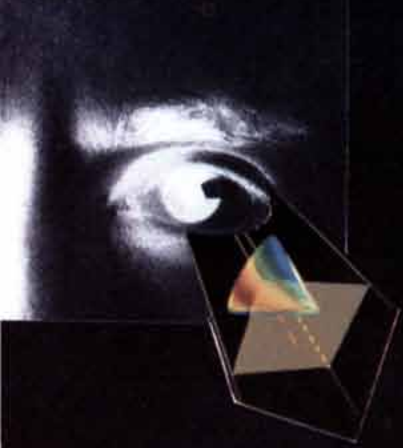
The speed of the PXCL 5500 workstation isn't limited to graphics. Featuring a 10-MIPS CPU engine, with up to 16Mb of main memory, and using the most advanced RISC technology, the PXCL 5500 system can easily run the most demanding engineering and analysis programs. And all that power is housed in a slim 26" cabinet that fits easily under a desk.

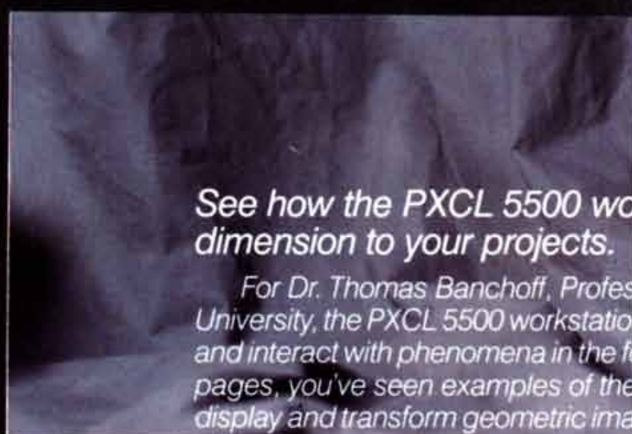
Imagine the possibilities.

With an operating system based on UNIX™ System V.3, the PXCL 5500 system makes it easy to develop or port your own software or run a wide array of applications ranging from industrial design to computational fluid dynamics to computer-aided manufacturing. And with Ethernet - TCP/IP communications, this workstation will fit easily into any engineering and scientific computing environment.

Whether you're a software developer or an end user, you can easily imagine the benefits you'll reap from this perfect balance of speed and power. Working with the Prime PXCL 5500 workstation, your possibilities are truly endless.

Prime





See how the PXCL 5500 workstation can add a new dimension to your projects.

For Dr. Thomas Banchoff, Professor of Mathematics at Brown University, the PXCL 5500 workstation is a means to calculate, render and interact with phenomena in the fourth dimension. On the previous pages, you've seen examples of the PXCL 5500 system's ability to display and transform geometric images. Now you can read how this ability is transforming the way we see our world. As well as the other worlds that lie beyond our vision.

To receive Dr. Banchoff's fascinating report, "Discovering the Fourth Dimension," simply send in the coupon below.

- ☐ *Please send me a free copy of Dr. Thomas Banchoff's report, "Discovering the Fourth Dimension."*
- ☐ *Please send me additional information on Prime's PXCL 5500 workstation.*
- ☐ *Please have a sales representative contact me.*

*Name*____

*Title*____

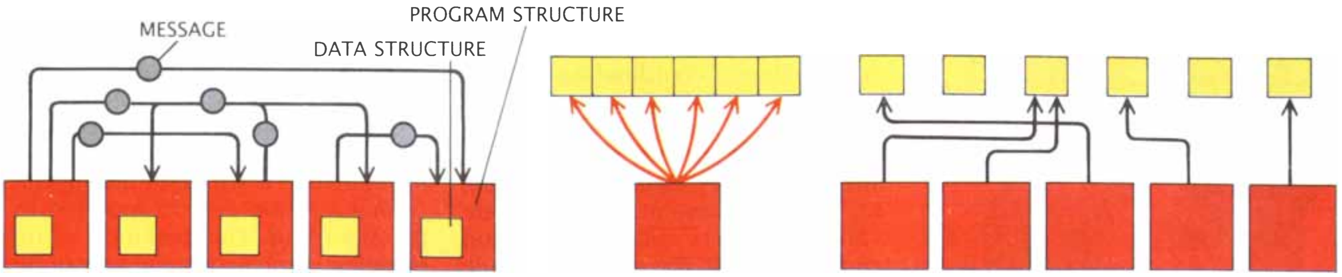
*Company Name*__________________________________

*Address*______________________________________

*City, State, Zip*________________________________

*Phone*____

Now you can read about what the PXCL 5500 workstation did for Dr. Banchoff. And imagine what it can do for you.



STYLES OF COMMUNICATION distinguish parallel-program models. The message-passing model establishes many subprograms (*left*), each with its own data structures (the arrays, lists and other patterns of data that are transformed by active program structures). The subprograms are barred from one another's data structures; they can communicate only by sending messages. In a second approach, seen in programs for a computer known as the Connection Machine, many identical program structures carry out a single instruction, issued by a "boss" program, to transform many parts of a data structure (*middle*). A related approach allows simultaneous operations on the same data structure by unsynchronized program structures (*right*).

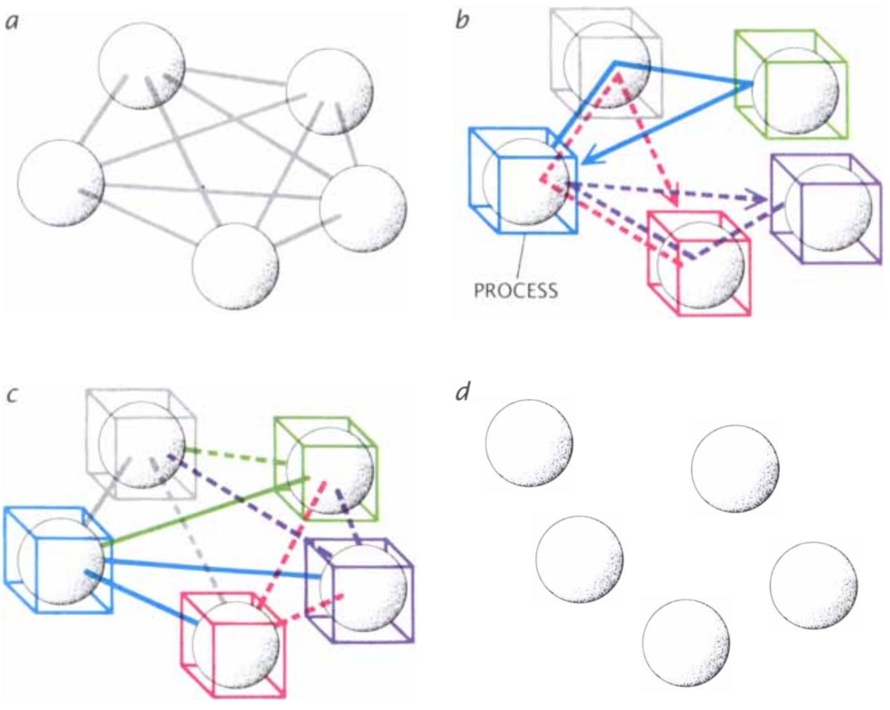
Thus (in a sense) the computer itself and not the program sets up parallel threads of activity. The languages that are usually associated with dataflow machines, "functional languages," make it easier for the machine to decompose the program into parallel activities. These languages simplify the data relations among statements by allowing a programmer to set the value of a variable once only. After a variable's value has been set, other parts of the program can use the variable freely and simultaneously without worrying about whether and when its value will change.

Work now under way (at the Massachusetts Institute of Technology, the University of Manchester and the University of Utah, among other institutions) on dataflow machines and functional languages will shed light on the workability of this automatic approach to parallelism. It should be noted, too, that functional languages are not strictly wedded to dataflow machines. Some investigators feel that their simplicity and closeness to purely mathematical models of programming make them interesting in their own right, regardless of parallelism. Others (Paul Hudak of Yale University is a notable example) are using them as clean and elegant bases for explicitly parallel languages—our next topic.

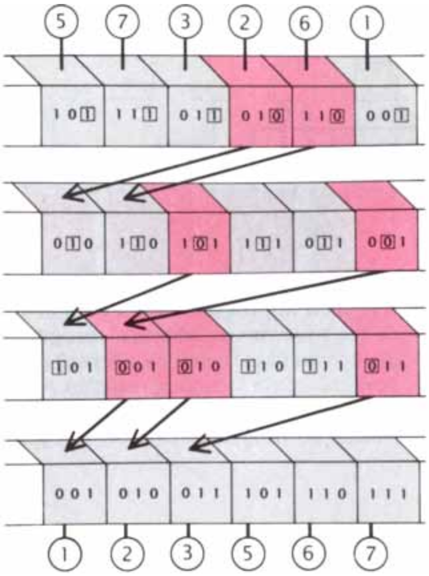
If parallelism can be found automatically, why go any further? Why bother to create new, explicitly parallel software models? In fact automatic transformations work well only for certain kinds of programs. A parallelizing compiler, for example, rewrites a sequential program in parallel machine code on the basis of the program's expected behavior. Many programs—numerical computations in particular—do unfold in a predictable way, and so a compiler can reliably determine which parts of the task can be done in parallel. Some programs, though, have execution patterns that are complicated or unpredictable, and here the implicit parallelism is harder to discover. Many artificial-intelligence programs, for example, seem to fall into this group.

Automatic transformations have a more basic drawback: they can discover the parallelism latent in an existing algorithm (the procedure by which a program solves a problem), but they cannot invent new, fundamentally parallel algorithms. Programmers who are freed to think in parallel sometimes invent entirely new ways of solving problems. In order to embody their inventions in working programs they need languages that allow parallelism to be expressed—languages based on parallel software models.

The second and third answers to my initial question, then, are embodied in two broad approaches to ex-



MESSAGE-PASSING PROGRAM computes the relative motions of several objects interacting by gravity (*a*). One process is in charge of each object. In each computational cycle each process circulates a message giving the position of its object to half the other processes and is visited by similar messages from the remaining processes (*b*). On receiving a message, each process computes the force between its object and the object whose process dispatched the message and adds the result to the message. Thus each process learns about half the forces affecting its object from the returning message and computes half itself (*c*). Each process can then update the position of its object (*d*).



PROGRAM for the Connection Machine sorts a sequence of numbers, each of them assigned to a single process. Each process simultaneously checks the last position (the least significant bit) in the binary version of the number. Numbers with a 0 move to the head of the sequence: they are reshuffled among the machine's memory elements. The procedure is repeated for each bit position; numbers with the same bit retain their relative positions. After a number of steps equaling the number of positions in the largest value, the line is sorted, smallest value first.

plicit parallelism. They can be distinguished by the mechanisms they establish for communication among the parallel threads of activity.

The first school envisions a collection of parallel activities, each resembling a self-contained, conventional program. Each of these "processes," or subprograms, consists of program structures and data structures—arrays and lists, for example—just as a conventional program does. The program structures are active: they execute instructions that transform the passive data structures. (The algorithm is embodied in the program and data structures jointly.)

Processes that are collaborating on a problem will ordinarily need to share data. But here data structures are sealed within processes, and so processes cannot draw on one another's data directly. Instead they exchange messages. When one process has data for another, it generates a message and hands it to a process of a different kind, a message-delivery program. The message-delivery program routes the message to its destination. The scheme adds complexity to the program as a whole: it means that each process must know how to generate messages and where to send them. But it also solves the difficult problem of coordinating communication among the processes.

Suppose processes could refer directly to one another's internal data structures. They would be free, then, to read inconsistent or incomplete values that were still in preparation and not ready for use by outside processes. Requiring all communication to take place through messages solves such coordination problems; messages are generated only when complete and self-consistent data are available.

The complexity of most message-passing schemes means they are best suited to problems that break down into relatively big, largely independent units whose communication needs can be reduced to a simple and regular exchange of messages. Consider, as an example, simulating the movements of n objects as each object exerts gravitational attraction on the rest. The task can be stated simply, but it requires much computation; simulations of large systems unfold slowly on conventional computers. A good way to make them run faster is to compute the forces on all n bodies in parallel, update their positions and repeat the computation.

One message-passing approach to the problem, described by Charles L. Seitz of the California Institute of Technology, creates n processes, one for each object. During each computation cycle each process circulates a message containing the current position of its object to half of the other processes and receives similar messages from the other half. Suppose a message from P_j (the process in charge of object j) makes its first stop at P_k (the process in charge of k). P_k then computes the current forces between k and j, adds this information to the message and relays it to P_l. The message eventually returns to P_j carrying a record of the forces between j and half of the other objects in the system. Similar messages from the remaining objects in the system have meanwhile visited P_j, carrying the information it needs to compute the rest of the forces affecting j. P_j can now calculate the net effect of all forces on j's position.

The substantial work required of each component process means that this n-body solver, like most message-passing software, is best suited to a fairly coarse-grained machine—one with powerful processors, but fewer of them than are found in a "fine-grained" computer. Systolic programming, a radical variant of message passing, is an exception to the coarse-grained rule. A systolic program dispenses with the message-delivery process; it runs on a machine that is wired for the specific communication pattern required by the algorithm. The geometry of the machine mimics the geometry of the program. Data are pumped rhythmically ("systolically") through the fixed network. The fixed communication patterns mean that systolic computers are usually fine-grained machines that are best suited to solving problems with many simple parts. A hard-wired communication scheme also limits them, ordinarily, to special purposes.

In message-passing programs every data structure is sealed inside some process. In the other large and diverse set of approaches to parallel programming many processes can share access to the same structure. In one instance of this approach every element of a large data structure can be modified simultaneously by a string of identical processes taking orders from a single "boss" process. In another, rules are established by which many processes can gain access to common data on their own initiative.

The first approach is tied to specific hardware; it is embodied in several programming languages designed for a parallel computer called the Connection Machine [see "The Connection Machine," by W. Daniel Hillis; SCIENTIFIC AMERICAN, June]. The Connection Machine is a fine-grained parallel computer that can be described as an "active memory." The machine's memory is divided (in what is now the largest model) into roughly 65,000 small pieces, each one controlled by its own simple processor. Connection Machine programs mimic this structure; they rely on data structures that are distributed over the active memory, one data element per processor. Such data structures in effect transform themselves as each processor carries out the same instruction in lockstep synchronization. Thus a program might, say, multiply a set of 10,000 values by 3 simply by storing each value in a separate active-memory element and issuing the instruction "Everyone multiply your contents by 3."

A surprisingly large number of procedures work well in this environment; one example is sorting numbers, a basic operation throughout computer science. One sorting program for the Connection Machine is based on an algorithm called radix sort. In the Connection Machine's version (which is derived from a related sequential version), the unsorted numbers are arranged one to a processor—that is, one to an active-memory element. Every process simultaneously checks the least significant bit in the string of 0's and 1's that represents the number in binary form. Numbers with a 0 in this position move forward in the sequence; numbers with a 1 move back. (A number can hop directly from one active-memory element to any other.) The relative positions of numbers with the same bit are maintained during the reassortment. The procedure is repeated for the next bit, and after a number of iterations equal to the number of bits in the largest number an ordered sequence is produced, smallest number first. Up to a limit set by the machine's 65,000 processors, the procedure is equally fast no matter how many numbers are sorted.

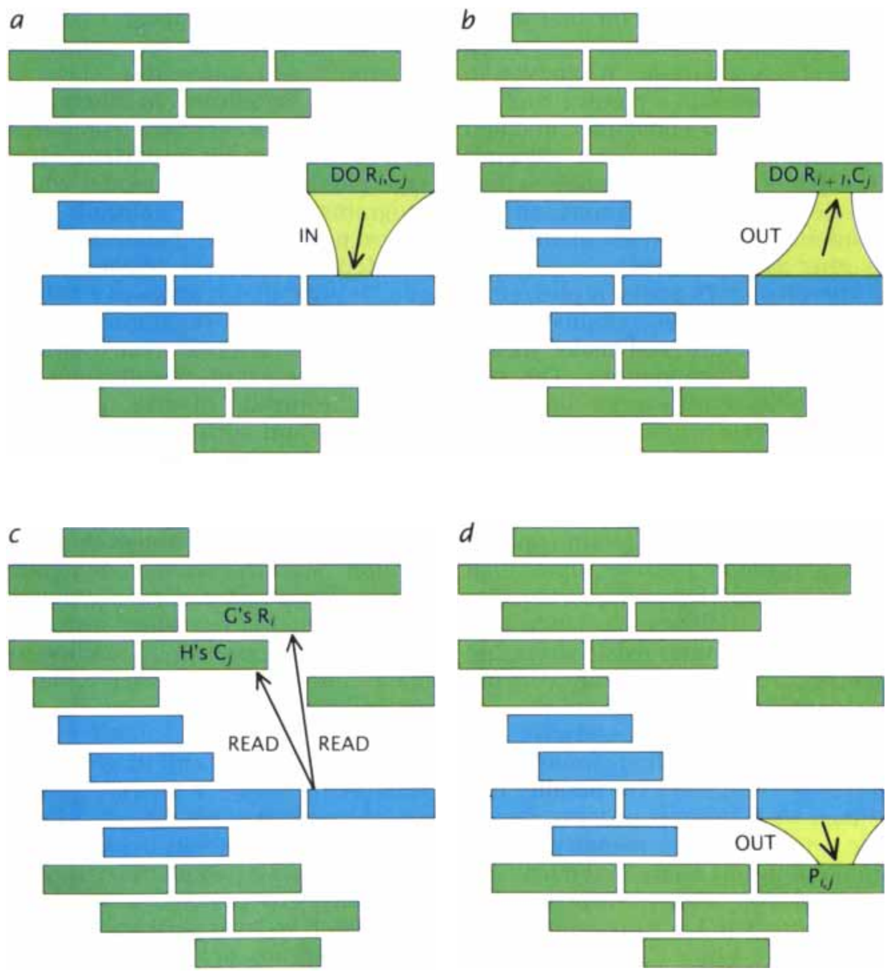
The Connection Machine allows many things to happen to the same data structure at the same time, and it coordinates all this activity by providing a single boss process and a chorus line of faithful workers to carry out orders. Another shared-data-structures approach handles the coordination problem by allowing a roving process to seize a temporary monopoly on the chunk of data it is transforming, thereby preventing other processes from acting on the same data at the same time. In contrast to the neatly regimented style of Connection Machine programs, this second approach is adapted to more loosely organized, heterogeneous problems. It is based, moreover, not on a specific machine but on a construct unrelated to hardware called tuple space. The approach is embodied in a language called Linda, developed by Nicholas Carriero, me and others in our group at Yale.

Picture tuple space as a formless bag filled with interacting objects called tuples—where a tuple is simply an ordered set of elements. Tuples come in two kinds. Active tuples play the role of processes in the message-passing model: they do the work of computation. They communicate by generating, reading and consuming passive tuples. But unlike messages in a message-passing program, passive tuples are integral parts of the tuple-space program: they hold its shared data structures.

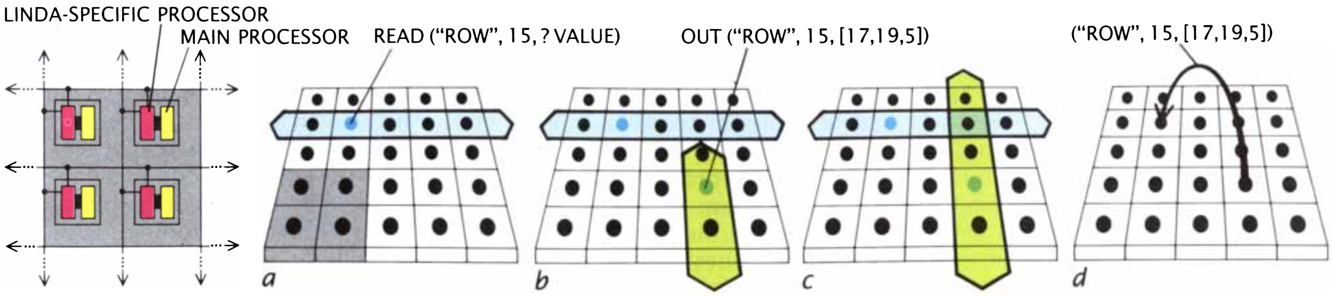
Passive tuples come and go as the program executes, like notes tacked on a bulletin board and then untacked again. Active tuples that need data can either read passive tuples or consume them; as active tuples develop results, they generate new passive tuples (they tack up new notes). An active tuple that has finished computing turns into a passive tuple holding the computation's final result. In Linda, then, a parallel program is a swarm of simultaneously computing active tuples surrounded by and ultimately blending into a haze of passive tuples.

A Linda program for matrix multiplication might store the rows of one matrix and the columns of the other one in passive tuples, one row or column per tuple. Active tuples repeatedly read a row, read a column, compute their inner product and dump the result in a new passive tuple. The more active tuples there are, the faster the program runs (up to a limit determined by the size of the hardware and the size of the problem). A different solution might take advantage of the fact that active tuples turn into passive ones when they have finished computing: the program would represent the result matrix as a series of active tuples, each of them responsible for computing one inner product. The active tuples simultaneously calculate one inner product each and then turn into the result.

Because tuple space makes it easy for tasks to be divided on the fly and for a program's shape to change as



TUPLE-SPACE PROGRAM for matrix multiplication divides the data (the contents of *G* and *H*, the matrices to be multiplied) among objects called passive tuples (*green*) and divides the work of computing inner products among active tuples (*blue*). To find out which inner product should be computed next, an active tuple consumes a passive tuple specifying the row and column to be multiplied (*a*). It updates the data in preparation for the next active tuple looking for work and regenerates the passive tuple (*b*). It then finds and reads the tuples holding R_i and C_j, the row and the column (*c*). After computing the inner product, $P_{i,j}$, the active tuple generates a new passive tuple to hold the result (*d*). (Other active tuples are busy computing other inner products.) IN, OUT and READ are operations in Linda, a language developed by the author and his colleagues.



CREATING A TUPLE SPACE, in which any active tuple can gain access to any passive tuple, is a difficult task on a parallel computer in which the memory is divided among the processors. In the Linda Machine, a parallel computer being built at the AT&T Bell Laboratories specifically for the tuple-space language Linda, the memory is distributed among many separate nodes, or subcomputers, wired together in a grid (*left*). To read a passive tuple containing, say, the row of values identified as row 15, an active tuple projects a READ request accompanied by a partial description of the passive tuple to all the other nodes in the same row (*a*). Another computation may still be generating the tuple; when it appears on some node, that node projects its own signal down its column (*b*). The node where the read request and the signal from the passive tuple intersect (*c*) then routes the data (*d*).

it executes, it is a promising basis for representing complex and rapidly evolving systems in microcosm. Consider, strictly by way of illustration, a parallel program that maintains a real-time model of the current state of U.S. air traffic. Whenever an airplane takes off, a passive tuple is created to hold constantly updated information reflecting the system's best guess about the plane's status. Data on weather conditions, airport conditions and so on are maintained in other passive tuples.

The program's active tuples, computing in parallel, evaluate this changing picture and make traffic-control recommendations. The system's ability to refocus its concentration is crucial: when some sector is particularly busy, many active tuples can crowd around to help out. When traffic is light, active tuples can turn their attention to background tasks—sifting through passive tuples that have "landed," for example, in order to compile statistics. No Linda program this ambitious exists so far, but simpler programs of this kind are under development.

How can such fluid software machines inhabit the rigid circuitry of a physical computer? On parallel machines in which all processors are wired to a common memory, tuples are kept in this shared memory. Active tuples are assigned to processors as necessary to keep them busy, and each active tuple has access to all passive ones. Creating a tuple space on a machine with "distributed memory"—memory that is divided among the processors, with none shared in common—is a harder problem. But tuple-space systems do exist on a variety of such machines, and a new distributed-memory machine designed specifically for Linda is now being built in Sudhir R. Ahuja's group at the AT&T Bell Laboratories.

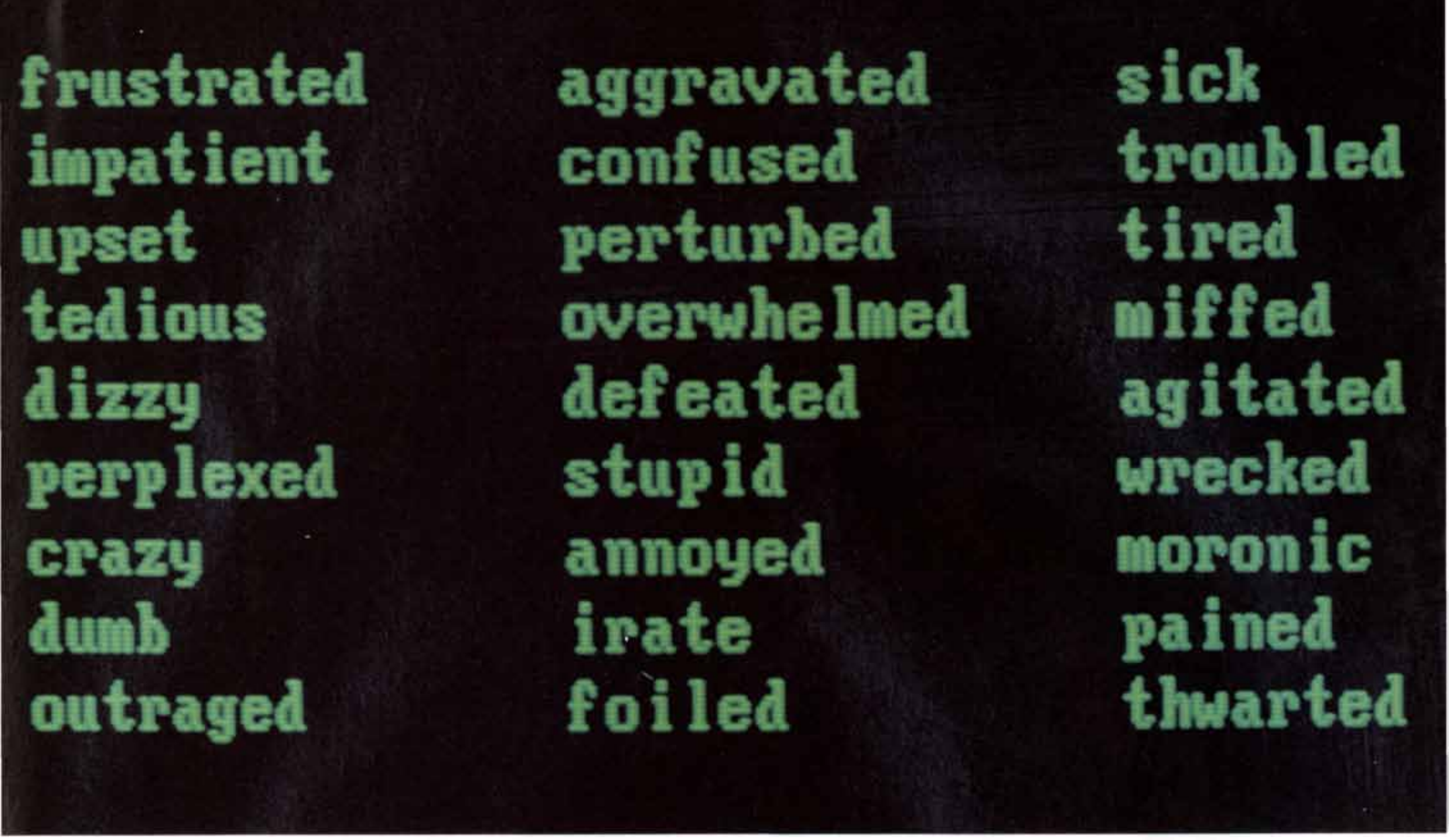
Each of the competing software models has its strong points. Parallelizing compilers appear to work best for a fairly narrow range of programs, but since this class is extremely significant to numerical computing, parallelizing compilers will continue to be important. Message passing is a good model when programs decompose naturally into logical networks of separate and fairly independent pieces; opponents argue, though, that it imposes a stilted and unnatural programming style. Linda shares (and may even improve on) message passing's tolerance for unsynchronized, diverse program structures and adds the naturalness of shared data structures, but experience with Linda is not yet sufficient to prove the tuple-space system's power in the large. Although the Connection Machine model demands synchronism and a certain amount of homogeneity, few other program models can approach its ability to make the most of massive, fine-grained parallelism.

Even as research and tinkering on working software models continues, the pursuit of parallelism is leading computer scientists to ask fundamental questions about the nature of programs. The parallelism we have considered so far is peripheral to the real nitty-gritty of computing. It is more of an expedient to be called in when a big, slow program needs a kick in the pants than it is a basic thought tool. But there is another vantage point. If we clamber up to this out-of-the-way spot and look down on computing from exactly the right angle, we find that parallelism of a certain kind is central and omnipresent in all of programming.

In a sense it is impossible to think of a computation that is not fundamentally parallel. Consider by way of analogy the recipe for a chocolate-chip cake. Computer programs are often likened to recipes: both are designed to transform input into some kind of output. Recipes are usually understood to be strongly sequential. The cake recipe, for example, might be expressed as follows: (1) Make the batter, (2) add chocolate chips, (3) dump the preparation into a cake pan and bake it, (4) make icing and (5) spread the icing on the cake. But the cake recipe is also implicitly—albeit passively—parallel. To carry it out a series of objects must exist simultaneously: a bowl, a mixer, a cake pan, an oven, a cook. For a bowl to exist and then an oven and then a cook in a sequence of nonoverlapping lifetimes is just as unsatisfactory as it is for the mixing and the baking and the icing of the cake to take place simultaneously.

Virtually any program is fundamentally parallel in the same sense: it consists of a collection of pieces—function definitions, data structures, executable statements, for example—all of which must exist in parallel for the program to work. Where does this observation lead? In one important sense, computations look just the same from this new angle as they do from the traditional point of view. In other words, there seems to be a basic symmetry in the structure of computations.

The symmetry amounts to the following. From the traditional viewpoint, most computations consist of a series of steps, each defined in terms of another series of steps, each of which consists of another series of steps and so on. On the largest scale



Don't hold back now. How do you really feel about working with columns?

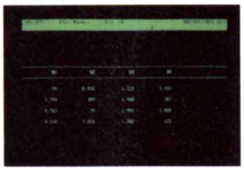
Columns. The black hole of word processing technology.

Try to move a column with your present word processor. Good luck. Try to add or delete a column. Sorry.

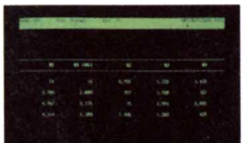
Lotus Manuscript™ can change how you feel about working with columns. Really. Because Lotus Manuscript thinks vertically as well as horizontally.

Consider this. On ordinary word processors, when you create a table, or put any information in columns, you work horizontally, using tabs to create a vertical effect.

With our sophisticated table editor, you can actually create each column independent of one another. You can change information, even delete, move or insert columns, without the usual litany of expletives deleted.



Inserting a column with your present word processor would be a true test of patience, right?

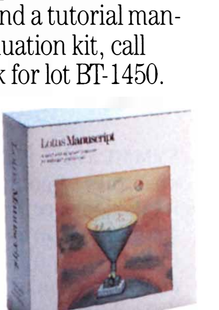


It's simple with Manuscript, because Manuscript thinks vertically as well as horizontally.

And it's not just limited to columns you create yourself. Take tables from Lotus® 1-2-3® or Symphony,® for instance. When you import that tabular information into Lotus Manuscript, it recognizes it as columns, so you can change, delete, move, insert or format this material as well. You can even dress up your tables with a variety of boxes and borders.

As you can see, Lotus Manuscript is not just another word processor. It's a complete document creation system, with more impressive features than we could ever go into here.

Lotus Manuscript is designed to work on most IBM® PCs and compatibles.* Its familiar 1-2-3 interface makes it easy to use. And our Manuscript evaluation kit makes it easy to try. For $10.00, you'll get a presentation disk, working software, and a tutorial manual. To get your evaluation kit, call 1-800-345-1043, ask for lot BT-1450. Or, for more information, see your Authorized Lotus Dealer, or write Lotus Development Corp., 90 Annex, Atlanta, GA 30390-03070.



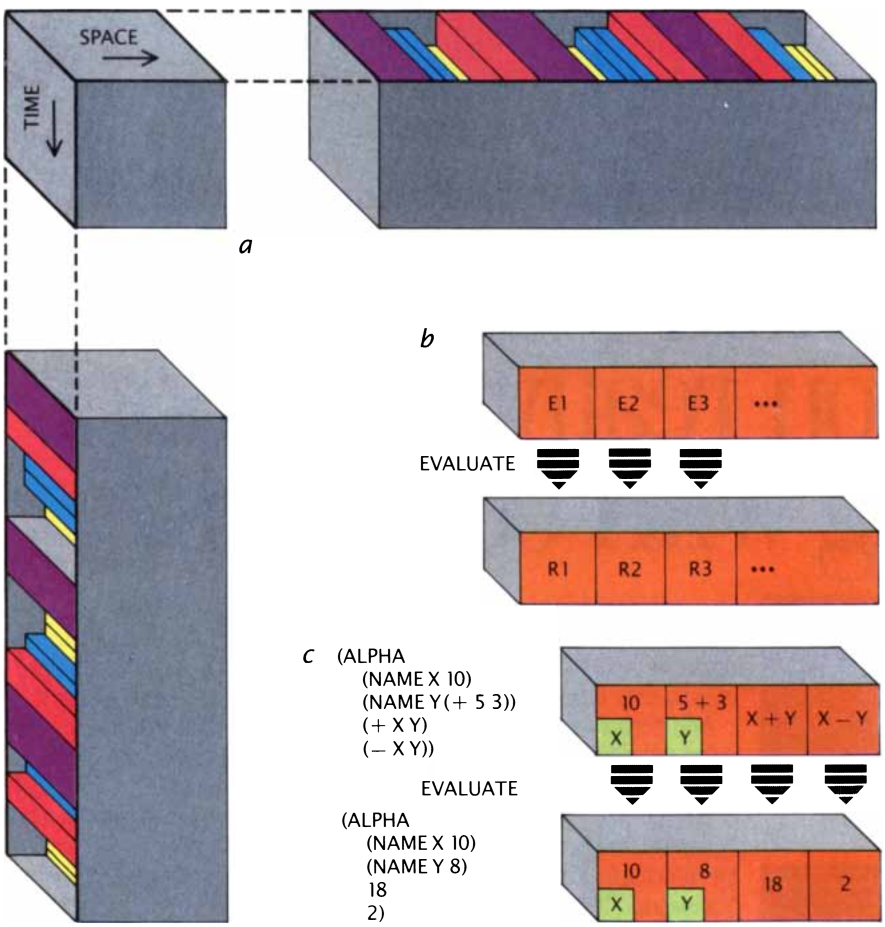
Lotus Manuscript™

each step might be a complex subroutine; these might execute simpler routines, which might evaluate sequences of statements, which resolve ultimately to primitive machine operations. From the sequential angle this makes a computation—up to a certain point—self-similar, like the coastline of Maine. The computation's texture is much the same on any scale. Large features resolve to medium-size features, which hold smaller features and so on.

From the symmetric, spatial viewpoint, programs are constructed in much the same way. On the coarsest level they are often collections of code files or modules, which probably contain subroutine definitions and data structures, which in turn may contain finer-grained declarations, and so on down to primitive values and variables and executable expressions (and ultimately to single words of memory). Conventional programming languages take advantage of temporal self-similarity by allowing the programmer to treat steps of any size as being equivalent, so that two complex or two simple operations can be strung together in the same way and the same kind of loop can, for example, execute a million operations per iteration or a dozen. Parallel, or "spatial," self-similarity is hidden, however: programmers are forced to define objects differently on different scales.

A new class of programming languages supports self-similar structuring along both the spatial and the temporal axes. These "symmetric" languages, which our group is developing in partnership with Suresh Jagannathan of M.I.T., Thomas London of Bell Laboratories and other colleagues, enable programmers to decompose computations not only into temporal sequences of successively simpler steps but also into "parallel," spatial sequences of successively simpler objects. In a symmetric-language program the parallel objects that are hidden within a conventional program are laid out in the "space" within a computer. They are arranged in a definite spatial pattern (we know *where* each object is located, just as we know *when* each step in a temporal series takes place), and additional objects may be nested—again in a definite pattern—within each object. Objects can refer to each other either by position or by name. Thus an object containing the number 3 might be named *x;* another object's contents might be given as $x + 5$ or, perhaps, as "the contents of the 17th object from the left" + 5. The text of a symmetric-language program, then, describes the contents of each object and its position on a map.

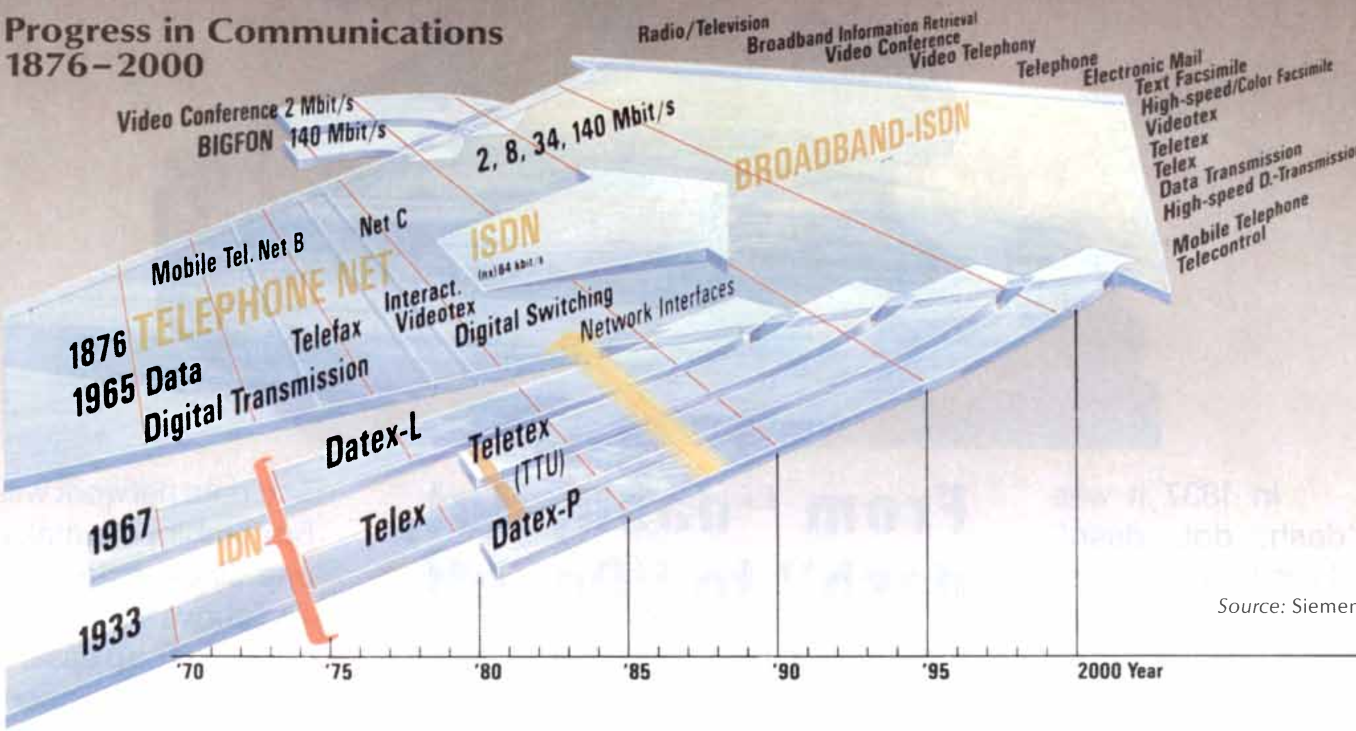
When such a program executes, all regions are evaluated simultaneously, and the result is a new map that shows the final state of each object and matches the original map of uncomputed expressions. The symmetric program in execution resembles a photographic negative in the process of developing.



SYMMETRIC PROGRAMMING LANGUAGE recognizes the basic symmetry in the structure of computations (*a*). A program is executed over a certain period of time, and it takes up a certain amount of "space" within a computer. Conventional programs treat a computation as a temporal sequence of steps, with smaller steps nested inside larger ones. A computation has much the same aspect when it is viewed from the spatial angle as a collection of coexisting objects (which might be modules, data structures or procedure definitions). A symmetric-language program (*b*) specifies a computation as a map of objects, which may be as simple as a value or a variable definition or as complex as a large subprogram. When the program is run, all the elements (*E*) are evaluated in parallel to yield a new map of results (*R*). A simple program in the language Symmetric Lisp (*c*) specifies a sequence of objects, which are transformed as the program is executed.

What we learn about building parallel software machines may ultimately change our understanding both of programming and of the limits of computing. Parallel programs are (potentially) vastly more powerful than old-fashioned sequential ones. Where does it all lead? Parallelism will transform many realms of activity, from graphics to real-time monitoring programs to artificial intelligence and beyond. Perhaps most important, parallelism will make it possible to model and analyze natural phenomena at a vastly more sophisticated level than we can attempt today. Martin H. Schultz of Yale describes the ultimate goal of research into parallelism as "the elimination of experimental science." Schultz's thought isn't meant literally—we do need a smattering of real data so that we can debug our computer models—but it should give us pause.

Communications: Towards the 21st Century



Progress in Communications 1876–2000

Source: Siemens

Part II of a Report by Andrew Hargrave

Poised For The Next Big Leap

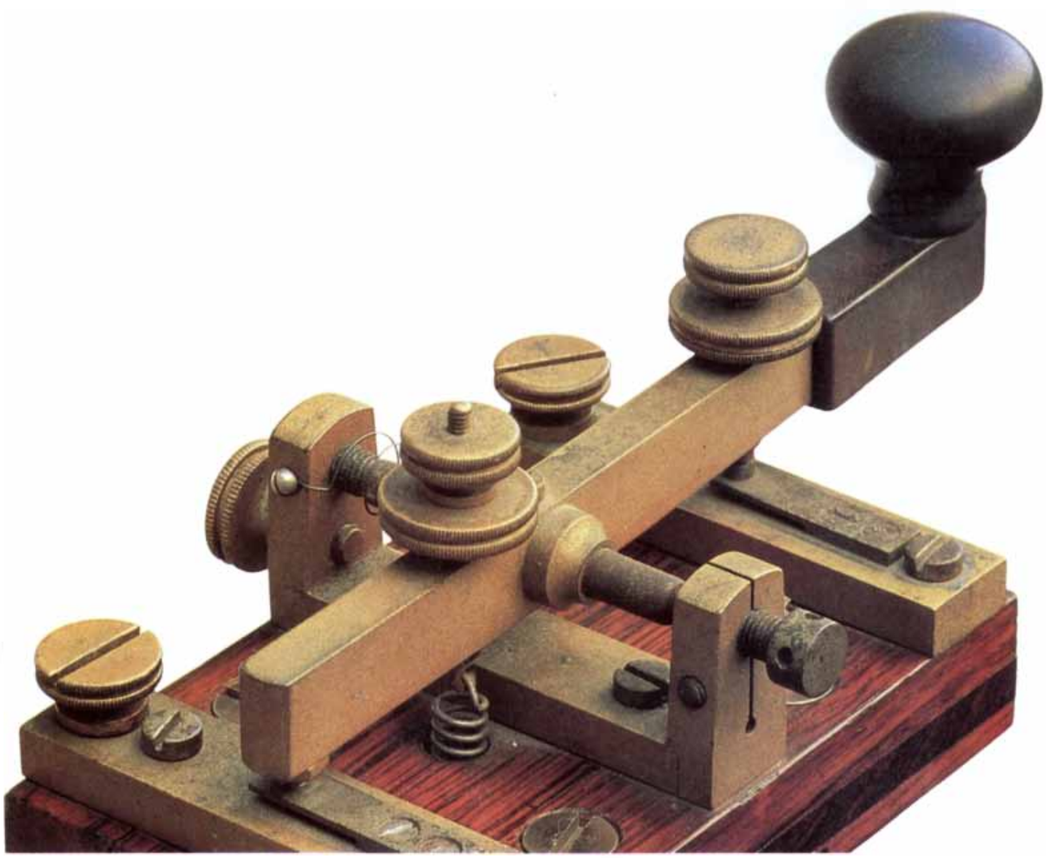
"The Year 2000 holds great promise in terms of markets, economics and technology . . . but it will also create great challenges for sovereign nations, corporations, social and political organizations and the individual." — Edmund B. Fitzgerald, Chairman and Chief Executive Officer, Northern Telecom Ltd.[1]

TELECOM '87 which opens in Geneva later this month will illustrate and underscore in great detail both the promise and the challenges. Technology and cost-effectiveness will run hand in hand throughout the exhibition and the forum — the latter with the global theme of "Communications Age: Networks and Services for a World of Nations" — for the benefit of both users and suppliers. The Geneva show will, among others, provide a bridge between the two sides, promote mutual understanding of one anothers' needs and aspirations in the development of products, systems, services and ideas, some on the market, some proven, some in the experimental stage, some only a distant possibility.

The changing nature and role of communications was underlined recently by Mr. Richard E. Butler, Secretary-General of the International Telecommunications Union which sponsors the four-yearly show. "We see the huge expansion and diversity of services which even today we are unable to predict effectively, nor to define . . . Once they were a useful adjunct to activity; today they are the integral part of many, and particularly large-scale activities, be it government, public service, manufacturing, commercial or servicing in any community."

Information management, a $800 billion-a-year industry worldwide, is one of the major factors in economic progress as well as a key to a high rate of employment, be it an advanced in-

[1] Telecommunications in the Year 2000 — speech to a business audience in Stuttgart.



TEAM/BBDO 6.581

From "dash, dot, dash" to "On, off, on". The development towards ISDN.

In 1837 it was "dash, dot, dash" which brought communication technology a decisive step ahead. The sequence of short and long electrical signals, the morse alphabet, made it possible to transmit messages by wire.

ISDN today's decisive step in telecommunications works according to a principle that is just as simple. The letters stand for "Integrated Services Digital Network" and that means that soon all telecommunication services will have one common language, namely a digital one. Or else "On, off, on".

This means that traditional technology will be replaced by future-oriented electronics. In the telephone network of the future speech, text, data and pictures will be transmitted digitally. The telephone works on the same intelligent technology as a computer does, namely digitally. Furthermore, the telephone network and the integrated text and data network will be combined to make one single extensive and above all efficient network. The pleasant result of this co-operation: The different types of terminal equipment, from the computer to the telephone, communicate via the same subscriber line. The Deutsche Bundespost developed the ISDN in two pilot projects until it was ready for mass production and furthermore fixed a clear extension concept for ISDN.

speech
text
picture
data
ISDN

Integrated Services Digital Network is the name of the idea for the future. This telecommunication network makes it possible to transmit speech, text, pictures and data via one uniform network.

Of course it also considered the possibilities of international co-operation. For the Deutsche Bundespost there is another standard for the future of new communication systems apart from economic efficiency, quality and progress – that is international communication worldwide.

Deutsche Bundespost

dustrial country or a less developed one. Put another way while, for example, in an advanced country like the United States, information management itself accounts for "only" 6—7 percent of the gross national product, "information dependency" (according to Mr. Fitzgerald) exists in respect of 60—70 percent of the total US workforce — ten times that figure.

Boom conditions for the industry over the next decade and more are widely predicted. Mr. Butler estimates the number of telephones worldwide (650 million at present) should rise three to six-fold by the end of the century, with annual increases in employment averaging 230,000 to 400,000, half of them new jobs. At the same time, economies of scale should result in a 30 percent drop in production costs every time output doubles.

During the same period (Mr. Fitzgerald forecasts) the information management industry's sales worldwide should rise five-fold, to $4,000 billion a year, if present growth rates continue. FORUM '87, which runs right through the week-long exhibition, covers five major themes: telecommunications policy, services (including technology, innovations, new materials, etc.), the legal framework, infrastructure and economic growth, regional network development and international communications.

Part II of TELECOM '87 Preview (the first part was published in May) seeks to report, interpret and analyze some of the issues. It also attempts to provide a glimpse of future, even more revolutionary developments.

ISDN – Benefits and Pitfalls

The integration of dataprocessing and telecommunications within ISDN should benefit both the users and the suppliers.

- For the subscriber, especially the business ones, the main benefit lies in the simplicity and functionality of a single terminal combining a host of, hitherto separate services — a single switch and pair of wires (or eventually optic fiber) giving access to the whole world outside — with one, hopefully cheaper account to settle at the end of the payment period.
- For the PTTs, ISDN should mean higher revenues, improved networking capabilities for both residential and business users and, in the longer term, cost savings.
- For the manufacturers and associated providers of systems, software, materials, etc., ISDN is a supreme opportunity for growth in terms of technological advance, sales and profits.

Of course, there are always winners and losers in a technological revolution, even in the context of the same technology. As Mr. Eric Nussbaum, of Bell Communications Research (Bellcore), remarked (Quoted in Communications Systems Worldwide, May 1987 issue) at a recent symposium (he was referring specifically to switching, but it applies to other pieces of telecommunications equipment as well) high development costs, competitive multiple supplier situations, rapid technological obsolescence and unquantifiable customer needs and demands were "all taking their toll on the technologists' dream of creating truly new generation switches again."

DGT, the French PTT which pioneered digital exchanges in the early 1970s, has been publicly talking about the possibility of beginning to install new systems within the next 5—7 years. At present, the French network relies exclusively on the virtually interchangeable E10-B (CGE-Alcatel) and E10-MT (Thomson, CGCT) digital systems.

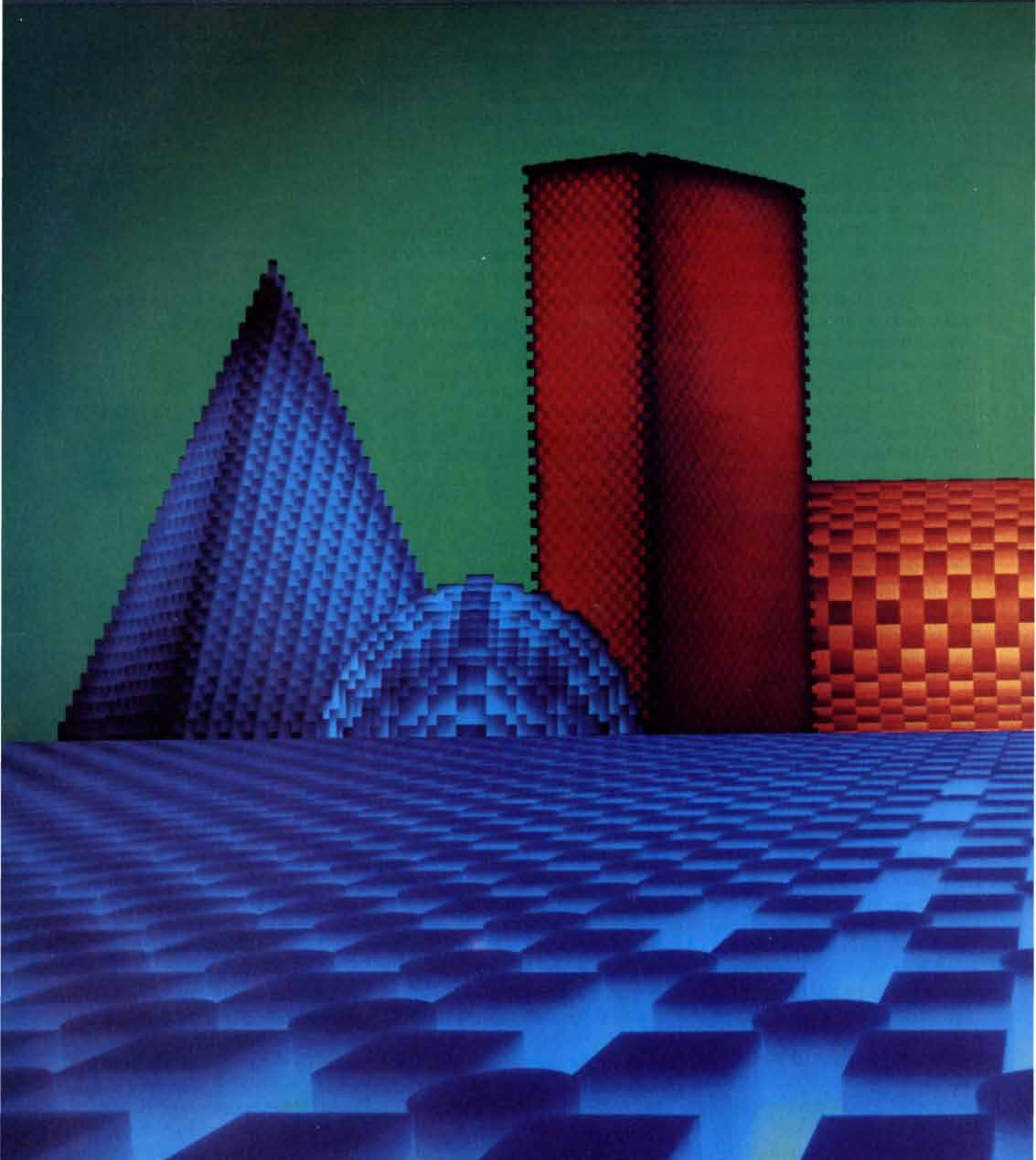
It may therefore be no coincidence that the move, merging CGE's telecommunications subsidiary Alcatel with ITT's telecommunications interests, originated in France. Alcatel NV, the joint group and second largest public switching supplier in the world (after AT&T) has now two switching systems — the older E10-B and ITT's unique, fully distributed System 12, the newest digital switching system in the world. According to Mr. Jo Cornu, Alcatel's Vice President of public network systems, (a former ITT executive), although the new group's two fundamentally different systems will continue to be marketed, "identical solutions" to problems are being sought.

Standardization is, of course, vital to progress to link smoothly and efficiently a diversity of networks and systems as well as their nuts and bolts. The opening up of national markets to real international competition is also considered important, particularly for manufacturers such as Ericsson and Philips which operate from a narrow domestic base. But the dismantling of national barriers is also vital for other leading suppliers for whom volume sales are essential to spread the huge cost of developing existing systems as well as initiating new ones.

The Market: Who is doing what and where

One of the outstanding features of the convergence of dataprocessing and telecommunications has been the proliferation of mergers, takeovers, alliances, partnerships and joint ventures across the whole spectrum of the information technology industry. But while the pooling of resources — human, technological and financial — has been a significant factor in such moves, it is by no means the only one. Access to hitherto jealously guarded and politically motivated national

WE BELIEVE THAT THE GENIUS OF THE FUTURE LIES NOT IN TECHNOLOGY



Visit us in Geneva at TELECOM '87 for hands-on software demonstrations and an introduction to our

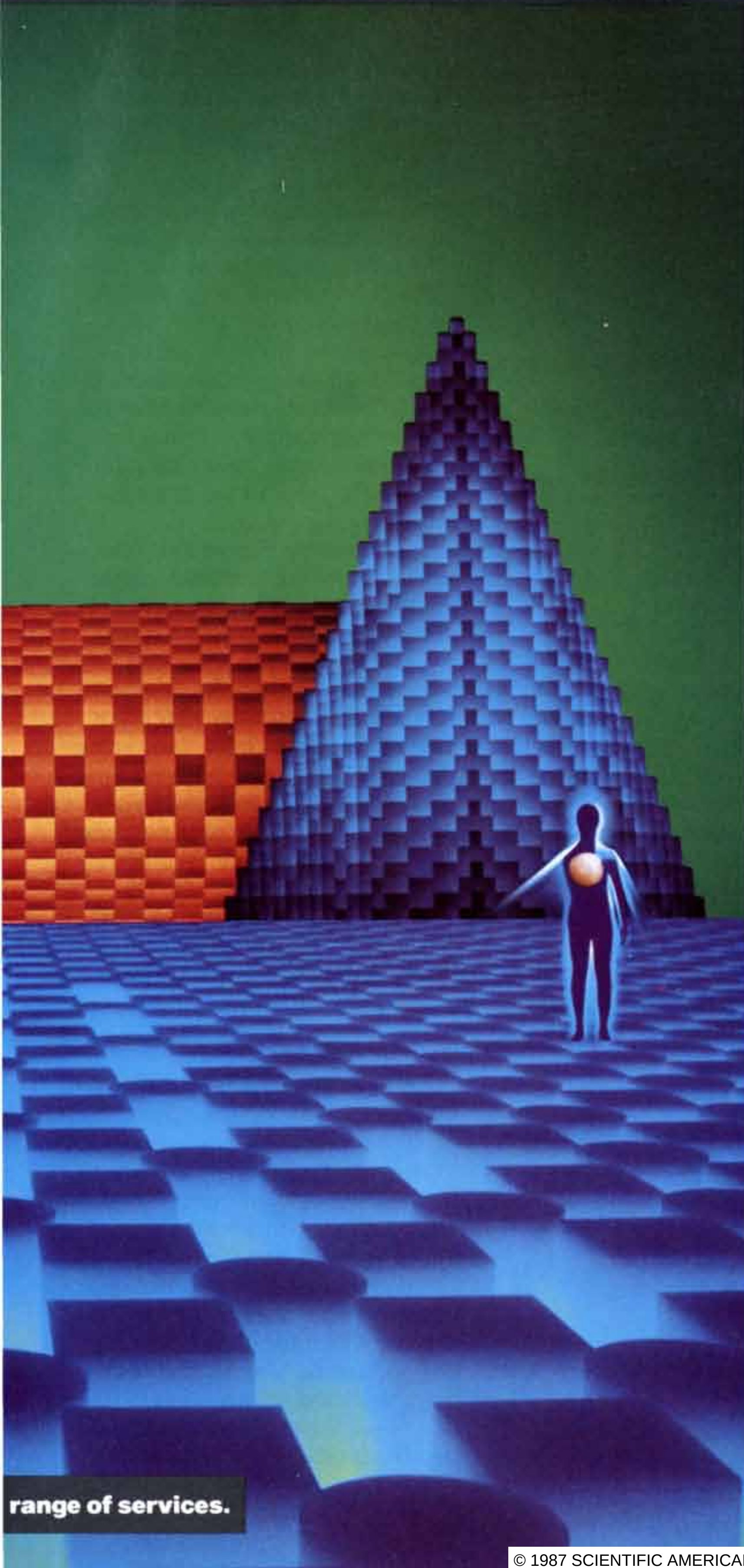

range of services.

Table I
Europe: Estimated Public Switching Market Shares by Country for Leading Suppliers — 1985

Country	Alcatel[1] %	Siemens %	L M Ericsson %	Others %
Austria	25	25	—	Kapsch 25[2]) Schrack 25[2])
Belgium	78	22	—	—
Denmark	2	10	86	B&O 2
Spain	72	—	28	
Finland	4	18	36	Nokia 42
France	84	—	—[2])	CGCT 16[3])
Great Britain	—	—	10	GEC 40, Plessey 40 STC 10[4])
Ireland	50	—	50	
Italy	20	12	18	Telit 50[5])
Norway	60	—	40	
Netherlands	7	5	18	APT (AT&T/Philips) 70
Portugal	50	50	—	
West Germany	32	50	—	Bosch (Telenorma) 7, IBM 5
Sweden	—	—	20	Televerket 80[6])
Switzerland	35	23	—	Hasler 42
Total: Europe	37	21	11	Telit 6, GEC 5, Plessey 5

Source: Alcatel
1) Alcatel (includes ITT's merged telecommunications activities).
2) Northern Telecom licensees.
3) CGCT (now Ericsson-Matra).
4) STC (former ITT subsidiary, not included in Alcatel deal).
5) Telit (combined Italtel and Telettra, now negotiating merger).
6) Swedish PTT manufacturing its own equipment.

markets has certainly been another.

The recent, often bitter battle over the ownership of CGCT, with a 16 percent share of the public and 7 percent share of the private switching market in France has been a case in point. It had been coveted by several of the world's leading telecommunications suppliers, mostly with French partners — the main contenders including APT, the AT&T-Philips group (with SAT) and Siemens (with Jeumont-Schneider). Northern Telecom and Italtel were also among those who publicly expressed interest.

In the event the winner was Ericsson, in some ways a "neutral" in the battle, in partnership with Matra. The latter had earlier acquired the private switching arm of CGCT, thus providing Ericsson with a further outlet in the French market.

So there can be little doubt that in spite of the trend in recent years of liberalizing national telecommunications markets — a trend encouraged by the ITU as well as right-wing governments in major European countries — there is still a long way to go in most of Western Europe. As Table I indicates, established native suppliers still have the edge in the public switching field, even in some smaller countries. Others are captive markets for the giants.

Europe represents about one-third of the total world market in telecommunications equipment, although European companies capture only around 10 percent of the sales. (This trend is, of course, accentuated by the US and Japanese domination of the dataprocessing end of the market.)

Table II
Telecommunications Equipment Sales Worldwide: 1985

Company	% Share
AT&T	16
Alcatel	12
Siemens	11
Northern Telecom	6
Ericsson	5
NEC	4
IBM/Rolm	3
APT/Philips	2
GEC	2
Others[1])	39

1) include Plessey/Stromberg-Carlson, Nokia, Telit, STC, Pirelli, Fujitsu, etc.

In terms of the global market, Europe is somewhat overshadowed by North America (including the US), with its 40 percent share of the total. (Like Europe it, too, has a substantial trade deficit in information technology products.) This explains why, in their search for volume sales, most European companies, especially Siemens and Ericsson, are doing their utmost to gain a foothold in the US market. The acquisition of Stromberg-Carlson, a minor but significant public switching supplier, by Plessey had the same purpose.

Conversely, the size and importance of the European market in telecommunications, both public and private, is acknowledged in moves by North American companies to break into Europe, usually in partnership with established European firms. The setting up of APT by AT&T and Philips to develop a European version of the former's 5ESS digital switching system — 5ESS-PRX — is only one spectacular example.

Table II, compiled by Alcatel, may serve as a rough guide to the overall state of the worldwide telecommunications market.

Business Systems — A Growth Area

While public switching is more of a headline-catcher through the involvement of governments, national carriers and the use of political pressure by leading domestic suppliers, it is the private business sector which provides far more opportunities, and a somewhat less restricted environment.

Nevertheless Table III, which embraces the four largest European countries as well as the US

It isn't size.....it's skill.

Pirella Göttsche

As with so many activities, results are determined by the most skillful. On the playing fields and in the market those who are quick and responsive triumph over the large and ponderous. And so it is with Italtel.

Today the world is growing smaller. The "Global Village" is happening and we are an integral part of it.

Italtel is designing, manufacturing and delivering systems and equipment both for public and private telecommunications as well as for office automation.

Italtel knows that a serious commitment to research and development is of paramount importance to its success.

There are more than 2,000 people, more than 10% of the staff engaged in research and development.

Almost 11% of the sales revenue is committed to research and development.

Italtel's culture is based on ideas and on the knowledge that the customer must be served quickly and with intelligent responses.

Italtel's UT digital public switching system is just one example. We have many more, like the entire range of up-to-date digital multiplexers, fiber optic systems and rural radio.

To get the full story, make a skillful decision: talk to Italtel. For more information write to: Italtel, International Sales Organization, via A. di Tocqueville 13 - 20154 Milano, Italy.

Italtel
IRI-STET GROUP

TELECOMMUNICATIONS TODAY
TELECOMMUNICATIONS TOMORROW

Table III
Office Systems (PABX) Market Shares in Europe and the US[1])

West Germany	%	France	%	Great Britain	%	Italy	%	US	%
Siemens	36	Jeumont-Schneider/Telenorma	35	British Telecom	29	Italtel	29	AT&T	23
Telenorma	20	Alcatel/Telit	30	Mitel (BT)	16	Olivetti	25	N Telecom	23
SEL (Alcatel)	17	Thomson	21	Plessey	25	FACE (Alcatel)	16	Rolm (IBM)	15
Nixdorf	10	CGCT (Matra/Ericsson)	7	STC	12	FATME (Ericsson)	6	Mitel	9
DeTeWe	8			GEC	9			NEC	7
Tekade (Philips)	7							Others	23

Sources: Olivetti, others (including Office Automation, May 1987 for US figures).
1) All, except US figures (1985), are for 1986

is a pointer that even in the private business switching (PABX) market, the conditions applying to public switching, including partnerships, acquisitions and major adaptations of systems to suit local circumstances cannot be ignored altogether.

It may be noted that all the major companies supplying digital systems for public networks are also competing in the substantial and fast-growing private switching market. In addition, one finds companies until comparatively recently associated with dataprocessing, such as Nixdorf; or office machinery (Olivetti, DeTeWe); or telephony (Telenorma); or business communications (Mitel, now a British Telecom subsidiary) increasingly involved in this highly competitive market.

Some, including major mainframe computer manufacturers have followed the IBM/Rolm route to capture a slice of the business communications market, especially with the advent of ISDN and the integration of dataprocessing and telecommunications.

Progress in digital PABX technology is one major reason for both AT&T and Northern Telecom to take more than a passing interest in the European private switching market. (As Table III indicates, while both have a substantial share in the US office communications market, neither can claim a significant slice of the major European ones.)

For AT&T, the natural European ally in private switching is Olivetti in which it already has a 23.5 percent stake. The Italian company, second only to IBM in Europe's personal computer market and also a leading supplier of office systems and machinery, has in the last four years risen to second place (see Table III) in its domestic PABX market with the ICS 6000 system for larger business users. AT&T has carved itself a modest niche in the UK with its System 85, used mainly by US branches operating there. However, the joint AT&T/Olivetti assault on Europe's business systems market is only just beginning.

A joint design team has been working for some time on an "international version" of an AT&T business system called System 75. It is to be introduced in Italy (largely complementary to the ICS 6000) and in the UK in the first place. As Mr. Gianfranco Casaglia, an executive of Olteco (Olivetti's telecommunications arm) points out, System 75 is "tailored to the standards and requirements of business users in several European countries." It has an AT&T microprocessor and an ISDN interface, complies with CCITT standards, but its digital switches are adapted to each country's requirements, with "application packages" — both hardware and software — for individual types of customers, such as hotel or retail chains.

Not that AT&T, Olivetti or others underestimate the difficulty of penetrating "tight, already over-supplied national PABX markets." Northern Telecom, leader in both public and private switching in North America alongside AT&T, is also making determined efforts to break in a major way into Europe, too.

Mr. Fitzgerald is hopeful that Northern Telecom would raise the present 4 to 5 percent share of revenues from non-American sources to 15 percent by the early 1990s, with Europe as "one of our major markets." Its strategy is partly to import into Europe systems such as "Centrex,"popular in North America but virtually unknown in Europe (Centrex enables the business subscriber to use the public network for his specific needs) and to form alliances with European manufacturers (AEG, in Germany, for packet-switching, Kapsch and Schrack in Austria, for public networks, among others) or public administrations such as Mercury and, for some items, BT in the UK, Televerket in Sweden, etc.

According to a recent Dataquest survey, expenditure on public telephone exchange equipment in Europe will decline from about $3 billion in 1986 to $2.4 billion in 1990, with all the major countries registering a drop. Table IV has a breakdown of sales as between various sub-sectors of telecommunications equipment and the leading suppliers.

Managing Transmission

As noted earlier (see also Table IV), transmission and cables already form around 40 percent of all expenditure on telecommunications equipment in Europe, while the proportion spent on switching (about a third of the total) is on the decline. According

Table IV
Telecommunications Europe: Estimated Revenues and Market Shares by Individual Items and Major Companies: 1985

Sector	Total Market in $ million	Alcatel %	Siemens (GTE) %	Ericsson %	APT/ Philips %	Bosch Telenorma %	Telit %	Plessey %	GEC %	Pirelli %	Others[1] %
Switching	4,330	37	21	11	1	1	6	5	5	–	13
Transmission	3,200	29	21	2	14	8	6	3	4	–	13
Telecom Cable	1,900	19	10	5	12	–	–	–	2	15	37
Business Comm.	2,100	19	20	5	7	11	2	8	4	–	24
Subscriber Terminals	1,700	20	16	7	2	9	2	3	3	–	38
Totals	13,200	25	20	7	6	5	4	4	3	2	24

Source: Alcatel
1) "Others" include companies already mentioned in previous tables as well as a large variety of component and accessory suppliers. As the table shows, up to 40 percent of all telecommunications expenditure now goes into transmission, including the links between subscribers and exchanges of all types; between countries and continents, on the ground, under the sea, by microwave or satellite.

to Dr. Peter Radley, Technical Director of STC Telecommunications, the drop in switching outlay may be as much as 30 percent by the turn of the century.

Within five years, says Dr. Radley (quoting US sources), the present approximate parity between switching and transmission may have altered to nearer one-third/two-thirds, partly as a result of cost-conscious, liberalized markets, like the US and the UK.

STC, the former ITT subsidiary in the UK which now includes ICL, the UK's largest native computer manufacturer, "rejects the idea that the switch is the center of the telecommunications universe. Our answer (Dr. Radley adds) is the management of the transmission network for the subscriber."

"The cost-conscious subscriber will appreciate that transmission costs have declined by a factor of almost 40 in the past decade while those of digital switches by not nearly as much."

One of the major areas of growth in telecommunications, the spanning of oceans, seas and straits by submarine cable, is in fact largely the result of optic fiber development. (Leading companies such as Alcatel — which now includes Les Cables de Lyons and ITT's transmission as well as switching interests — expect 60 percent of its sales to come from optic fiber within the next 2-3 years.)

A recent study by Kessler Market Intelligence (Worldwide Markets for Undersea) refers to plans for five transatlantic and three transpacific networks as well as the spanning of the Indian Ocean, the Mediterranean and North Seas. It predicts worldwide markets in submarine cabling to increase from a relatively modest $117 million in 1986 to $2 billion in 1989 and almost $4.8 billion by 1995.

The first transatlantic fiber optic cable TAT-8 is due to enter service next year, with a five-fold reserve capacity through multiplexing. A second cable TAT-9 (by the same consortium, led by AT&T and British Telecom and comprising other PTTs) should be completed by 1991.

There are, however, even more ambitious plans ahead. A $600 million project called "Market Link", sponsored by Nynex (a regional Bell operating company) and Britain's Cable and Wireless, provides for two transatlantic cables (PTAT 1 and 2), a 7,000 kilometer link between Weston-super-Mare in the UK and Long Island in the US, with a spur to Bermuda. PTAT-1 is due to enter service by mid-1989, PTAT-2 by 1992.

STC which prides itself on having won 85 percent of all submarine cable projects open to international tender (STC Annual Report 1986), is the „sole prime contractor" for PTAT-1. The rest in the hunt include Alcatel and Pirelli, the Italian tire-and-cable company which is tendering for several of the submarine contracts, having already completed one spanning the Straits of Messina.

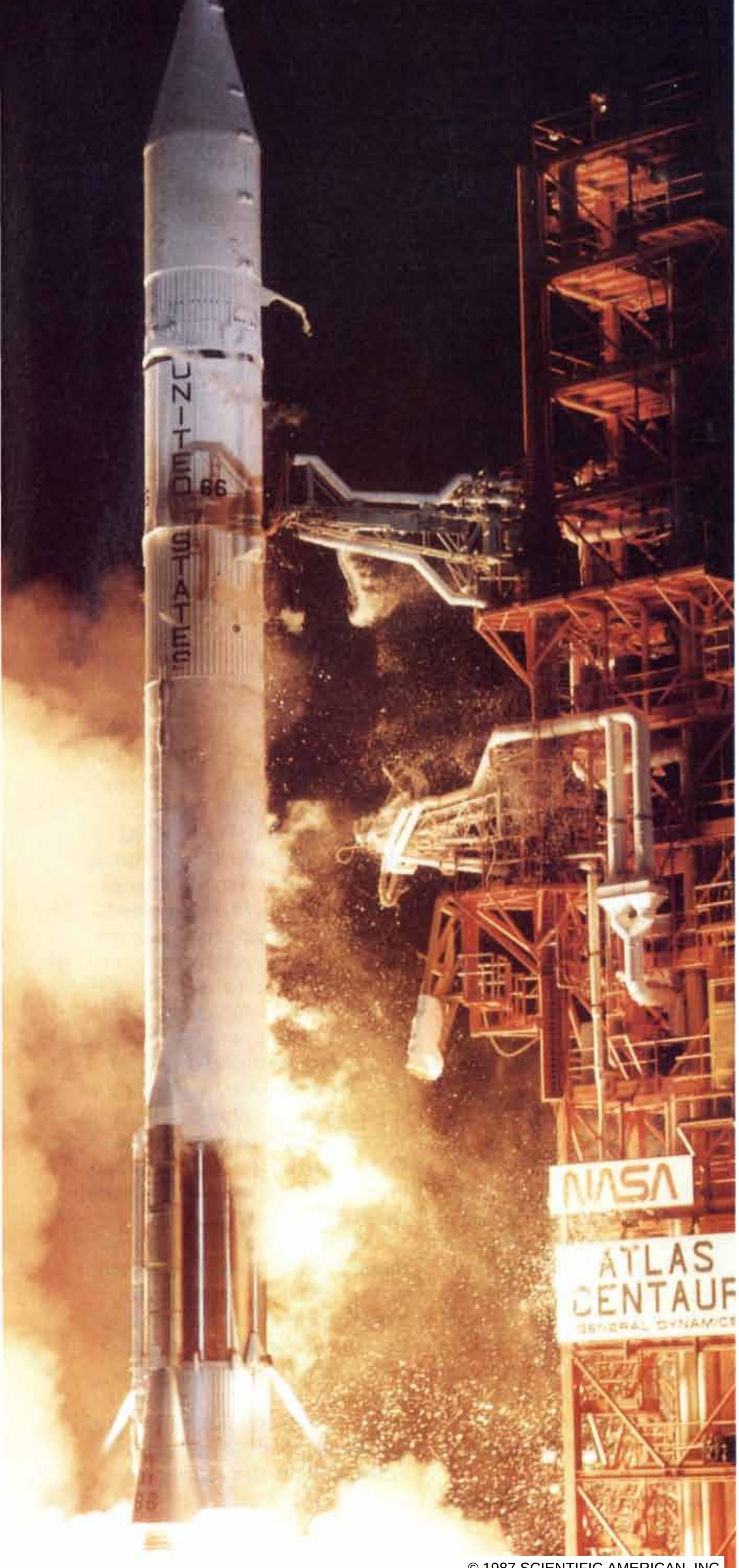
Cable & Wireless, a leading transmission systems company, is the prime promoter of the proposed global "digital highway" linking the major business capitals of the world on land and by sea (of which PTATs-1 and 2, PPAC, a Pacific-Indian Ocean link, and others are some of the components).

At the time of writing, C & W and its US and other partners were involved in fierce controversy with the Japanese Government over being allowed to compete for a second national telecommunications network. The Japanese Government would prefer C & W and its partners joining a Japanese-run consortium — a move strenuously resisted by the company, with backing from the UK Government.

Major schemes to develop optic fiber networks are in progress all over Europe as well as in Japan.

Is there not a danger of market saturation? Mr. Jonathan Solomon, Director of Corporate Strategy at C & W (in an interview with Dennis Gilhooley, Telecommunications, March 1987), was quoted as saying that "competition and choice necessarily involve a degree of overcapacity: the question is when and how this capacity can be filled." Mr. Solomon seemed confident that the developing ISDN and additional value-added services would swell the need for ample circuit capacity.

Satellites are an important segment of the worldwide communications network, competing in certain areas with land and submarine cable links. To anticipate


UNITED STATES
NASA
ATLAS
CENTAUR
GENERAL DYNAMICS

this to some extent, says the Italian carrier Telespazio, the trend for networks is to be tailored to the specific needs of user categories — mobile, business, television, etc.

While satellite and cable services are, to some extent, complementary, they inevitably compete in certain areas, particularly in long-distance communications. However, in the opinion of Mr. Maschio of Pirelli, "cabled transmission is generally quicker and more reliable for ordinary voice-data transmission."

The Chip — A Key to ISDN

The development of computer and telecommunications technologies — miniaturization of computer power culminating in the microchip on the one hand and digital switching and transmission on the other — ran along parallel lines but separately for the best part of the 1970s. The major telecommunications suppliers, some of whom like AT&T, NEC, Siemens or Ericsson, had already been involved to varying degrees in dataprocessing, first tended to create their own customized chips to meet requirements: but as competition sharpened, the major semiconductor manufacturers, led by Motorola and Intel Corporation of the US, responded by marketing their own "telecom chips." Motorola's first telecom-specific IC (integrated circuit) dates back to 1976: the company claims to be the first in the world to announce, in 1985, "a complete ISDN product portfolio."

The cry from telecommunications suppliers is for more and more power and speed to be squeezed into the microprocessor, to service ever more transmission lines from ever smaller surfaces. The 32-bit microprocessor — said to be equal in performance to the larger minicomputers, but at only a fraction of the price — is already having trial runs and is about to enter the market.

Motorola's policy is to design general-purpose microprocessors for all switching systems, with special features added on as and when required. To absorb changes in systems which are now coming much faster than before — the cycle is said to be 10-15 years rather than 20-30 years before the advent of digitalization — the company has evolved the "serial processing core" concept.

Mr. Ray Burgess, Motorola's European Microprocessor Marketing Manager, lists the company's processing technology "family" areas under five headings: microcontrollers, which aim at compressing more and more circuitry into smaller and smaller areas; 16/32-bit integrated processors for ISDN terminals; digital signal processors for modems and multiplexers; very high speed processors for regional exchanges — usually clusters of microprocessors or a single customized one; and, finally, specialized high-speed communication controllers, such as the "multi-link" LAPD (link access procedure) protocol for both signaling and data transfer in ISDN configurations. (This last mentioned should hit the market in the last quarter of 1987). Its interface has both Motorola and Intel "family options."

Software — Panacea or Bottleneck?

Although ISDN, the single-route transmission of differing signals — voice, data, graphics and images — is only beginning its long journey, few doubt that its success depends increasingly on the quality, reliability, productivity and cost of software. (As already noted, software may amount to as much as 70-80 percent of the cost of an entire telecommunications system.)

It is a classic chicken-and-egg situation. At the seminar mentioned earlier (Communication Systems Worldwide, May 1987) there seemed to be general agreement about software being seen as the panacea to enable the rapid introduction of new and innovative services into an evolving network. On the other hand, as Mr. Robert Martin, Vice President Software Technology and Systems at Bellcore pointed out, the growing complexity of the software business was leading to bottlenecks.

"Sadly, there have been no major software break-throughs, and the most optimistic people in the industry estimate that productivity is increasing at the most by a meager 7 percent a year."

Mr. Martin called this "a frightening scenario" and suggested that the communications industry would have to look at new ways to develop and deploy software. "Architecture" would have to be changed to give "distributed intelligence with functional components" which could then be "glued together" like building blocks, under centralized control.

Unless such a major change took place, Mr. Martin warned, "software will indeed be *the* impeding force in getting new services into the network." Mr. Martin's remedy, including the need for industry cooperation, building blocks and centralized control, has of course been in the air for some time. It is firmly associated with the issues of standardization, OSI (Open System Interconnect), the CCITT recommendations and the ITU's own initiatives.

Already at the World Telecommunications Forum at Singapore in May 1985, Dr. T. Kobayashi, then President and Chief Executive Officer of the NEC Corporation, called for "a strategy for software", including "optimum production technologies and tools for effectively developing and maintaining better quality software."

He also advocated "modularization", each software product interfacing with all the others as a means of overcoming what he described as the "cottage industry" type of software development. It is a method being increasingly adopted in the industry.

Dr. Kobayashi's warning that the scale of programs required for the next generation of switching systems in the ISDN age would be "ten times greater than those of the latest systems" was linked with the plea for "joint support for software development" by the telecommunications and computer industries as these share the same users.

In the meantime, however, companies are still struggling with the consolidation of the two technologies — telecommunications and dataprocessing — into a single one and (to use Dr. Kobayashi's phrase) with the problem of supporting joint software development, in their own interest.

There are certainly plenty of attempts to relieve the "software bottleneck" especially among the computer manufacturers. One way of achieving it, is a reversal of the process of constructing software packages. Instead of adjusting it to suit the hardware, the package precedes it or is at least constructed in the early stages of hardware manufacture.

All in all, systems producers are eliminating bottlenecks wherever possible — and wherever recognized at an early stage of hardware development. But there are still no universal panaceas . . .

Towards The 21st Century...

As the last decade of this century approaches, the main problem for suppliers of the information technology industry seems to be two-fold. Dr. Nasko, of Nixdorf, has summed it up thus:

- to integrate in time the technological innovations into standardized hardware and software environments, so that these new technologies are converted into substantial efficiency gains; and
- to move these products into solutions tailored to the customer's needs.

Several of these innovations have been touched upon in this report, as have the attempts to tackle the problem as Dr. Nasko has formulated it. There are, however, yet more innovations in the pipeline that are still largely in the laboratory stage but which may have a profound impact on the industry as a whole.

Perhaps the most significant of these (because it directly challenges silicon, itself a break-through, making today's chip possible) is gallium arsenide as a basic material for integrated circuits.

Most of the major IT suppliers are experimenting with GaAs, a compound of gallium, a by-product of aluminium and arsenic, a by-product of copper or lead refining. According to Bell Northern Research (Northern Telecom's research subsidiary/Telesis (BNR) 1987/1), chips based on GaAs could be three to ten times as fast as silicon chips, or alternatively consume one-tenth of the power; they can detect, emit and convert light into electrical signals, opening up possibilities of optoelectronic properties on a single chip; can resist 10,000 times the radiation; can withstand operating temperatures of 200°C; and have higher electronic mobility.

If this is the case, does it mean that GaAs will supplant silicon as the base material for ICs in the foreseeable future? Not so, in the view of the BNR researchers for reasons that are both economic and technological. Although silicon requires extensive refining to high-purity standards, it is an abundant natural substance found in sand, while gallium arsenide (also in need of extensive refining) forms less than 0.1 percent of the earth's crust. Moreover, while silicon is a single, pure element, homogeneous and stable, GaAs wafers are affected by impurities, reducing production yields.

However, once demand for high-speed chips rises to a level making GaAs chips economical, there are a number of applications which could once again revolutionize the industry.

Perhaps the most important one is in integrated optoelectronics, opening up new possibilities in the design of high-speed systems.

How far in the future all this is seems to depend less on technology but on the ability of the suppliers to finance it and, above all, of the market to absorb it. The suppliers are not, of course, a homogenous entity. The mega-mergers such as CGE-ITT (now Alcatel) or AT&T-Philips (APT) are largely based on the combined objective of greater financial strength to fund development and a larger market share to guarantee survival.

However, the smaller, more specialized companies represent a radically different outlook. To quote Dr. Nasko once more: "Competitive advantage will not come from accumulating different products under a single roof but from an ability to offer expertise and responsive support to end users . . . They represent a very different approach from that of marketing IT under a single brandname."

But whichever the approach, the decisive factor in any technological advance is the reaction of the market. And on this matter most suppliers agree: the key to consumer acceptance of the digital environment is "user-friendliness" — that the equipment should be not only reliable and cheap but also easy and practical to use. And the means to achieving this are, as noted so often in this report, in the main, integration and standardization.

WHEN COMMUNICATIONS HELP IS NEEDED, AMERITECH HAS THE SURE HANDS FOR THE JOB.

It takes more than state-of-the-art technology to keep communications moving on the best course for our customers. It also takes knowledgeable people working together to help others send, receive and use information in any form, any time and anywhere.

Ameritech is a leader in communications. We're the Ameritech Bell companies of Illinois, Indiana, Michigan, Ohio and Wisconsin. And we're much more.

We provide the latest in voice, image and data products and services wherever they are needed by business, and we arrange competitive financing packages to meet our customers' requirements.

Ameritech pioneered cellular mobile phone service and publishes the Ameritech PagesPlus® phone directories and a growing number of specialized publications. We're also leading creators and marketers of computer software, and we help develop new technologies that advance the future of communications.

Our tradition of helping customers goes back more than a hundred years. We have the talent, the technology and the team to keep that tradition shining brighter than ever.

AMERITECH

AMERICAN INFORMATION TECHNOLOGIES

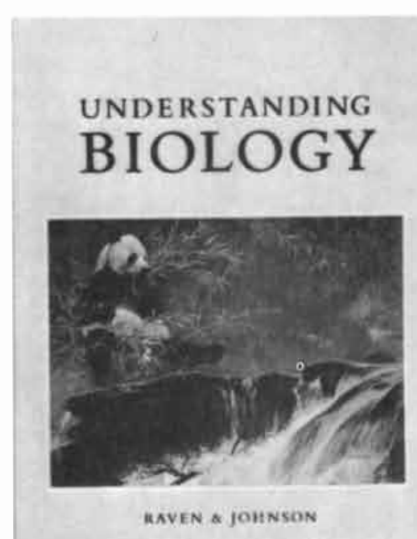
Helping you communicate℠

The companies of Ameritech: Illinois Bell · Indiana Bell · Michigan Bell · Ohio Bell · Wisconsin Bell · Ameritech Communications
Ameritech Mobile Communications · ADR · Ameritech Services · Ameritech Credit · Ameritech Development · Ameritech Publishing

A UNIQUE BREED,

A NEW ENVIRONMENT . . .

. . . your classroom, the perfect place to nurture the minds of introductory biology students. Two renowned scientists and biological scholars, Peter Raven and George Johnson, offer you the finest of the species—*Understanding Biology.* From a distinct evolutionary perspective, this unique creation moves logically through the general fundamentals in the first 23 chapters into the realm of biological diversity, with plant and animal biology in the remaining 20 chapters. Superlative, articulate, and beautifully illustrated, *Understanding Biology* captures nature's workings and works of art in one comprehensive volume.



TIMES MIRROR
MOSBY

For more information, call Times Mirror/Mosby College Publishing TOLL-FREE in the U.S., (800) 633-6699. In Canada, call (800) 268-4178.

SC2073

Data-Storage Technologies for Advanced Computing

In five years magnetic devices will "write" and "read" data twice as fast as today's systems on a disk holding five times more data. Magneto-optical technology offers an even higher storage density

by Mark H. Kryder

The fact that the Santa Clara valley in California, regarded as the center of the U.S. electronics industry, has been dubbed "Silicon Valley" piques manufacturers of magnetic data-storage devices. After all, the companies based in the valley actually derive more of their revenue from magnetic devices than from semiconductor devices. A more appropriate sobriquet, the manufacturers suggest, would be "Iron Oxide Valley," after the commonest material from which magnetic-recording mediums are made.

Although their peeve is intended as a joke, it does make a point: magnetic data-storage technology as embodied in the familiar tape and floppy-disk devices and in more sophisticated rigid-disk devices is as critical for today's computers as semiconductor technology. Indeed, tomorrow's more powerful computing systems will be useful only if they can access data in larger volumes and shorter times than they can today. Hence the faster speed of computation offered by advances in semiconductor devices and in the way the devices are interconnected and programmed must be matched by an increase in capacity and performance of data-storage devices.

Magnetic devices lend themselves to the permanent storage of massive amounts of data, because they offer much greater memory capacity at a lower cost per bit of data stored than semiconductor devices. A typical rigid-disk device stores 800,000 characters (the equivalent of about 400 typewritten pages) in a square centimeter of recording medium; some of the larger, multidisk devices store a total of more than five billion characters. The data can then be found in 15 thousandths of a second and retrieved at a rate of three million characters per second. Moreover, because some disks and all tapes can be easily removed and replaced without destroying the data stored on them, they provide a virtually unlimited reservoir of information.

Not only are the memory-and-recall capabilities of magnetic data-storage technology impressive but also some of the equipment represents consummate engineering as well. A rigid-disk device consists of an extremely smooth plate coated with a magnetic-recording medium and set spinning at 3,600 revolutions per minute. A microscopic recording head—literally flying on a cushion of air 10 millionths of an inch thick—skims over the plate's surface at a relative velocity of about 100 miles an hour, unswervingly following a circular track only half a thousandth of an inch wide. The head is subjected to accelerations many times that of the earth's gravity in moving swiftly from track to track, and yet it manages to find and position itself unerringly over the proper track. If the head were to "crash" into the medium at such a velocity, the disk and all the information stored on it would be destroyed. Nevertheless, these devices typically operate flawlessly for more than 10 years.

As superbly designed as such magnetic data-storage devices are, even better performance is envisioned in the future. New magnetic materials and more advanced recording heads are expected to yield a fivefold improvement in storage density over today's products within five years. Most of the devices will continue to "write" (store) and "read" (retrieve) data by means of electromagnetic heads like those in audio tape recorders, but devices that apply lasers to write, read and erase information on thin magnetic films will come to gain a significant share of the data-storage market. Such magneto-optical recording can reach even higher data-storage densities and may well replace the more traditional magnetic recording for many applications.

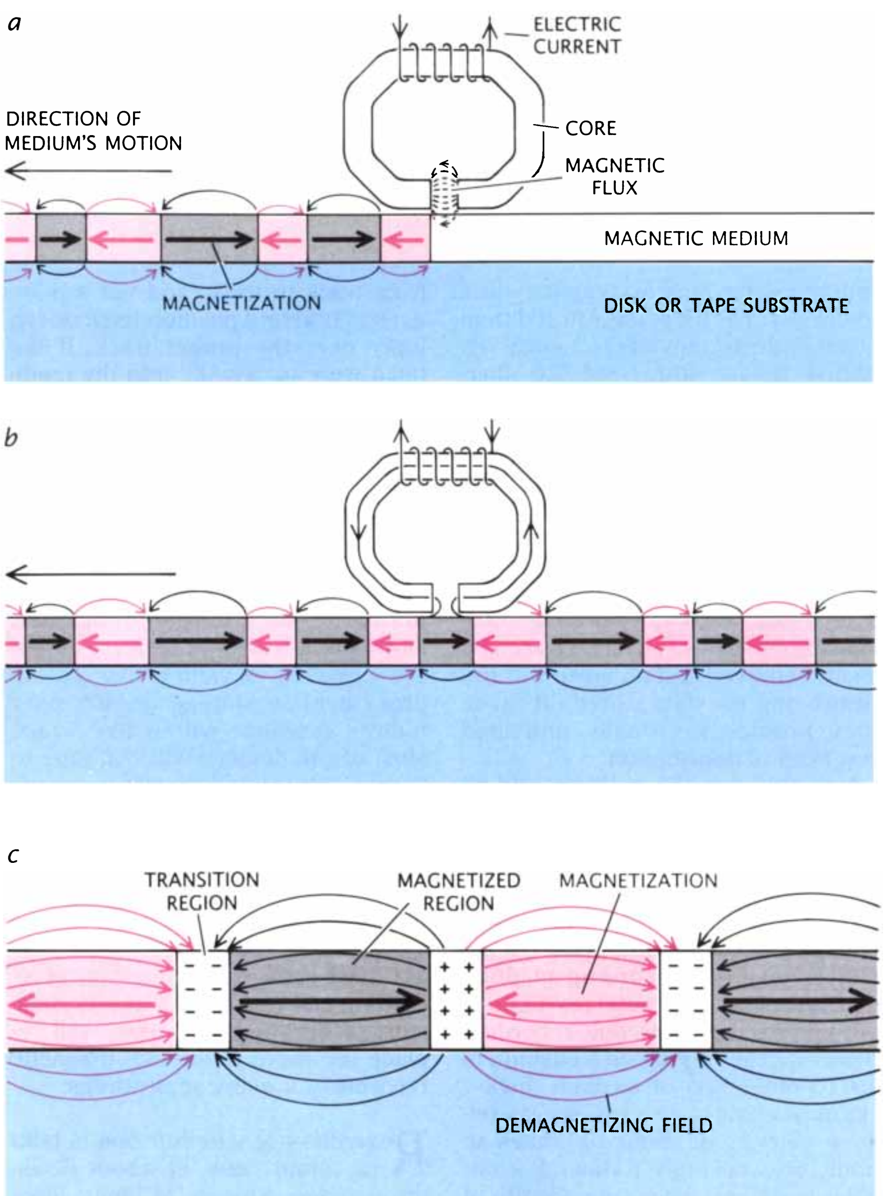
Regardless of whether one is talking about tapes or about disks, the magnetic storage of binary data essentially amounts to impressing a pattern of magnetization on a medium. This is accomplished by means of an inductive recording head, consisting typically of a few turns of wire wrapped around a ring-shaped core that can be easily magnetized. An electric current in the wire induces a magnetic flux in the core; reversing the flow of current reverses the direction of the magnetic flux. Because lines of magnetic flux spread out as they bridge a gap in the core ring,

TWO TYPES OF THIN-FILM HEADS are integrated into one microscopic device made by PCI of the Control Data Corporation. The brown coiled structure, like the coil of wire in a traditional inductive head, carries a current that establishes a magnetic flux across a gap to write data on a medium. Data are read by an underlying magnetoresistive head, which detects magnetic fields as changes in resistance. Such a head is sensitive to the fields just above a data track. Because the read element can be made narrower than the width of a data track, it is also less susceptible to interference from adjacent tracks.

they can be applied to magnetize a medium brought near the gap. The direction of magnetization depends on the direction of the magnetic-flux lines bridging the gap.

Hence a pattern of magnetization can be impressed on the medium as it moves under the gap simply by repeatedly reversing the flow of current in the recording head. Because computer-generated data are stored as a string of binary digits, the pattern of current reversals (and therefore the pattern of magnetization reversals) corresponds to a succession of 0's and 1's that constitute the individual bits of data.

To read the data encoded in a medium's pattern of magnetization, advantage is taken of the magnetic fields established by each magnetized region. These fields actually oppose the region's magnetization [*see illustration on this page*] and are therefore generally called demagnetizing fields. Because the demagnetizing-field lines extend somewhat beyond the medium's surface, they can intersect the head and induce a magnetic flux in the core—as long as the head is brought close enough to the medium. Reading is therefore accomplished by essentially reversing the recording process: when the head passes by a transition region, one in which the medium's magnetization changes direction, the corresponding change in the amount of flux through the core registers as an induced electric current in the wires around the core. The variation in induced electric current can be detected and interpreted as bits of data.



MAGNETIC DATA STORAGE is achieved by magnetizing regions of a medium on a moving disk or tape substrate. The regions are magnetized in one of two opposite directions by an electromagnetic head consisting of an easily magnetizable core wrapped with wire (*a*). When current flows through the wire, a magnetic flux is set up in the core. As the flux spans a gap in the core, the flux lines spread and magnetize the underlying medium as it moves past. If the current is reversed, the flux and the medium's direction of magnetization are reversed. The data thus stored can be "read" because the magnetic fields set up by the magnetized regions extend beyond the surface of the medium. As the medium sweeps by, the core of the head is exposed to a changing magnetic flux that induces a current in the coil of wire (*b*). A closer look at the pattern of magnetized regions and their fields (*c*) shows that the fields actually start and end between the regions, in the so-called transition regions. Because the fields counteract the magnetization of the medium, they are referred to as demagnetizing fields.

On a disk, data bits are normally recorded in concentric circular tracks. To a certain extent the number of bits that can be stored along any one track (the linear bit density) can be increased by making the transition regions smaller, thereby making it possible to fit more magnetized regions in the track. Yet there is an effective limit to how small a transition region can be. Magnetized regions in a track can be visualized as a train of bar magnets laid end to end. A transition region separates each of a magnetized region's poles (north and south) from the like poles of neighboring regions. Making the transition region smaller therefore implies that like magnetic poles would have to be brought closer together, but this is resisted because like poles repel one another. Moreover, if the magnetized regions are made too small in relation to the size of the transition regions, the demagnetizing fields may in fact be strong enough to reverse a region's magnetization. The way out of the dilemma is to ensure that the medium has high coercivity: a measure of the magnitude of the magnetic field that must be applied in order to cause the magnetization in the medium to reverse direction.

Another way to increase the linear bit density in a track is to make the layer of magnetic medium thinner. Because demagnetizing fields arise from transition regions, making the regions thinner weakens the fields. Yet if the medium is made too thin, the demagnetizing fields that extend beyond the medium will be too weak to be detected, making it impossible for the data encoded in the medium's pattern of magnetization to be read.

Here again careful choice of the medium material is one way to alleviate the problem. If a material can be made strongly magnetic, relatively little of it needs to be deposited on the substrate. It is therefore possible to have a thinner medium layer while preserving the same readback-signal amplitude in the head. Materials that are strongly magnetic are said to have a high magnetization. Hence both high magnetization and high coercivity are desirable properties of a magnetic medium if the signal is not to be lost when the linear bit density is increased.

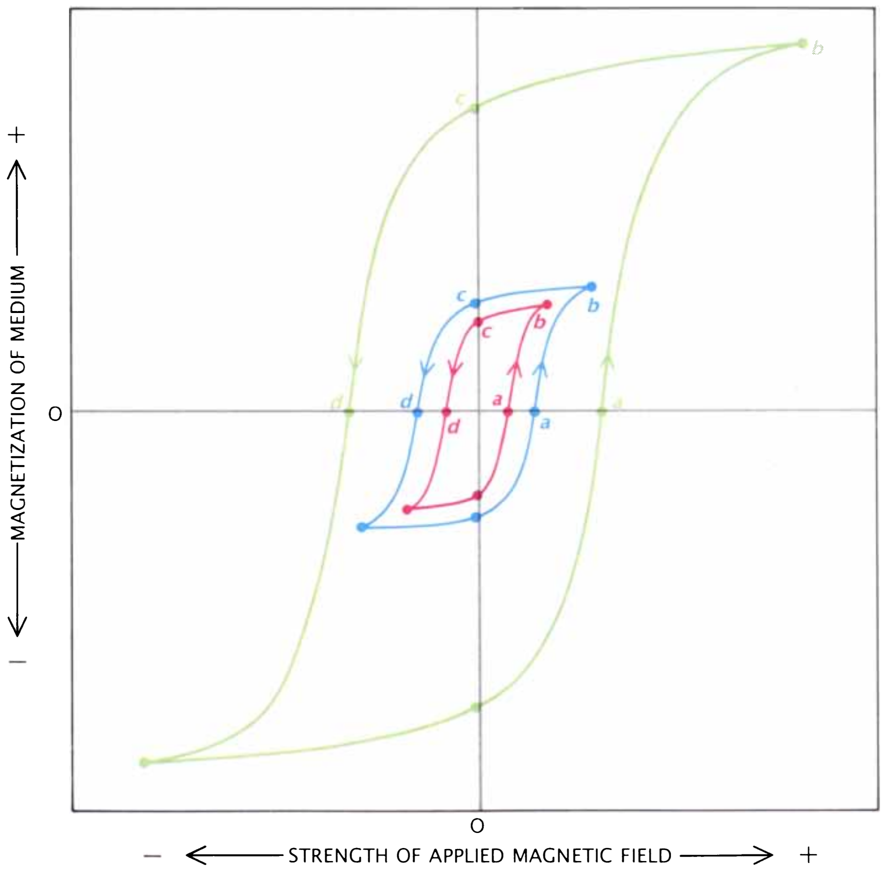
Because needlelike particles of iron oxide (in a form known as gamma iron oxide) that have been immobilized in a binder matrix exhibit the desired properties of high magnetization and high coercivity, this composite material has become the most widespread recording medium. To raise the coercivity of the iron oxide medium further the particles are often coated with cobalt. Another popular high-coercivity medium consists of particles of chromium dioxide in a binder matrix. The chromium dioxide particles have a coercivity similar to that of cobalt-coated iron oxide particles, but they are more regular in shape. Their more uniform shape is thought to make it easier to disperse them evenly in the binder, resulting in a better data-recording medium.

Particles of a pure metal, such as iron, are being developed for advanced high-density recording, since they offer even higher coercivity and magnetization than the particles in current magnetic mediums. A problem with particles of pure iron is that they are so small and so magnetic that they are hard to separate from one another, making it difficult to disperse them uniformly in a binder. In addition pure iron oxidizes easily; fine iron particles will actually burn when they are exposed to air. Consequently the particles must be handled with exceptional care during manufacturing. Moreover, the binder must not only hold the particles to the substrate but also act as a barrier against water vapor and oxygen. Finally, pure iron is mechanically softer than iron oxide and therefore more susceptible to wear.

Although particulate mediums will no doubt continue to be applied in magnetic data-storage devices, thin films of metallic alloys containing mainly nickel and cobalt have recently had a significant impact in the market for advanced magnetic data-storage devices. Thin-film mediums are not particulate but polycrystalline: they are made up of grains, or tiny crystals, of magnetic material. In effect the grains behave as if they were tiny magnetic particles, but they are packed together much more densely than the particles in ordinary recording mediums. Thin-film mediums can therefore be made to have higher coercivity than particulate ones and to offer higher magnetization as well.

Most thin-film mediums are deposited on a disk substrate by means of a plating process. The disk is immersed in a chemical bath containing, among other substances, the material that is to be the medium. The composition of the bath, its temperature and its agitation affect the final composition and properties of the plated medium.

Another process by which thin-film mediums are built up on a substrate is sputtering. Sputtering is done in a low-pressure chamber containing an inert gas such as argon. In the chamber there is a negatively charged electrode, called the cathode, and an electrode kept at or near ground potential, called the anode. The material to be sputtered onto a substrate is placed in electrical contact with the cathode. The substrate (generally a disk) is usually kept near zero potential and placed in close proximity to the charged material. A hot filament causes the argon gas in



MAGNETIZATION AND COERCIVITY of iron oxide particles (*red*), chromium dioxide particles (*blue*) and a thin-film alloy (*green*) are indicated respectively by the height and width of each hysteresis loop: the closed path traced by each medium's magnetization as a function of an applied magnetic field. The hysteresis loops show that as the field strength is increased, the magnetization also increases from zero (*a*) to a maximum value (*b*). If the field is then eliminated, the magnetization does not drop to zero (*c*); it can be brought to zero only if the magnetic field is reversed and increased to a value known as the coercivity of the material (*d*). Coercivity is a measure of the material's resistance to reversal of its magnetization. To serve for high-density data storage a material must sustain high magnetization and have high coercivity; in other words, the hysteresis loop of the material has to be long and wide. By this criterion thin-film alloys make the best high-density medium, followed in order by chromium dioxide and iron oxide.

the sputtering chamber to ionize, and the high negative voltage (typically 1,000 volts or more) of the cathode and the material attached to it attracts the positively charged argon ions. The argon ions impinge on the material with such high energy that atoms or clusters of atoms are knocked free from the material and collect on the nearby substrate.

So far plated thin films have been more popular than sputtered ones. This can be attributed in part to the fact that industrial sputtering systems are more expensive than plating systems. Sputtered mediums may eventually replace plated mediums, however, since it is easier to add trace amounts of elements such as chromium or rhenium, which increase the medium's coercivity.

In addition to the careful selection of a medium's magnetic properties, changes in the design of the read/write head also offer some potential as a way to improve the performance of data-storage systems. In particular, positioning the head closer to the medium—reducing the head's so-called flying height—would allow the registering of stronger magnetic signals in both the medium and the head. Stronger signals in turn would make it possible to operate the device faster without error. Yet reducing the flying height in magnetic-disk devices is extremely difficult to do in practice. Indeed, the spacing between the head and the medium in rigid disks has already become so small that the slightest contamination by dust, fingerprints or smoke particles will cause the head to crash, destroying the data on the disk. (For this reason rigid disks are designed so that they cannot be removed from their airtight containers.)

Even in tape recorders and floppy-disk drives, which are generally regarded as "contact" recording technologies, there is a clearance between the recording head and the magnetic medium. A coating of lubricant, for example, is often applied on the medium in order to avoid excessive wear of the head and the medium. Moreover, the unevenness of the medium's surface and its relative velocity ensure that the head actually "bounces" along the medium, making only occasional contact with it. In principle the clearance could be reduced in contact-recording devices by applying a thinner coat of lubricant, making the medium smoother and applying more pressure between the head and the medium, but these measures would result in greater wear on both.

Another change in head design that can be made in order to increase the linear bit density is to narrow the gap in the core of the head. Narrowing the gap width reduces the size of the magnetized regions in the medium, making it possible to record more of them on a track. The core of most magnetic-recording heads is made from ferrite, a material containing oxides of iron and other metals such as manganese, nickel and zinc. Over the past five years gap widths on ferrite heads have been reduced from about 2.5 micrometers to .5 micrometer. Moreover, it has recently become possible to fabricate advanced recording heads on ceramic substrates by borrowing techniques applied in the manufacture of microscopic semiconductor devices. Such miniaturized heads, made from a thin film of Permalloy (an alloy of nickel and iron), currently have gap widths of about .5 micrometer; in future thin-film heads the gaps might be half that value.

Other changes in head design, although not increasing the bit density per track, can nonetheless result in a greater overall data-storage density. The most obvious change would be to reduce the width of the head, since this determines the track width. Making the tracks thinner allows more of them to be fitted on a disk, but it also makes it harder to keep the head properly on the track and to avoid interference from the magnetic signals of adjacent tracks. Hence although reducing the width of the track is the likeliest avenue for achieving enhanced areal (as opposed to linear) recording density, it will come about only with the development of more complex heads and

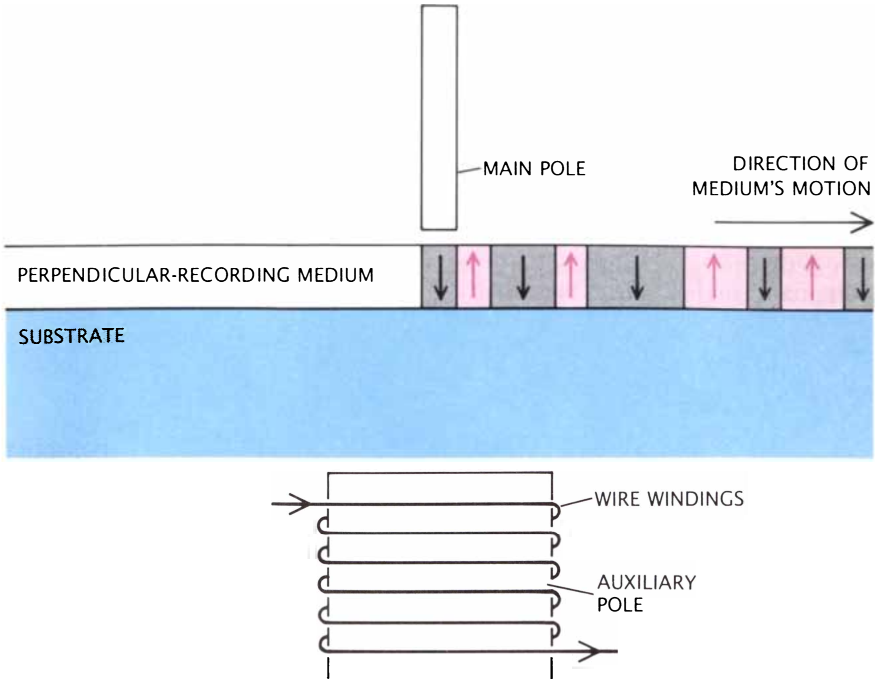
MAGNETIC PARTICLES embedded in a gluelike binder are the commonest magnetic medium for data storage. The most popular particulate mediums contain the gamma form of iron oxide (*left*). Some manufacturers of magnetic mediums maintain that because chromium dioxide particles (*middle*) have a more uniform shape, they can be spread more evenly throughout the binder. Both compounds exhibit a preferred direction of magnetization along the particle's axis. Barium ferrite particles (*right*) are shaped like hexagonal plates. Because the preferred direction of magnetization of the particles is at right angles to the plate, the compound is a candidate for application in perpendicular-recording mediums (*see illustration on opposite page*).

tracking mechanisms. The trend toward more complex head structures will accelerate the change from ferrite heads to thin-film ones.

One step in that direction has recently been taken by the IBM Corporation and the Control Data Corporation with the introduction of a device that combines a normal, inductive head for writing with a new kind for reading: a magnetoresistive head. A magnetoresistive material varies its electrical resistivity as a function of its magnetization. Therefore a change in a medium's magnetization can change the resistance of an overlying head made from such a material. If a constant current is applied to the head, the change in resistance can be readily detected as a voltage change. Magnetoresistive heads offer at least an order of magnitude more sensitivity than inductive ones to the magnetic flux emanating from the medium. Furthermore, their output signal depends on the total flux, not the rate of change of that flux, as is the case with an inductive head. Whereas the electrical signal in an inductive head depends on the speed with which the medium sweeps by it, the signal from a magnetoresistive head is independent of the velocity of the medium with respect to the head. Hence magnetoresistive heads are particularly advantageous whenever the relative velocity of the medium is low or variable, as it typically is in tape drives.

Because the magnetoresistive head cannot generate an external magnetic field, it can only read data. As a result a magnetoresistive head must always be coupled with an inductive head that writes data. Yet the magnetoresistive read head need not be a separate structure. Since magnetoresistive heads are very thin (between .03 and .05 micrometer), they can be inserted in the gap of a ferrite or even a thin-film inductive recording head. Integrating a magnetoresistive read head with an inductive write head not only improves the readback sensitivity but also solves track-interference problems during reading, since the width of the read element can be made narrower than the track width.

Strictly speaking, what I have discussed so far applies to longitudinal magnetic recording: the writing and reading of data recorded as changes in magnetization parallel to the plane of a disk or tape. Certain mediums can also be made to have their preferred direction of magnetization perpendicular to the plane of the medium. The most popular perpendicular-recording mediums are alloys of cobalt and chromium deposited by sputtering. These polycrystalline mediums tend to have all their grains oriented so that the net preferred magnetization direction is perpendicular to the plane of the film. Unfortunately such metallic thin films do not exhibit good wear characteristics. That is why barium ferrite—a particulate medium—is also being investigated as a medium for perpendicular recording. The platelike particles of barium ferrite can be mixed with extremely wear-resistant particles as well as lubricants to minimize wear of the medium.



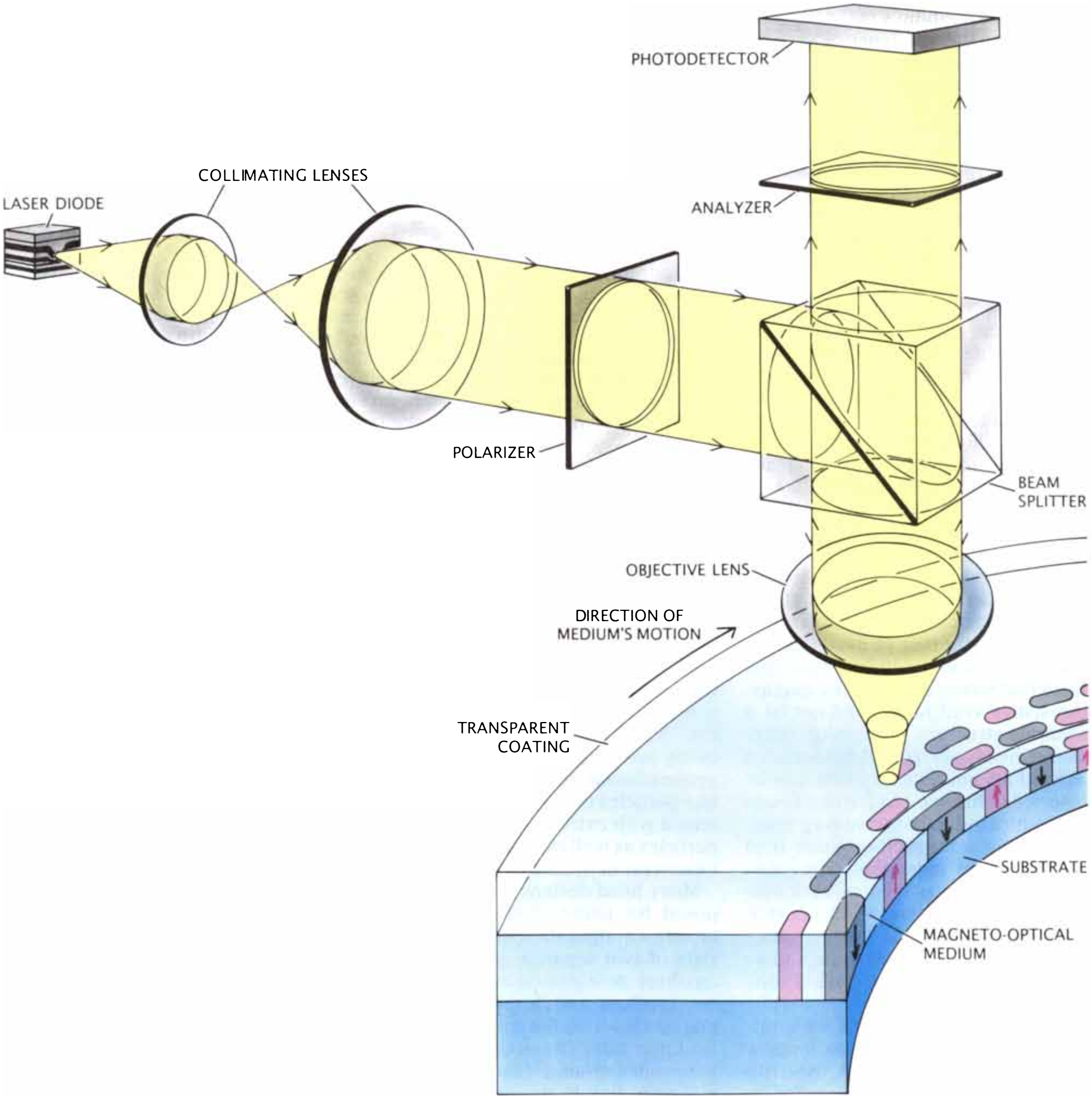
PERPENDICULAR RECORDING is accomplished by means of two poles on opposite sides of the magnetic medium. When current flows through the wire wrapped around the larger auxiliary pole, a magnetic flux is induced in the pole. This magnetic field is not strong enough to magnetize the medium, but it can magnetize the main pole on the other side of the medium. The medium is then magnetized perpendicular to its direction of travel by the combined fields of both poles. Since the magnetized regions do not lie end to end as they do in ordinary data recording, the transition regions between them are thinner, allowing more magnetized regions to occupy a track on the medium.

Many head designs have been proposed for perpendicular recording. In one configuration the head consists of two separate poles: a large auxiliary pole placed on one side of the medium and a thin main pole placed closer to the medium but on the other side. The auxiliary pole has wire coiled around it that generates a magnetic flux in the pole whenever current flows through it. The auxiliary pole's magnetic field, although it is inadequate to magnetize the medium on its own, can magnetize the main pole on the other side of the medium. The two poles' combined magnetic field is then strong enough to cause the medium to be magnetized perpendicularly to the plane of the medium; the medium's magnetized regions are, so to speak, stacked on end. The encoded pattern of such perpendicularly magnetized regions can then be read by the double-pole head much as a pattern of longitudinally magnetized regions is read by a conventional head. Changes in the direction of the medium's magnetic fields induce a varying magnetic flux in the auxiliary pole. The varying flux is then detected as current reversals in the pole's windings.

Because oppositely magnetized regions in perpendicular-recording mediums are not placed end to end (as they are in longitudinal-recording mediums), the demagnetizing fields in perpendicular-recording mediums tend to become weaker as the linear bit density is increased. Another advantage of perpendicular recording arising from the different arrangement of the magnetized regions is that the medium can be made thicker and therefore more resistant to defects without making the transition regions bigger. In addition, because the demagnetizing fields are weaker, there is less variability in the width and spacing of magnetized regions, resulting in fewer errors in writing and reading.

In spite of its inherent advantages perpendicular recording still has not been incorporated into a significant commercial product. Part of the reason is that to take advantage of the potentially higher linear bit density of perpendicular-recording mediums it probably will be necessary to get the main pole closer to the medium than is the case now for longitudinal-recording heads. Yet, as was mentioned above, this is the most formidable barrier to achieving higher linear bit density in any type of magnetic data recording.

The main hindrance to the commercialization of perpendicular recording is that it cannot be introduced piecemeal by adapting one part of the data-storage system to it at a time. Not only must the medium be changed but also the recording head and the signal-processing electronics must be changed. Although several companies have been aggressively studying perpendicular recording for a decade, the data-storage industry has at the same time become more



MAGNETO-OPTICAL DEVICE employs a laser to write and read data. To write, the laser heats a spot on a medium whose coercivity drops with increasing temperature, making it possible to magnetize the heated spot easily with a weak applied magnetic field. To read the data, the laser beam is switched to a lower intensity and polarized. Because the plane of polarization is rotated when polarized light is reflected off a magnetized medium, a second polarizing filter (called the analyzer) can convert the change in polarization of the reflected beam into a change in light intensity. A photodetector registers changes in intensity.

entrenched in longitudinal-recording technology. Since it appears possible to achieve higher bit densities by continuing to advance the established technology (chiefly by developing thin-film mediums), companies lack the incentive to make the investment necessary to change recording technologies.

It appears that longitudinal recording will continue to dominate, at least in rigid-disk drives. On the other hand, thin-film mediums are unlikely to cope with the abrasion they would suffer in tape recorders and floppy-disk drives, where the head makes contact with the medium. In the case of contact storage devices perpendicular recording might well be incorporated—if a suitable wear-resistant medium can be found.

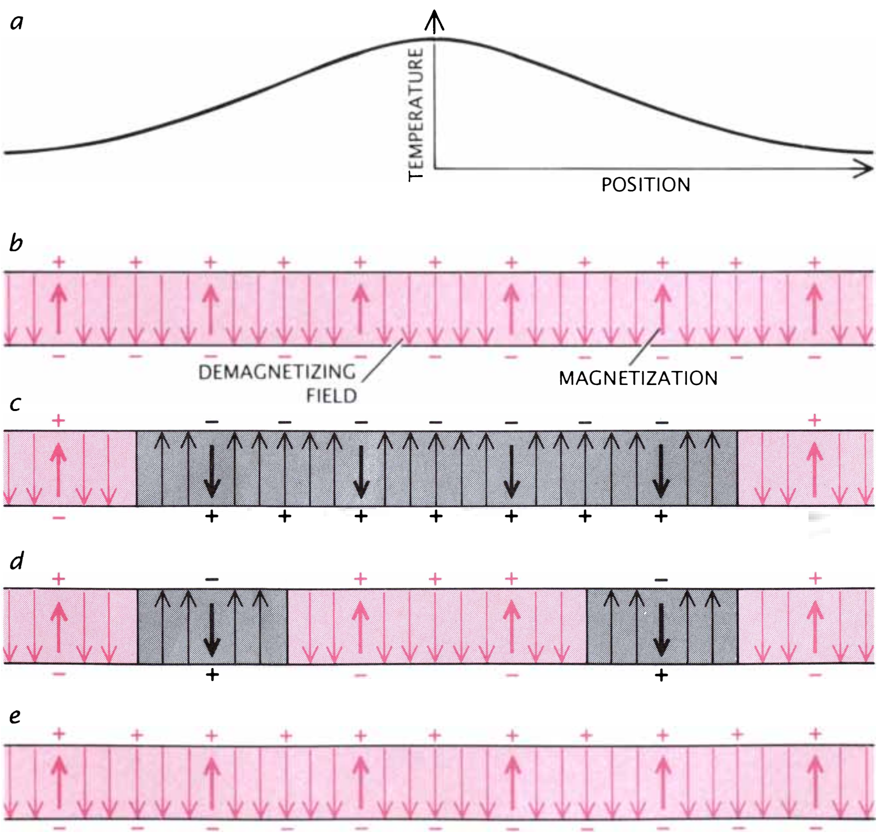
The rapid pace of improvement in magnetic data-storage devices and the great size of the industry that manufactures them make it virtually impossible for any other data-storage technology to secure a major part of the market in the next five years. Nevertheless, other technologies are threatening to become competitive. For example, optical-recording technology, which dispenses with magnetic mediums, offers a higher linear bit density and lower cost per bit stored than magnetic-recording technology does. Yet current optical drives offer considerably less performance than magnetic drives in terms of data-recording and -retrieval rates and data-access time. Furthermore, they are not erasable: data can be written into an optical disk until it is full and then it can be repeatedly read, but new data cannot be written in place of the old.

The limitations of an optical data-storage system can be circumvented while its advantages are preserved in magneto-optical recording—a hybrid system that relies on a laser beam to write and read data on a special magnetic medium. Like perpendicular-recording mediums, the medium on magneto-optical disks is made to have its preferred axis of magnetization perpendicular to the disk's plane. The key difference between the two mediums is that the magneto-optical medium is designed to exhibit high coercivity at room temperature but low coercivity at higher temperatures. Hence a laser beam can be focused on a small region on the magneto-optical medium, heating it to a temperature at which the region's magnetization can be easily established by an applied magnetic field. As in magnetic-recording mediums, the pattern of magnetization encodes binary data in the magneto-optical mediums.

To read the recorded data a phenomenon known as the Kerr effect is applied. The Kerr effect is a rotation of the plane of polarization that a light beam undergoes when it is reflected off a magnetic medium. Depending on whether the magnetization in the magneto-optical medium is pointing up or down, the rotation will be either clockwise or counterclockwise. The data can then be read from the magneto-optical disk by detecting the rotation in polarization that occurs when a polarized laser beam is made to reflect off the medium. Reading data from a magneto-optical disk can be done with the same laser that wrote the data but operating at a fraction of the power applied for writing.

The most promising mediums for magneto-optical recording consist of an alloy of rare-earth elements (such as gadolinium and terbium) and transition-metal elements (such as iron and cobalt) sputtered onto a disk. The magnetic moments of the individual rare-earth and transition-metal atoms are oppositely directed, so that the net magnetization in the bulk alloy is actually the difference in magnetization between the two elements. Because the magnetizations of the two types of elements vary differently as the temperature changes, the net magnetization and coercivity of the alloy can be made to change with temperature.

In particular, the magnetization of the rare-earth element in the alloy dominates at lower temperatures and the oppositely directed magnetization of the transition metal dominates at higher temperatures. At a certain intermediate temperature, which is referred to as the compensation point, the net magnetization of the alloy is zero. If the temperature is raised high enough, however, the net magnetization again approaches zero because the thermal motion of the atoms becomes strong enough



DEMAGNETIZING FIELDS in a magneto-optical medium can be applied to write and erase magnetized regions, eliminating the need for an external magnetic field. A laser pulse gives the medium a characteristic temperature distribution (*a*) that reduces the coercivity in a uniformly magnetized region (*b*), enabling the internal demagnetizing field to flip the magnetization of the heated region (*c*). The demagnetizing field of the flipped region is also reversed, so that a later pulse can cause the magnetization in the center of the flipped region to revert to its original direction (*d*). The outer parts of the twice-flipped region are unstable and collapse, leaving the medium as it first was (*e*).

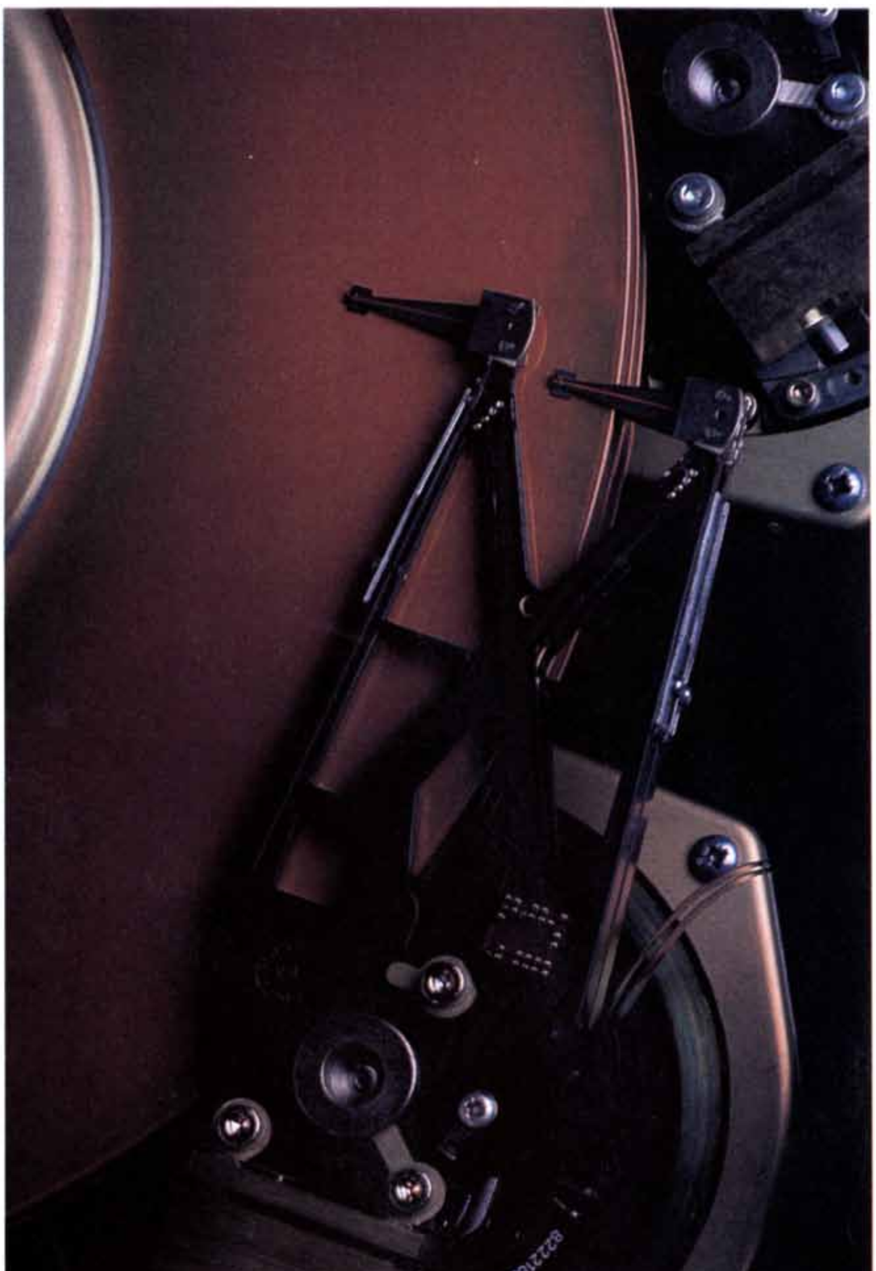
to randomize the orientation of the atomic magnetic moments. The temperature at which this takes place is called the Néel temperature.

At the compensation point the coercivity tends toward infinity. The reason is that an applied magnetic field has no net magnetization with which to interact. On the other hand, just below the Néel temperature a weak magnetic field is sufficient to determine the direction in which the transition-metal atoms are magnetized. As a consequence a magneto-optical medium exhibits high coercivity near the compensation point and a decreasing coercivity as it approaches the Néel temperature.

Although it had long been recognized that magneto-optical disks are in principle erasable and rewritable, many investigators believed until recently that it was impossible to reverse the magnetization in a magneto-optical medium at the rate at which data could be written on a virgin medium. The fields needed to reverse the magnetization of the medium—even when it is heated by a laser—were thought to be too strong. A head capable of generating such field strengths and reversing them in step with a reasonable data-recording rate would have been very difficult to achieve. For this reason investigators thought that in order to rewrite on a magneto-optical disk it would be necessary to spin the disk first through one revolution under the optical head just to erase old data and then through another revolution to write new data, effectively uncoupling the data-erasure rate from the data-recording rate. Such a design, however, would add several milliseconds to the time required for overwriting a disk.

It has now been shown that direct overwriting of a magneto-optical medium is possible. The trick is to make the demagnetizing fields in the disk's medium strong enough (or, equivalently, make the medium's high-temperature coercivity low enough) so that they flip the magnetization of a region whenever the medium is heated; no external magnetic field needs to be applied [*see illustration on preceding page*]. In one possible arrangement the direction of magnetization of a region on the disk is read by a low-power laser system and then compared with the direction the new data assigns to the region's magnetization. If the magnetization needs to be reversed, a second laser system directs a pulse of laser light onto the region, heating it so that the demagnetizing fields can flip the direction of magnetization. Alternatively, the leading laser heats the medium in the region to be overwritten with a laser pulse of sufficiently short duration for regions with reversed magnetization to revert to the original state while leaving the others unaffected, thereby ensuring that the magnetization of all regions point in the same direction. The trailing laser then emits longer light pulses to flip the magnetization of those regions that need to be reversed in accordance with the new data. In either case both laser beams can be focused through one lens and their focal points need be only a few micrometers apart.

Because the read/write laser can be focused on a small spot (on the order of half a micrometer) and can follow a track to an accuracy of .1 micrometer, extremely thin tracks of data are possible. Indeed, experimental magneto-optical data-storage systems have demonstrated that areal bit densities more than an order of magnitude higher than that of current magnetic disk drives are possible. Moreover, the system's focusing



MAGNETIC DISK DRIVE made by the IBM Corporation applies thin-film electromagnetic heads (at the tips of the triangular arms) to write data quickly and at high density.

lens can be kept several millimeters from the medium and still maintain a sharp resolution through a transparent layer. The large spacing between the lens and the medium ensures that there is no concern about head crashes, and the overlying transparent layer protects the medium from environmental degradation. Furthermore, since foreign matter on the layer is not in the focal plane of the lens, it does not cause errors as long as it is not very large. (It is for this reason that the familiar digital audio compact discs, which also read data optically, can be made removable.)

In addition to the optical components associated with the task of writing and reading, a magneto-optical head also contains components that maintain the laser beam in focus and make it possible for the lens to track the data on a spinning disk properly. All these parts are typically fitted into a head measuring only a couple of centimeters on a side and weighing about 150 grams. Yet as small as this may seem, it is large compared with a conventional magnetic-recording head. Because of the larger mass of the optical heads, they cannot be rapidly moved from track to track, and access times in optical drives have typically been about .1 second—an order of magnitude larger than the times for magnetic drives.

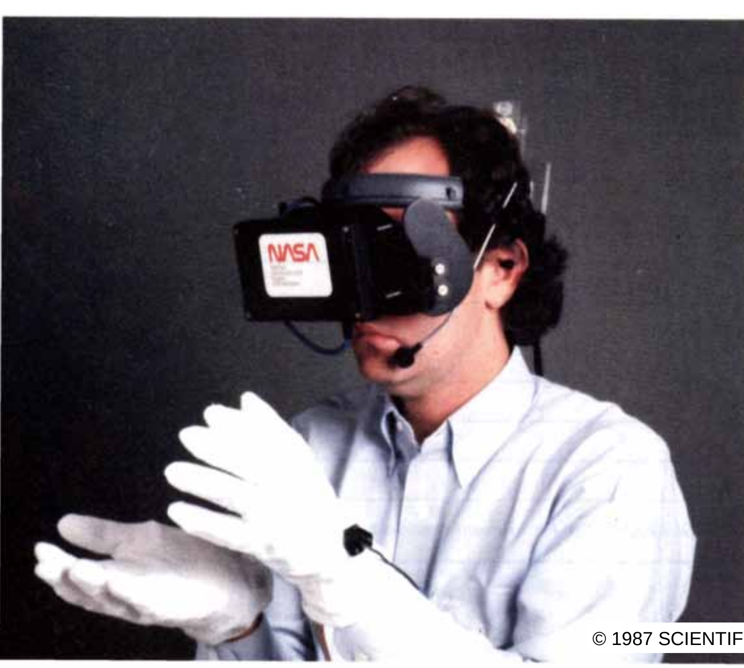
Nevertheless, it is likely that the access times of magneto-optical drives will be significantly reduced. Some investigators, for example, have demonstrated that optical fibers can link the lens with the other optical components. In such a system only the structure associated with the lens needs to be moved to access a track on the disk, and this structure can be made to weigh about as much as a magnetic head. Other workers have proposed integrating optics and the necessary electronic devices in a single device by means of microfabrication techniques. It seems likely that these innovations as well as others will make it possible to reduce the access times of magneto-optical drives so that they are comparable to the times of magnetic drives.

The capability of overwriting old data directly with new data, along with the very high data-storage densities (on the order of 10 million characters per square centimeter), the absence of concern about head crashes and the fact that optical mediums are removable, makes magneto-optical recording a strong competitor of conventional magnetic recording for data-storage applications.


NASA

NASA

NASA

NASA

NASA

NASA

Interfaces for Advanced Computing

Why should sophisticated computers be difficult to use? The coming generation of supercomputers will have the power to make elaborate "artificial realities" that facilitate user-computer communication

by James D. Foley

A flight simulator exemplifies the ability of modern computer technology to mimic reality. Computers orchestrate the sound, force and motion that approximate the aerodynamic behavior of an airborne plane, and specialized supercomputers provide the imagery. The convincing visual displays are particularly difficult to generate; supercomputers can be masters of illusion. Yet this mastery is rarely exploited in ordinary scientific applications.

Why confine the simulation capabilities of computer technology to the cockpit? Might not the machine that can re-create the sensations of flight also synthesize familiar contexts for scientific problems? Might it not build a communicative environment more natural than the customary typed commands and keyboard? In short, should it not be possible to program a computer to construct an "artificial reality" with which a user could interact?

For many computer scientists and engineers the answer to that question is an emphatic yes. Blueprints for artificial realities even more complex than flight simulations have already been drawn up. Interface technologies are being developed that will make supercomputers more responsive to human modes of communication including touch, gestures, speech and even a kind of eye contact. In addition to more realistic graphics displays, the next generation of supercomputers may feature hands-on manipulation of computer-generated images along with tactile sensations and force feedback. Sensors will measure the position of a user's head and track the movements of his eyes; voice-recognition programs will allow computers to interpret spoken language.

Researchers hope that artificial realities will make learning about and exploiting a supercomputer's capabilities more efficient and enjoyable. These elaborate interfaces might do for scientists and engineers what spreadsheet programs such as Lotus 1-2-3 have done for accountants. Of course, artificial realities hold more promise for some problems than for others; it is hard to imagine, for instance, how desktop publishing would benefit from three-dimensional representation and wall-size screens. On the other hand, many scientific problems, particularly those that can be represented in three dimensions, call out for a greater degree of interaction between man and machine.

The interface between the user and the computer may be the last frontier in computer design. In the past few years hardware costs have fallen dramatically; software costs are also decreasing, albeit less rapidly. Techniques for maximizing computer efficiency and minimizing the use of memory have largely been established, although they are constantly being refined. What remains to be addressed is the maximization of user efficiency.

Scientific computing in particular has been forced to focus on hardware performance because its computational needs are so great. With the coming generation of supercomputers, that focus may finally be allowed to change. A Cray-2, currently one of the fastest computers in the world, can perform 1,000 million floating-point operations per second (MFLOPS). (FLOPS are a standard industry measure of computational speed; personal computers run at speeds of anywhere from 1,000 to 100,000 FLOPS.) The coming generation of Crays, slated to appear next year, will run 10 times faster. Even the workstations that provide graphics for supercomputers are becoming more powerful: a workstation Stellar Computer Inc. expects to release by January will operate in the 50-MFLOPS range.

How can this additional power make the interaction between user and computer more like cooperation than confrontation? Obviously an increase in the speed of computation is a boon in itself, decreasing and in some cases eliminating the waiting period during which the computer generates results. Less obvious are the means for coupling supercomputer users more tightly to the problem-solving process.

Currently, using a typical supercomputer is a bit like consulting a machine-age oracle. There is a period of preparation during which a problem is defined and its parameters are specified on a workstation. The workstation organizes the problem in a form appropriate for the supercomputer; its ensuing compu-

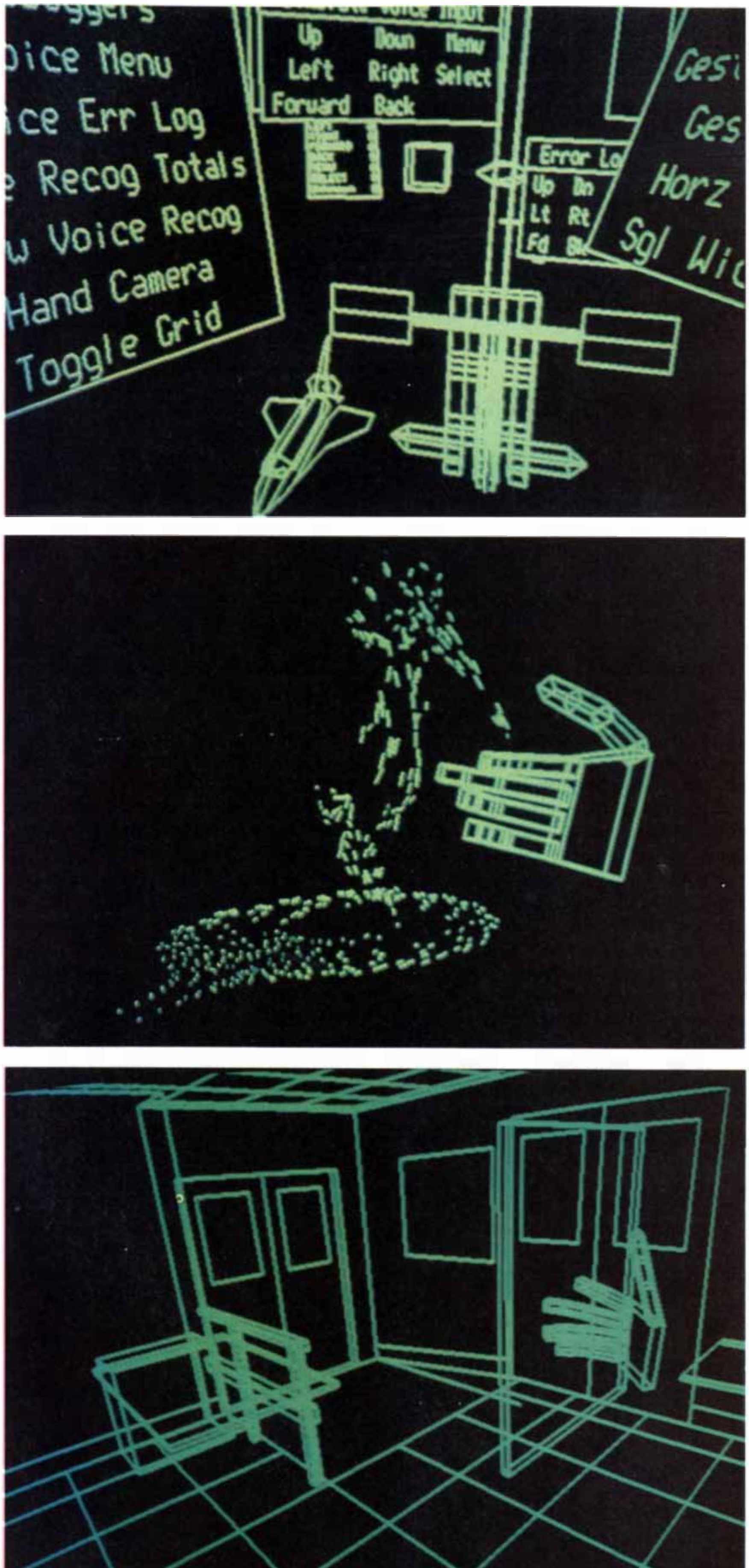
HEAD-MOUNTED MONITOR with a position and orientation sensor, gloves that track hand and finger movements and a microphone wired for voice recognition transport the user to a computer-generated reality. The user issues instructions to the computer by pointing, talking, gesturing and actually handling graphics images. Workers at the National Aeronautics and Space Administration's Ames Research Center have constructed several artificial realities for use in this system (*see illustration on next page*).

tations can take seconds, hours or days. In many cases the user cannot interrupt or alter the computations once they have begun, and if the results suggest checking alternative parameters, the ritual must be repeated from the start.

The advent of artificial realities, foreshadowed by flight simulation, will fundamentally change the way a person works with a supercomputer. Artificial realities allow the user to interact with the computer in an intuitive and direct format and to increase the number of interactions per unit of time. The ultimate objective of artificial-reality research is to develop a simulated environment that seems as "real" as the reality it depicts. The profoundest strength of the interfaces, however, may lie in their ability to go beyond reality itself, by modeling in concrete form abstract entities such as mathematical equations and by enabling users to surmount problems of scale in manipulating atoms and galaxies alike.

Artificial realities have three components: imagery, behavior and interaction. Realistic visual *imagery* helps the user to interpret the information being presented by the computer. The images may represent real objects, such as building frames, or abstractions, such as patterns of fluid flow. These images *behave* the way the objects or abstractions they represent would behave. Behavioral modeling exacts the heaviest computational toll, because it often entails solving extensive sets of equations over and over again. Finally, the user *interacts* with an artificial reality in much the same way as he interacts with the three-dimensional world: by moving, pointing and picking things up, by talking and observing from many different angles.

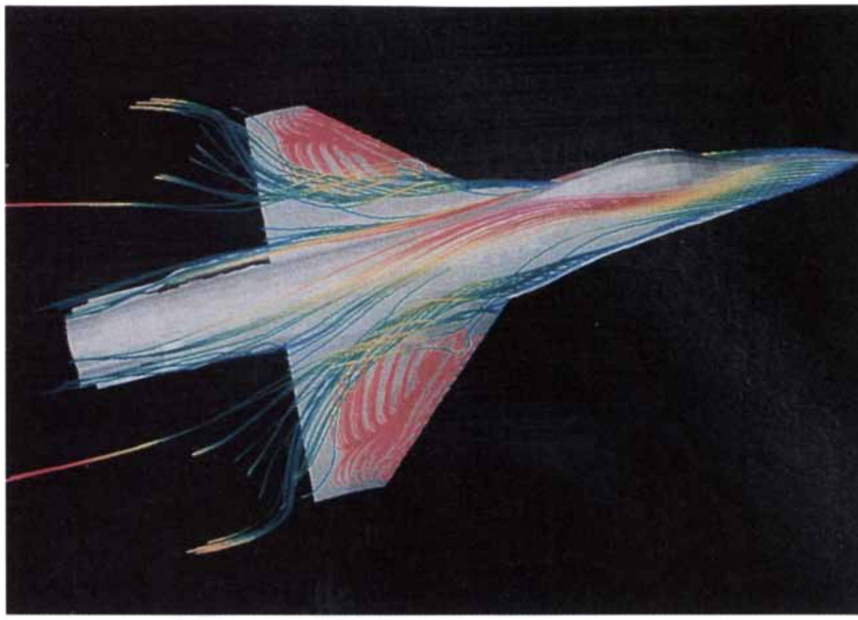
The interactive component of artificial realities lags behind the other



GRAPHICS DISPLAYS projected inside the helmet shown on page 126 include menu interfaces (*top*), airflow patterns (*middle*) and the laboratory housing the system (*bottom*). These photographs only hint at the realism of actual displays, which provide depth cues by showing each eye a view from a slightly different perspective and also allow the user to pan across the computer-generated environment by turning his head. The user can select a menu option with a word or a gesture, turn an airflow model to look at it from another angle or reach out and "touch" the laboratory's walls and desks.

components. Commercial technologies for viewing and manipulating in three dimensions are still relatively primitive. What kinds of interactive devices will restore the balance?

I shall begin by addressing the most familiar component of the computer interface: the display monitor. The typical computer workstation is equipped with a 19-inch-diagonal color monitor; viewed from a distance of two feet, the display subtends 37 degrees of visual field horizontally and 28 degrees vertically. Yet the visual field of one eye, assuming the head is fixed, spans 180 degrees horizontally and 150 degrees vertically. Displays that fill the visual field give the observer a sense of being part of a scene rather than on the outside looking in: witness the IMAX and OMNIMAX wide-screen motion-picture systems. Hence wall-size projections of computer screens could help to immerse the user in an artificial reality.



SIMULATION OF AIRFLOW PATTERNS generated by an F-16A fighter jet exemplifies the sophistication of current supercomputer graphics. The path of air and its elevation above the body of the aircraft are indicated by colored streaks: blue signifies low elevations, red high. Red swirls on the wings reveal areas of shear stress. The information this image provides helps engineers to design aircraft that have less drag and more lift. Equally refined graphics are already standard features of many artificial realities.

Although a larger display better fills the visual field, it will show no more detail if it uses the same number of pixels, or picture elements. Pixels are the discrete points of light that make up a cathode-ray-tube image; they define the resolution of an image. A typical workstation monitor consists of roughly a million pixels arranged in a 1,280-by-1,024-pixel grid. At viewing distance each pixel subtends about two minutes of the visual field, but the human eye can distinguish detail down to one minute. Currently a 20-inch-square color monitor is available with approximately four million pixels in a 2,000-by-2,000 grid; viewed from a distance of two feet, each pixel subtends an angle of roughly 1.4 minutes. A monitor that fully exploits the acuity of human vision at normal viewing distance has yet to be constructed.

Given these limitations, how else can realism be improved? One alternative is the head-mounted display. This display can facilitate depth perception in much the same way as a Vu-Master achieves its stereoscopic effect: each eye is provided with a slightly offset view of the same image. Other depth cues come from motion parallax, the phenomenon that describes the shift in background that occurs when an observer looking at a point in space changes position. This effect is achieved with the aid of a sensor that registers head position and orientation. Furthermore, because the sensor recognizes gross head movements, the user can enjoy the illusion of scanning an artificial panorama as he turns his head. The images he sees depend on the direction in which he is facing.

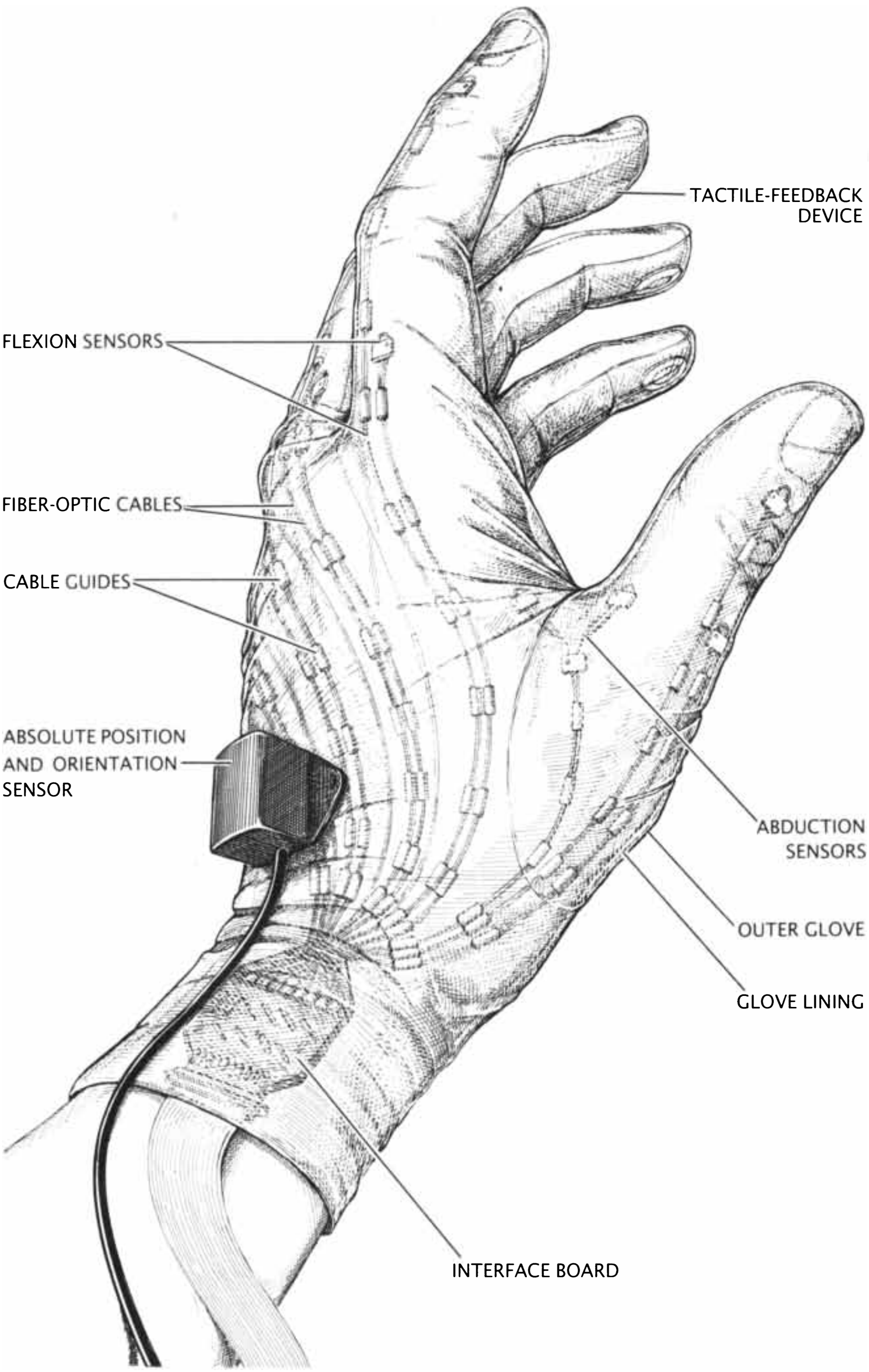
In the first head-mounted display, built by Ivan E. Sutherland in the 1960's, miniature cathode-ray tubes acted as displays and mechanical linkages relayed head position and orientation to the computer. Today lightweight liquid-crystal monitors and electronic sensors have made implementation more practical. The most advanced system incorporating these features not only creates artificial realities but also replaces one reality with another. At the Ames Research Center of the National Aeronautics and Space Administration, Scott S. Fisher, Michael W. McGreevy and James C. Humphries have constructed a helmet to be worn in the space station that would project to an astronaut inside the station what a robot operating outside the station "sees." When the astronaut's head turns, the robot's camera eyes swing in the same direction.

The electronic sensor that registers head position and orientation in the NASA system figures importantly in many other interface systems. Manufactured by the Polhemus Navigation Sciences division of the McDonnell Douglas Corporation, the sensor works by sending electromagnetic pulses from three transmitter antennas to three receiver antennas. In both the transmitter and the receiver units the antenna coils are at right angles to one another, forming a Cartesian-coordinate system.

The transmitter is a box roughly two inches on a side that must be placed within five feet of the receiver. It emits three pulses in sequence, one from each antenna. The pulses induce a current in the coils of the receiver, a cube less than an inch on a side that is placed on the object being tracked. The strength of the current depends both on the distance the receiver is from the transmitter and on the relative orientation of the transmitter and receiver coils. A computer can calculate the three-dimensional position of the receiver unit from the nine current values resulting from three successive pulses. The three pulses are repeated about 40 times per second and the resulting images move somewhat erratically; smooth simulated motion probably will not be possible until the sensor can produce 60 pulsed triplets per second.

Installed on the NASA helmet, the Polhemus sensor would be rather adept at determining the direction of a user's gaze except for one catch: the eyes can and often do move independently of the head. To surmount this difficulty engineers are exploring a technology borrowed from experimental psychology. Psychologists have long used devices called eye trackers to gather data on how people read and examine pictures. Eye trackers bounce a beam of light off the cornea of the eye. The direction in which the light is reflected indicates where the user is looking: the point of regard.

Eye trackers are still quite new to the computer scene. Trackers that attach to eyeglasses can be had for a few thousand dollars, but they are not very accurate. A more elaborate system projects a pinpoint of infrared light onto the cornea and detects its reflection with a wide-angle television camera placed approximately three feet from the user. The camera stays locked on the eye in spite of considerable head movement unless the movement is quite rapid; it might be baffled, for instance, by a sneeze.



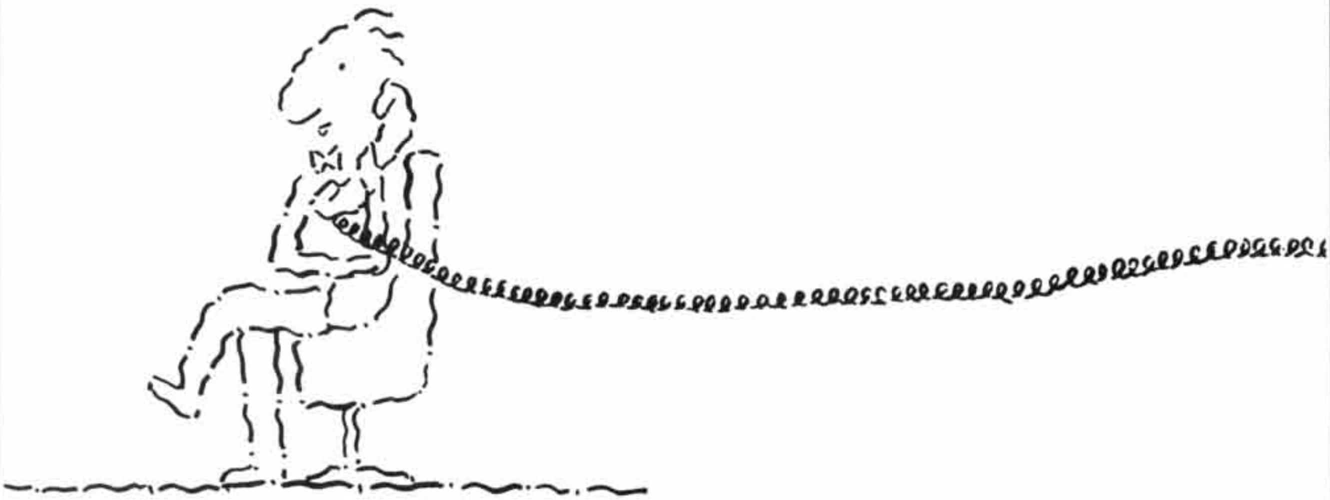
DATAGLOVE developed by VPL Research, Inc., translates hand and finger movements into electrical signals. Between two layers of cloth, fiber-optic cables anchored at both ends to an interface board run the length of each finger and double back. Each cable has a light-emitting diode at one end and a phototransistor at the other. Cables are treated so that light escapes when a finger flexes; the phototransistor converts the light it receives into an electrical signal. The position and orientation sensor is made by the Polhemus Navigation Sciences division of the McDonnell Douglas Corporation.

Wall-size screens, head-mounted displays, position sensors and eye trackers can improve the credibility of an artificial reality by broadening the visual field and by enhancing detail at the point of regard. The displays can present actual images, as NASA's space-station helmet will, or the artificial images generated by a supercomputer. How will the user interact with such displays?

Most current interaction devices are limited to two dimensions. Even manipulations of three-dimensional computer simulations must be specified through a two-dimensional medium, either a mouse or a joystick. Worse yet are dials that separately control the three axes. Suppose an interaction device could be made that combined the precision, control and agility of the human hand. Actually such a "device" is available: it consists of the hand itself, equipped with a Polhemus sensor and a special glove that can record hand and finger movements.

The DataGlove was developed in the past three years by Thomas G. Zimmerman and L. Young Harvill at a small California company called VPL Research, Inc. Fiber-optic cables sandwiched between two layers of cloth run the length of each finger and thumb. Both ends of each cable are anchored in an interface board near the wrist. An LED (light-emitting diode) at one end sends light down the shaft of the cable to a phototransistor at the other end. The phototransistor converts light into an electrical signal a computer can recognize; the signal travels from wrist to terminal through electrical wire. Although ordinary fiber-optic cables will transmit light when they are bent, the cables in a DataGlove are treated at the sites where fingers flex so that light escapes when a finger is crooked, or when, say, the user moves his thumb toward his forefinger. The greater the movement is, the more light is lost.

Coupled with a Polhemus sensor, which can be mounted on the back of the hand, the DataGlove has exciting potential. Fisher's group at Ames uses the glove in conjunction with the space-station helmet; NASA hopes that one day a robot outside the station will be able to carry out complex maneuvers and repairs by mimick-



ing the hand movements of an astronaut inside the station. VPL is also incorporating DataGlove principles in a DataSuit that covers the entire human frame.

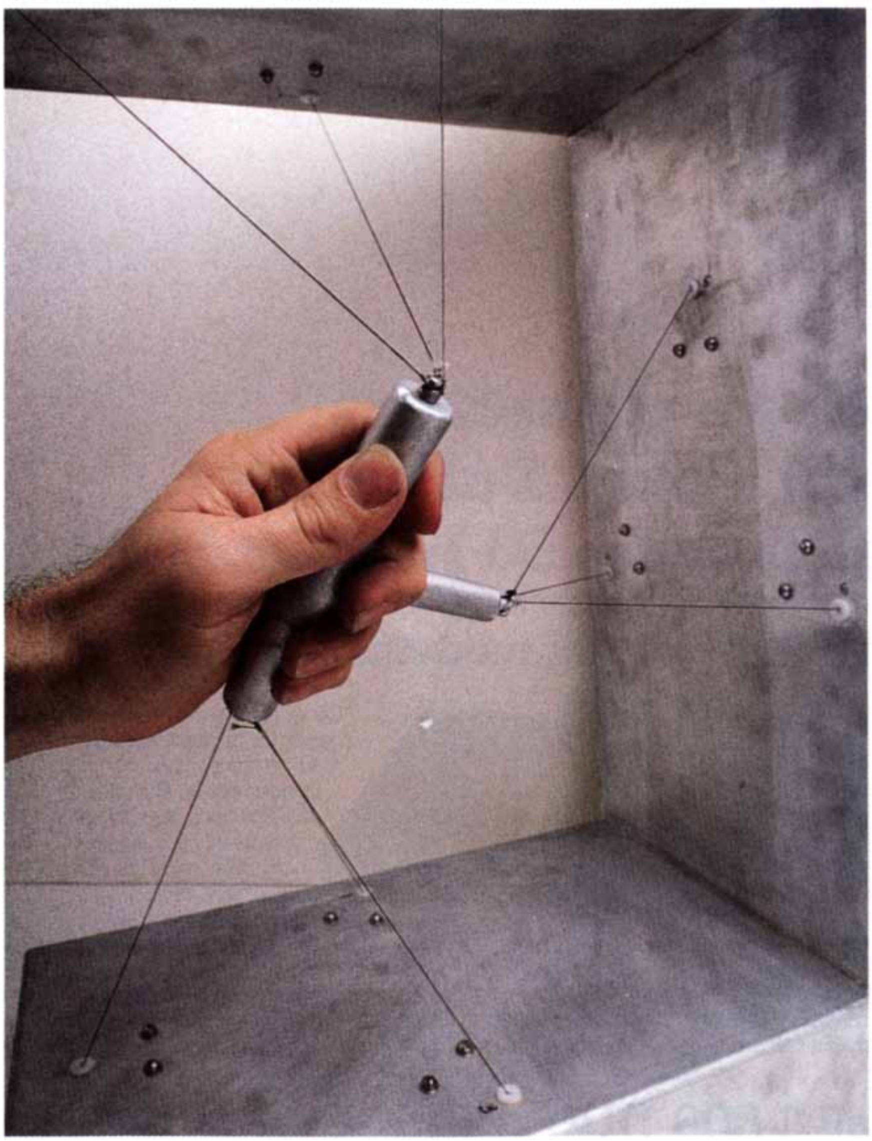
An interface as complex as the DataGlove may not make sense for every kind of problem, but many of the scientific equations that are presented to supercomputers model systems that are easier to deal with directly than by typed commands. Furthermore, artificial realities can add a tactile element to systems that cannot ordinarily be touched. For instance, imagine a biochemist examining two molecules: an enzyme and the substrate to which it binds. He knows the structure of both the enzyme and the substrate, and both molecules are displayed on a computer monitor. The chemist wants to find out what part of the enzyme interacts with what part of the substrate. Armed with a DataGlove, he could quickly manipulate both molecules like two pieces of a jigsaw puzzle to see what parts fit together.

Now imagine being able to feel the topography of the enzyme molecule: its crevices and projections, its smooth edges and sharp corners. Imagine probing with a finger the enzyme's active site, which exerts a strong chemical attraction on the substrate. Imagine maneuvering the substrate close to the active site and feeling the pull of the interatomic forces that join the two! In order to achieve these effects engineers are exploring methods for creating tactile and force feedback.

Three technologies for tactile feedback are being tested that could be incorporated in the DataGlove design. One technology is adapted from a tactile-feedback device for the blind in which small solenoids push blunt wires against the skin. This type of actuator, however, is probably too large: each is about a third of an inch thick. Piezoelectric crystals can also be employed in conjunction with the glove. The crystals vibrate when they are activated by an electric current, and the mind interprets their vibration as pressure. A third approach exploits the new "memory metals" that change shape with temperature. Small insulated pieces of this metal could be oriented to push against the skin when they are heated by an electric current.

The sensation of force is more difficult to convey in a glove device than tactile sensation, although memory metals offer some promise here as well. As long ago as 1968 a group directed by Frederick P. Brooks, Jr., of the University of North Carolina at Chapel Hill adapted for force feedback a remote manipulator device of the kind used to handle radioactive materials. Today the most effective force-feedback system is the "joystring" built by Richard J. Feldmann of the National Institutes of Health. Named after its predecessor, the joystick, Feldmann's joystring is a simple rigid T about three inches long connected at each end to three taut wires. The wires are in turn connected to shaft encoders and servomotors. The user grips and manipulates the joystring T; the computer reads the user's hand movements through the shaft encoders and generates force and torque feedback by means of the servomotors.

In addition to the joystring, only a handful of force-feedback devices have been made, each one unique. The systems have been limited to specialized applications in research laboratories. Simpler designs can respond to force input but cannot generate force feedback. At George Washington University my colleague John L. Sibert and his group have developed a novel application of such a force-input system. They have devised a paint system for artists based on a commercially available data tablet that senses not only the position of a stylus but also the orientation of and force applied to the stylus. A computer uses this information



JOYSTRING is one of the more effective force-feedback devices. Designed by Richard J. Feldmann of the National Institutes of Health, the apparatus relays to a computer the position of a hand grasping the suspended T; the supercomputer in turn directs servomotors to exert force through differential tension on the nine wires connected to the T.

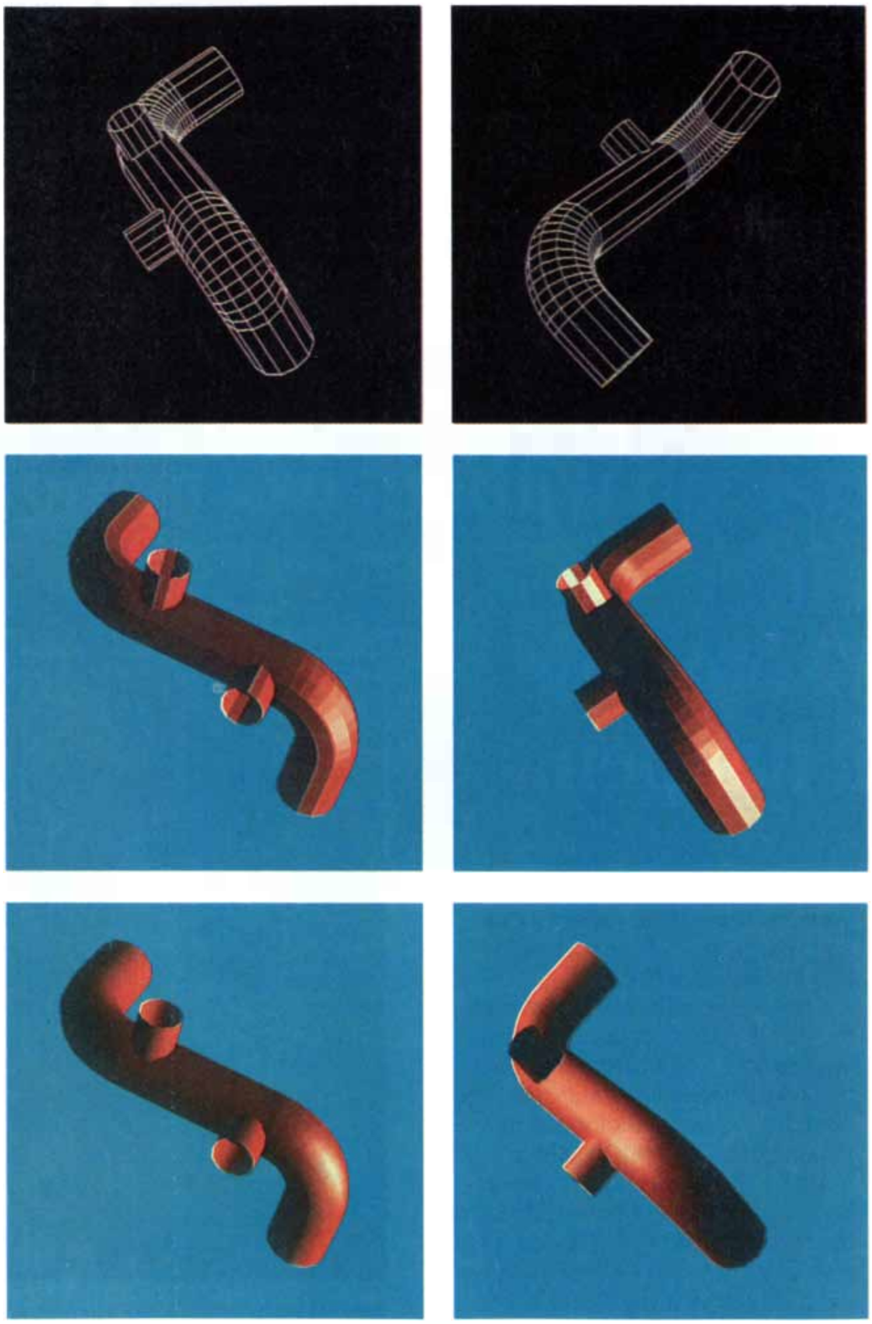
to simulate the behavior of a paintbrush. As downward pressure increases, the "brush" spreads out and the width of the line being drawn changes with the orientation of the stylus. Sibert's artificial reality reflects to some extent the artist's traditional tools, but it also creates artistic tools that until now have not existed.

In what other ways might a user want to interact with a supercomputer? In many cases talking or gesturing to a computer may be more appropriate or convenient than nonverbal manipulations of symbolic images. Both voice- and gesture-recognition systems are further advanced than the other components of artificial realities. More than 10 years ago Nicholas Negroponte and Richard A. Bolt of the Massachusetts Institute of Technology demonstrated the feasibility of voice recognition in computer interaction. Now machines with vocabularies of several hundred words are routine co-workers in jobs requiring use of both hands, such as the testing of electronic assemblies.

The technology for voice recognition simply has not yet been integrated into most computer problem-solving environments. This unfortunate procrastination in adopting new technology has occurred in the past: the mouse, developed in the late 1960's, was not commercially available until early in the 1980's. Meanwhile voice-recognition systems are becoming more sophisticated. Kenneth Davies of the IBM Corporation's Thomas J. Watson Research Center recently demonstrated to me an experimental system with a vocabulary of 20,000 words—about 98 percent of the typical English speaking vocabulary. The machine can even interpret phonetically abstruse phrases such as "Write Ms. Wright a letter right away." IBM expects to make its system affordable within a few years.

The technology for gesture recognition has been languishing for several years with some difficulties waiting to be ironed out. Systems must be taught to sort through ambiguities in strings of gestures and to discern when one command ends and the next begins. James R. Rhyne of the Watson Research Center has developed a system that recognizes hand gestures made in two dimensions with a penlike positioning device. In a spreadsheet application the user can total two groups of numbers by circling each group on a data tablet and making a summation sign. The computer enters the total at the position on the spreadsheet indicated by the summation sign. Although this technology is still in its infancy, I can envision a day when a biochemist clad in DataGloves gestures at molecular conformations on a supercomputer display and says, "The phenylalanine 221 on this helix [pointing] doesn't interact properly with this [circling] glutamine 57. Change it to a histidine."

Artificial realities are well on their way—in fact, one of the most difficult hurdles remaining involves integrating technologies that have already been established. Once this hurdle has been cleared, the gadgets themselves will require a relatively small investment; a DataGlove, for instance, costs about $8,000. Indeed, the research and programming effort currently expended on advanced interfaces will probably be the most expensive aspect of their implemen-



ROTATED PIPES helped the author to explore how realism affects a user's performance. Subjects shown two images side by side were asked whether or not one image was a rotation of the other. The author found that users compare pipes that are colored in and highlighted (*middle*) about 20 percent faster than they compare outlined figures (*top*), but that further refinement (*bottom*) does not significantly improve comparison.

tation. Will the benefits of artificial realities justify the cost?

Instinctively it would seem that a system inviting user interaction and presenting information in accessible formats would be much faster, more instructive and easier to learn than conventional interfaces. This kind of hunch is very difficult to quantify. Would a materials engineer understand a stress analysis better if he could apply the stress with his own hand? Will molecular interactions become more obvious if their forces can literally be felt? And how "real" do artificial realities need to be to accomplish their purpose?

I began to ask this last question as it pertains to visual imagery some time ago. Working with my colleague Woodrow Barfield and a graduate-student team led by James W. Sandford, I examined experimentally the effect of increased realism on the speed with which computer users carry out a simple mental manipulation of two images. The task, called mental rotation, is commonly presented to subjcts by experimental psychologists studying how people represent images in their minds. My subjects were shown two different images of pipelike structures side by side and asked to decide whether or not one image was a rotation of the other. In some tests the images were only outlined; in another set they were colored in and highlighted, and in a third, colored and highlighted images were given smoothly shaded surfaces. We found that users can compare colored-in images about 20 percent faster than outlined figures, but that the more sophisticated shaded representations did not improve comparison time any further. Hence there may sometimes be a limit to how "real" artificial realities must be.

As for questions about the ultimate value of artificial realities, the answers are still to come. Nearly 20 years ago Brooks's graduate student James J. Batter noted that some students studying graphic displays of two-dimensional force fields gained a better understanding of the concepts involved if they could not only see the force vectors but also feel them. Batter's study, which used a simple two-dimensional force-feedback device, is the most recent example of research assessing the worth of artificial realities. Along with my experiments at George Washington University, it illustrates the type of research needed to determine whether the promise of artificial realities is more than an illusion.

Networks for Advanced Computing

Computers linked in networks can interact and can share programs, data and expensive hardware. The physical link is not enough: one computer must be able to understand what a different one is saying

by Robert E. Kahn

Why are computers connected to networks? Why do people talk? The answers are similar: to share information, to cause actions or events—in general, to collaborate in some way. Much has been written about efforts to create intelligent computers that reason like human beings, but intelligent behavior is not exclusively an individual matter. It is in large measure founded on communication: the web of connections, both lasting and transient, that human beings maintain with others of their kind. Computer networking is the science of getting computers to communicate, which is as important, if not as challenging, as getting them to think.

There are two essential aspects to the networking problem. First, one must establish an electronic link between computers, along with an orderly process for transferring information over the link, intact and error-free. Second, one must establish a common language, so that one computer can understand what the other is saying over the link.

In what follows I shall touch on both subjects, but I must begin with a qualifier. There are many kinds of networks. The computerized reservations systems operated by airlines are one kind; indeed, the SABRE system introduced by American Airlines was one of the earliest examples of computer communication by way of telephone lines. Another early example of a network is the early-warning system for air defense. These networks are essentially single-purpose systems involving little direct interaction among users. (Indirect interaction in the reservation system occurs, for example, because of finite available space on each airplane.) The networks I shall discuss, on the other hand, are the ones over which many users and machines having many purposes interact directly and rapidly. Although great progress in the development of such systems has been made in the past decade, the field is still in its infancy.

Hooking up computers in a network does not guarantee that they will be able to accomplish anything, any more than an international telephone connection is a sufficient condition for meaningful communication between people. The machines must be able to understand one another. Not surprisingly (and much the same could be said of human communications) that problem is hardest to solve when the differences in hardware and software are great.

Computer networks are not limited to communicating conventional data and programs. For example, voice input may be digitized and stored for later retrieval and delivery. A network can be equipped with a voice-recognition system for control or to provide input for programs. In many cases a synthesized voice may convey output to the user. Facsimile input may be entered into a computer to be incorporated with other documents, or for transmission, storage or processing. Video input and output, graphics and computer-generated images can also be part of the interaction mediated by a network.

The origins of computer networking can be traced to the development of the first time-sharing computers in the early 1960's, a time when a computer was an expensive and scarce resource. The idea behind time-sharing is simple. As many tasks, notably program development and debugging, require only a small fraction of the capacity of a large computer, it makes economic sense for the computer to serve not just one user at any one time but many. This objective can be achieved by switching among many user programs every few milliseconds or so. The switching is managed by a program called an operating system, which also controls and manages devices such as disk drives and printers and mediates all resource allocation within the computer.

From time-sharing to networking was a short intellectual step. Once it had been shown that a relatively small group of users could share a single computer, it was natural to ask whether a large, scattered community of workers could share the resources—data bases and even specialized programs—available on their respective time-sharing computers. With this idea in mind, the Advanced Research Projects Agency of the U.S. Department of Defense (ARPA, now known as DARPA) began in 1969 to link research centers around the country in a computer network called the ARPANET. At about the same time, Tymshare saw the economic value of efficient networking to support interactive communication between remote terminals and its time-sharing computers. The telephone system could also provide communication between computers; it was already widely used to connect terminals to remote computers. The telephone companies responded to the emerging requirement for wide-area computer networks by providing point-to-point circuits for dedicated use with the networks.

The ARPANET, a wide-area network built by Bolt, Beranek and Newman, Inc., was the first interactive computer-to-computer network. As such it had to overcome major technical hurdles, the most significant of which was the need for a new high-bandwidth and low-delay switching strategy. The telephone network employs the strategy of circuit switching: each call is routed through the network along a dedicated path, or connection, that is reserved for the duration of the conversation, just as if the two communicating parties were directly connected by a single circuit. A current problem with this approach, as anyone who makes long-distance telephone calls can attest, is that it takes several seconds to set up a connection. Such a delay may be insignificant in the case of normal voice communications, but it would be intolerable when interposed between the rapid, staccato exchanges of data that often take place between computers. A fast circuit-switching network could reduce this delay to a fraction of a second, a tolerable level for many uses.

A dedicated circuit, however, can be wasteful. For example, if circuit switching were applied to roads, and you wanted to drive, say, from New York to Washington, you would call the highway system and ask them to close one lane on the entire stretch of Interstate 95 to all other traffic. The result would be a lot of reserved but empty highway.

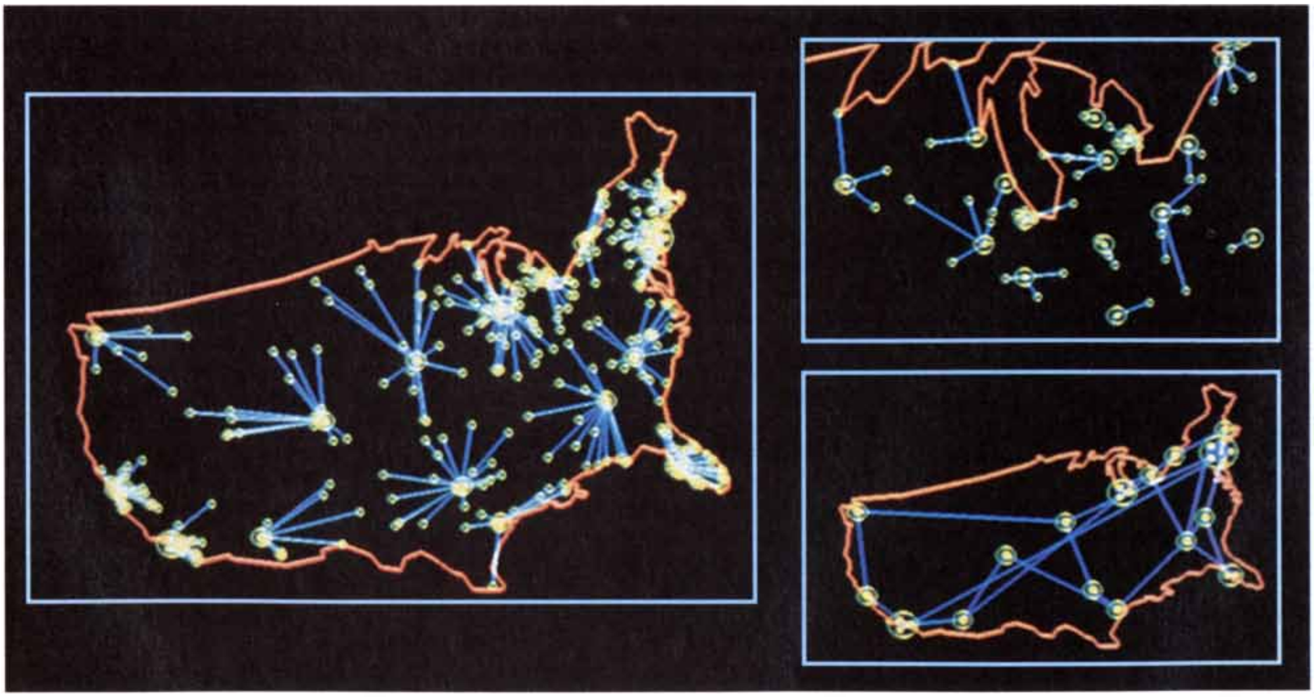
The solution pioneered in the ARPANET was a technique called packet switching. In packet switching the information being sent from one computer to another does not travel in a continuous stream through a dedicated circuit. Rather, each user computer is connected to a node (which in the original ARPANET was simply a minicomputer) that subdivides messages from the computer into a series of data packets.

A message contains an arbitrary sequence of binary digits preceded by some addressing information whose total length is no longer than the maximum size allowed by the network. A "header" containing at least the destination address and a sequence number is attached to each packet, and the packets are then sent across the network, which acts like a rapid version of the postal service, reading the addresses and delivering the packets within a fraction of a second. There is no circuit setup delay because no circuit is set up. Consequently there is also no wasted capacity since there is no individual connection between two machines. A small percentage of the communication capacity, however, is used for routing information, headers and other control information.

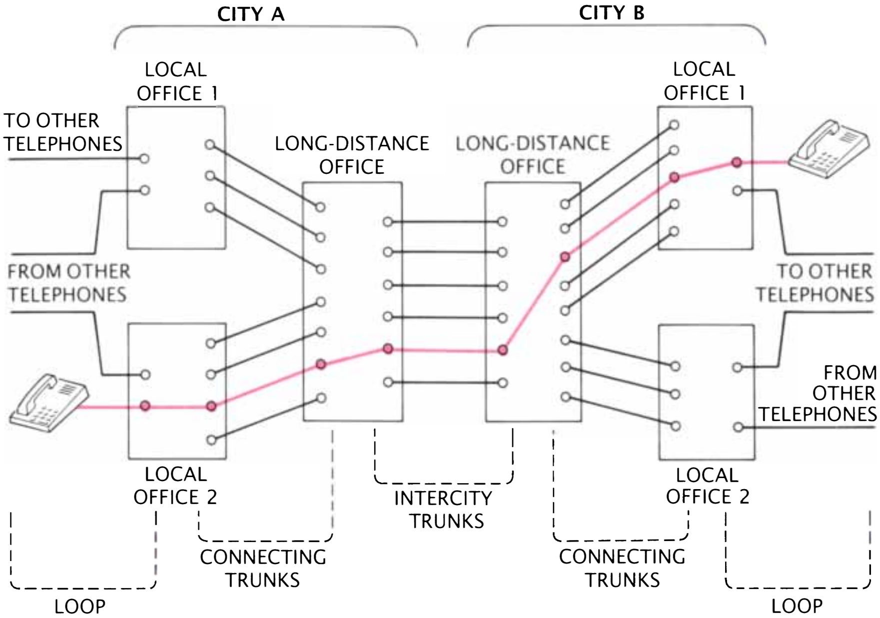
Indeed, successive packets from the same message may well take different paths to reach the same destination. The quickest path through the net at a given time is determined jointly by the nodes according to a distributed-routing algorithm. Every node estimates a "distance metric" from its location to every destination on the network, taking into account which lines are down, which nodes are busy and so on. The estimates are updated frequently (for example, every half second or so) and transmitted to neighboring nodes. On the basis of this information a node can route an arriving packet to the next leg of its journey. The original ARPANET routing algorithm accomplished this without knowing the overall topology or condition of the network. The current algorithm distributes full topological information to all nodes.

When communication proceeds in isolated short bursts (for instance, when a terminal operator types commands to a remote mainframe), packet switching can lead to more efficient utilization of the network's capacity. Since no transmission capacity is specifically reserved for an individual, time gaps between packets can be filled by packets from other users going to the same or other destinations. In effect, packet switching implements a kind of distributed multiplexing system by enabling all the users to share all the lines on the network all the time.

To link a number of nodes (n) in a packet-switched network, one needs a minimum of $n-1$ lines; if one additional line linked the two



WIDE-AREA NETWORK is shown in the design stage, with several layers of the system seen in different windows at the same time. The largest map shows host computers and PADS, or packet assemblers/disassemblers, linked to nodes called packet switches. The map of the Great Lakes area shows individual terminals linked to PADS. On the small U.S. map the packet switches are connected by trunk lines. The display was generated by the BBN Communication Corporation's DESIGNet expert system.



CIRCUIT SWITCHING is employed in the long-distance telephone network. For each call the switching offices set up a circuit (*color*) between two telephones. Part of the circuit's transmission capacity must be reserved for the user for the duration of the call.

ends, the topology of such a net would be a simple ring. It is desirable to add cross-connecting lines to the ring. That way if a single node or circuit fails, the impact on the network is reduced. A highly reliable network can be built with a number of lines that is only slightly greater than the number of nodes. In fact, the failure of a particular line in the network need not interrupt a communication between two sites, as it inevitably would in a circuit-switched network; the packets can simply be routed along an alternative path. Of course, the network must be designed to avoid or rapidly dispense with any traffic congestion.

Since the packets may follow different paths, they may arrive out of sequence at the destination. That is unacceptable for many applications, and so the destination node is generally programmed to reassemble the packets in proper order before delivering the original message. In that case the receiving computer will be unaware that the message had been split into packets. The sending computer too is unaware of the underlying switching technique: it merely supplies the address of the destination at the beginning of the message, and the first packet-switching node attaches the address to each packet of the message. To a user it seems to be an ordinary dedicated circuit: the network supplies a "virtual circuit."

The idea of a virtual circuit has its roots in the telephone culture. And yet, for some applications, network subscribers might be just as happy or happier with the equivalent of postal rather than telephone service. For example, if an image is to be transmitted from one computer to another, and the data are divided into packets, it really does not matter in what order the packets arrive as long as they all do arrive with enough information to assist in assembling all the packets to reconstruct the desired image. In that case single-packet "datagrams" could be transmitted through the network like letters through the mail—except that the datagrams would arrive in a few milliseconds rather than a few days.

Since the feasibility of packet switching for wide-area computer networks was demonstrated by the ARPANET, a number of commercial packet network services have been offered, including U.S. Sprint's Telenet, the McDonnell Douglas Corporation's TYMNET and a variety of networks in other countries such as P.S.S. in the U.K. All are common carriers, operating over commercial dedicated lines (the network companies install their own packet switches), and are available to the public. For the most part these wide-area networks offer virtual-circuit service.

The design of the communication network will depend on the distances involved, the type of devices to which it will connect and the applications to be supported. In the case of two or more nodes in relatively close physical proximity, the transmission medium will typically be a high-speed bus or a local-area network. For wide-area networks the medium will typically consist of multiple transmission links along with the associated switching technology.

Local-area networks have become extremely popular of late. They connect two or more computers that typically are in a single building or are grouped in a campuslike setting; usually they serve a single organization. Not only do they enable members of the organization to send messages to one another and to access the same data bases and programs (as is also the case for wide-area networks) but also they allow different workstations to share such expensive equipment as storage facilities and printers, whose cost has not fallen as rapidly as the cost of computers themselves. It is estimated that several hundred thousand local-area networks have been installed worldwide. These networks, based on the use of shared coaxial cable or twisted pairs of wires (rather than elaborate switches), generally provide a datagram service. Reordering of datagrams into sequence is done by the destination computer, if needed.

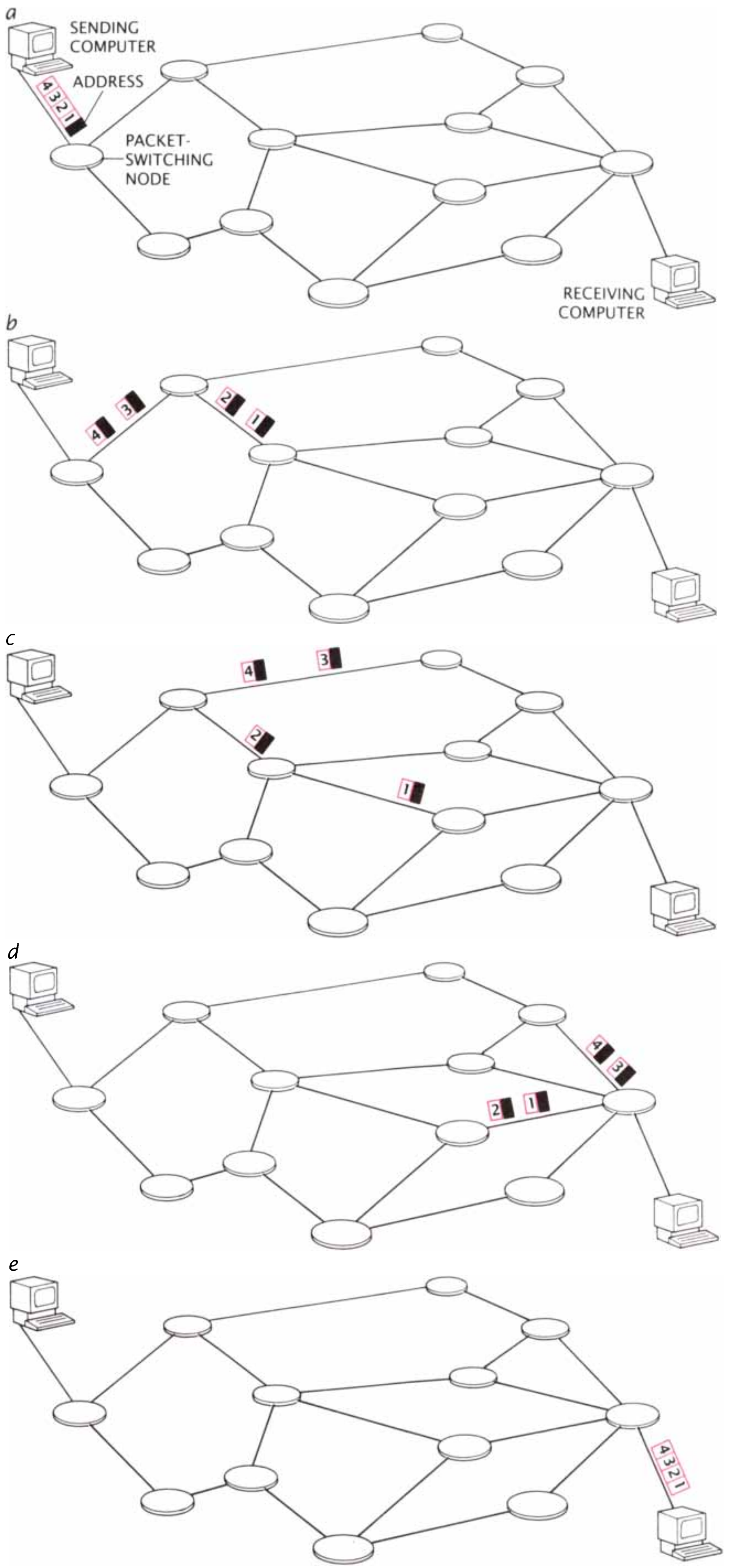
The two commonest local-area network topologies are the tree, of which the Xerox Corporation's Ethernet is the most familiar example, and the ring, which has been commercialized as the token ring by the IBM Corporation. Both are packet networks. The Ethernet consists of a single channel, usually made of coaxial cable, to which computers, printers and other machines are attached. Each machine is connected to the cable by an interface that is both transmitter and receiver: it splits messages from the machine into packets and sends them out over the cable, and it scans the cable for packets addressed to its machine.

The Ethernet is a broadcast network: messages from a particular machine propagate throughout the net in both directions along the cable. Hence a conflict can arise when two machines on the network transmit at the same time. In the Ethernet each interface has a mechanism that detects packet collisions, aborts the transmission and reschedules it for a short time later; under these conditions there can be variable delays in communication. The system works well, but at high rates of data transmission it may become inefficient if the number of collisions grows too large. Today's Ethernets carry about 10 megabits of data per second.

The token-ring network is also a broadcast network. Unlike the Ethernet, however, the token ring organizes access to the net by means of a token (a short sequence of bits) that circulates continuously around the ring. In order to transmit data, a machine interface must first wait for the token to pass. It then removes the token temporarily, inserts an addressed packet (of a specified maximum length) onto the ring and reinserts the token behind the packet. The packet is removed by the intended recipient before the token makes a full circle. Thus collisions do not occur. Depending on the traffic loading, however, it may take longer than usual for the token to pass and variable delays may occur.

Today's token rings also handle a transmission rate of 10 megabits per second of data. Experimental rings built of optical fiber, through which information is carried by pulses of light rather than by electronic signals, have achieved rates of 100 megabits per second or more. Optical technology could well expand the scope of local networks even further.

Communication between identical types of computers has certain natural advantages since the system software running on the two machines can be identical. One important challenge in getting different kinds of machines to work together is to find a way to mask differences in the underlying hardware so that a common software environment can be provided for the machines. Much progress has already been made in developing techniques to exchange data between machines, although many barriers still remain. More diffi-



PACKET SWITCHING was developed for computer networks. A transmitting computer sends out an addressed message (*a*). At the first packet-switching node (or at an interface between the computer and the node) the message is divided into individually addressed packets (*b*). The packets travel through the network independently. Each node chooses the next node on a packet's path, taking into account information received from neighboring nodes on traffic, line failures and so on. As a result the packets may follow different paths (*c*), and they may arrive out of sequence at the destination node (*d*). Here the destination node restores them to their original order before delivering them to the receiving computer (*e*).

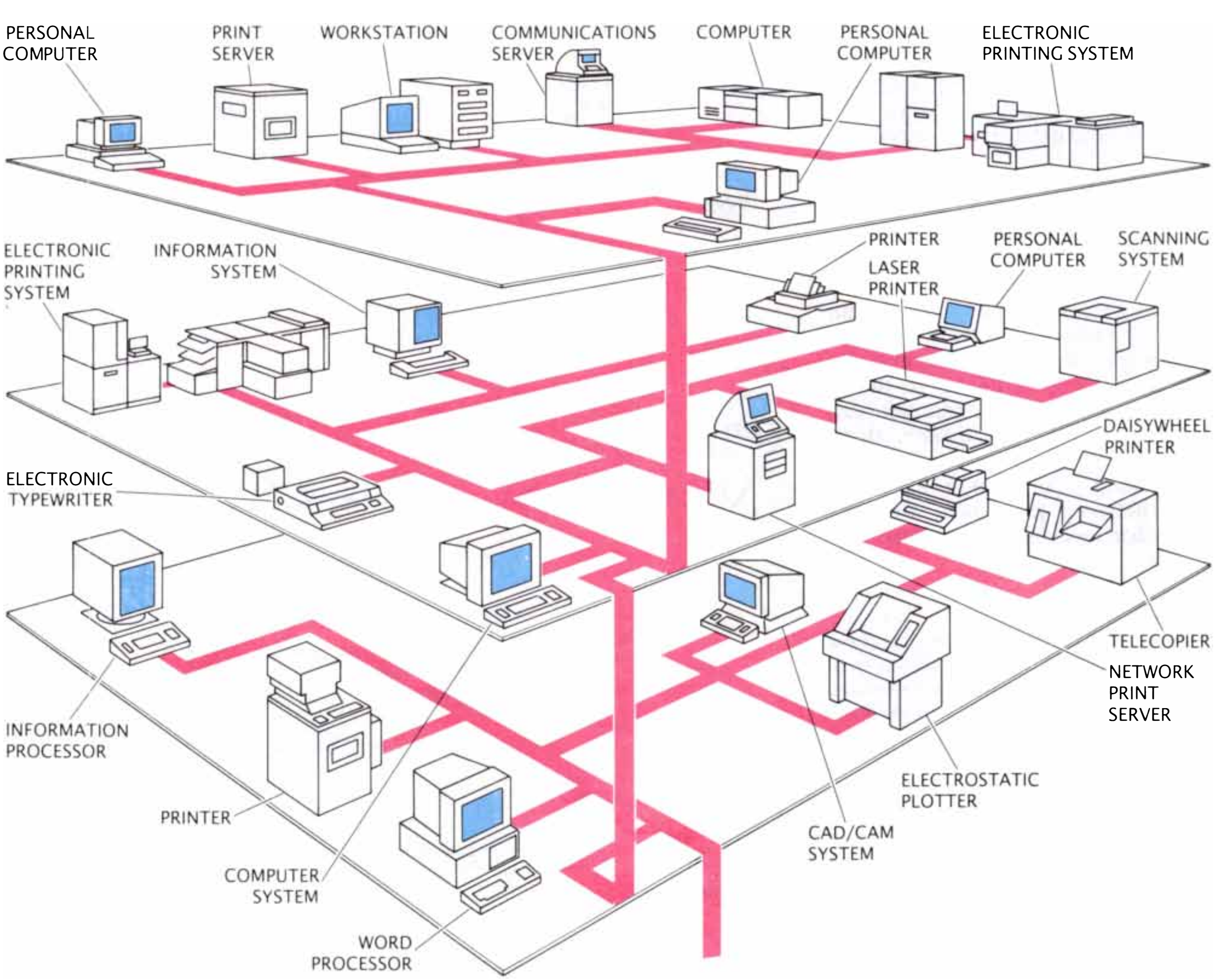
cult still is the ability to exchange software between heterogeneous systems.

Programs that are exchanged between identical types of machines start out with a much better chance of running since the underlying instruction sets are the same and the same operating system software environments can be provided. Yet even if a program could run without compilation or modification of any kind at a remote site, there may not be sufficient memory at the destination to support it. The compilation requirements for a high-level language program may be adequately met on one machine and not on another. Only when the machines have identical software environments is the likelihood of success in transferring and running a program reasonably high.

The most powerful network functions have generally been limited to commonly administered systems of homogeneous machines. Examples of such functions are distributed file systems (which can give a workstation ready access to files independent of their location in the network) or migrating programs (which travel over the network to the location where they are to be executed). In contrast, most interactions achieved between heterogeneous machines are still fairly simple ones: sending messages or files from one machine to another, running a program from a remote terminal and so on. Yet heterogeneity is the rule among computers; even an individual organization often buys equipment from several manufacturers. Hence there is a strong incentive to overcome the barriers that prevent communication between heterogeneous machines.

Consider a very simple example of such a barrier. One computer stores data in 32-bit words; in another the word length is 16 bits. If the two machines are to communicate, the receiving computer must know how to arrange the data in its memory. Since the word length in both cases is an integer multiple of eight-bit characters (bytes), the task is comparatively easy. It gets more complicated when a 32-bit computer is to communicate with a 36-bit computer: in that case certain bytes must be split up or empty spaces must be left in memory. Furthermore, data cannot simply be transmitted in isolation, without the equivalent of a "cover letter" to describe or otherwise interpret them.

When a program rather than nonexecutable data is to be transmitted, the need to supply context information that enables the program to run at the remote site may be even greater. "Context" may mean the entire software environment in which the

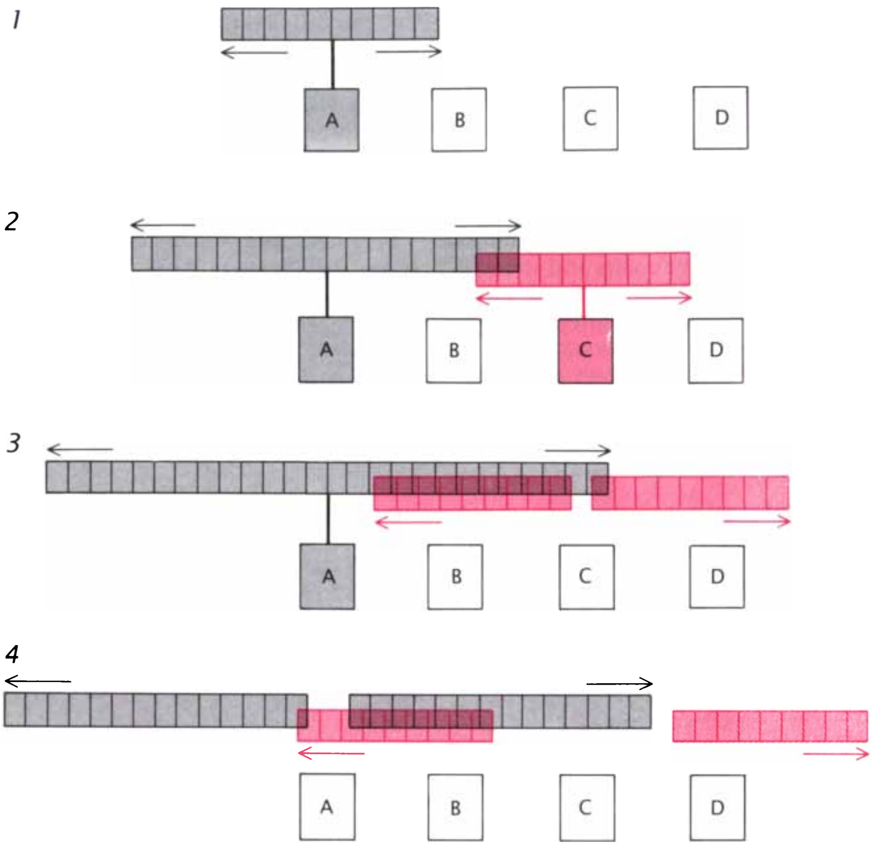


LOCAL-AREA NETWORK enables workstations in a building or a small area to communicate and to share resources. This is the Xerox Corporation's Ethernet, in which machines are connected by a coaxial cable or by several cables linked by repeaters.

program is designed to function, including most notably the operating system. In principle one could transmit the entire software environment along with the program, but in practice that is difficult to manage: many operating systems are themselves designed for particular hardware, and in general they cannot be installed on the machines of another manufacturer. An operating system such as UNIX, written in a higher-level language, is portable between some machines since the appropriate compilers are widely available. Still, some tailoring or installation is needed to adapt UNIX to specific machine characteristics. One solution might be to create a "virtual machine architecture" that enables heterogeneous machines to mask the differences in their hardware. A standard microcode architecture would make it possible to explore various virtual machine architectures.

A number of computers working together form a kind of distributed system. The exchange of context information between these computers implies a series of protocols that must be strictly adhered to by the various participating machines to indicate the kind of context information and how to interpret it. These protocols are to distributed systems what programming languages are to software. They enable one to specify the overall architecture and operation of the system's critical elements.

A desirable approach being explored internationally is to try to create a common communication environment for distributed systems by getting hardware and software vendors and network designers to adhere to industry-wide standard protocols. The protocols establish one form of "open" communication architecture for the network.



COLLISION DETECTION is an essential feature of the Ethernet. Since packets travel in both directions along a coaxial cable, two stations may start to transmit simultaneously. To prevent confusion each station on the Ethernet has a collision-detection mechanism. In the sequence shown here, station *A* begins transmitting a packet to station *D* at a time when the network is quiet (*1*). Before the packet can reach *C*, however, *C* starts to send its own packet (*2*). The two packets collide, and when *C* detects the collision, it stops transmitting (*3*). Soon after that *C*'s packet reaches *A*, which thereupon also stops transmitting (*4*). Both stations wait for a random short interval and then try again.

For some time now the International Standards Organization has been engaged in drafting network protocols to support a concept called Open Systems Interconnection (OSI). As a way of organizing the problem, it has come up with a conceptual model that divides the workings of a network into seven layers, each of which is to be associated with a specific protocol or protocols. For example, the bottom layer is the physical one that contains specifications for signaling conventions, transmission rates and so on. Near the top of the model a "presentation" layer stipulates how data are to be interpreted for presentation; a computer with the right translation software will be able to converse with computers that format their data differently. Finally, the protocols in the top layer basically specify the types of messages that can be sent over the network to achieve various applications.

The interactions made possible by the present OSI architecture are limited primarily to the transfer of data. It does not prescribe how computers on the network should collaborate on high-level applications. Yet in spite of its ambitious albeit limited goals, the open-architecture strategy has been difficult to implement. Essentially the problem is that protocol specifications are inevitably ambiguous and are interpreted in different ways by different vendors. A single difference in interpretation can be (and has been) enough to make the resulting products incompatible.

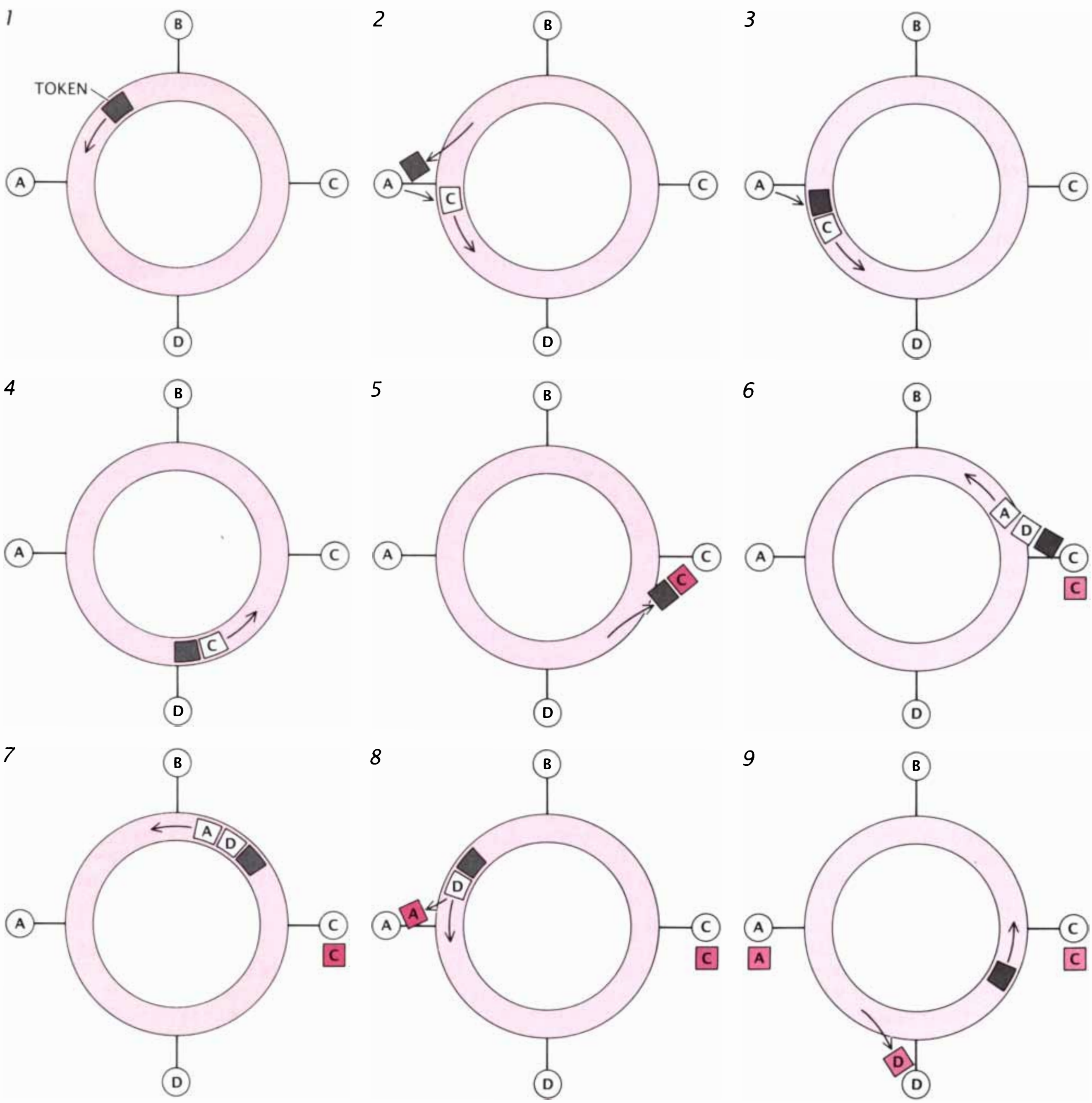
In 1985 a group of vendors established the Corporation for Open Systems (COS) to help resolve such incompatibilities in the OSI protocol implementations. The corporation is the equivalent of an underwriter's laboratory; its purpose is to test various products and give them a seal of approval if they are compatible with other certified products. The tests are not foolproof, of course. Two products that are both compatible with the COS benchmarks may still not work together properly. Hence before a product can be certified with reasonable confidence it has to be tested with at least several other certified systems. In contrast, the widespread popularity and usage of the TCP/IP transport protocol in the U.S. (which preceded the OSI transport protocol) is due to its widespread implementation in tested installations.

Meanwhile networks proliferate, and the cacophony grows. The compatibility problem recurs at another level: not only must the machines on an individual network be able to understand one another but also the networks themselves must be able to communicate. Organizations often want to connect their local networks to wide-area networks or to link several local networks together. But the various nets may have different data formats, packet lengths, transmission rates (local networks are typically much faster than wide-area ones), addressing mechanisms and so forth.

The original ARPANET host transport protocol was designed with the idea that the ARPANET would function somewhat like a perfect computer peripheral. Clearly this assumption is no longer tenable, since many new kinds of networks (such as radio networks, which may not always be able to deliver all packets) have become available.

The solution the DARPA researchers adopted was to develop an internet architecture. This architecture has come to be heavily used for local-area networks and has been widely emulated. The CCITT X.75 network-interface protocol specifies an alternative for network interconnection that presumes a virtual-circuit setup between networks. It is normally used to connect networks (virtual circuits) with similar data rates and so does not require a separate gateway.

The internet architecture includes the transport protocol TCP mentioned above and also gateways to link networks. The gateway is a special node designed to encapsulate the text of messages from one network into the form required for another, and vice versa. It uses an internet protocol (IP)



TOKEN RING is another common design for local networks. Access to the network is controlled by a circulating sequence of bits called a token. To send a message a station must wait for the token to pass, remove the token from the ring, put an addressed packet (or sequence of packets) on the ring and place the token behind the packet. A station can remove a message addressed to it without removing the token. Here *A* sends a message to *C*, which receives it and then sends its own messages to *A* and to *D*.

to interpret internet-wide addresses, to fragment packets if necessary and to otherwise facilitate the transport of packets between possibly dissimilar networks. For example, if a 500-byte-long packet arrives at the gateway to a network whose maximum packet length is 250 bytes, the gateway cuts the packet in two, attaches addresses to both halves and routes them (possibly independently) toward their destination. Gateways can also communicate with other gateways, forming an internet system that routes messages among multiple networks. OSI now also incorporates IP as an internet mechanism.

A fundamental assumption in developing the internet architecture was that no internal changes would be required to enable any network to participate. This implied that the gateways had to have the following key properties: the ability to pass packets between networks with different internal and external properties such as data rates, maximum packet sizes, formats, timing and error messages, and the ability to communicate with other gateways for the purpose of internetwork routing.

A major potential networking area for the future is portable communications and computing. Although most requirements today are mandated by the office or home environment, a significant potential exists for transforming society if equivalent communications can be extended to the mobile environment. There is already a basis for the ability to start a computation in one location, perhaps even continue it with a laptop computer on an airplane and complete it at a different location. Mobile communication by means of cellular telephone, packet radio networks, civilian-band and amateur radio and paging devices offers a start. Because of technical limitations such as multipath propagation, maximum data rates in the mobile environment are unlikely to exceed a few megabits per second. This is acceptable because many important applications of computers and communications require only moderate computation and limited communication. On the other hand, many new applications for higher-speed services are also being explored.

Increases in network capacity have often been motivated by the need to support more users. Although each user's demand for communications capacity may remain relatively constant, the aggregate of all user demands increases. This has been the basis for the evolution of the analog telephone system. Individual voice channels were aggregated into groups of voice channels and these groups in turn were aggregated into still higher-level groupings. Likewise, individual user demands for digital transmission can be aggregated by the carriers or by the user. In fact, this is the basis for the nascent Integrated Services Digital Network (ISDN) in which a user will be able to decide how to multiplex digital data onto one or more circuits. Multiplexing, however, is only one dimension of the capacity issue.

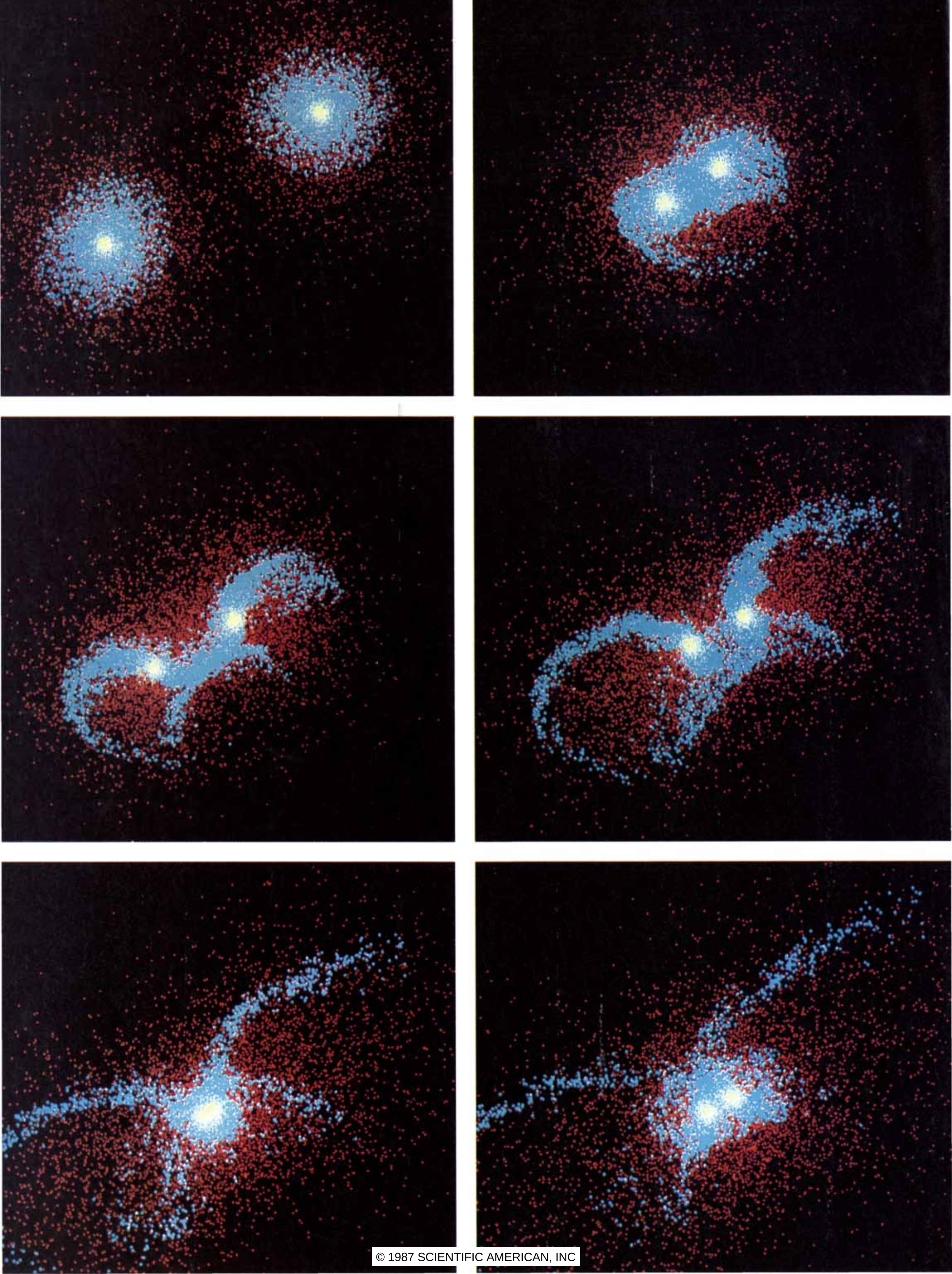
In the future there will be fundamental changes in the types of services networks offer and in their cost. The changes will be made possible in part by a vast expansion in the transmission capacity of communication lines. Today's tariffs are still largely based on the cost of individual voice channels. If a line can handle 24 voice channels and the bulk of the carrier's revenue is determined by sales of voice channels, it will be natural to price the line somewhere between the cost of one voice channel and 24 channels. At these multiples, optical circuits will be all but unaffordable to most users unless there is a significant tariff restructuring. Restructuring is possible only if the communication system evolves so that the bulk of its revenue is based on a higher-speed requirement such as video. If the carriers can sell about as many video circuits as voice channels at about the same price, the carriers will lose no revenue and the basis for very-high-speed circuits will be established. This is no small undertaking, however, and success presupposes that society generates significant demand.

When the original ARPANET was built in 1969, it relied on special high-speed dedicated lines that carried 50 kilobits per second. The National Science Foundation has recently established a supercomputer network, NSFNET, that has similar data rates. NSF plans call for leasing higher-speed lines in the near future as it seeks to meet the scientific community's long-term needs for network support. Today digital lines operating at about 1.5 megabits per second are available, and 45-megabit-per-second service has recently been introduced. Once the current rewiring of the U.S. optical fibers has been completed, those numbers will seem small. Transmission rates in excess of one gigabit per second have been achieved in field trials, and speeds of 10 gigabits per second might be attained at the century's end. "Multiplexing" signals of different wavelengths may make transmission rates in excess of 100 gigabits on a single fiber both feasible and economic.

It is not hard to think of things to do with such capacity. A large data base could be rapidly communicated to other computers. A dynamic video simulation of the type that is becoming common in scientific research could be transmitted from the supercomputer to remote workstations. Images, voice and data could be sent simultaneously between collaborating offices.

How much transmission capacity would such activities require? In a computer terminal with a square display that is 1,000 pixels (picture elements) on a side, each pixel in a black-and-white picture is encoded by one bit. Transmitting, say, 30 frames per second (for flicker-free use) would entail an information flow of 30 megabits per second. Color images might require 16 bits of information per pixel, and so 480 megabits per second would be needed to transmit a color motion picture in real time. When larger screen displays (say four feet by six feet) become available, the transmission requirements may increase by another factor of 10 or even 100 if real-time requirements exist. To be sure, data-compression techniques can be used to reduce the requirements somewhat. Nevertheless, it should be clear that computer networks have many potential applications for the information-carrying capacity offered by optical fibers.

The point becomes even clearer when one considers the function of computers. They are not mere displayers of information: they can process or even "reason" about the information as well. To reason about a picture may require considerably more information than is contained in the picture itself, and this contextual information must be transmitted with the picture. The same thing is true of information that is not in the form of a picture, such as a document or a VLSI design: to reason about it, a computer must have considerable contextual knowledge. The promise optical fibers hold for computer networks is that they will make possible the levels of information flow needed for the networks to fulfill their ultimate purpose, which is to enable computers to collaborate intelligently on solutions to human problems.



Advanced Computing for Science

Computational experiments are enriching scientific investigation. They are now becoming as important as theory, observation and laboratory experiments

by Piet Hut and Gerald Jay Sussman

High-speed computation is dramatically changing the way science is done. Traditionally investigators have developed idealized models from which predictions can be made that are tested by observation or experiment. In complicated systems it is hard to find simplifications without taking the risk of throwing out the baby with the bathwater, because some of the most important phenomena can emerge from the unforeseen combination or amplification of small effects.

Computers are enabling investigators to cope successfully with such phenomena, which range from the diffusion of charge carriers in a semiconductor to collisions between galaxies containing millions of stars. By using models of such systems it is increasingly possible to work out the consequences of simple theories in complex situations. In addition, with sufficient computation it is possible to determine the consequences of a theory without invoking questionable approximations or simplifications. An accurate computational model allows for measurements to be made in much the same way as they are in laboratories. But a computational model is better in that infeasible experiments can be done and parameters inaccessible to experiment or observation can be measured. For example, because Newton's law of gravity is well understood and accurate computational models of the solar system exist, it can be determined how the orbit of the earth would differ if Mars did not exist.

Theorists have often been forced to oversimplify so that approximate consequences can be computed and compared with the real world. One result of the increase in computational power available to scientists is a shift in the balance away from such reductionist methods to "analysis by synthesis." The synthetic approach is called for when the fundamental processes of the interactions among the parts of a system are known, but the detailed configuration of the system is not. One can attempt to determine the unknown configuration by synthesis: one can survey the possible configurations and work out the consequences of each. By carefully matching the observable details of the experimental situation with these consequences, one can choose the configuration that best accounts for the observations.

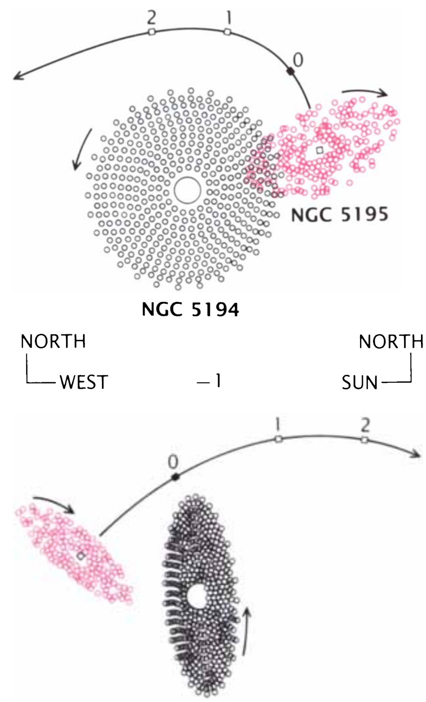
A famous example of the synthetic approach from the 19th century is the attempt that was made to understand the observed but unexplained perturbations in the orbit of Uranus. Investigators added a hypothetical planet to the solar system and varied the parameters of its orbit until a satisfactory reconstruction of the perturbation was found. The work led directly to the discovery of Neptune, found near the predicted position. In the past the synthetic approach was limited to comparatively simple situations. The availability of high-speed computation has allowed synthetic methods to take their place firmly next to the traditional methods of reductionist analysis.

Analysis by synthesis has recently shed new light in an old field of astrophysics, gravitational dynamics, for here the increased speed of computers has enabled investigators to tackle problems that have long eluded other approaches.

On a grand scale the universe is populated by swarms of galaxies, some floating in relative isolation, others clumped in clusters and in clusters of clusters. Each galaxy consists of huge numbers of stars and vast amounts of gas and dust. The distances between the stars in a galaxy are sufficiently large compared with the diameters of the stars for collisions between stars to be extremely rare, even though the relative velocities of stars are typically tens or hundreds of kilometers per second. On a larger scale, collisions between galaxies are more frequent because the distances between galaxies are only 10 or 100 times larger than their diameters. On even larger scales, clusters of galaxies are separated from one another by distances comparable to their own diameters; interactions between clusters of galaxies take place over periods comparable to the present age of the universe and therefore cannot be studied as isolated events.

Collisions between galaxies make for some of the most spectacular traffic accidents in the universe. The view of the sky revealed by telescopes provides ample evidence for such violent encounters. There are paired galaxies that are close to each other in the sky and appear to be con-

SEQUENCE OF COMPUTER IMAGES simulates a collision between two spiral galaxies; the image at the bottom right resembles a pair of galaxies known as the antenna galaxies (also called galaxies NGC 4038 and NGC 4039) in the constellation Corvus. The yellow dots represent the central bulges of the galaxies and the blue dots the surrounding disks of stars. The red dots represent the dark halo of matter in the galaxies; the halo cannot be seen but is manifested by its gravitational effects. The simulation, covering about half a billion years, is by Joshua Barnes at the Institute for Advanced Study.



COLLISION BETWEEN GALAXIES was simulated in a classic experiment carried out on a computer in the early 1970's by Alar Toomre of the Massachusetts Institute of Technology and Juri Toomre, then at New York University. The simulation stages the encounter of the "whirlpool nebula" (also known as galaxy NGC 5194, or M51) and galaxy NGC 5195 (*left*). The Toomres chose a number of initial conditions after a careful analysis of the available observational data. For each set of initial conditions they

nected by bridges. Whereas most galaxies have either an approximately symmetrical spiral design or a simple spherical or elliptical shape, the galaxies that appear to be involved in collisions are often deformed or irregular. In the 1950's it was suggested that the deformations in these interacting galaxies are caused by the gravitational forces galaxies exert on each other. In the 1960's more exotic suggestions, involving magnetic fields and explosions, became popular. The motivation for rejecting the simpler gravitational explanation was that it was hard to see how purely gravitational interactions could result in sharp features such as the complex and detailed shapes of the tails and bridges emanating from, and sometimes connecting, pairs of interacting galaxies.

In the early 1970's Alar Toomre of the Massachusetts Institute of Technology and Juri Toomre, then at New York University, carried out the first systematic computer modeling of galaxy encounters. Their results were surprising: using only gravitational forces, they managed to reproduce the shapes observed in a number of systems of interacting galaxies. The Toomres approximated the gravitational field of a galaxy by the field of a point mass in the galactic center. They modeled the disks of the galaxies, which include many tens of billions of stars, with a mere few hundred particles that moved in the field of the model galaxies without contributing to that field. Even with these crude approximations, the investigators found that encounters of two model galaxies can produce a dazzling variety of splatter patterns that often include sheets and ribbons. When viewed edge on, the sheets and ribbons provide an explanation of some of the thin and sharp features that had driven other astronomers to invoke complicated magnetic-confinement processes.

The work of the Toomre brothers provides a clear example of how new information is obtained from synthesis. The sequence of computer "snapshots" on these two pages, for instance, shows that the streamer apparently connecting the "whirlpool nebula" (also known as galaxy NGC 5194, or M51) with galaxy NGC 5195 is not really a bridge at all. The connection is simply an illusion, a consequence of the positions of the galaxies as they are seen in the sky.

Many problems in physics are still too complex to be tackled by direct synthesis of an all-encompassing model. Instead a two-pronged approach that starts with simulations of microscopic parts of a large system may be possible. Synthetic methods can often provide quantitative insights into the statistical behavior of very small subcomponents of a macroscopic system. A higher-level model that combines these statistical rules as building blocks can then be used to predict the behavior of the system as a whole.

An example of a problem that currently exceeds today's computational limits is the detailed modeling of the evolution of a star cluster consisting of a million stars—the size of some of the globular clusters orbiting our galaxy. It has proved possible to gain understanding of the clusters by exploiting a microscopic approach in which interactions between a single star and a double star (a pair of stars bound together tightly) are studied.

The gravitational interaction between two stars or planets is easy to calculate using pencil and paper. The

developed a sequence of "snapshots." One of the sequences yielded a snapshot (*right*) that closely resembles the observed configuration of the galaxies. The two rows present two different viewing angles of the same event; 100 million years have elapsed between successive frames. The Toomre experiment shows that the streamer apparently connecting M51 and NGC 5195 is actually an illusion—a projection effect. At this time NGC 5195 is well behind M51 and the collision is a matter of history.

equations of motion become intractable, however, for interactions among three bodies. What happens, for instance, when a single star collides with a double star? As long as the single star is far from the double one, the two stars making up the double star revolve around each other in an unperturbed elliptical orbit. If the third star passes by at a modest distance, the internal orbit of the double star will be perturbed. If the single star penetrates the double-star system, the encounter may be much more complicated.

The ultimate outcome of such an encounter can be classified into three qualitatively distinct types. The encounter may yield three isolated, unbound stars. In that case, in analogy to atomic physics, we say the double star is ionized. Alternatively, the encounter may yield a double star and a single star. If the resulting double star is made up of the intruder and one of the members of the original double star, we say we have an exchange reaction. If the members of the resulting double star are the members of the original double star, we classify the event as a flyby. Sometimes, during a near head-on encounter, the double star will temporarily capture the single star. The resulting three-body system can remain bound for a long time—unstable relationships lasting for hundreds or even thousands of orbital periods are not rare—but ultimately the system will decay again into a single star and a double star.

The type of outcome is quite sensitive to the initial conditions. A slight change in encounter velocity or orientation angles is enough to alter the complicated three-body dance drastically. An analytic description of this behavior is out of the question.

The number of parameters needed to describe the initial conditions for such scattering experiments is nine, a number that is far too large to allow a systematic survey of all possible combinations. The way the outcome of the experiments depends on the precise values of the parameters is extremely complex. Even if seven of the parameters are held fixed and only two are varied, a complicated picture emerges [*see bottom of illustration on next page*].

One way to gain quantitative insight into three-body scattering processes is to simulate a large number of individual scattering experiments and analyze the results statistically. Values for most parameters are chosen at random, subject to constraints that make sure the experiments lie in domains of interest to the astrophysical application at hand. When enough experiments are done, the uncertainties that are introduced by this "Monte Carlo" method of sampling the initial conditions become small enough so that interesting conclusions about the average outcome of three-body scattering experiments can be drawn.

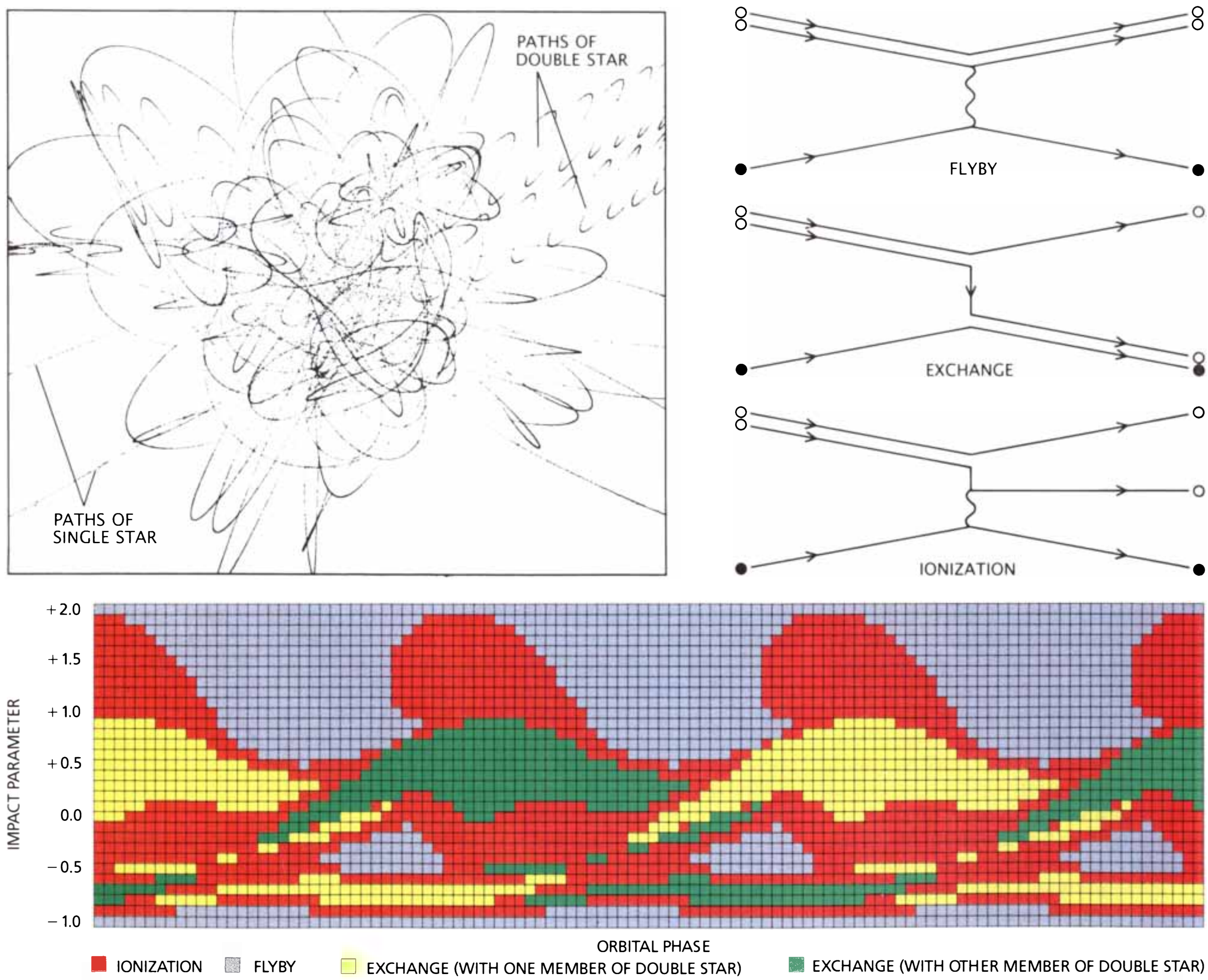
Why are astronomers interested in this kind of collision? The answer lies in the role of double stars in generating "heat." In a collision between a double star and a single star, the double star can shrink, transferring energy to the single star and thereby heating the pool of stars around them. This process is analogous to nuclear fusion, wherein atomic nuclei collide and fuse into heavier nuclei, releasing energy. Nuclear fusion is the same phenomenon that makes the stars, including the sun, shine.

Similarly, orbital shrinkage of double stars induced by encounters can heat the core of dense star clusters. This heat can balance the losses at the surface of star clusters, where stars boil off continuously. In constructing detailed models of globular clusters the statistical behavior of encounters between single stars and double stars must be known. Just as nuclear reaction rates are an essential ingredient in calculations of the structure and evolution of individual stars, so gravitational stellar reaction rates are essential in calculations of the structure and evolution of star clusters.

The problem of determining the gravitational reaction rates is similar to the one a nuclear physicist faces when asked to determine a nuclear reaction rate. A beam of high-speed nuclei of one type can be aimed at a target plate containing nuclei of another type. By counting the number of interactions, measuring their properties and analyzing the data statistically, the physicist can describe the average behavior during the encounter of two nuclei.

Gravitational reaction rates can be measured in a similar way, using a computational laboratory. One of us (Hut) has participated in a project to investigate the dynamics of star clusters by employing a statistical characterization of three-body scattering. Part of the project required doing millions of Monte Carlo three-body scattering experiments in which two of the parameters were held fixed and seven were varied in a complete statistical survey.

The three-body scattering experiment was quite feasible with available computer resources; it took less than a year of computer time on a Digital Equipment Corporation VAX 11/780. A similar scattering experiment with galaxies is much more difficult. Just as particle physicists



ENCOUNTER between a single star and a double star (a pair of tightly bound stars) is complex, as the computer-generated simulation at the top left shows. Attempts at calculating such three-body scattering events using only pen and paper would be futile. Fortunately the essence of the process can be distilled by overlooking the details and classifying only the eventual outcome, just as the scattering of particles in subatomic physics is studied. In this case there are three possible outcomes (*top right*). In a flyby the double star remains intact even though its orbital parameters may be affected. In an exchange the intruder replaces one of the original members of the double star. In an ionization event all stars become unbound. The number of parameters needed to describe the initial conditions for such scattering experiments is nine—far too many to allow a systematic survey of all possible combinations. In the graph at the bottom, which summarizes the outcome of more than 1,000 different scattering experiments, seven of the parameters are held constant and only two are varied: the impact parameter (the point of closest approach of the single star to the double star) and the orbital phase (the relative positions of the members of the double star).

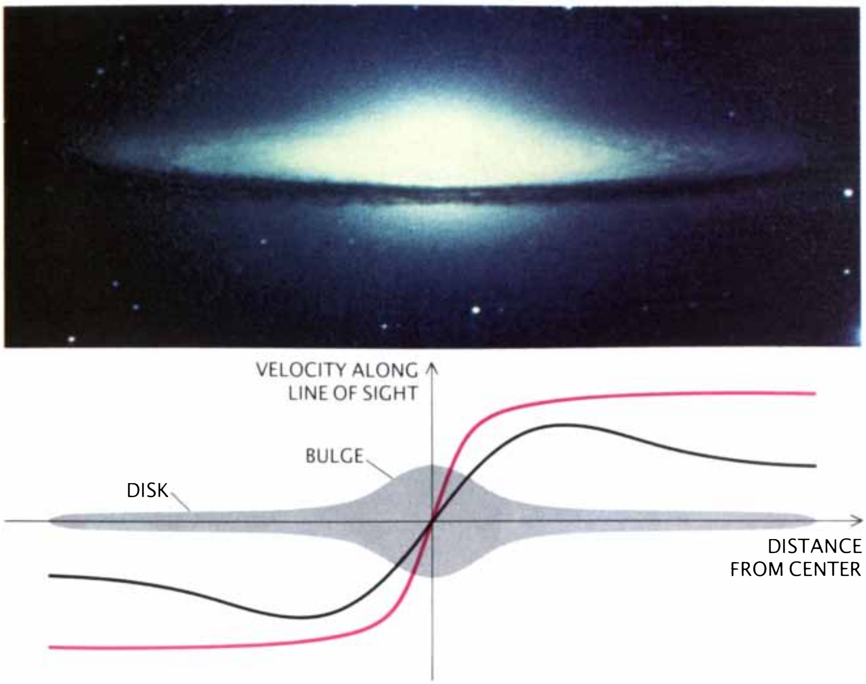
At Lufthansa, we understand the pressures involved in maintaining high standards. But we also know the rewards. For us, they include an outstanding on-time arrival record, a meticulously maintained fleet and friendly, courteous service.

Perhaps that's why in 82 countries Lufthansa is considered by many to be the very best in the world.

People expect the world of us.

Lufthansa
German Airlines

Lufthansa is a participant in United's Mileage Plus program and Delta's Frequent Flyer program. See your Travel Agent for details.



EXISTENCE OF DARK MATTER, or invisible material in the universe, is inferred from dynamical arguments. In a typical spiral galaxy, such as the "sombrero galaxy," or M104, the stars revolve around a central bulge (*top*). The bulge is surrounded by a disk (*bottom*) and an invisible halo. According to Newton's law of gravity, if the only matter between a star in the disk and the center is what is visible, the rotational velocity of the star should fall off rapidly with increasing distance from the bulge (*black curve at bottom*). Actually extensive measurements show that the rotational velocity does not fall off; rather, it remains constant for a great distance (*colored curve*). The constant velocity implies that invisible matter is found in a large volume around the visible shapes.

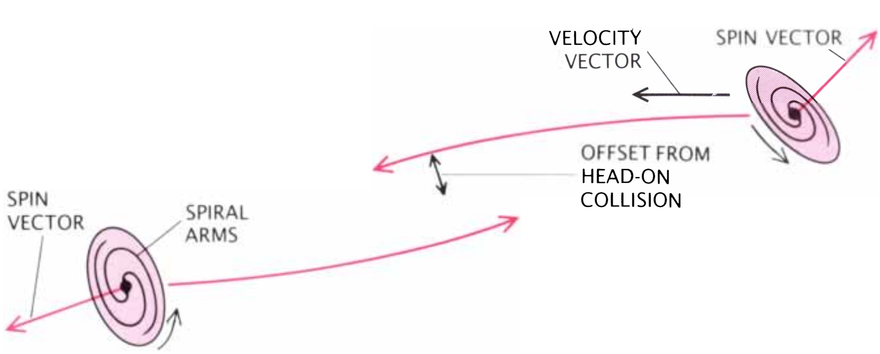
throw atoms and subatomic particles together and examine the shrapnel produced in order to learn about the particles' internal structure, so galactic traffic accidents could provide important clues about what kind of matter permeates the universe and how it is distributed. It is embarrassing that although a great deal has been learned about the universe in the past few decades, it is still not known what most of the universe is made of. Stars, gas and dust—everything visible either with the unaided eye or with radio, infrared, X-ray and gamma-ray telescopes—make up only a small fraction of the total matter in the universe. The nature of the invisible matter remains one of the central unsolved problems of astrophysics.

There is good dynamical evidence that most galaxies incorporate large amounts of unseen material. In a typical disk galaxy the stars revolve around a common center. According to Newton's law of gravity, the orbital speed of a star in the disk should fall off rapidly with increasing distance from the center—if the only matter between the star and the center is the visible matter. Actually extensive measurements show that the rotational speed does not fall off with increasing distance; rather, it remains roughly constant for a considerable distance. The constant rotational speed implies that most of the mass is not visible and that this invisible matter is distributed in a large volume around the visible shapes.

Galactic collisions afford opportunities to study the distribution of dark matter. The splatter patterns produced sample the gravitational field of the invisible material. In effect, star-based paint is used to trace out what otherwise would have remained invisible. The problem of interpreting these tracings can be attacked with massive computer simulations. What is unknown here is the distribution and amount of mass in any particular galaxy. Given sufficient computer power and software support, one could simulate many collisions having different parameters and mass distributions. One could then look for splatter patterns that most closely resemble the actual patterns observed. The unknown masses and their distributions could be deduced if the models are sufficiently sensitive.

A typical galaxy has roughly 100 billion stars, each of which interacts with all the others through the gravitational force. Computing the interactions among so many particles is outside the range of any computer currently envisioned. On the other hand, the behavior of such a system can be approximated by a system that has far fewer particles if every million or so particles in the actual galaxy is represented by a single particle in the model.

What is the minimum number of particles in a scattering experiment of two spiral galaxies necessary to make the model accurate enough to allow detailed quantitative comparison with the real world? A real galaxy can be thought of as consisting of a centrally concentrated bulge of stars, which may be flattened somewhat by rotation, a disk of stars and gas in predominantly circular orbits and a more extended, roughly spherical halo of dark matter whose existence is inferred on dynamical grounds. These components exist in stable, stationary equilibrium in their mutual gravitational field. Since spiral galaxies have a disk diameter that is nearly 100 times larger than the thickness of the disk, at least a few hundred thousand particles must be sprinkled in to represent the disk;



DISTRIBUTION OF DARK MATTER within galaxies could be deduced by extensive computer simulations of galactic collisions. Streamers of stars would sample the gravitational field of the unseen matter and trace out what would have remained invisible.

any smaller number would not allow the finite thickness of the disk to be resolved. A comparable number of particles in each of the components of both colliding galaxies requires that a realistic scattering experiment involve somewhat less than a million particles.

Integrating the equations of motion for a system consisting of between half a million and a million interacting particles presents a formidable task, requiring the best in numerical algorithms and the most powerful computers. But these are only part of the problem: there is also an enormous range of parameters that describe the possible initial conditions for a run of such a system. To set up an experiment one must choose distributions for each component and assemble the model galaxies. Values for all the initial parameters of the collision must also be chosen.

The most challenging aspect of the numerical simulation is the analysis of the results. Part of the analysis must be concurrent with the running of each galaxy encounter to determine whether it has run far enough to be terminated. Analysis is also needed to ascertain which initial conditions should be considered next to obtain the most interesting results. Finally, analysis must be done to abstract the qualitative structure and model parameters needed to characterize the results of the ensemble of experiments.

When has a galaxy scattering experiment been run long enough to be terminated? Suppose one were to look at a motion picture of the scattering process. Galaxies approach each other, crash with a great splattering of stars and either merge into one galaxy that soon settles down or separate and go their way with some damage to each of them. After some time one can determine whether the central regions of the galaxies merge or separate. At this point one just waits until measurements can be made of how much and what kind of damage was done.

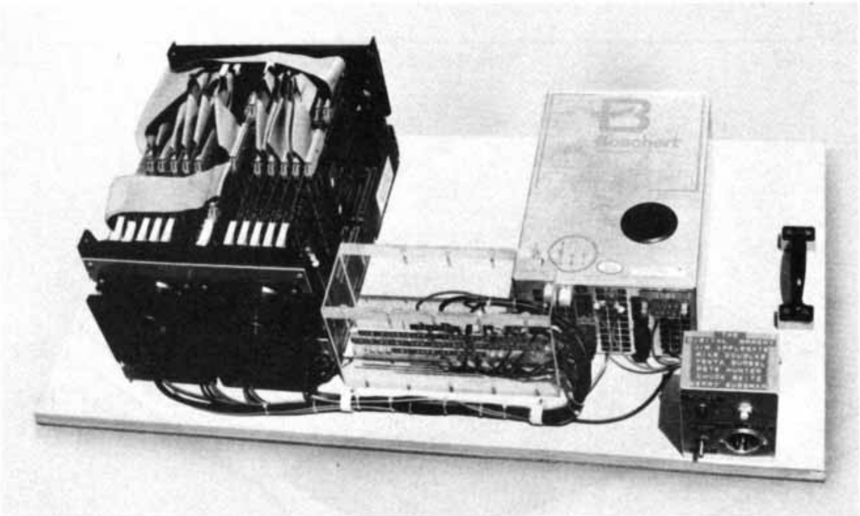
The recognition of the termination and the classification of the outcome of a galaxy-collision experiment are a qualitative pattern-recognition task that is hard to automate but not impossible. The initial system consists of two hulks, each one gravitationally bound to a small, dense core. As the collision proceeds, the cores either separate altogether on unbound orbits or eventually merge. If they merge, there may be several passes of decaying, elongated orbits before they become tightly bound. In the process some of the particles from each hulk will become unbound; some will form bound or unbound transient shells, streamers and tails.

The required recognition is quite coarse. One must be able to follow a central core that is moving in a smooth way. It is not hard, computationally, to find the position of a central core in a motion-picture frame. It is not hard to predict the path from the positions of the cores in preceding frames. By following the cores in this way one can see whether they are becoming bound or whether they will separate.

We are currently embarking on a long-term project to model galaxy collisions. Several years may pass before we have definite results, because of the amount of computation required. Parallel processors could help in our simulations [see "Advanced Computer Architectures," by Geoffrey C. Fox and Paul Messina, page 66, and "Programming for Advanced Computing," by David Gelernter, page 90].

The scientific computation needed to extend the frontiers of knowledge often requires substantial resources. The reason is partly that the models are computationally complex and partly that numerous experiments are necessary in order to characterize a class of phenomena. Scientists have traditionally obtained these resources either by acquiring large-scale computers or by renting time on them. Both routes are expensive. Large-scale computers not only carry a large price tag but also entail a huge administrative burden. Furthermore, because large computers are often not well organized for a particular problem, the construction of appropriate software can be a long and complex task. Of course, much of the problem is determining exactly what algorithms are needed to investigate the phenomena, but the architecture of the computer can either help or hinder this development. Software for three-dimensional hydrodynamics that runs well on a conventional computer takes years to write.

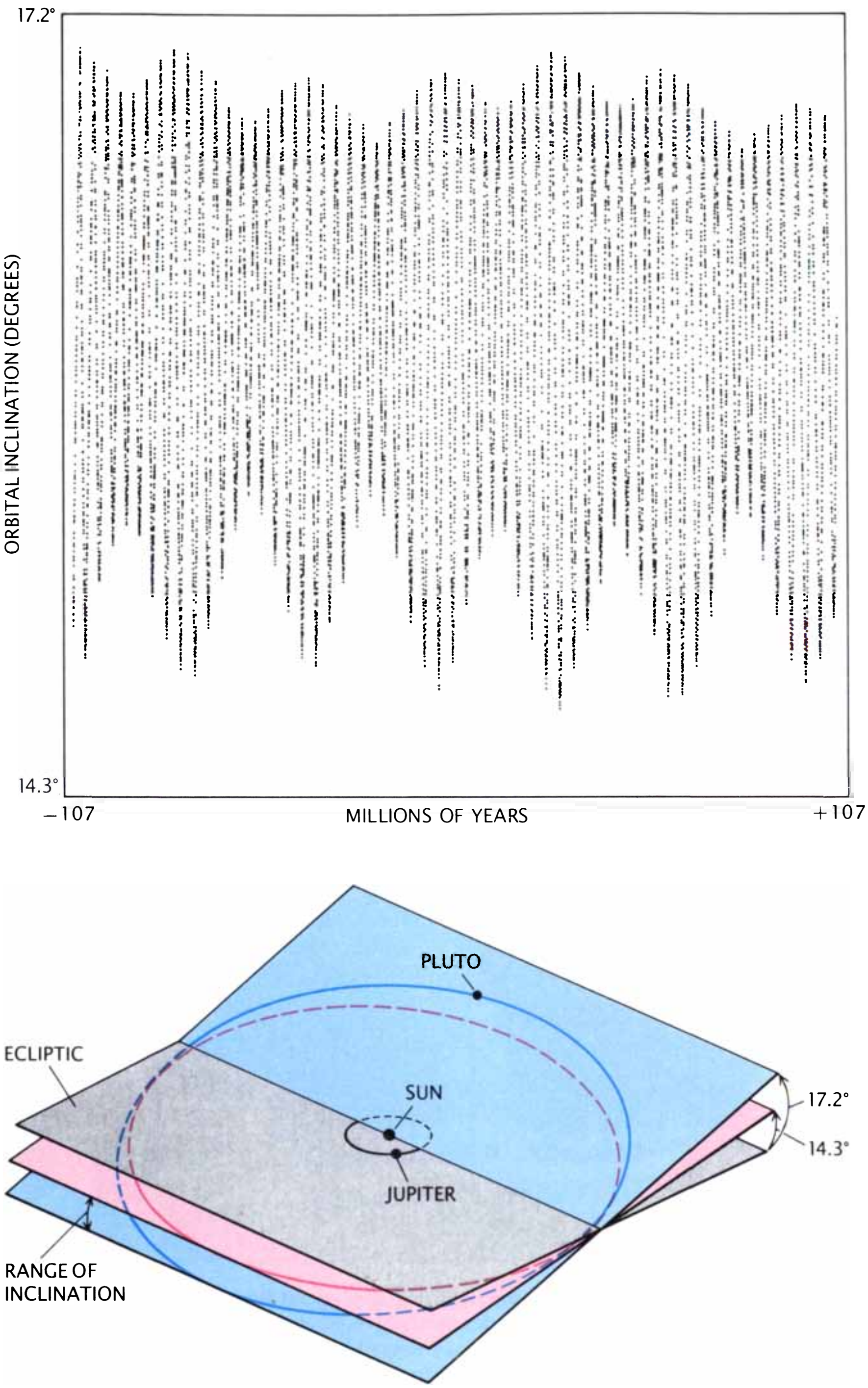
A significant part of the size, expense and difficulty of programming a large computer comes from the machine's generality. It is designed to perform well on a large variety of problems. In contrast, a specialized computer can be simple and physically small. Indeed, such hardware can be easier to design and build than software. One would expect to find scientists and computer designers cooperating in the development of computers tailored to specific applications in scientific comput-



DIGITAL ORRERY is a special computer for carrying out high-speed, high-precision simulations of planetary orbits. The device consists of about a cubic foot of electronic equipment and dissipates the power of a 150-watt light bulb. For the problems it was designed to solve, the computer does approximately 10 million floating-point operations per second, which is 60 times faster than a VAX 11/780. The specialized parallel architecture of the orrery, which is well matched to orbital-mechanics problems, is the key to obtaining such high performance. The orrery has been used to determine what the orbits of the five outer planets in the solar system must have been for the past 100 million years and what they are expected to be for the next 100 million; the period is 40 times longer than any previous precision integration of the motions of the planets.

ing. For instance, a machine designed to solve particular partial differential equations of mathematical physics could be built in about the same time, and at the same expense, as it would take to prepare programs for this application to run well on a supercomputer. Moreover, the specialized computer can become an ordinary experimental instrument belonging to the research group that made it, thereby avoiding the administrative burden and the scheduling problems associated with expensive, shared resources.

The cooperative design of specialized computers will not simply result in a few new machines for solving a few problems. Such projects will fundamentally change the culture of scientific computation. Traditionally, although scientists have been involved intimately in the design of their instruments, computers have been treated differently. Scientists are primarily users of computation; their needs are filled by a few remote vendors, most of whom are concerned primarily with the business market. Scientists must be enlisted in the design of their computing instruments. In fact, designing a computer is to a large extent a software-engineering problem, and the key to making the design of special-purpose computers an ordinary part of scientific-instrument building is simply to bring into the culture the attitude that such activity is feasible.



INCLINATION OF PLUTO'S ORBIT over a period of 214 million years has been determined with the Digital Orrery (*top*). The inclination of a planetary orbit is the angle between the plane described by its orbit and the ecliptic, which is the plane of the earth's orbit (*bottom*). In the graph a 34-million-year modulation of a basic 3.8-million-year oscillation can be discerned; there is even some evidence for a cycle of 137 million years.

The kind of attitude change we foresee is precisely analogous to the change in attitude that took place a few years ago regarding the design of very-large-scale integrated (VLSI) circuits, in which tens of thousands of components are imprinted on a single semiconducting chip. In 1980 Carver A. Mead of the California Institute of Technology and Lynn A. Conway, then at the Xerox Palo Alto Research Center, codified a set of design rules for VLSI. By doing so they removed much of the mystery from VLSI design and made it possible for computer architects to design their own special-purpose chips. The MOSIS manufacturing service, which grew out of the work of Mead and Conway, separates the design of a circuit from its manufacture. It is now possible to design a circuit and ship the specification to a central service for manufacturing. The service has recently been expanded to include the specification of layouts for printed-circuit cards, making it possible for most of a computer's components to be manufactured at a distance.

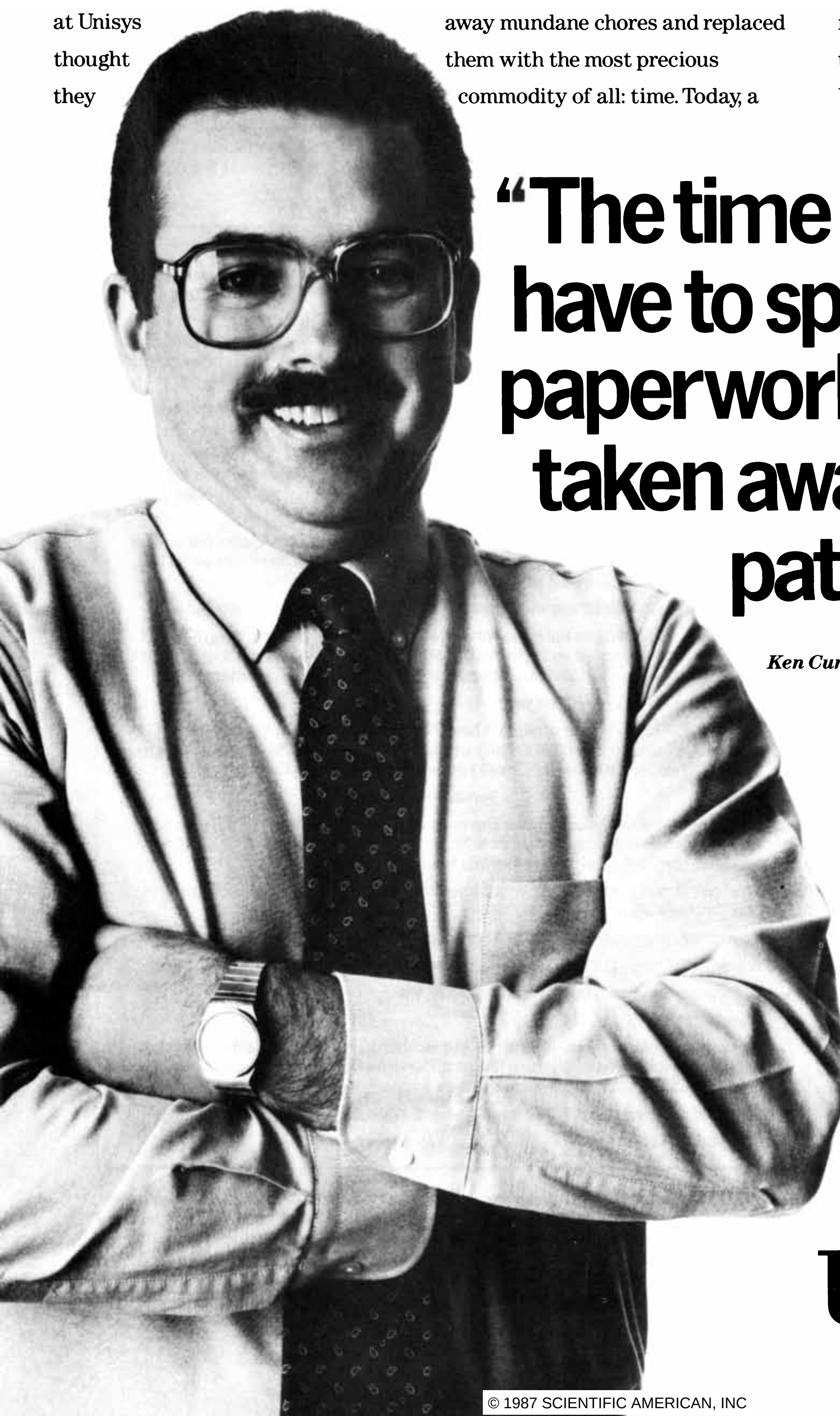
An example of a special-purpose scientific computer that was created by such a technique is the Digital Orrery, which was built by one of us (Sussman) and a group of collaborators. The Digital Orrery was designed for investigating the long-term stability of the solar system. The computer is optimized for high-precision numerical integrations of the equations of motion describing small numbers of gravitationally interacting bodies. The actual design and construction of the Digital Orrery was carried out in one year by six people (three theoretical physicists, two computer scientists and a technician). The device

MARYLAND'S Calvert Memorial Hospital wanted the same things all hospitals want today. Less time spent on paperwork and more time spent with patients.

Ken Cunzeman and his crew at Unisys thought they could help. The Unisys team was composed not only of computer experts, it included nurses and hospital managers as well.

Together with Calvert, they implemented a system that took away mundane chores and replaced them with the most precious commodity of all: time. Today, a nurse can order a lab test, inform other departments, and make sure that all charges have been recorded automatically.

Needless to say, there was an increase in productivity. Perhaps more important for Calvert Memorial, there was an increase in smiles. Unisys and healthcare. The power of ².



"The time nurses have to spend on paperwork is time taken away from patients."

Ken Cunzeman, Marketing Manager, Unisys.

UNISYS

The power of ²

consists of about one cubic foot of electronic equipment and dissipates the power of a 150-watt light bulb. For the specific problem it was designed to solve, the computer is 60 times faster than a VAX 11/780 with floating-point accelerator. The machine is currently serving in a study of orbits in the asteroid belt, aimed at understanding how empty spaces called Kirkwood gaps were formed. The Digital Orrery is also being employed to investigate the long-term stability of the orbit of Pluto. It has been used to determine what the orbits of the outer planets must have been for the past 100 million years and what they are expected to be for the next 100 million.

Before the Digital Orrery was built, high-precision integrations over simulated periods of millions of years were prohibitively expensive. As a result there were only a few small experiments making use of carefully scheduled resources. In spite of the obvious advantages of a special-purpose machine, the notion that one could be designed and built was simply not part of the cultural outlook of the astrophysics community. In fact, a preliminary proposal for construct-

Scientist: Let's consider the polar-ring galaxy A0136–0801.

Computer: I have photometric profiles and rotation curves for this galaxy from Schweizer, Whitmore and Rubin (Astron. J., Vol. 88, No. 7, July, 1983). Here they are. [We assume a graphic display of the data.]

Scientist: Let's model the mass distribution of the galaxy as a combination of a Miyamoto disk with a truncate isothermal halo. First let's fit the measured photometric profile of the spindle to the Miyamoto disk.

Computer: There is no good least-squares fit. This is the best I can do. [Again, display should appear.]

Scientist: The problem is that the polar ring passes over the northwestern part of the spindle, decreasing the brightness there. See how the profile dips around −3 arc seconds? You should eliminate these anomalous data points from your data for this fit.

Computer: O.K. I can least-squares fit the measured photometric profile with a Miyamoto disk with a scale-length of .68 kiloparsecs and a scale height of .92 kpc. Look at how good the fit is until we get out past 10 arc seconds.

Scientist: Next, let's fit the measured rotation curve to the mass distribution.

Computer: How should I use the rotation-curve data for the polar ring? The circular velocities at equal radii [the outer edge of the disk and the inner edge of the ring] appear to be similar—do you think the data sets can be merged?

Scientist: Yes, let's try it that way.

Computer: This will take a while. I have three parameters to vary: the mass of the spindle's disk, the central density of the isothermal and the isothermal's core radius. Do you want to wait? I'll call you when I have a good answer.

[. . . Some time later.]

Computer: I have a good match. The mass of the disk is 1.75×10^{10} solar masses. The isothermal halo has a mass of 9.26×10^{10} solar masses within 16.9 kiloparsecs. The core radius of the isothermal is 8.17 kiloparsecs.

Scientist: I don't understand that large a core radius. . . . How do you compute your isothermal?

Computer: My potential is proportional to log $(R^2 + B^2)$ for a core radius of B.

Scientist: That doesn't sound right; my potential has an arc tangent in it.

Computer: So your density function is just the $\rho_0/(1 + R^2/B^2)$, right?

Scientist: Yes.

Computer: That explains it. Both of our isothermals have the same asymptotic behavior, but I measure my core radius differently from the way you do. In your units the core will be smaller, about three kiloparsecs.

Scientist: Sounds good. What mass-to-light- ratio do we have for the spindle?

Computer: About five.

HYPOTHETICAL DIALOGUE between a computer and an astrophysicist illustrates how interactive systems in the future could help to construct, simulate and analyze models. The authors envision a system that can carry on a conversation similar to the one shown here, although probably in a more formal language. The most important feature of this hypothetical dialogue is the conceptual level of the interaction with the computer: the program understands a considerable amount about galaxy modeling.

ing the Digital Orrery submitted to the astronomical instruments division of the National Science Foundation was rejected on the grounds that the project would be infeasible. Such an attitude is striking in view of the immensely complex engineering projects that have been accomplished by this same community when building telescopes. Building computers is just not that difficult.

Improving the effectiveness of scientific computation requires going beyond incremental improvements in hardware and algorithms. The efficiency of the experimenter must be improved by providing him or her with the computational equivalent of a toolbox. The toolbox will be equipped with tools for building as well as manipulating computational models that enable investigators to specify and control computational experiments in terms of the underlying qualitative phenomena. The tools must do more than set up data for a numerical computation, run it and analyze the results; they must also understand how the computational models work and how they relate to the other descriptions of the phenomena being studied.

We envision systems that combine powerful numerical methods with sophisticated symbolic manipulation of both algebraic and algorithmic descriptions. An advanced environment will provide symbolic mathematical support for the automatic preparation, execution and analysis of the results of numerical experiments. Symbolic methods can be used to help manipulate the dynamical equations into a suitable form for simulation and to construct high-quality simulation programs, optimized for the particular situation from more general algorithmic descriptions. A scientist should be able to discuss a galaxy model with a computer in much the same way as one might describe it to a human assistant. We believe such computer systems are within reach even though none has yet been built.

New tools have often changed civilization in ways not envisioned by the inventors; automatic computation has already outgrown purely numerical calculation. Symbolic and qualitative computations are becoming an integral part of the scientific environment. Computers will soon provide the scientist with an effective laboratory assistant rather than a number-crunching slave.

Advanced Computing for Medicine

In time computers may be as basic to medicine as the stethoscope. Medical systems can store data and retrieve it selectively; soon they will be "smart" enough to advise on diagnosis and treatment

by Glenn D. Rennels and Edward H. Shortliffe

For several weeks Ms. Jones has had low-back pain. Bed rest and aspirin have not helped. Perhaps a prescription analgesic would do better, she thinks, and she goes to a family doctor in the city to which she has recently moved. She tells the doctor the pain started soon after she played tennis for the first time in a year. Examination reveals only some limitation of motion in the left elbow and tenderness in the lower back.

The doctor turns to his personal computer, which has access to a nationwide computer network, and he links to a hospital computer in the city from which Ms. Jones has moved. The hospital computer has electronic medical records to which authorized medical workers can gain access with a patient's permission. Within seconds the doctor finds that Ms. Jones underwent a lumpectomy to remove a small tumor of the breast five years ago. She has been perfectly healthy ever since and neglected to mention the episode in discussing her back pain.

The doctor disconnects from the network and calls up QMR, a diagnostic system. QMR, which has comprehensive knowledge of almost 600 diseases stored in its memory, suggests that Ms. Jones may have arthritis, muscle strain or recurrence of the cancer with metastasis to the base of the spine. The doctor, reassured to see that QMR's "second opinion" agrees with his own concern, orders a bone scan to eliminate the possibility of recurrent cancer.

Such a scene is no longer just futuristic dreaming. We think the time is not far off when physicians will consider the computer to be as essential a medical instrument as the stethoscope. The medical information systems that are now emerging from research laboratories can be grouped into two categories. "Communication systems" store medical information, retrieve it selectively and transmit it; "advice systems" apply the information to help doctors diagnose a patient's condition, as QMR did in the situation described above, or to propose, monitor and help to manage a course of treatment.

Advances that will make the systems much more effective are described in some of the preceding articles in this issue. Although hardware improvements will make computer systems more effective, the emphasis in a number of laboratories, including our own at the Stanford University School of Medicine, is on the development of software methods that can represent medical information and expertise in a machine and make that information and expertise readily available to a physician. Here we shall first describe, by example, the capabilities and promise of communication systems and advice systems. We shall then discuss some fundamental problems that still confound software designers' attempts to develop computer programs that can "reason" about a case much as a doctor might.

The need for communication systems arises in part because it is increasingly difficult for a physician (or a biomedical investigator) to read, memorize and remember all the information needed to solve a particular problem. A physician may not have encountered the patient's complaint or disease for a period of months or years—during which diagnostic and therapeutic technologies may well have changed.

Physicians have accordingly started to use literature data bases, or bibliographic retrieval systems. MEDLINE is the best-known example. A large computer at the National Library of Medicine in Bethesda, where the system was developed, holds the titles of essentially all articles from the world's biomedical literature of the past 25 years and the full abstracts of the recent ones. Each article is indexed with carefully chosen key words and phrases, so that it is relatively easy to call up a list of relevant titles and then ask to see the abstracts of the most appropriate papers. Bibliographic retrieval systems have been introduced that also store the full text of papers and enable one to locate pertinent sentences within any paper.

As more capabilities are added to its repertoire the computer will become the principal means of access to the biomedical literature and to disease data bases for which libraries will be responsible. Libraries will be viewed not as physical places—buildings filled with books and journals—but as distributed sources of information. This view is not really in conflict with the traditional model of medical libraries as information centers. It is simply that new technologies allow the model to escape the confines of physical structures.

Communication systems provide access to information; now advice systems are being developed that can apply stored information to the solution of particular real-world problems. For example, we and our colleagues have been working with an experimental advice system we call ROUNDSMAN. More particularly, it is a critiquing system: the physician describes a clinical problem and a proposed plan of action and ROUNDSMAN thereupon delivers a critique of the plan.

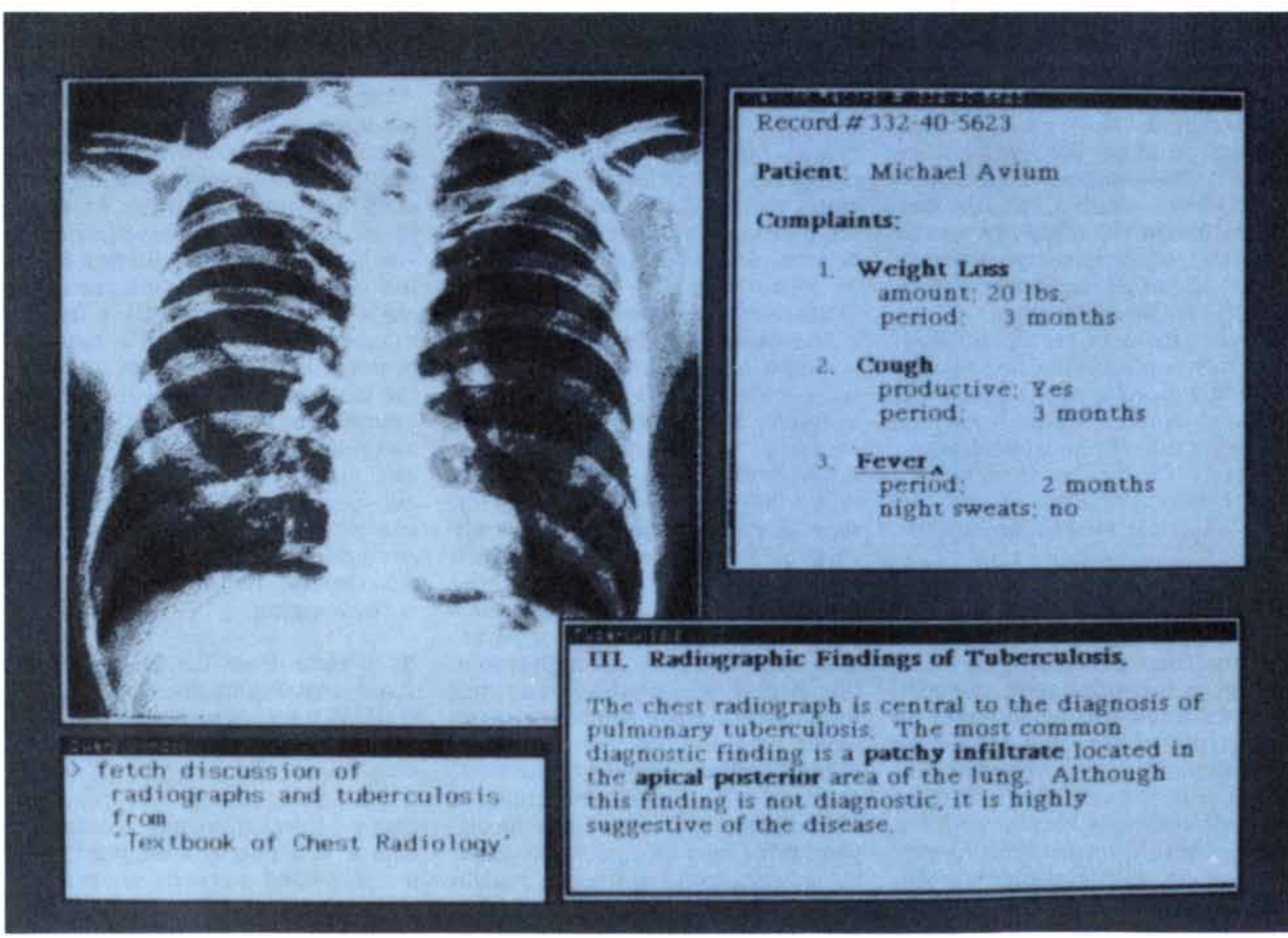
A central objective of this system is to provide access not only to the data in articles but also to the *meaning* of the data in the light of expertise. At this early stage ROUNDSMAN can produce patient-specific analyses of various options in breast-cancer management based on 24 articles. To achieve this capability we first encode information from the articles. Bits and pieces are stored: for example, "Doe and Roe *NEJM* 1986...in all cases admitted to study, lymph nodes were removed and documented to contain tumor...473 cases... randomized to treatments...236 cases by mastectomy...237 by lumpectomy..." Many more details of study design and results are entered.

Then we elicit subjective judgments about the information from an expert—the kind of doctor a fellow worker might ask for advice about a difficult case. Enough judgments are stored with the data to cover a host of different situations ROUNDSMAN might be asked to consider. Suppose the expert tells us the study conclusions must be applied cautiously to the typical stage II cancer patients because the study patients (who have positive lymph nodes) are sicker; the expert's advice and his way of phrasing the issue go into the system's memory.

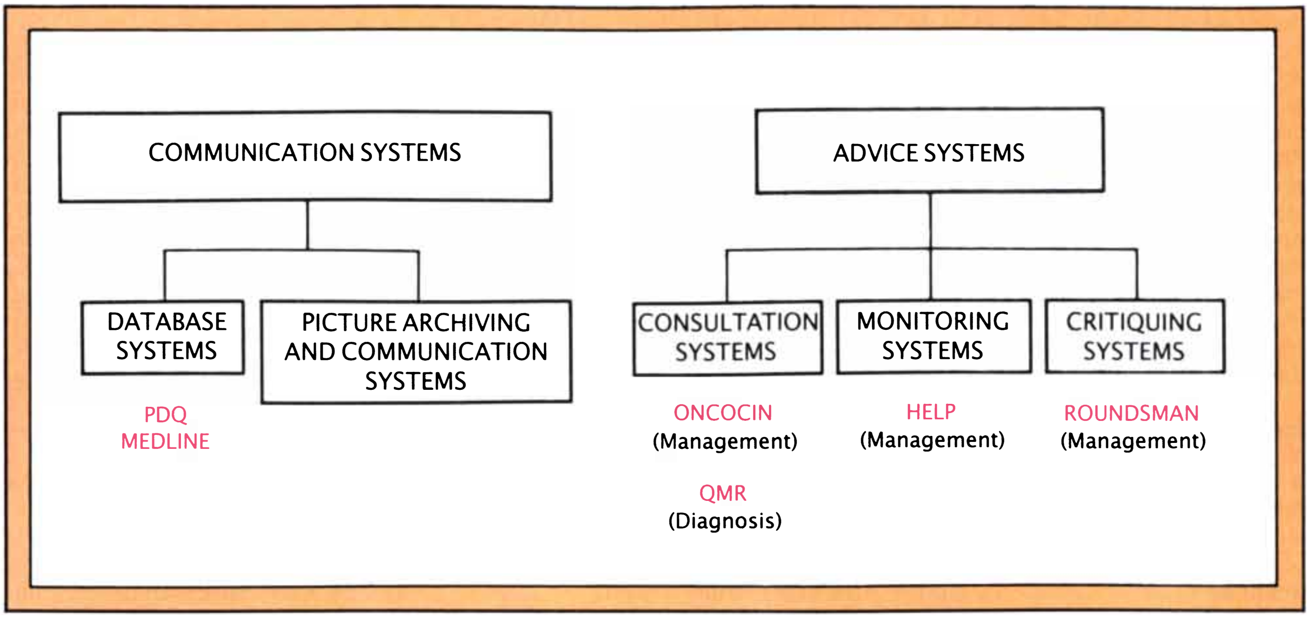
Consider what might have happened when the hypothetical Ms. Jones first discovered a lump and was referred to a surgeon. He examines her, calls up ROUNDSMAN, enters his patient's characteristics and proposes to do a lumpectomy. ROUNDSMAN responds: "Doe and Roe reported in the *New England Journal of Medicine* in 1986 that lumpectomy was effective in such cases, but note that the study population was in a worse prognostic stratum than your patient." After bringing this consideration and others to the surgeon's attention, ROUNDSMAN suggests that in spite of the discrepancies the results of the study generally support the plan of lumpectomy.

The distinctions between software for communication systems and software for advice systems can be understood if one looks at the respective capabilities and objectives of two experimental systems, Physician Data Query (PDQ) and ONCOCIN. Both systems deal with cancer-treatment protocols: carefully specified plans for conducting clinical trials in which alternate approaches to treatment are compared. A protocol sets forth particular surgical procedures, chemotherapeutic or radiation regimes, dosages and treatment schedules. It also outlines the appropriate responses to more subtle issues, such as what to do when a patient has toxic reactions to treatment or when to conclude that a therapeutic approach is not working.

PDQ is a communication system: a database, operated jointly by the National Cancer Institute and the National Library of Medicine, that can provide physicians with information about the protocols for a large number of formal clinical trials currently under way. With a personal computer and a modem a doctor can gain access to a central computer at the li-



MEDICAL WORKSTATION might bring the radiology department, the patient-records department and the medical library to a doctor's office. For this hypothetical display the physician called up Michael Avium's chest X ray and medical record. The patient's complaints were suggestive of tuberculosis, and so the doctor asked the system to search a textbook for a discussion of radiographic evidence of TB (*bottom left*). The computer thereupon found and displayed one such discussion (*bottom right*).



MEDICAL COMPUTER SYSTEMS can be categorized as communication systems and advice systems. Communication systems store, retrieve and transmit bibliographic material, patient records (including up-to-date test results) and other data. Advice systems participate actively in diagnosis or in management of patient care. Systems discussed in the text are shown in color.

brary and, with the help of specialized software, identify protocols for which particular patients are eligible. The protocols typically present new approaches to difficult cases—approaches the doctor may be able to adopt (unless the protocol involves administration of an experimental drug). Alternatively, the doctor can obtain the names of nearby hospitals and physicians participating in trials of a course of treatment and refer the patient to them.

Suppose the bone scan ordered for Ms. Jones confirms the presence of cancer at the base of the spine. Her physician could elicit from PDQ the names of medical centers working on advanced treatments for metastatic breast cancer. The doctor would then be able to refer her to one of those centers.

The PDQ system is intended to display a wide range of information, not to help its user apply the information to a particular patient. We and our colleagues have designed ONCOCIN, a system that complements PDQ by helping physicians who are caring for individual cancer patients after they have been enrolled in a protocol and their treatment is under way. ONCOCIN is both a medical-record system and an advice system (which today deals with only a small number of treatment protocols). It keeps track of the course of treatment and of the patient's progress, but it also has enough knowledge to assume an active role in suggesting how the protocol should be adapted to the needs of a specific patient.

Suppose that at the regional cancer center Ms. Jones enrolls in a clinical trial of a new drug regimen. On each visit the physicians carrying out a series of complex chemotherapy treatments can be guided by ONCOCIN. It stores the results of each test, keeps track of her condition and proposes how the dosage of each drug should be adjusted on each visit according to her response to treatment.

The protocol's details are represented within the computer in structures known as inference rules. For example, a rule might say: "*If* the patient's white-blood-cell count is lower than normal but still greater than 3,000 cells per cubic millimeter, *then* administer drug *A* but give only 75 percent of the usual dose." Such rules allow ONCOCIN to advise about the proper therapy to administer on a particular visit and to call for particular tests. The system's high-resolution display screen simulates a paper form that is already familiar to oncologists, who describe their patient by filling in certain sections of the form. ONCOCIN offers advice by filling in blank spaces in other areas of the display. If the user has questions about the advice, ONCOCIN tries to explain it on the basis of the protocol's logic as stored in the program's rules. The user is, of course, still free to override the advice: no computer program can have access to the full range of variables the treating physician should consider in making a patient-care decision. It is assumed the user may have additional information that calls for adjusting ONCOCIN's advice.

The principal research questions for systems such as ROUNDSMAN and ONCOCIN are software issues (some of which will be explored below), but such hardware advances as faster processors and larger memory chips will certainly facilitate routine use of the systems. For example, memory chips can now hold only a limited number of protocols. If a patient's protocol is not in internal memory, the machine must fetch the needed information from storage on a disk, swapping it for information currently held in memory. This is a slow process—perhaps too slow for routine use in a clinic. The high-capacity memory chips now under development will substantially reduce the amount of such swapping.

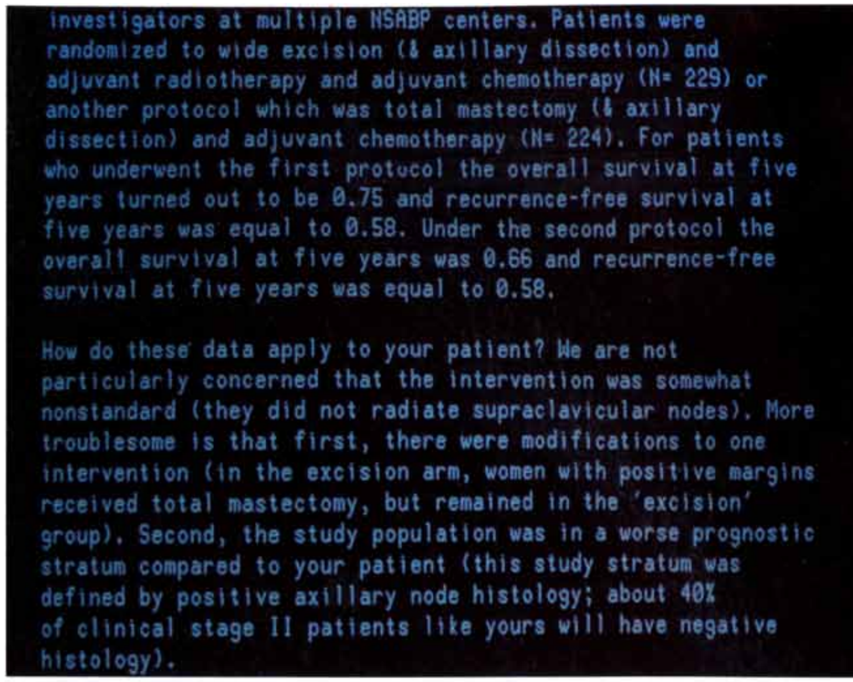
It is clear from the discussion of ROUNDSMAN and ONCOCIN that the style of interaction between an advice system and its users can vary. ROUNDSMAN, as we mentioned above, is a critiquing system: it reacts to the user's current thinking—a proposed course of action or a diagnostic hypothesis—and suggests alternatives if necessary. A consultation system, on the other hand, generates independent analyses and recommendations, which users can compare with their own thoughts. Whereas ONCOCIN consults on the management of a patient's illness, the advice system called QMR (for Quick Medical Reference), mentioned at the beginning of this article, consults on diagnosis.

QMR is a microcomputer adaptation of INTERNIST-1, a large diagnostic program developed at the University of Pittsburgh School of Medicine. It has knowledge of 577 diseases and of their interrelations with 4,100 signs, symptoms and other patient characteristics. It can serve health professionals in three modes. In its basic mode QMR is an expert consultation system that provides patient-specific diagnostic hypotheses. It does so by asking itself, in effect: How often will patients with a given characteristic have a particular disease? How often will patients with a certain disease manifest a given characteristic? The system can request information it needs that was not volunteered and can suggest tests and point out areas that need further investigation. It analyzes the information obtained and comes up with the most probable diagnosis or diagnoses.

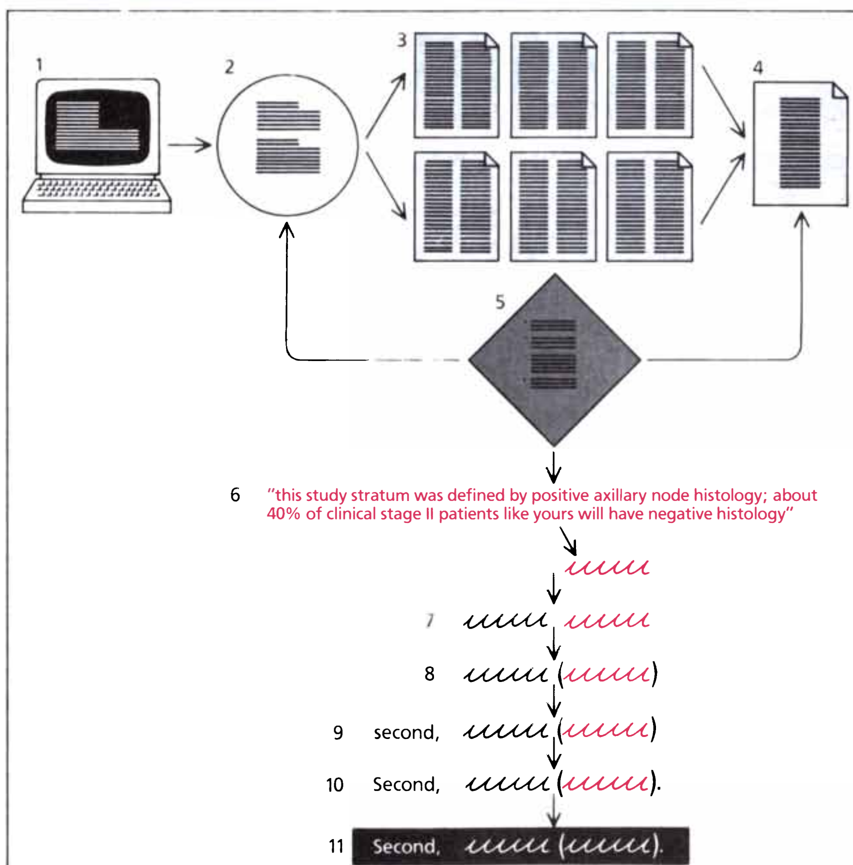
Second, QMR can be an electronic textbook, listing the patient characteristics reported to occur in a given disease or, conversely, reporting which of its 577 diseases can be associated with a given characteristic. Third, as a medical spreadsheet it can combine a few characteristics or diseases and determine the implications. For example, one can specify two apparently unrelated medical problems and obtain suggestions about how coexisting diseases could, under the right circumstances, give rise to both problems.

INTERNIST-1, the predecessor of QMR, was shown to perform almost as well as academic physicians in diagnosing difficult cases. QMR has the same broad knowledge base and diagnosing capability, to which the electronic-textbook and spreadsheet modes have been added. QMR is now being field-tested at Pittsburgh and at collaborating institutions.

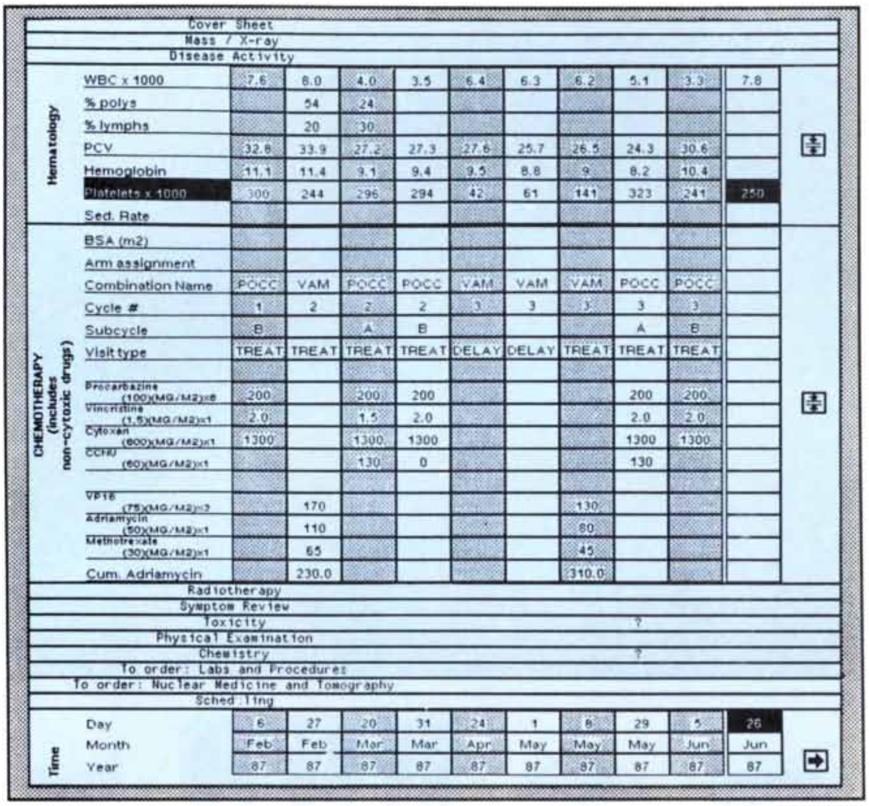
When an advice system such as QMR stands alone, it must generally wait for a user to seek advice. Moreover, the user of an advice system normally must manually enter patient information that may already reside in a nearby data-management communication system. Might an advice system automatically tap such a data source in order to volunteer advice whenever it may be appropriate? One of the first integrated hospital information systems able to do that, called HELP, has been developed


investigators at multiple NSABP centers. Patients were randomized to wide excision (& axillary dissection) and adjuvant radiotherapy and adjuvant chemotherapy (N= 229) or another protocol which was total mastectomy (& axillary dissection) and adjuvant chemotherapy (N= 224). For patients who underwent the first protocol the overall survival at five years turned out to be 0.75 and recurrence-free survival at five years was equal to 0.58. Under the second protocol the overall survival at five years was 0.66 and recurrence-free survival at five years was equal to 0.58.

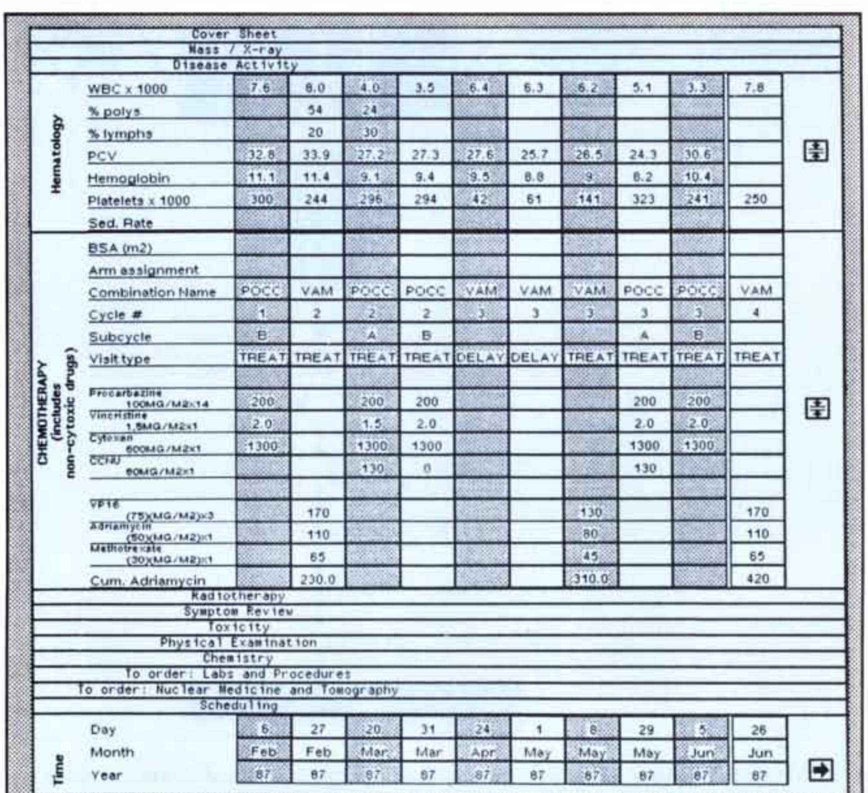
How do these data apply to your patient? We are not particularly concerned that the intervention was somewhat nonstandard (they did not radiate supraclavicular nodes). More troublesome is that first, there were modifications to one intervention (in the excision arm, women with positive margins received total mastectomy, but remained in the 'excision' group). Second, the study population was in a worse prognostic stratum compared to your patient (this study stratum was defined by positive axillary node histology; about 40% of clinical stage II patients like yours will have negative histology).



ROUNDSMAN generates coherent comments on a proposed plan of treatment, as in the excerpt giving data from a published paper and commenting on it (*top*). The process by which the last sentence in the example was generated is shown (*bottom*). The doctor enters a description of his patient and proposed treatment (*1*). The system organizes his input into a "clinical context" (*2*), searches for a relevant article (*3*) and extracts pertinent data (*4*). Then it calls up its knowledge base of expert judgments (*5*) and uses it to contrast the clinical context with the pertinent data. One contrast statement (*6*) becomes an element of the emerging prose. The system precedes the statement with a generalization (*7*), parenthesizes the particulars (*8*), adds "second" to position the statement in a larger list (*9*), capitalizes and punctuates (*10*) and displays the result (*11*).



	1	2	3	4	5	6	7	8	9	10
Cover Sheet										
Mass / X-ray										
Disease Activity										
Hematology										
WBC x 1000	7.6	8.0	4.0	3.5	6.4	6.3	6.2	5.1	3.3	7.8
% polys		54	24							
% lymphs		20	30							
PCV	32.8	33.9	27.2	27.3	27.6	25.7	26.5	24.3	30.6	
Hemoglobin	11.1	11.4	9.1	9.4	9.5	8.8	9	8.2	10.4	
Platelets x 1000	300	244	296	294	42	61	141	323	241	250
Sed. Rate										
CHEMOTHERAPY (includes non-cytoxic drugs)										
BSA (m2)										
Arm assignment										
Combination Name	POCC	VAM	POCC	POCC	VAM	VAM	VAM	POCC	POCC	
Cycle #	1	2	2	2	3	3	3	3	3	
Subcycle	B		A	B				A	B	
Visit type	TREAT	TREAT	TREAT	TREAT	DELAY	DELAY	TREAT	TREAT	TREAT	
Procarbazine (100)(MG/M2)x8	200		200	200				200	200	
Vincristine (1.5)(MG/M2)x1	2.0		1.5	2.0				2.0	2.0	
Cytoxan (600)(MG/M2)x1	1300		1300	1300				1300	1300	
CCNU (60)(MG/M2)x1			130	0				130		
VP16 (75)(MG/M2)x3		170					130			
Adriamycin (50)(MG/M2)x1		110					80			
Methotrexate (30)(MG/M2)x1		65					45			
Cum. Adriamycin		230.0					310.0			
Radiotherapy										
Symptom Review										
Toxicity								?		
Physical Examination										
Chemistry								?		
To order: Labs and Procedures										
To order: Nuclear Medicine and Tomography										
Scheduling										
Time										
Day	6	27	20	31	24	1	8	29	5	26
Month	Feb	Feb	Mar	Mar	Apr	May	May	May	Jun	Jun
Year	87	87	87	87	87	87	87	87	87	87



	1	2	3	4	5	6	7	8	9	10
Cover Sheet										
Mass / X-ray										
Disease Activity										
Hematology										
WBC x 1000	7.6	8.0	4.0	3.5	6.4	6.3	6.2	5.1	3.3	7.8
% polys		54	24							
% lymphs		20	30							
PCV	32.8	33.9	27.2	27.3	27.6	25.7	26.5	24.3	30.6	
Hemoglobin	11.1	11.4	9.1	9.4	9.5	8.8	9	8.2	10.4	
Platelets x 1000	300	244	296	294	42	61	141	323	241	250
Sed. Rate										
CHEMOTHERAPY (includes non-cytoxic drugs)										
BSA (m2)										
Arm assignment										
Combination Name	POCC	VAM	POCC	POCC	VAM	VAM	VAM	POCC	POCC	VAM
Cycle #	1	2	2	2	3	3	3	3	3	4
Subcycle	B		A	B				A	B	
Visit type	TREAT	TREAT	TREAT	TREAT	DELAY	DELAY	TREAT	TREAT	TREAT	TREAT
Procarbazine 100MG/M2x14	200		200	200				200	200	
Vincristine 1.5MG/M2x1	2.0		1.5	2.0				2.0	2.0	
Cytoxan 600MG/M2x1	1300		1300	1300				1300	1300	
CCNU 60MG/M2x1			130	0				130		
VP16 (75)(MG/M2)x3		170					130			170
Adriamycin (50)(MG/M2)x1		110					80			110
Methotrexate (30)(MG/M2)x1		65					45			65
Cum. Adriamycin		230.0					310.0			420
Radiotherapy										
Symptom Review										
Toxicity										
Physical Examination										
Chemistry										
To order: Labs and Procedures										
To order: Nuclear Medicine and Tomography										
Scheduling										
Time										
Day	6	27	20	31	24	1	8	29	5	26
Month	Feb	Feb	Mar	Mar	Apr	May	May	May	Jun	Jun
Year	87	87	87	87	87	87	87	87	87	87

ONCOCIN consults about chemotherapy by way of an electronic spreadsheet, two "snapshots" of which are shown: a data-gathering stage and an advice stage. Windows for hematology and chemotherapy are open; others are available. The doctor enters hematology data for June 26 (*top*). Working with those data and with knowledge of the applicable protocol and of the patient's condition and earlier treatment, ONCOCIN applies rules to recommend dosages for the patient's June 26 visit (*bottom*); under different conditions the system might have modified the dosages or advised delaying treatment.

over the past 15 years at the LDS Hospital in Salt Lake City.

HELP runs on a central computer that is connected with terminals and printers throughout the hospital. There are now at least four terminals and a printer in each nursing division, for example, and a terminal at each bedside in intensive-care units; the plan is to have a terminal beside all 520 beds in the hospital. The central computer is first of all a data-management facility, storing the admission history of each patient and keeping track of each patient's medications, laboratory results and current status. In addition to serving as a data-management system, the central computer incorporates thousands of special logical subprograms called HELP sectors. Their purpose is to monitor all available data for each patient and keep watch for any of a number of specified alert conditions. When any such condition is identified, HELP issues a warning message to the appropriate staff members.

When the hypothetical Ms. Jones was hospitalized for her original breast surgery, a system such as HELP might have monitored her medications, laboratory-test results and vital signs. Unlike a physician, it can review test results within seconds, day or night. Suppose Ms. Jones was taking the drug digoxin for a heart condition. If her blood-test results revealed a low potassium level, HELP would have warned the floor nurse that low potassium is dangerous to a patient taking digoxin and that Ms. Jones might need prompt treatment with an oral solution of potassium chloride.

HELP is a superb example of how the flexibility and effectiveness of an advice system can be heightened by integration with a communication system. And HELP is just the beginning. Local-area-network technology should make it possible to add to hospital information systems a specialized advice system such as ROUNDSMAN, a diagnostic system and even a facility that would bring X rays, sonograms, CAT scans and other imagery from the radiology department to a terminal in every operating room and at every bedside.

Considering their power and their availability in experimental versions, why are advanced medical systems not in widespread clinical use today? The fact is that the methodology for constructing advice systems is not yet mature: the design of effective software still poses some fundamental theoretical and practical challenges.

A basic element of most advice systems is a model of certain clinical skills, notably the process of decision making. How is one to model such an eminently human and specialized capacity? There are many strategies, based on such varied approaches as pattern recognition, statistical regression, branching algorithms and decision theory. Examination of another approach, that of artificial intelligence (AI), will highlight some of the difficulties encountered in developing advice-system software.

Investigators into medical AI have been influenced by psychological research suggesting that what distinguishes experienced physicians is primarily the richness of their knowledge—that is, their store of information about diseases, symptoms and treatments—rather than any particular way of processing that information. Computer programs that use AI techniques to represent and manipulate detailed knowledge of a field of expertise are known as expert systems. In developing an expert system medical AI investigators seek first of all to develop an optimal representation of medical knowledge within the computer. Much of the leverage of a medical knowledge-based system derives from its store of knowledge and the way it is structured, rather than from any novel processing techniques (such as an algorithm the computer might handle better than a human being can).

One of the primary challenges in constructing a knowledge-based system is to cope with the sheer quantity of information to be represented and manipulated. Take, for example, QMR's 577 diseases and 4,100 patient characteristics. When these 4,677 entities are linked and weighted (to show that certain symptoms suggest certain diseases, that one disease might cause another and that some patient characteristics are more important than others), the QMR knowledge base may contain some 250,000 items of knowledge. Many of them are not explicitly recorded in any textbook. It took 10 years of collaboration between computer specialists and doctors to make the items sufficiently concrete and explicit for representation in the system. Moreover, QMR covers only a subset of the important known diseases. A central task in expert-system design, then, is to identify key knowledge elements and their interrelations and then to encode them in data structures and linkages that make it possible to apply the knowledge effectively to solve problems.

A second design challenge is to ensure that the data structures representing knowledge elements will enable people to add to and correct the system's knowledge. Medical discoveries change medical practice, and so a knowledge-based system must accept incremental changes in its knowledge base. Tinkering with the entire program to accomplish each updating is impractical. The data structure (whether it is lists, a hierarchy, a network or some other structure) that represents medical knowledge should therefore ideally be independent of the algorithms that manipulate the knowledge. This independence is hard to achieve because certain medical knowledge is itself somewhat algorithmic, in that it specifies a strategy for problem solving rather than an actual fact. For example, diagnosticians often seek to establish the general scope of an illness—for instance, "liver disease"—before trying to pin down a particular kind of liver disease.

In addition, each element of medical knowledge should ideally be represented in modular form, so that the alteration of one element leaves the others undisturbed. For example, two data structures intended to achieve such modularity are rules and frames. A rule is an inferential statement of the form "If *A*, then *B*" by means of which a system can reach a conclusion through the logical technique known as modus ponens: If *A* implies *B*, and if *A* is known to be true, then *B* can be inferred. The inferred results of one rule can be used by another rule to make a second inference; that is, the rules "chain together" dynamically.

A frame, on the other hand, represents an entity, such as a disease, that is characterized by "slots" defining prototypical features of that entity, such as the symptoms caused by the disease. It is also possible to define methods for inferring the values of slots from other aspects of the frame or from other frames.

The modularity offered by rules and frames is inherently difficult to attain, in part because medical knowledge is more than the sum of

```
Associations List

Pulmonary Disease  and  DIARRHEA Chronic

Pairs of Diseases consistent with Entered Finding and Topic

 Atelectasis
 caused-by Carcinoid Syndrome Secondary To Bronchial Neoplasm

 Eosinophilic Pneumonia Acute <LOEFFLER>
 caused-by Hookworm Disease

 Pulmonary Legionellosis
 predisposed-to-by Immune Deficiency Syndrome Acquired <AIDS>

 Pleural Effusion Exudative
 caused-by Pancreatic Pseudocyst

 Pneumococcal Pneumonia
 predisposed-to-by Carcinoid Syndrome Secondary To Bronchial Neoplasm

 Pneumocystis Pneumonia
 predisposed-to-by Immune Deficiency Syndrome Acquired <AIDS>

 Pulmonary Hypertension Secondary
 caused-by Progressive Systemic Sclerosis
 or co-occurring-with Schistosomiasis Chronic Hepatic

 Pulmonary Infarction
 predisposed-to-by Carcinoma Of Body Or Tail Of Pancreas
 or predisposed-to-by Carcinoma Of Head Of Pancreas
 or caused-by Hepatic Vein Obstruction

 Pulmonary Lymphoma
 coinciding-with Lymphoma Of Colon
 or coinciding-with Small Intestinal Lymphoma

 Pulmonary Interstitial Fibrosis Secondary Diffuse
 caused-by Progressive Systemic Sclerosis
```

QMR is asked (*color*) to list diseases that might account for the presence of two apparently disparate problems: pulmonary disease and chronic diarrhea. The list seen here is a subset of QMR's complete response, generated by the system on the basis of its knowledge base of 577 diseases, 4,100 patient characteristics and their interrelations.

RULE
To determine the dose of methotrexate administered in VAM chemotherapy in protocols 20-83-1 and 2091:
If: the serum creatinine level (in mg per dl) exceeds 1.5
Then: do not give methotrexate

FRAME		
Name: Tularemia		
Symptoms and Signs:	Evoking Strength	Frequency
FEVER	0	5
SKIN Lesion Culture Francisella Tularensis	5	4
LYMPH Node Enlarged	1	4
SKIN Ulcer	1	4
EKG Sinus Tachycardia	0	4
TULARENSIS Skin Test Positive	4	3
EXPOSURE To Rabbits Rodents Small Mammals	2	3
TICK Bite Recent History	2	3
HEADACHE Severe	1	3
Can cause:		
Pyrogenic Shock	1	2
Appendicitis Acute	1	1
Predisposes to:		
Endocarditis Acute Infective Left Heart	1	1

RULES AND FRAMES are the elements of two approaches to representing medical knowledge. Rules have the form "If condition *X*, then action *Y*." When rule-based ONCOCIN (*top*), in determining the proper dosages for VAM chemotherapy, establishes that the "if" part of its methotrexate rule is true, it concludes that the "then" part is indicated. Frames are descriptive details of such objects as diseases and symptoms. In frame-based QMR the disease frame for tularemia has weighted links to various signs, symptoms and other diseases. "Evoking strength" weights reflect the likelihood that the disease is present when a certain symptom is observed: 0 means the symptom is nonspecifiic and 5 means the disease is the only cause of the symptom. "Frequency" weights, on the other hand, indicate how likely the disease is to give rise to a particular symptom: 1 means it does so rarely, whereas 5 means it does so in almost every case.

its parts. For example, ROUNDSMAN represents each journal article as a distinct object in its knowledge base. When a new article is published, one would like to augment the knowledge base in a modular fashion. Yet the new information in an article often casts a new light on previous articles, altering their interpretation. The more one tries to accommodate this nonmodular aspect of medical knowledge—for example by inserting links between articles to represent their interrelations—the more one creates an entangled knowledge base of tightly coupled elements that is practically impenetrable when one tries to update it.

A third research challenge arises because medicine has only limited models of the human body or of disease. This is in contrast to systems that are designed, say, to facilitate a manufacturing process. In manufacturing there is clear understanding of the function of individual components and of how they are assembled; production problems and fault diagnosis can often be approached by dealing with the combinatorial characteristics of an assembly sequence or of the ways components might malfunction.

Few components of the human body and few of their functions are completely understood at this mechanistic level. Physicians therefore need to make use of a wide range of knowledge—including rules of thumb learned from the best clinicians, empirical evidence derived from statistical analysis of clinical trials, causal mechanisms inferred from anatomy, physiology and molecular biology, and social issues pertinent to good medical care. A major task for medical computer scientists is to develop techniques for encoding and applying such different kinds of knowledge in a coordinated way.

The conceptual and design tasks facing developers of advice systems can be contrasted with the practical problems that already challenge developers of communication systems. These problems can be appreciated if one considers the barriers encountered in introducing a picture-archiving and communication system, or PACS. Such systems are well beyond the research stage but are still generally unavailable because both logistical and economic barriers thwart their implementation.

A PACS stores X-ray and other images in digital form, delivers them to where they are needed and displays pictures of high photographic quality on screens. At the present time doctors walk to the radiology department to examine an X-ray picture; patients must often carry their X rays from office to office. It would be much more efficient if X-ray images could be distributed directly to hospital wards and clinics, particularly if the same film could be viewed simultaneously by different physicians at different sites.

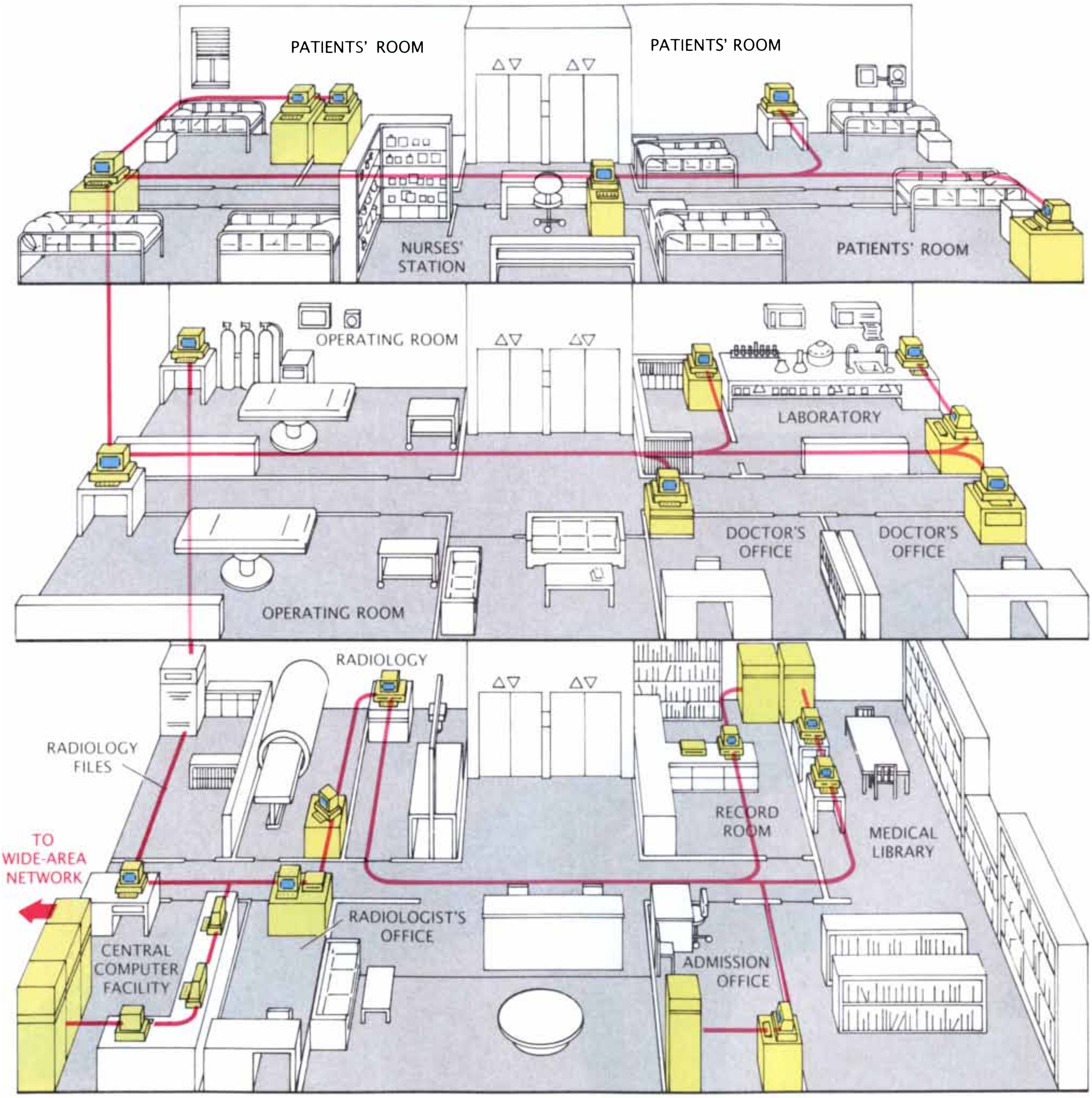
Software for creating images and displaying them in this way is already well in hand; fiber-optic communication lines exist that can support the necessary signaling rate of 200 megabits per second; terminals that display the images with sufficient resolution are available (and indeed are in use in the motion-picture and television industry). To be sure, digital imaging systems will benefit from further research, but even present technology could dramatically improve the distribution and interpretation of medical images. Yet in today's competitive financial climate for medicine it is hard to justify an expensive new technology.

What would it take to introduce a PACS into a busy modern hospital? Laying the fiber-optic network would be disruptive as well as expensive. To distribute high-resolution images widely the hospital would need a large number of workstations, each with a display resolution of 2,048 by 2,048 pixels (discrete picture elements), with each pixel able to register from five to eight shades of gray.

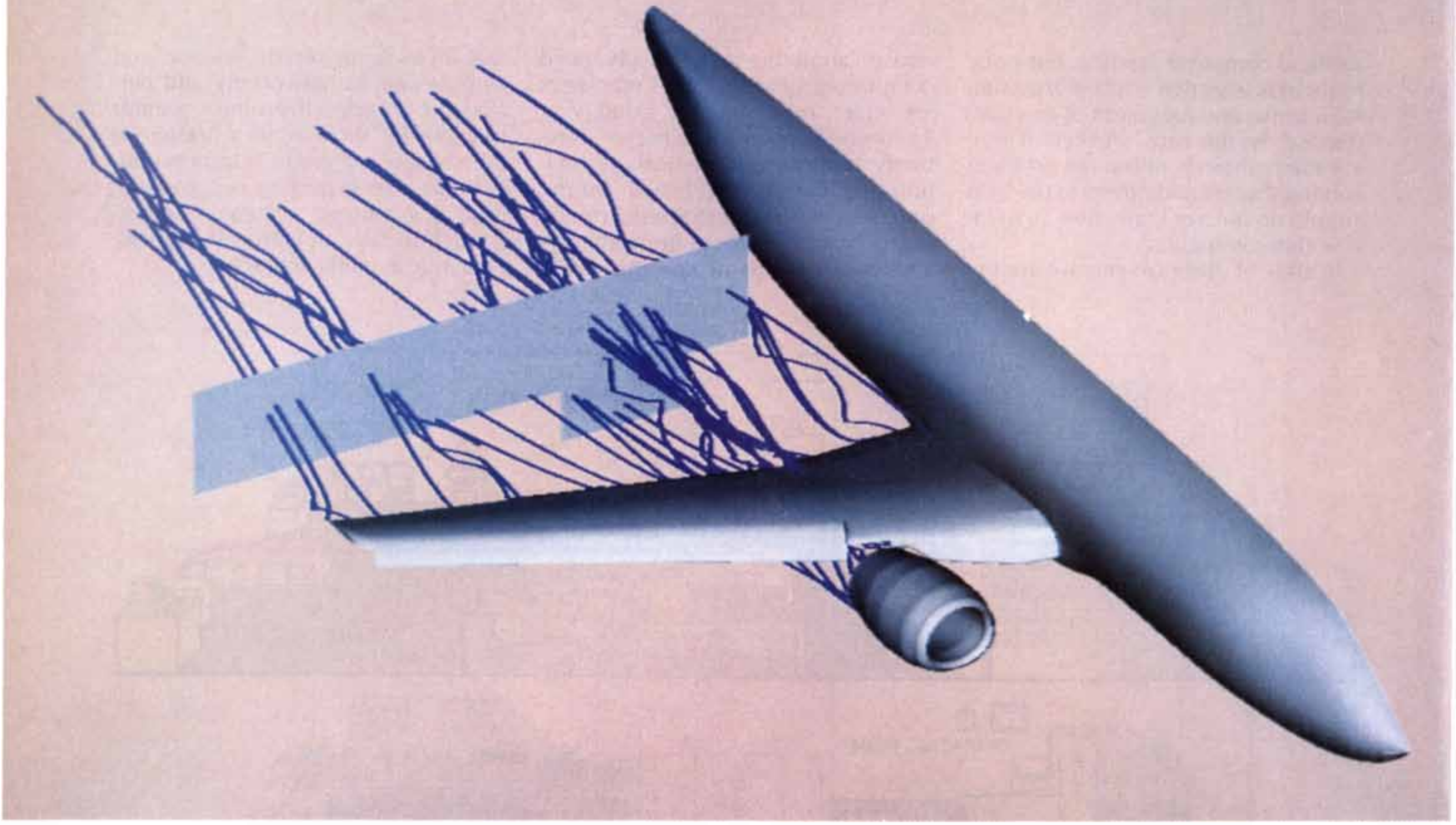
Given the expense of such a system, some institutions have been adopting digital imaging equipment piecemeal: buying a digital machine when a conventional machine needs to be replaced. Gradually enough digital equipment and workstations may be acquired for integration into a PACS network. This is just one example of how financial and logistical barriers can inhibit the dissemination of an available technology.

Computer systems do not now—nor will they soon—have a sufficiently complete understanding of medicine's technical, clinical and social considerations to approach the richness and flexibility of human expertise. The advice systems we have described do represent advances in medical computer science, but none of them is effective without the common sense and judgment of an experienced health-care worker. There are many aspects of human problem solving that we and others in the field simply do not yet know how to model within a computer.

In spite of these caveats we are optimistic about the future of advanced computing for medicine. Computers are more than number crunchers. Their ability to store, retrieve selectively and transmit medical information is now well established. On the horizon are systems that will provide reasoned advice about the diagnosis and management of specific cases. Advances in materials science and chip design, in networking and parallel processing all promise greater freedom for those of us working on the software of medical information systems. We expect to see growing use of advanced systems "smart" enough to take an active, if subordinate, role in medical practice.



HOSPITAL INFORMATION SYSTEM envisioned for the future will use a local-area network to transmit not only text but also images to workstations throughout a hospital. Patient information will be entered from—and transmitted to—the admitting office, record room, laboratories, operating rooms and bedside. Digitally stored results of X-ray, CAT-scan, ultrasound and other examinations will be transmitted by fiber-optic cables to wherever they are needed. Library material will be accessible too, and a wide-area network will bring in distant data bases and medical advice systems. Unlike current systems driven from a central computer, future systems will decentralize much of the memory and processing to individual workstations. A system like this would be a more comprehensive version of the HELP system that has already been developed at the LDS Hospital in Salt Lake City.



Advanced Computing for Manufacturing

Supercomputers may assume a major role in industry. They have already greatly influenced the design of such aerodynamically efficient products as airplanes and cars

by Albert M. Erisman and Kenneth W. Neves

By making it possible to process large quantities of arithmetic operations very quickly, the supercomputer has provided design engineers in industry with new insight into computational analysis. In so doing, the machine has opened the door to improved product performance and to major savings in the design and development phases of industrial products.

In spite of these successes, however, there are a great many problems associated with supercomputers, and the complete adoption of them by both the aircraft industry and manufacturing at large remains a serious challenge. For example, certain problems relevant to industry are too complex for even today's advanced computers to solve. In addition the many limitations of existing software have severely curtailed the value of the supercomputer to industry. Finally, the transition from computational analysis on the supercomputer to the full integration of computer-aided design (CAD) with computer-aided manufacturing (CAM) is largely an unsolved problem. Meeting these challenges is a goal that must be achieved. A closer look at airplane design will explain why.

The field of aerodynamics was one of the first to employ computers to solve computational problems. As early as 20 years ago aerospace engineers were using computers to model aerodynamic flow around an airfoil (the cross section of a wing) and, to a limited degree around the body of an airplane. Today supercomputers enable aerodynamicists to model entire planes and to dissect the various components of drag passing over the individual parts of the plane.

According to our colleague, Paul E. Rubbert of the Boeing Company, computational aerodynamics has revolutionized the aerodynamic desig.. process and the supercomputer now stands alongside the wind tunnel in its importance to the aircraft industry. Moreover, the complexity of computational research that is carried out worldwide in aerodynamics is among the largest of any scientific field.

The supercomputer has afforded aerodynamicists the power to examine details of aerodynamic-flow patterns that are impossible to measure in a traditional wind tunnel and infeasible using traditional computers. Moreover, complex data entered into a supercomputer can be converted into a three-dimensional graphics display at a workstation where it can be visualized from a variety of angles and perspectives. For this reason aerodynamicists at Boeing routinely carry out computational analyses of regions of an aircraft, including the body, the wing, the strut that holds the engine (called a pylon) and the casing around an engine (called a nacelle). Wind-tunnel testing may indicate the presence of drag, but the computer can dissect the individual components of drag and give precise information about their relative contributions. For the first time computational analysis can provide insight otherwise missed in physical experimentation.

Supercomputers played a critical role in the design of the Boeing 737-300, which entered service in 1984, 17 years after its predecessor, the very successful Boeing 737-200, was delivered to its first customer. Responding to the demand for more fuel-efficient aircraft, Boeing decided to modestly extend but otherwise retain the basic body design of the 737-200 and to replace the engines with newer, more efficient models. A technological problem presented itself, however: the new engines were much larger in diameter than the ones they were going to replace.

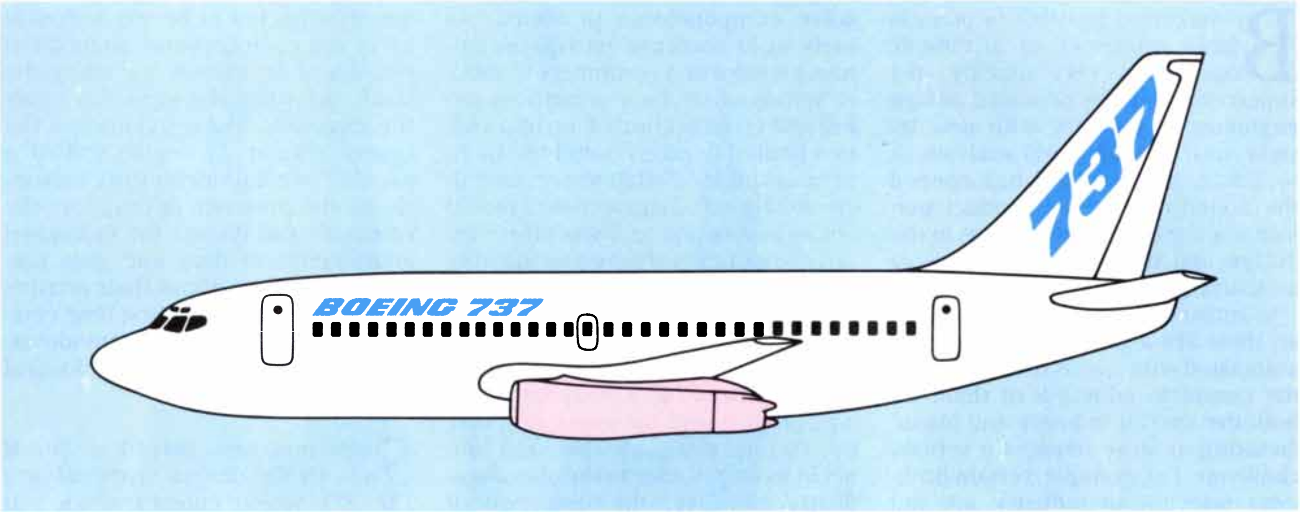
It was clear that the larger-diameter engines could not be attached to existing struts under the wings of the 737-200. To ensure adequate ground clearance with the engines in this position, the landing gear would have to be lengthened, adding excessive weight, requiring extensive structural modifications and adding greatly to the cost of the aircraft, a solution that was untenable. Boeing engineers reasoned that if the engines were mounted in front of the wing instead of under it, adequate ground clearance could be obtained without sig-

AIRFLOW OVER THE SURFACE of an airplane can be modeled on a supercomputer and displayed as a three-dimensional image. Particle traces emanating from both the engine exhaust and the trailing edge of the wing on a Boeing 737-300 are simulated here (*top*). Aerodynamicists can chart the exact path of particle flow by constructing an artificial planar surface at the rear of the wing to measure the coordinates of particles at various points along the airstream. The supercomputer can also model overall airflow and corresponding pressure levels affecting the plane (*bottom*). Here air pressure is modeled for the Boeing 767-200 during takeoff and landing (when the landing flaps are extended). Regions where air pressure is lowest (and airflow is highest) are indicated in red; areas of highest pressure (and lowest airflow) are indicated in purple.

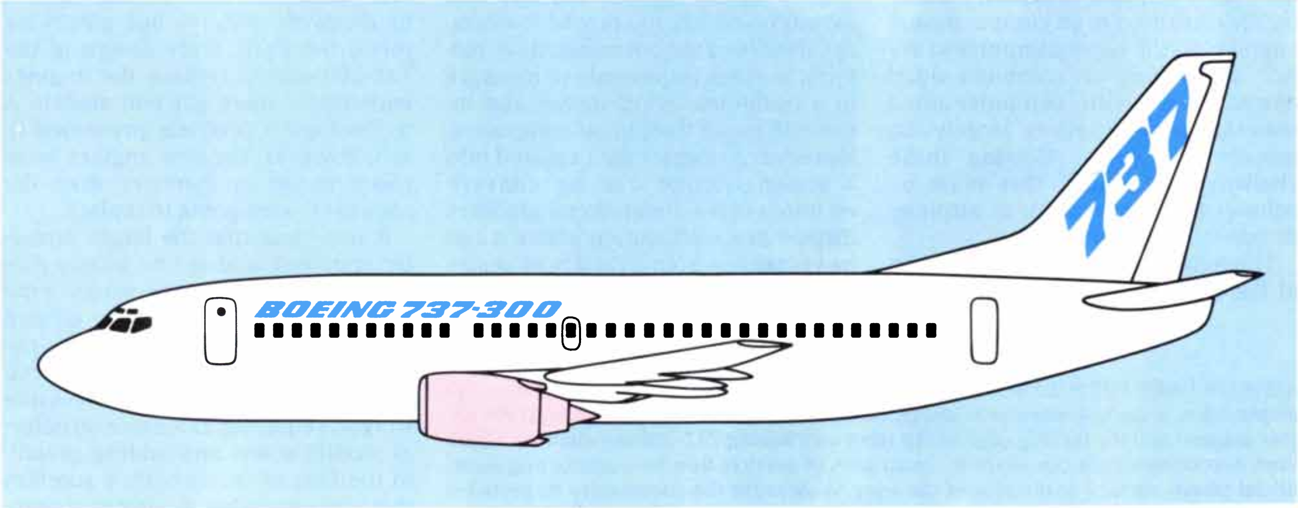
nificant additions of weight and cost. There was, however, one overriding obstacle to the plan: wind-tunnel testing had repeatedly shown that unacceptable levels of drag are created when the engines are mounted near the leading edge of the wing. Rubbert says that during the previous 20 years aerodynamicists had tried many times to develop such a closely coupled installation by modifying the design of the nacelle in the wind tunnel but had failed repeatedly.

As expected, computer analysis indicated that close-coupling the engine to the wing created excessive drag. Unlike the wind tunnel, however, computer analysis revealed the actual flow mechanism by which drag was created—a secret that had eluded discovery during two decades of wind-tunnel testing. By manipulating variables on the computer Boeing aerodynamicists could evaluate several alternative designs and at a greater level of detail than would be possible in the wind tunnel. Based on these computational studies they were able to devise an innovative solution to the placement of the engine and the design of the nacelle. By carefully modeling the upper part of the nacelle they were able to alter the design, making it asymmetrical and somewhat pear-shaped, a configuration that was shown to all but eliminate interference with normal airflow over the wing, resulting in a close-coupled engine with greatly reduced levels of drag.

An equally good example of the power that has made the supercomputer an important tool in aircraft design can be found in the development of the Boeing 757 airplane, delivered to its first customer in 1982. Because its development was scheduled about six months behind that of the wide-bodied 767, the company chose to incorporate identical cockpits in both. The reason for this was purely pragmatic: pilots are required to undergo a special certification program for each type of cockpit (and each new instrument panel) they fly, which usually means a sepa-



TWIN ENGINES THAT POWER the Boeing 737-200 are encased in small-diameter nacelles mounted under its wings. These engines are less fuel-efficient than those contained in larger-diameter nacelles because friction on the airflow is relatively high.



DEVELOPMENT OF LARGER, more fuel-efficient engines led to the redesign of the 737-200. The result is the 737-300, shown here with the engine mounted in front of the wing, a position that had been rejected because traditional wind-tunnel analysis revealed extremely high drag when engines were placed there. With the help of the supercomputer, aerodynamicists were able to dissect the different forces acting on an engine modeled in this position and redesign the nacelle to minimize drag.

rate program for each model of airplane. If the 757 and the 767 could be designed with the same cockpit, the pilots would have to undergo only one certification program to be qualified for both planes. This translates into considerable savings of money and time for the commercial airline companies. Boeing was faced with a major difficulty, however: the 757 is a narrow-bodied plane and smaller than the wide-bodied 767. How could the architecture of the modern-technology cockpit, designed first for the 767, be retained in the 757, given the different sizes of the two planes?

The flow of air around the plane at the point where the cab, which houses the cockpit, and body intersect is critical to performance; too much drag in that area can significantly increase fuel consumption and interior noise. Detailed aerodynamic analysis of the 757 (with a new cab wide enough to house the critical cockpit components of the 767) was considered essential to determine if and how such an airplane could be constructed.

Standard protocol called for building various mock-ups of these sections with different attachment sites and testing them in a wind tunnel to find the best (that is, the most aerodynamic) solution to the design problem, but time was limited and the company was faced with production deadlines. Instead these design alternatives were analyzed computationally on the supercomputer. In a matter of a few days (as opposed to the several months required for wind-tunnel analysis) a cab design was found that housed the cockpit of the 767 within the 757's narrow body while successfully meeting the criteria for size and airflow efficiency. In fact, Boeing engineers were sufficiently confident in their results that they recommended fabrication of the new plane begin prior to wind-tunnel verification of the results.

The Boeing company's newest airplane, the 7J7, still under development, opens a new era in jet propulsion. The jet engines of this plane are near the tail of the aircraft and drive twin sets of curved fan blades attached to the rear of the nacelle. The external blades are more efficient than the enclosed fans of ordinary jet engines and lead to considerable savings in fuel while providing the plane with speed capabilities comparable to existing planes. Its radical design would not have been possible without the supercomputer. Other aspects of airplane design that rely on complex mathematical modeling and the supercomputer include structural analysis of individual components, electromagnetic analysis and the design of control systems.

Virtually all the major automobile manufacturers are now turning to the supercomputer for help in the design of aerodynamic cars. The Ford Motor Company, for example, relied heavily on the supercomputer for the design of its 1986 Taurus car. In addition to fuel savings for the consumer, the supercomputer is estimated to have saved Ford several million dollars in design costs, primarily because it eliminated the need for multiple prototypes. Car manufacturers also use supercomputers extensively in crash simulations. By modeling collision damage on a computer screen they are able to save vast sums of money that would otherwise be spent on the destruction of costly prototypes.

The petrochemical industry is perhaps the foremost example of industrial use of the supercomputer; since the late 1970's it has relied heavily on the supercomputer for both oil exploration and recovery. Seismic monitoring tapes can be entered into the computer and used to identify specific geologic features associated with oil reserves. This enables petrochemical engineers to identify profitable drill sites with a significantly greater degree of accuracy than is possible using traditional methods, thereby increasing the probability of striking a major deposit. Because the cost of drilling is so high (up to $15 million per well on land and considerably more than that offshore), the costs of computing are relatively insignificant compared with the benefits. Furthermore, supercomputers play a role in oil extraction; by modeling the fluid dynamics of steam and chemical extraction they enhance recovery rates.

Supercomputers have a number of applications and can generally be applied to any problem where complex mathematical models can provide new insight into product design or productivity enhancement. They have been used to model molecular structures and radar patterns as well as to create film images (scenes in the motion picture *The Last Star Fighter*, for example, were generated entirely on a supercomputer). They are also employed extensively in flight simulation and pilot training programs, and most recently they have been adopted by large financial companies that utilize them to analyze investment portfolios.

In construction work, supercomputers are used to model stress and earthquake resilience on such major projects as bridges, nuclear-reactor containment vessels, large spectator arenas (such as the Calgary Saddledome in Alberta) and skyscrapers. They have also begun to change the face of weather forecasting: detailed data on climatic and geographic conditions can be combined with actual storm readings and entered into the computer to produce a three-dimensional picture of a moving storm. Meteorologists can predict the path of the storm and issue early warnings to the population areas most likely to be affected by it with much greater accuracy than was possible before the supercomputer.

We have described some of the successes associated with the use of the supercomputer in manufacturing and other industries. These successes have not come easily nor are they as extensive as they might be. In order to understand why that is so, the limitations of the supercomputer must be addressed.

The term supercomputer is used in two ways: to refer to those computers that have the greatest raw power, measured in terms of speed and memory size, or to refer to those that are defined by their inner architecture, that is, the hardware components on which they are built. In this article we use the term supercomputer to refer to computers in the first category: those that have maximum power capability. Such computers usually have an advanced architecture as well; computers in the second category—machines that are somewhat less powerful but nonetheless are architecturally advanced—we shall call near-supercomputers.

Because supercomputers were developed in order to solve numerically complex problems, the speed at which they operate is considered a crucial measure of their performance. A prototype of the first supercomputer, made in the early 1970's, was the Control Data Corporation's CDC 7600 model, which operated at a speed of five million floating-point operations per second, or five megaflops (MFLOP's). Today supercomputers operate 200 times faster than that, at speeds of approximately one billion floating-point operations per second, or one gigaflop (GFLOP). Near-supercomputers are considerably slower but still have impressive performance peaks on the order of from 20 to 100 MFLOP'S.

A problem that might take a year to complete on a traditional mainframe can be solved in slightly less than an hour on a supercomputer. This breakthrough in computing power is made possible by the unique vector and parallel architecture that forms the computer's hardware. The concepts of vector and parallel architecture are relatively new to the field of computer science, but they are based on two approaches: pipelined arithmetic units and parallelism (that is, the replication of computational units). [For a more detailed discussion see "Advanced Computer Architectures," by Geoffrey C. Fox and Paul C. Messina, page 66.] Both approaches can make substantive increases in computer speed possible but yield poor results when they are improperly matched by software.

A drawback to vector architecture, which operates much like a conveyor belt or an assembly line, is that the average speed at which operations move through the assembly line depends on the length of time needed to complete the first computation entering the system and the number of computations that can be processed in the pipeline. This is a subtle but important point. Computer programs that minimize the total number of calculations necessary for an analytical problem but do not have a long stream of computations to perform (called long vectors) sometimes fall short of performance expectations.

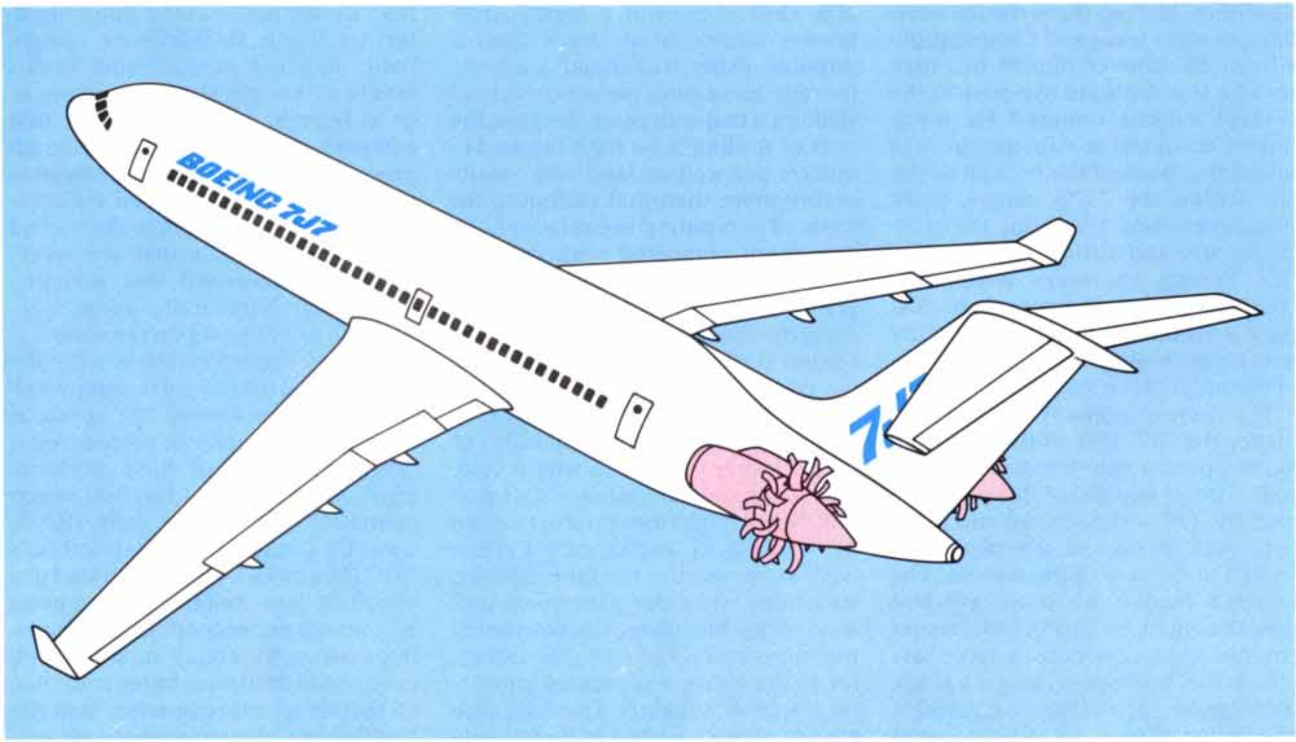
In theory parallel processors, in which a number of interconnected computers attack a problem simultaneously, have unlimited potential for speed. That is, *n* computers can solve a problem in one *n*th of the time it would take a single computer. But like the "million-monkeys principle" in business management (a project that normally would take one person 10 years to complete cannot be done by one million people in one minute), parallel hardware has also fallen short of performance expectations.

Both vector and parallel architectures were introduced as a means by which significant breakthroughs in computational power could be achieved. Evaluating a supercomputer's performance solely in terms of operations per second is much like measuring the quality of different sports cars solely on the basis of maximum speed, without taking into account factors such as acceleration, cornering and the conditions under which they can be driven.

Computer performance is based on a range of factors, including memory size, the ability to handle small arrays of numbers (known as small vector size), the speed at which discrete quantities of data can be processed (known as the machine's scalar performance) and the capabilities of the software with which it is compatible. Computers that perform well at the high end of the scale (they can process large amounts of data very quickly) do not always perform well at the low end of the scale when small discrete quantities of data must be processed. Unfortunately most applications have both types of computing requirements.

Speed, then, is not the sole determinant of a computer's performance capabilities. Computer performance is also dependent on two other factors: memory size and the speed with which data flow from the computer's memory to the computational units inside the machine.

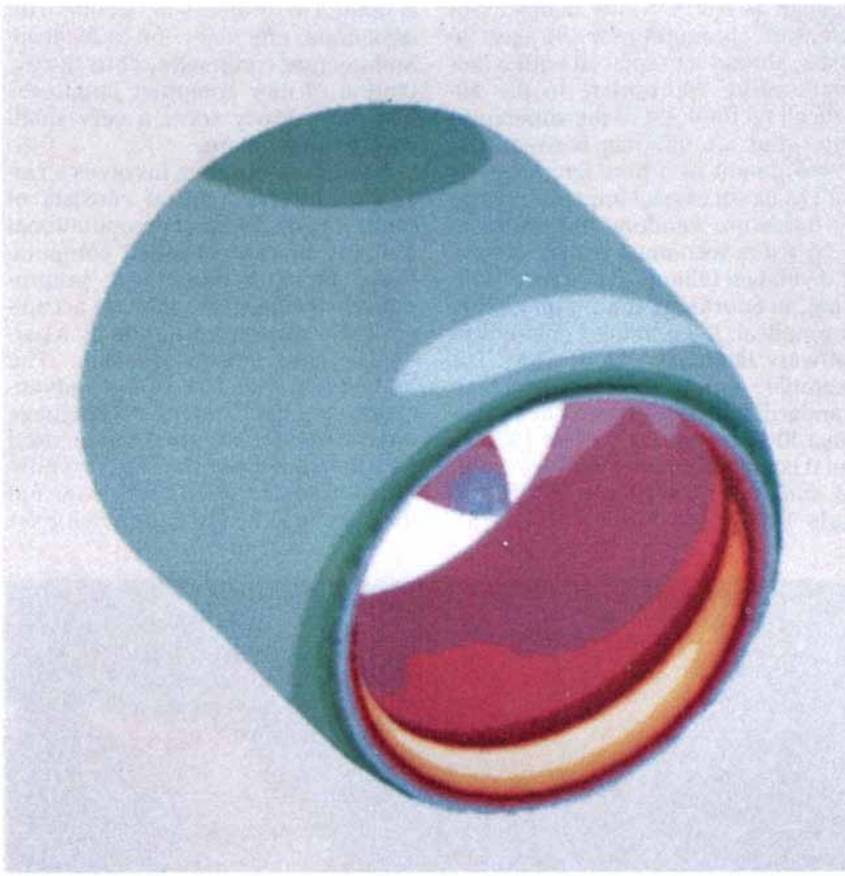
Memory is a major factor in determining the performance of a computer, and the speed at which a calculation is carried out is frequently dependent on the size of a given machine's memory banks. Bottlenecks often result from the sheer volume of data that must be manipulated. Indeed, moving large quantities of data from secondary storage to



SCHEMATIC drawing of a new plane, the Boeing 7J7, represents a new era in jet propulsion. The plane's engines, mounted near the tail, drive exposed sets of fan blades, which are more efficient than the enclosed fans of conventional jet engines and result in sizable fuel savings. The design would not have been possible without supercomputer analysis of many variables.

memory and back again may be the most time-consuming aspect of any problem in computation. Although supercomputer memories are growing at a rate comparable to computational power, in some cases this is still inadequate. The CDC 7600 of the early 1970's had a memory capacity of 500,000 15-digit numbers. The Cray-2, built almost 15 years later, has a capacity 512 times greater: it can hold 256 million 15-digit numbers in its memory. But as calculations become increasingly complex, even a memory size of several billion words may be insufficient for future computational needs. In order to tap a memory bank this large, extensive changes in software design may be necessary.

Matching the capabilities of a supercomputer to the needs of a particular user is very important. For example, the peak speed of a supercomputer is of little relevance to an aerodynamicist measuring stress on a wing; what matters to that user is the speed with which a complicated mathematical model of the wing can be created and put through a series of stress tests. Therefore the speed with which a computer can solve a complex problem involving many variables may be more important than how fast it can process a carefully constructed array of numbers. This is much like assessing the performance of a sports car over a winding mountain road compared with its peak speed on a straight, flat stretch of highway.



ISOLATED NACELLE is modeled in three dimensions on the supercomputer. The colors represent different velocity levels of air flowing in and around the nacelle during flight. The speed of the airflow is highest at the inside leading edge of the nacelle (*yellow*), where air is channeled through the casing and into the engine of the plane. It is lowest toward the rear of the nacelle (*purple*) and intermediate on the outside (*green*).

In either case, the sheer number of calculations produced on a supercomputer can be staggering. Standard computers can provide numerical values for 50 separate points on a grid sheet (measured at 20 different points in time), but they are dwarfed by supercomputers that can generate data for 50,000 grid points at 200 points in time. Fortunately, coincident with the supercomputer, the computer industry has developed advanced workstation capability to convert data into graphic images; without that capability it would be virtually impossible to decipher the quantity of data produced. Moreover, a time sequence of graphic designs can be put together and viewed in motion-picture form that visually displays data as a realistic simulation of a physical process. The flow of air particles over a wing, for example, can be seen as a continuous stream, enabling aerodynamicists to visualize the precise forces acting on it. Equally important, they can change the design on the screen by manipulating variables at the workstation and thereby pinpoint specific components of turbulence.

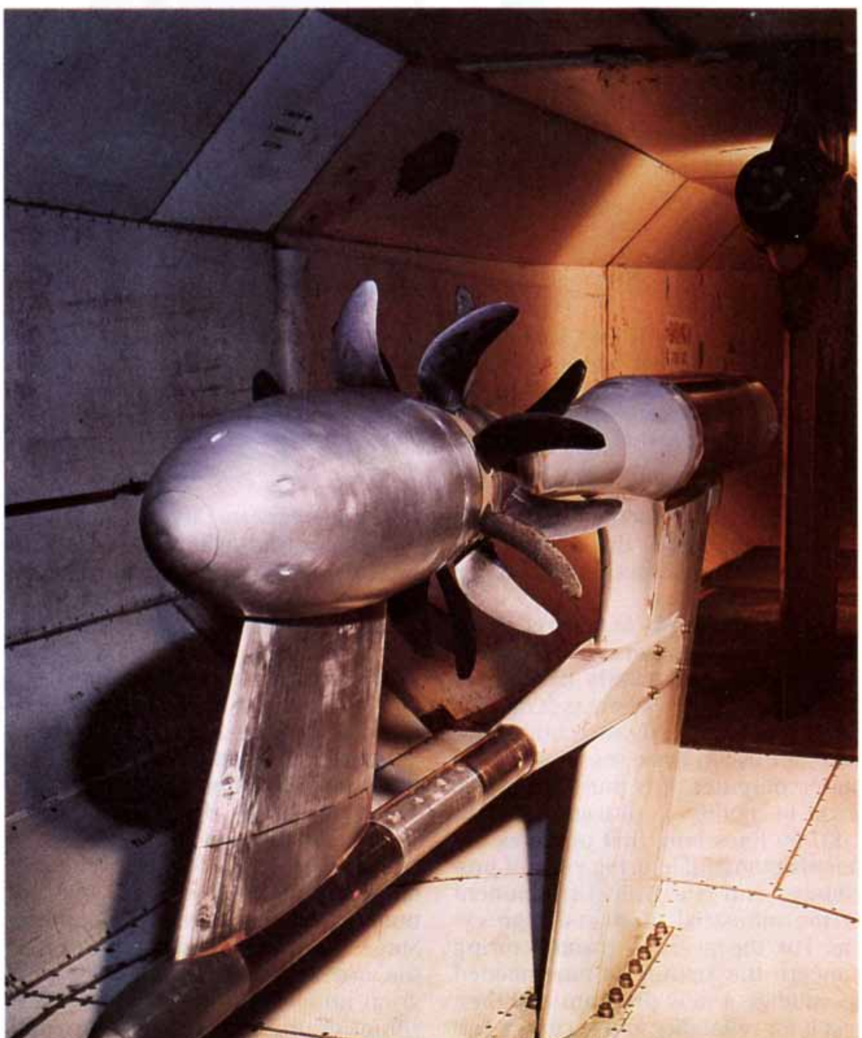
The effectiveness of supercomputers, like all computers, is dependent on software, particularly in the area of manufacturing. Whereas it may be fairly easy to modify or rebuild software for use in basic research on the supercomputer, it is much more difficult to modify a software program 100,000 lines long that operates under strict manufacturing control procedures and is a critical component of the industrial product-design cycle. For the average manufacturing concern the amount of time needed to validate a new program and then test it for reliability and accuracy can be prohibitive. It is therefore desirable that computer software be capable of making the transition from one type of computer architecture to another. Realistically, however, that is not always possible. Consequently, just as factories have been bulldozed to make room for automation and modernization, some antiquated software will have to be bulldozed to make room for a new generation of programs in keeping with the new generation of computers.

Attempts have been made to design tools to support the transition of software from traditional to advanced-generation computers, but these have had limited success. One such tool is the "smart compiler," which can automatically vectorize codes, that is, transform high-level language instructions into instructions specific for vector computers. Since compilers can only rearrange the order of a given software program rather than identify new algorithms or problem-solving approaches more compatible with the supercomputer, this approach is limited.

The only answer to the software problem, according to many computer engineers, is to write software programs in a new language specifically designed for vector or parallel machines. More traditional languages such as FORTRAN, the mainstay of scientific computing for the past 20 years, should be replaced with a language more appropriate to the advanced technology of the supercomputer. But the time lag between the development of a new language, as well as its successful implementation by hardware vendors and ultimately by users who must rewrite, verify and validate billions of dollars of software, is enormous if not completely impractical. Even limited changes in software language can be slow. For example, a new FORTRAN language standard, FORTRAN 8X, was begun almost 10 years ago (in the late 1970's), but it is unlikely that it will be accepted and widely available before the early 1990's. Even then FORTRAN 8X is limited in its ability to operate with maximum efficiency on advanced-architecture computers. Thus the evolution of new computer languages will only slowly solve a very small part of the problem.

Another alternative involves a category of software that consists of programs made up of computational building blocks. So-called computational building blocks are preprogrammed algorithms that are accepted as the standard program by a particular community of users. The building blocks have three advantages: they have a standard language interface (therefore they can be used with any machine); they are carefully tuned to hardware architecture, and they are specifically designed for the most commonly required computations such as matrix operations, Fourier transforms and sorting—the type of operations that form the backbone of many complex analytical problems. Because a major chunk of computing time is often spent on these kinds of calculations, building blocks (which are relatively cheap to construct) can lead to tremendous savings of time and money. Boeing's two building-block software "libraries," BCSLIB Mathematical/Statistical Library and Vectorpak Subroutine Library, are used hundreds of millions of times per year as part of operations on Boeing's Cray X-MP supercomputer and other advanced, industrial computers.

Building blocks have an advantage beyond their ability to handle complex computations: they allow one to assess the feasibility of a particular approach to modeling prior to the development of a program necessary for its application. By comparing a number of different approaches to the same problem prior to their execution, they save the user considerable amounts of time and money.



HIGH-SPEED WIND-TUNNEL modeling of aerodynamic components, although largely complemented by computational models generated on a computer, still plays an important role in the final stages of design. This unducted-fan engine (encased in a nacelle), which drives counterrotating fan blades, is mounted on a strut, where it can be tested at air speeds as high as 600 miles per hour. Probes are placed in the tunnel during testing to measure the velocity of air over the nacelle; the probes extending from the ceiling have microphones that monitor noise levels generated by the fan blades.

In spite of the power and tremendous growth of supercomputers, there remains yet another hurdle to their widespread adoption: the difficulty of integrating three components, namely the user, the graphics workstation and the supercomputer itself. Traditional computers are normally connected to their users by telephone lines that accommodate data exchange at a rate of 1,200 baud, also known as bits per second. But for supercomputers that are capable of handling enormous volumes of data, communicating at a rate of only 1,200 baud is a little like trying to drain the ocean through a straw; large quantities of data simply cannot be transmitted over long-distance wires very quickly.

Understanding the limits this imposes is critical to supercomputing: an engineer might enter data that take only five minutes to process on the supercomputer (as opposed to 12 hours on a traditional computer) but be hindered by a 12-hour delay in obtaining the output. Long-distance lines now have the capacity to process 56 kilobits of information per second; even at that speed, however, they are too slow for many problems run on supercomputers. The creation of motion-picture sequences (depicting particle flow over the wing of an airplane, for example) is even more time-consuming: instead of analyzing a single set of variables just once, the supercomputer must be instructed to process a series of computations and then output a series of graphic illustrations—all of which add to the flow of data over the network. For these reasons some users have eliminated telephone-line communication, instead finding that it is faster to send and receive data by an air-express courier service. Although not ideal, it is a solution that is becoming increasingly commonplace.

Another solution has been adopted by Alabama Supercomputer Network Services in Huntsville, scheduled to begin operations in January, 1988. The supercomputer there is expected to transmit information between campuses in the University of Alabama system at a speed of 1.5 million bits per second. On a larger scale, such as the continental U.S., a system of that type would be prohibitively expensive, but because the Alabama computer network consists of only a small number of sites, the operation is expected to work well.

The National Aeronautics and Space Administration's Ames Research Center at Moffett Field, Calif. (the site of the world's most powerful supercomputer), has taken an alternative approach to the problem of long-distance communication—a solution that works for some of their research problems. Officials there have placed powerful workstations in close proximity to the mainframe supercomputer, completely eliminating the need for long-distance lines. The headquarters of the Boeing Computer Services Company in Bellevue, Wash., has a similar arrangement. A corps of computational aerodynamicists is connected to the Cray X-MP by a series of graphics workstations housed in a laboratory not far from the supercomputer. This arrangement has reduced to less than several hundred feet the distance over which data must be transmitted.

A second step toward integration focuses on the workstation as a means of enhancing productivity rather than on the supercomputer. This is a "user-friendly" arrangement in which the scientist or engineer works at a personal computer or workstation using standard applications software. All communications with the supercomputer are handled directly by the workstation, utilizing software that communicates with the user in his or her application language. Data are entered directly into the supercomputer, where they are converted by the computer into vectorized codes; the person at the terminal is not constrained by high-level language. The fact that the analysis is done on a supercomputer is obvious only from the complexity of the results.

For most applications, user-friendly software is hampered by the slowness of network communication and the limited power of the personal computer. Boeing has implemented such a service, called The Petroleum Gallery Services, as part of an integrated suite of products designed for the petroleum industry. The Petroleum Gallery Services offers a single source of data, applications and computer power for geoscientists mapping and analyzing reservoirs.

In spite of these limited advances in integration, the true potential of supercomputing in manufacturing will be realized only by integrating computational analysis, at which supercomputers excel, with the manufacturing process where they have so far had little direct impact. There are pressing reasons why the integration of product-design analysis with manufacturing is important.

Information critical to the manufacturing process, such as design tolerances, manufacturing cost tradeoffs and long-term reliability concerns, is rarely taken into consideration during the computer analysis of a given design. By integrating manufacturing needs such as these with the initial analysis of a product design, compromises can be made that take both performance and manufacturing requirements into account.

Integration has not yet occurred for a number of reasons. Supercomputers are not very old: they made the transition from laboratory to industry a little less than 10 years ago and are therefore too recent to have had any impact on many products currently rolling off assembly lines. The design process mandates that the CAD/CAM relationship remain stable over the often lengthy design period. There must be strict configuration control of software, hardware and data over many years. Rapid changes in advanced computer architectures and the associated software changes are not compatible with this stability requirement. A product such as a commercial airplane, for example, undergoes a long transition between its original formulation and its final production. Because its components are highly interdependent, a change in the design of one component is likely to elicit change in others. Therefore various stages of the design process are "frozen" at certain points; further changes are not possible and the entire CAD/CAM process is "set in stone."

This may change as more complex models still in the developmental stage move into production, but a second concern needs to be addressed, namely that in spite of the tremendous benefits of CAD/CAM integration, the two components are not always compatible with each other. For example, aerodynamic modeling may stipulate that a wing have a certain shape for maximum efficiency, yet that design may clash with manufacturing requirements. The metal simply may not bend in accordance with the supercomputer design, or it may be too heavy or too brittle for a given design.

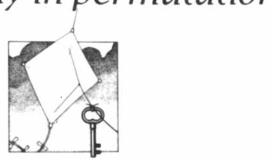
Finally, computational models for detailed aerodynamic-flow analysis strain the capabilities of even today's supercomputers; processing the enormously complex computations, involving various uncertainty and tolerance parameters, that are necessary for complete CAD/CAM integration is still beyond the scope of existing computers. In view of these limits, supercomputers have not been adopted for certain computational problems; however, this overlooks one of the better attributes of the supercomputer—the speed with which data can be manipulated.

Ultimately supercomputers may become integrated with artificial-intelligence computation, which is currently dominated by symbolic rather than numerical computing and is often done on specialized computers. In theory, artificial intelligence can someday be used to gain further insight into the large, numerically complex data sets currently processed by supercomputers.

It is safe to say that supercomputing is now in a transition period with change taking place simultaneously in many directions: power, architecture, analytical capabilities, competitive requirements and artificial intelligence. Supercomputers have revolutionized the aircraft industry to such an extent that the quest for faster and more powerful computers continues unabated. The next five years will be particularly challenging—not only because the opportunities for technological breakthroughs are many but also because in this era of stiff competition from abroad, technology in the U.S. may depend on the supercomputer.

THE AMATEUR SCIENTIST

Now there is Rubik's Magic, a new puzzle that provides a study in permutation operators



by Jearl Walker

Ernő Rubik, famous as the inventor of Rubik's Cube, recently produced a new and equally enchanting puzzle called Rubik's Magic. It consists of eight plastic squares aligned in a flat two-by-four array. The squares are connected by a nylon thread that runs along diagonal grooves in the plastic. The squares can be rotated about their edges.

One side of the puzzle displays three separated rings. On the back side the rings are in disarray. The challenge is to realign the squares and thereby form the rings on the back side. If you succeed, the rings interlock. The puzzle is then an incomplete three-by-three array missing one corner square.

One can go about solving the puzzle by random manipulation of the squares, flipping them in various ways as allowed by the thread hinges. A systematic study offers a better approach and displays several subtle mathematical properties of the puzzle. Recently Wolfgang Glebe presented such a study in *Spektrum der Wissenschaft,* the German edition of *Scientific American.* Here I repeat his analysis and give a solution to the puzzle that was found by one of his readers. It may be the shortest solution in the sense that it requires the fewest manipulations of the squares.

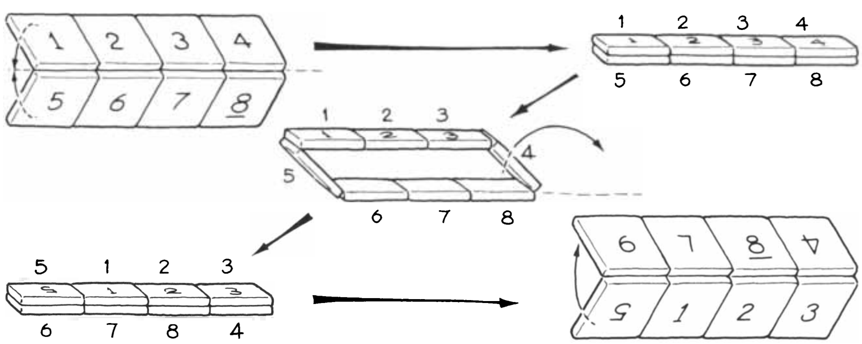
Turn the disarrayed side face up, with the long edge horizontal and the copyright symbol in the upper row. Attach a small label to each square. Beginning with the upper left square and moving to the right, number the labels. This arrangement of the squares and numbers is said to be state 0, symbolized as Z_0. In this state the squares are in a certain location with respect to one another and the numbers are in a certain orientation.

How many other unique states can the puzzle have when it is in a two-by-four array? You can change the locations of the squares while also reorienting the number on each square. Arrangements that amount to a simple rotation of the entire puzzle should not be considered unique. Mirror-image arrangements should also be excluded. Nevertheless, a calculation suggests there should be about 1.3 billion unique states.

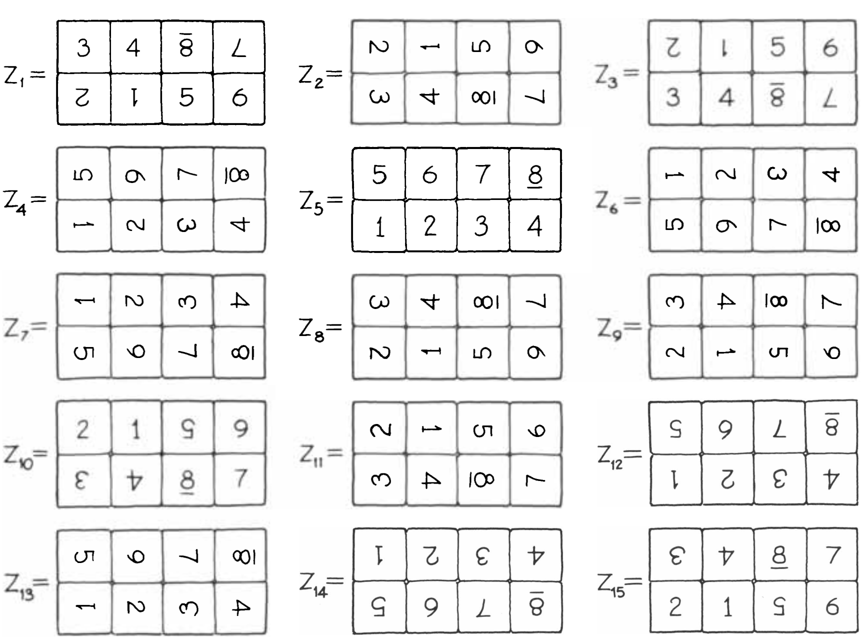
Glebe found that the calculation is misleading: there are only 32 unique states. The reduction in number results from the way the squares are attached by the thread hinges. A square is always attached to the same two neighboring ones, although their relative orientations can change. For example, square 1 is always attached to squares 2 and 5 and square 2 is always attached to squares 1 and 3.

The states can be categorized as either symmetrical or asymmetrical. In a symmetrical state, of which there are 16, the numbers 1 through 4 are lined up in a horizontal row or are grouped in a square on the left or the right side of the two-by-four array. Any other arrangement is said to be an asymmetrical state.

Begin with the puzzle in Z_0. How can you manipulate it into the other



Operation R *changes the symmetry of a state*



The symmetrical states other than Z_0

31 possible states? One transformation entails what Glebe calls operation *R*. Fold the puzzle away from you and around the horizontal center line. Make the upper row end up as the top layer in the resulting double layer. Now "roll" the double layer by moving the top layer to the right and the bottom layer to the left, each by one square. Next unfold the puzzle by rotating the bottom layer about the back edge of the top layer. Notice that it unfolds in only one direction.

Operation *R* can be employed with any state of the puzzle. The success of the roll depends, however, on which way the puzzle is initially folded. Sometimes you must fold it toward you for the layers to roll. The direction in which you finally unfold the puzzle also varies according to the initial state. The direction of the roll can also be varied by moving the top layer to the left and the bottom layer to the right.

Operation *R* transforms the puzzle from the symmetrical state Z_0 to an asymmetrical state with 6, 7, 8 and 4 (the 4 is inverted) on the upper row and 5 (inverted), 1, 2 and 3 on the lower row. Note that the companion squares are all still attached. From this state perform operation *R* again. The puzzle returns to Z_0. The rearrangement of the squares through an operation such as *R* is called a permutation. When an operation done twice returns the puzzle to its former state, it is said to be involutory. Symbolized, RR is equal to i, where i represents the identical state.

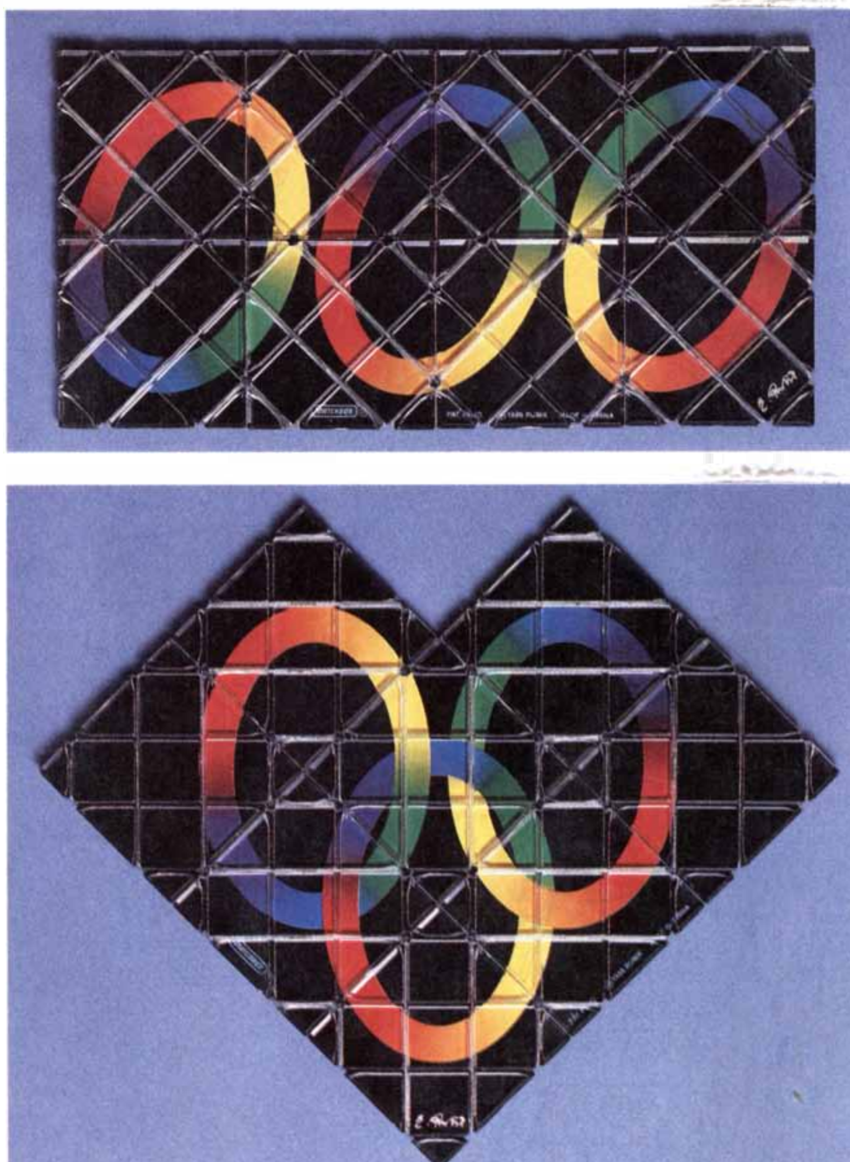
The next task is to find ways of transforming Z_0 into the other 15 symmetrical states. Glebe designates one method as operation *D*. Here again the first step is to fold the rows around the horizontal center line, either backward, which is appropriate when the initial state is Z_0, or forward. The success of the second step depends on the choice of folding direction. If the puzzle cannot be made to take the second step, return to the initial arrangement and do the folding the opposite way.

In the second step you expand the puzzle into a loop and then push in the left and right sides, allowing the top and bottom pieces to collapse onto them. The next step requires the stack to be opened. Depending on the state of the puzzle when you begin the operation, you may need to turn the stack over at this point. You cannot go wrong. If the stack cannot be opened, turn it over and try again. Finally, open the puzzle fully by folding the side flaps outward.

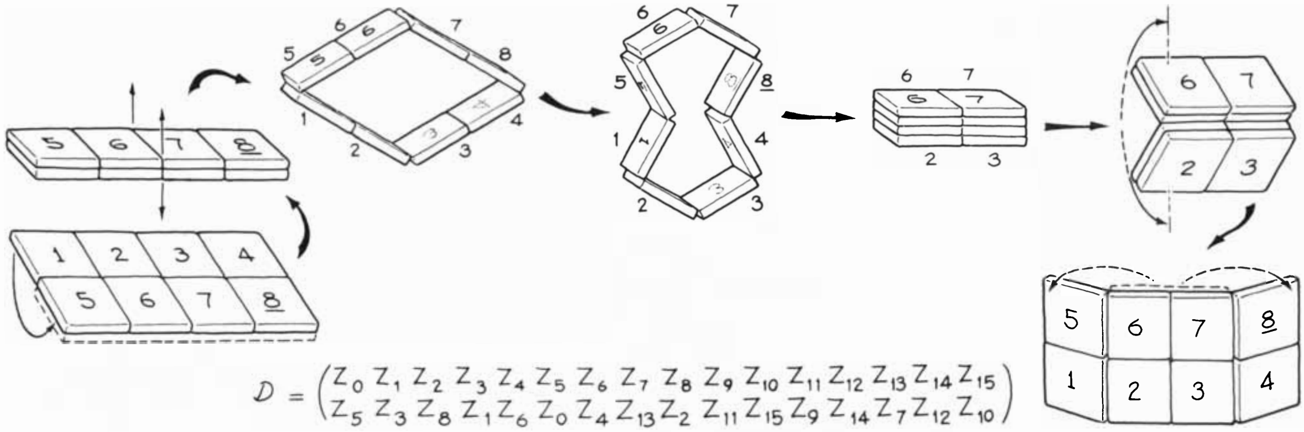
Operation *D* transforms Z_0 to Z_5, another symmetrical state in which the rows are interchanged with no reorientation of the numbers on the squares. This operation is also involutory. If you do operation *D* on Z_5 to return to Z_0, you will find that the folding and unfolding usually go in directions opposite to those you employed previously. The table in the top illustration on the next page indicates how operation *D* alters the other symmetrical states. For example, it transforms Z_1 into Z_3 and vice versa. A concise way of writing the transformation is $D(Z_1) = Z_3$.

Operation *E* is similar to *D*. You again fold the rows backward onto themselves and then pull the layers apart to form a loop. This time you make the top and bottom of the loop concave and then push the left and right sides inward to collapse on the top and bottom pieces. The next step requires that the group of squares be unfolded. For some initial states the step is immediately successful. For others you must turn the puzzle about a vertical axis for the step to work. You cannot go wrong here because the puzzle cannot be made to unfold erroneously.

Next reach behind the plane and unfold the top and bottom flaps. The final product is Z_3, another symmetrical state. Thus $E(Z_0)$ is equal to Z_3. If you transform Z_3 back to Z_0 with operation *E*, the first fold is opposite to the one shown in the second illustration from the top on the next page.

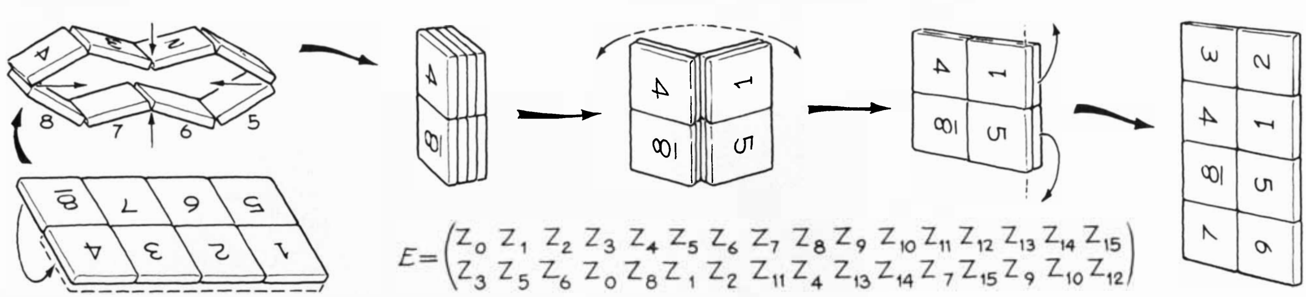


Rubik's Magic in its initial* (top) *and solved states



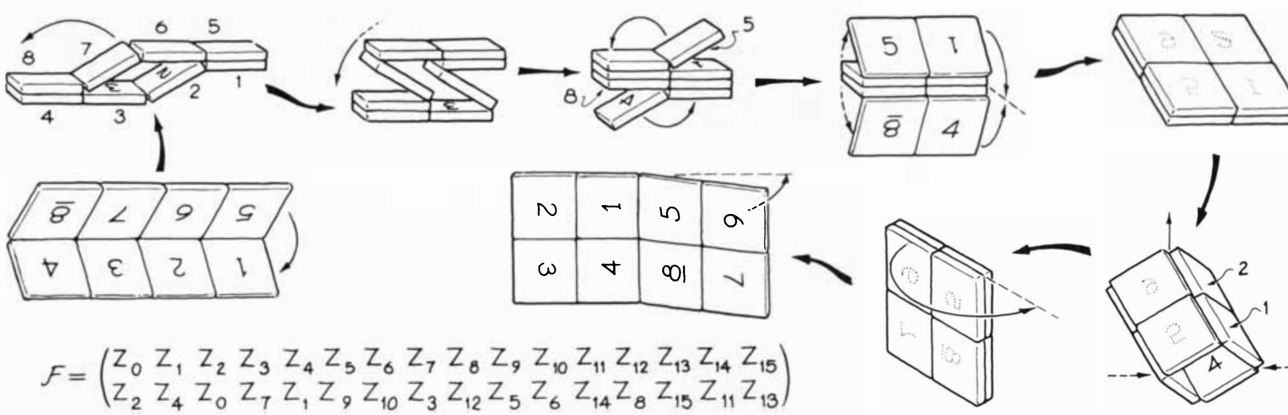
$$\mathcal{D} = \begin{pmatrix} Z_0 & Z_1 & Z_2 & Z_3 & Z_4 & Z_5 & Z_6 & Z_7 & Z_8 & Z_9 & Z_{10} & Z_{11} & Z_{12} & Z_{13} & Z_{14} & Z_{15} \\ Z_5 & Z_3 & Z_8 & Z_1 & Z_6 & Z_0 & Z_4 & Z_{13} & Z_2 & Z_{11} & Z_{15} & Z_9 & Z_{14} & Z_7 & Z_{12} & Z_{10} \end{pmatrix}$$

Operation D *interchanges rows*



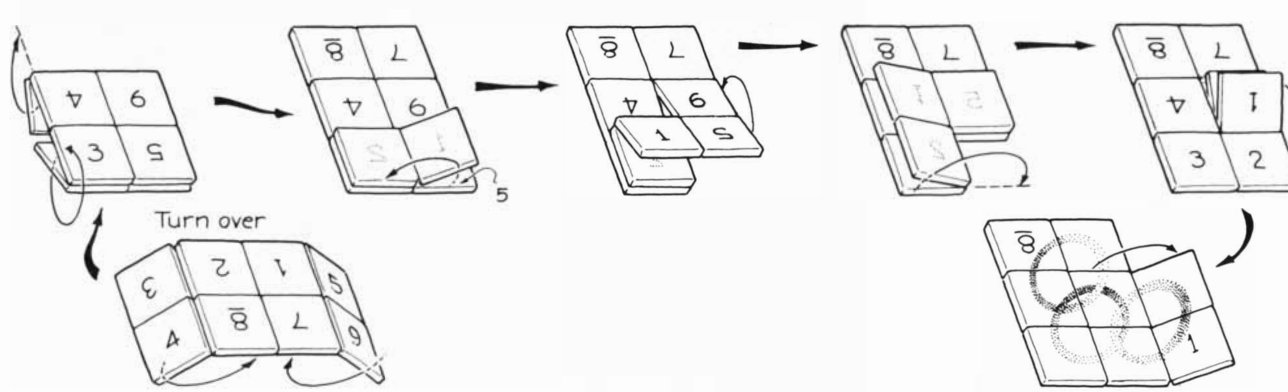
$$E = \begin{pmatrix} Z_0 & Z_1 & Z_2 & Z_3 & Z_4 & Z_5 & Z_6 & Z_7 & Z_8 & Z_9 & Z_{10} & Z_{11} & Z_{12} & Z_{13} & Z_{14} & Z_{15} \\ Z_3 & Z_5 & Z_6 & Z_0 & Z_8 & Z_1 & Z_2 & Z_{11} & Z_4 & Z_{13} & Z_{14} & Z_7 & Z_{15} & Z_9 & Z_{10} & Z_{12} \end{pmatrix}$$

Operation E



$$\mathcal{F} = \begin{pmatrix} Z_0 & Z_1 & Z_2 & Z_3 & Z_4 & Z_5 & Z_6 & Z_7 & Z_8 & Z_9 & Z_{10} & Z_{11} & Z_{12} & Z_{13} & Z_{14} & Z_{15} \\ Z_2 & Z_4 & Z_0 & Z_7 & Z_1 & Z_9 & Z_{10} & Z_3 & Z_{12} & Z_5 & Z_6 & Z_{14} & Z_8 & Z_{15} & Z_{11} & Z_{13} \end{pmatrix}$$

Operation F *transforms symmetrical states into other symmetrical states*



A solution to Rubik's Magic

The transformations of the other symmetrical states resulting from operation E are listed at the bottom of the illustration.

Operation F is quite different from D and E [*see third illustration from top on opposite page*]. Starting from Z_0, fold the rows forward. When you try to pull the double layer into a loop, the far right and left sides resist. Pull the right side over the left side to form a stack. Then flip the top right square over the top left square while also flipping the bottom left square under the bottom right square. Turn the stack so that you can unfold the top and bottom layers. If the pieces resist, turn the stack and try again. The top and bottom layers can be unfolded in only one direction.

Insert fingers into the near and far sides to separate the two layers of squares. The puzzle may resist to some extent until the threads adjust themselves. Guide the separation with your fingers. When the layers have been separated and the puzzle is tubular, push in the left and right sides of the tube to create two vertical layers. Unfold the layers by opening the near or the far end as is appropriate. Again the strings may resist slightly; you may need to guide the separation with your fingers.

F transforms Z_0 into Z_2, another symmetrical state. The transformation is involutory. If you transform Z_2 back to Z_0, the directions of folding and unfolding are the same as before.

So far the operations have produced Z_2, Z_3 and Z_5, each of which can be turned into asymmetrical states by R. You can produce the rest of the symmetrical states from Z_0 by employing two or more of the operations in succession. Try the permutation combination DE. (The notation implies that E is done first.) The result is Z_1. Since the permutation ED produces the same state, D and E are said to be commutative, that is, their order is unimportant.

F is a more powerful operation than D and E because it is not commutative. For example, $DF(Z_0)$ yields Z_8, whereas $FD(Z_0)$ yields Z_9. The table on this page indicates how all the symmetrical states can be obtained with various combinations of D, E and F operating on Z_0. Each product state can be turned into a corresponding asymmetrical state by R.

Suppose a series of operations is performed on one of the symmetrical states other than Z_0. Is the product a new state? It is not, as can be seen in one of Glebe's examples. Consider $DFEF(Z_{14})$. The first operation, F, produces Z_{11}, as determined from the table in the illustration of F: $DFEF(Z_{14})$ is equal to $DFE(Z_{11})$. Continue the procedure: $DFE(Z_{11}) = DF(Z_7) = D(Z_3) = Z_1$, which is a known state. Since no new states can be produced with any combination of operations D, E and F, the permutations listed in the table are said to be closed.

$$i(Z_0) = Z_0$$
$$DE(Z_0) = Z_1$$
$$F(Z_0) = Z_2$$
$$E(Z_0) = Z_3$$
$$DEF(Z_0) = Z_4$$
$$D(Z_0) = Z_5$$
$$EF(Z_0) = Z_6$$
$$FE(Z_0) = Z_7$$
$$DF(Z_0) = Z_8$$
$$FD(Z_0) = Z_9$$
$$FEF(Z_0) = Z_{10}$$
$$DFD(Z_0) = Z_{11}$$
$$FDF(Z_0) = Z_{12}$$
$$DFE(Z_0) = Z_{13}$$
$$FDFD(Z_0) = Z_{14}$$
$$DFEF(Z_0) = Z_{15}$$

The permutations

Glebe points out that in some cases a series of operations can be abbreviated. The permutation DE offers an example. Start with Z_0. Fold the rows back onto each other. Then pull at the center on the top and bottom sides of the double layer to separate the layers into a loop. Continue pulling until the left and right sides touch, forming a new double layer. When you unfold the double layer, you have Z_1.

All this work is a warm-up for solving Rubik's Magic. The trick is to find a means whereby one of the 32 two-by-four states can be changed into a three-by-three state with one corner square missing. Glebe discovered two ways, one of which requires 20 operations. A reader of his article found a simpler method, which is shown in part in the bottom illustration on the opposite page. Begin with operation R on Z_0 and then follow the instructions. The interlocked three rings are finally revealed when the flap on the right is folded to the right.

Many other techniques of ferreting out a solution to the puzzle can be found in the book by James G. Nourse listed in "Bibliography" [*page 183*]. The book also contains procedures for forming the puzzle into a variety of shapes.

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, Utah 84104. 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, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/87. $1.25 + 0.00.

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

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 953, Farmingdale, N.Y. 11737. Telephone inquiries: 1–800–227–5782. The date of the last issue on subscriptions appears at the 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

COMPUTER RECREATIONS

"After MAD": a computer game of nuclear strategy that ends in a Prisoner's Dilemma

by A. K. Dewdney

You are the leader of a nuclear superpower. A general, the chief of staff of the country's armed forces, has just handed you a sheet of paper bearing a somewhat confusing grid of numbers. There are two rows labeled "Us" and two columns labeled "Them." The general addresses you:

"This matrix contains the numerical essence of our best military, economic and political thinking. As you can see, there is a little box for each possible combination of actions. If we attack and the enemy does not, for example, our total gain is projected at negative 63 points—"

The general hesitates and then brightens somewhat.

"—but the enemy gain is projected to be negative 74 points."

As a leader you are used to making speeches and cutting ribbons. You have attended strategic briefings but confessed later that you found them rather complicated. And now you hear the chief of staff continue:

"In short, we have here a simple, two-by-two matrix showing the estimated payoffs and penalties if either side, neither side or both sides launch a nuclear attack."

		HE	
		COOPERATE	DEFECT
SHE	COOPERATE	(3,3)	(0,5)
	DEFECT	(5,0)	(1,1)

A matrix for the Prisoner's Dilemma

In reality you are in the initial stage of "After MAD," a computer game of nuclear strategy invented in 1985 at the Massachusetts Institute of Technology by Tad Homer-Dixon and Kevin Oliveau. The initials MAD stand for mutually assured destruction. The "after" refers to a point reached in play where the destruction is not so mutual: as the weapons deployed become increasingly accurate, it is eventually possible to launch a first strike that leaves the enemy too badly crippled to launch a serious nuclear counterblow. As a consequence each player becomes increasingly tempted to launch an unprovoked attack. The players have reached a situation known as the Prisoner's Dilemma [see "Metamagical Themas," by Douglas R. Hofstadter; SCIENTIFIC AMERICAN, May, 1983]. I shall explain below the peculiar ins and outs of "After MAD."

As I have already indicated, the key element of the game is a two-by-two matrix. The matrix is two-by-two because each side has two options. It can either attack or not attack. The contents of the matrix change from one step of the game to the next, so that each matrix summarizes the current strategic balance in terms of the numbers appearing in it. Each entry of the matrix consists of a pair of numbers, one pair for each possible combination of options taken by the two sides. The first number of each pair is the value of the combination's outcome for Us and the second number is the value for Them. If, for example, Us has decided to hold back and not attack while Them attacked, the first number would be low with respect to the second.

The subsequent game matrix depends, of course, on the combination of moves played by the two sides in response to the current matrix. Each game matrix leads to one of up to four other game matrices, depending on moves made by the players. In "After MAD" there are 110 matrices in the game, or 110 possible situations. Five distinct phases can be distinguished during play:

Early MAD. The strategic weapons of each side are effective only against "soft" targets such as centers of population and industry.

Second Phase (10 years after early MAD). MIRV's (multiple independently targetable reentry vehicles) are deployed on ICBM's (intercontinental ballistic missiles). Although a large number of weapons remain reserved for soft targets, each side has adopted a so-called counterforce strategy: its MIRV's are aimed exclusively at enemy ICBM silos.

Third Phase (20 years after early MAD). The accuracy of MIRV's has increased dramatically. MAD has lost credibility since a preemptive attacker is now assured of destroying 75 percent of the opponent's high-accuracy strategic weaponry.

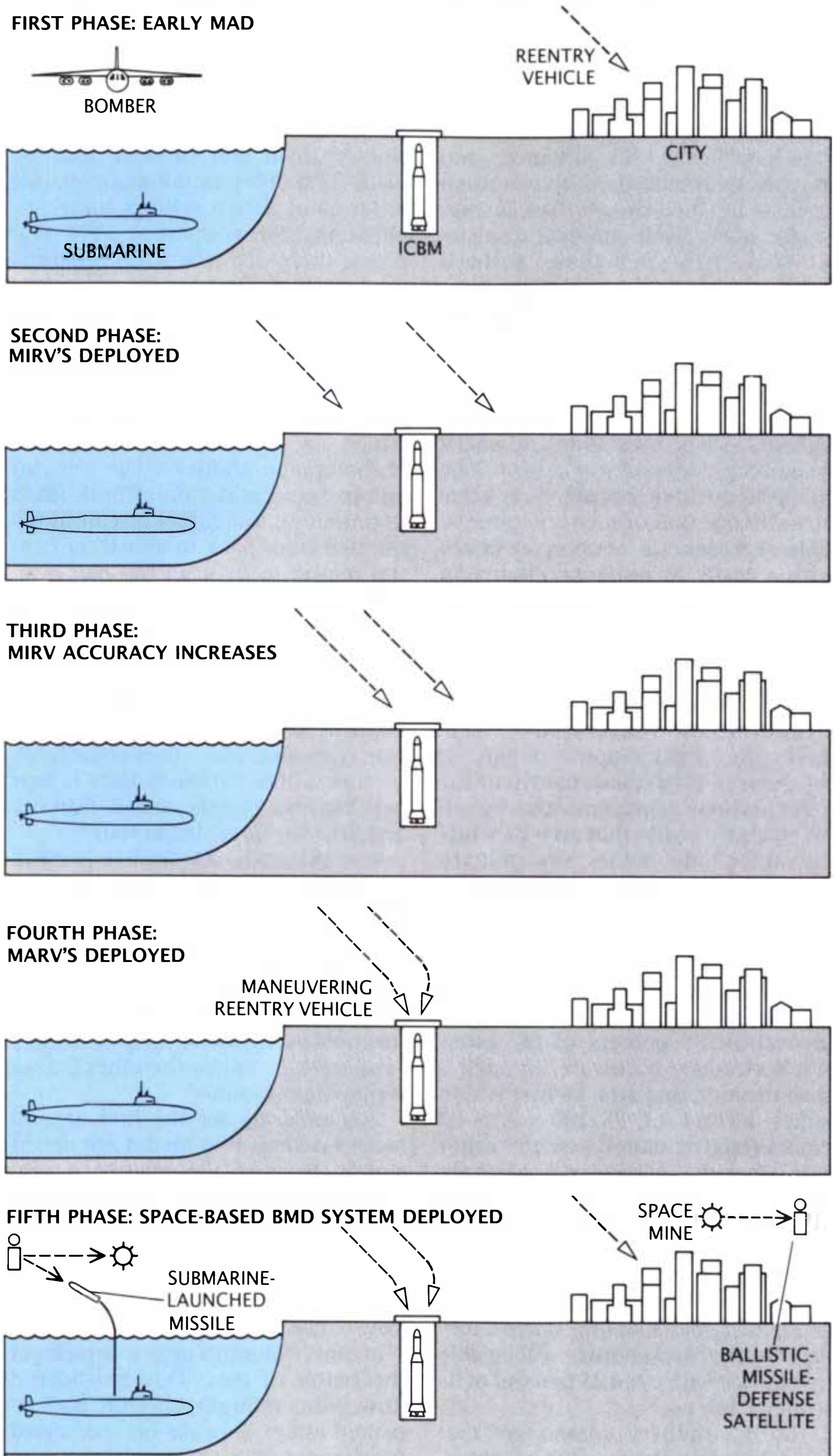
Fourth Phase (30 years after early MAD). Both sides have deployed MARV's (maneuverable reentry vehicles) that are capable of destroying virtually 100 percent of the enemy's ICBM's. MAD is eclipsed because a counterforce first strike is now militarily attractive; threatening to retaliate on population centers is not a credible deterrent to such an attack.

Fifth Phase (45 years after early MAD). Space-based ballistic-missile defense (SBMD) systems are able to destroy fully any ICBM attack involving up to 50 percent of the opponent's boosters. Each side also has space mines orbiting next to the enemy's SBMD-system components. The country launching a first strike could thus destroy the other's SBMD system, population centers and ICBM's. The attacker's SBMD would remain intact and be able to block any subsequent SLBM (submarine-launched ballistic missile) attack by the enemy.

Five central game matrices reflect the relative instabilities of these five phases [*see illustration on opposite page*]. Players progress from one central matrix to the next only if neither player attacks. Increasingly, however, they are tempted to attack by ever growing payoffs and the knowledge that the enemy faces the same temptation. The last two matrices are Prisoner's Dilemmas.

As Douglas R. Hofstadter described in these pages, a Prisoner's Dilemma begins when two people found near the scene of a crime are picked up by the police for questioning. As standard practice they are questioned separately to prevent them from coordinating their stories. It happens that the suspects are guilty. Neither one knows whether the other will confess or not. If both confess, things will go hard for the pair. But if only one confesses, he or she will look relatively good and will receive a much lighter sentence than the other.

In the language of the Prisoner's Dilemma, each prisoner may cooperate (with his or her partner) by denying involvement in the crime or, alternatively, may defect by telling all. There is a standard matrix associated with the Prisoner's Dilemma [*see illustration on opposite page*]. I have called the prisoners He and She in the belief that a romantic element often enliv-



		THEM	
		HOLD	ATTACK
US	HOLD	(3,3)	(−74,−63)
	ATTACK	(−63,−74)	(−74,−74)

		THEM	
		HOLD	ATTACK
US	HOLD	(3,3)	(−43,−19)
	ATTACK	(−19,−43)	(−43,−43)

		THEM	
		HOLD	ATTACK
US	HOLD	(3,3)	(−51,−5)
	ATTACK	(−5,−51)	(−45,−45)

		THEM	
		HOLD	ATTACK
US	HOLD	(3,3)	(−71,33)
	ATTACK	(33,−71)	(−62,−62)

		THEM	
		HOLD	ATTACK
US	HOLD	(3,3)	(−99,55)
	ATTACK	(55,−99)	(−90,−90)

The five phases of "After MAD" and their respective central game matrices ens a technical discussion. If both He and She cooperate, both receive a "payoff" of three points. If both defect, both receive a payoff of just one point. But if only one of the prisoners defects, He or She receives five points and the other receives zero.

What is the best way to play this game? Strategies for the Prisoner's Dilemma are evaluated by playing not one but many games. The average payoff becomes a figure of merit for a strategy.

In 1983 a Prisoner's Dilemma tournament was run by Robert Axelrod of the Institute for Public Policy Studies at the University of Michigan. Entrants in the tournament were computer programs embodying various strategies of play. Each pair of programs played 1,000 games. The clear winner was a very simple program called TIT FOR TAT. The strategy employed by the program was merely to echo what the opposing program had done on the preceding move.

The matrices in "After MAD" are similar to those in the Prisoner's Dilemma. I have substituted the words "hold" and "attack" respectively for "cooperate" and "defect." Also, the numbers in the matrices are usually larger, as in the central fourth-phase matrix of "After MAD" [*see illustration below*]. The matrix has a relationship among its payoffs that is characteristic of all Prisoner's Dilemmas: the "temptation" payoff (33), the "reward" payoff (3), the "punishment" payoff (−62) and the "sucker" payoff (−71) form a strictly decreasing sequence of numbers. One might hope players of "After MAD" would employ a tit-for-tat strategy, ensuring not only a reasonable score for each player but also, symbolically speaking, world peace.

Homer-Dixon and Oliveau have gone to some pains to make the game realistic, both in terms of the numbers in the game matrices and in terms of a printed scenario that accompanies each situation. They constructed the game in part as a research tool to probe the effects of technological change on the strategic relation between the U.S. and the Soviet Union, but their initial hypothesis about how students (among others) would tend to play the game was upset by actual play. Players generally did not wait for game matrices to become unstable before going to war. In 35 percent of the 100 games played at M.I.T. one side or the other launched a nuclear strike while MAD was still in effect.

It should be stated that the students playing the game had been carefully briefed in advance, and they were required to give written reasons for their moves. Yet, insofar as the world itself amounts to a single trial of the real thing, perhaps we have been lucky. I shall now describe how Homer-Dixon constructed the game matrices and how players responded to them.

Each payoff in each "After MAD" game matrix has three components: military assets, economic and social well-being and political power. The smallest positive payoff, 3, is composed of one unit of each component. This represents a country at peace with a small, incremental change in each of the three components over the short period represented by a move. The largest number of units, positive or negative, contributed by any one component is 33.

Here is how Homer-Dixon calculated the "temptation" payoff of the central third-phase matrix of the game, which represents the payoff received by a side that attacks while the other side holds. The military component is initially estimated at +5, but it has to be reduced by 3 to account for the cost of potential retaliation. The 3 was arrived at as follows. If one side launched an all-out counterforce attack, it would be able to destroy 75 percent of the other side's strategic hardware. In such a case the side that attacks first would inflict a loss of $75/100 \times 33 = 24$ points (approximately) on the other side. Since the attacked side has only 25 percent of its own weapons left after suffering the initial strike, it can therefore inflict on the attacker a loss of only $25/100 \times 24 = 6$ points. The loss of six points is discounted by a factor of two, reflecting uncertainty that the attacked country will be able to retaliate with even 25 percent of its strategic forces.

The +2 military component that remains when 3 (the cost of potential retaliation) is subtracted from 5 (the military gain) is then added to the contributions from the other two components: −2 for economic and social well-being and −5 for political power. Presumably the attacking country would lose some industry and population as a result of the enemy's response. Its government would (in a sane world) also be universally condemned for launching the attack in the first place. In sum, the preemptive attacker has earned $+2 - 2 - 5 = -5$.

The scenarios used in "After MAD" carefully spell out the strategic implications of the current situation. Homer-Dixon and Oliveau also explain each entry in the game matrix in terms of how it reflects those implications. Since there are 110 matrices, there are also 110 scenarios, each spelled out by text to the players arriving at that matrix.

A war develops whenever one (or both) of the players defects by attacking the other. At this point the game proceeds from the central matrix in which the attack developed to one of the branch matrices. The war can end in the destruction of both sides (terminating the game) or can move the two sides back to an earlier central matrix to pick up the pieces of their shattered existence.

Although players were instructed to do everything possible to maximize their point score, they seem to have taken many other factors into account. Unable to see or communicate with each other, they sometimes became a little paranoid. Here is how two players simply called Column and Row survived "After MAD":

Row played as a complete pacifist. As indicated in his computer log, Row was not trying to maximize his total score but attempting to follow moral principles instead. "The game I am playing will not be for points. I plan to play by my convictions for the most part. I shall probably always cooperate.... I hope the other player is playing my game."

Column's log for the first several moves reveals that he did not defect simply because the relative payoff was not large enough. Column, besides trying to maximize his score, was also trying to maximize the difference between his score and Row's. He expressed a fear early on that Row would attack him.

In time, Column's urge to attack got the better of him. This bewildered Row, who thought Column had attacked either because he was afraid Row would attack first or because he wanted to make the game more interesting. Row then hoped that Column

		THEM	
		HOLD	ATTACK
US	HOLD	(3,3)	(−71,33)
US	ATTACK	(33,−71)	(−62,−62)

Central matrix in "After MAD" fourth phase

would not attack again out of fear that Row was planning revenge.

For the next two turns both players cooperated. Row grew confident that Column would not attack again since he, Row, had proved his faithfulness. But Column's interpretation of these events was very different from Row's. Column in fact concluded that Row was quietly rebuilding his forces to the point where he could attack Column more effectively. Consequently, on the next move, Column attacked again. He had become somewhat nervous about Row's continuing cooperation: "Eventually I want to force him to such a low level that even if he does retaliate against an attack he will go below zero."

After the game Row reflected on his strategy and decided it needed some revision. His new strategy still follows moral principles, but it contains a distinct element of tit-for-tat: "Basically it consists of a policy of 'no first use' (NFU). In the event of an enemy defection I would retaliate immediately and continue to defect until my opponent either began to cooperate or made it clear that he would stop at nothing to win. In either case I would begin to cooperate myself since there is no sense in killing all his people too if he is determined to kill mine. Once my opponent had defected, I would then adopt a strategy of NFU until we reached the MARV stage. At this phase I would defect in order to escape the risk of a one-sided attack, which would leave me defenseless.... Although this approach may seem harsh, it appears to be the best way to ensure a minimal loss of human life."

From their experience with the student games Homer-Dixon and Oliveau concluded that players were interpreting the matrix payoffs not as they had been instructed but in terms of other criteria. Players sometimes attempted to maximize the difference between their own score and their opponent's. That may explain why so many defected from the start. Other players tried to maximize the sum of the scores. That might explain why 20 percent never defected. Still other players sought a position of strategic dominance or, conversely, adopted a purely moral stance.

One student summarized his feelings about the game in an essay. I quote some of his concluding remarks: "The temptations presented to the players in the simulation were compelling enough for several students to launch massive strikes against each other.... This does not bode well for the theory of balance of power, or balance of terror. If world leaders start viewing their playing field on video terminals, we could all be in for a game of PACMAN where all the dots get eaten and the change machine is out of quarters."

Homer-Dixon is currently an M.I.T. graduate student at the Center for International Studies in Cambridge. Oliveau is at Thinking Machines, Inc., also in Cambridge. "After MAD" was developed with funding from Project Athena, an ongoing program at M.I.T. sponsored by the IBM Corporation and the Digital Equipment Corporation to study the role of computers in education. Hayward Alker, a professor in the political-science department at M.I.T., helped to develop "After MAD" and has used it in his courses. Further information about the game and its availability can be obtained from Homer-Dixon by writing to him at the Center for International Studies, 292 Main Street, Cambridge, Mass. 02142.

I was overwhelmed by reader responses to the June column on algopuzzles. There were no programs to write, just some algorithms to design for the purpose of refueling trucks in the desert and switching trains. Several hundred one-track minds sent me their best efforts. I was pleased to have provided some vehicles for thought.

An algopuzzle is a puzzle that has an algorithmic solution: a recipe, or a procedure, for arriving at a specified goal. The train puzzles were the easier ones, particularly since I gave the solution to one of them in the article. It was naturally quite difficult to pick the best solution from the multitude of responses to the second puzzle, that of reversing an entire train using a single spur line capable of holding only one car at a time. Selected more or less at random from the earliest and best solutions to arrive is the following algorithm offered by Douglas A. Owenby of West Covina, Calif. The letter *A* designates the section of track from Problemtown to the spur line, *B* designates the spur itself and *C* designates the main track from the spur on to Solutionville. The *k*th car of the train is called *Pk*.

```
uncouple P train
forward to B
backward to C
for k = 1 to n
   forward to A
   couple Pk
   backward to C
   forward to B
   uncouple Pk
   backward to C
   forward to A
   backward to B
   couple Pk
```

Another solution starts with the last car of the train instead of the first car. As a number of readers noted, the above solution requires a total amount of work proportional to n^3, since a train of up to n cars must be dragged back and forth n times a distance of up to n cars in length.

The first desert-fox problem involved determining how far a patrol car could travel on n drums of fuel. Initially the fuel is stored in 50 gallon drums at a desert depot. The car can carry one drum at a time. How can the drums be moved into the desert in order to maximize the total distance traveled by the car? In the article I displayed an algorithm that guaranteed a trip of 600 miles on two drums. The optimum, as shown by William B. Lipp of Milford, Conn. (and many other readers), is 733.33 miles. In an amusing letter titled "Beat the Beast! Get 733 Miles on Only Two Barrels," Lipp exceeds the 666 miles I implied was possible. Here (lightly edited) is his algorithm:

```
Fuel vehicle from drum 1
Load drum 2
Forward 50 miles
Unload drum 2
Back to depot
Fuel vehicle from drum 1
Load drum 1
Forward 100 miles
Unload drum 1
Fuel vehicle
Back 50 miles to drum 2
Load drum 2
Forward 50 miles to drum 1
Fuel vehicle from drum 1
Forward 33⅓ miles
Unload drum 2
Back 33⅓ miles to drum 1
Load drum 1
Forward 33⅓ miles to drum 2
```

At this point the patrol car is 133⅓ miles from the initial depot. It now merely loads drum 2 (still full) and proceeds 600 miles farther.

That particular algorithm does not generalize to an optimum algorithm for more than $n = 2$ barrels. For example, it does not achieve the optimum of 860 miles for $n = 3$ barrels. I shall give a general algorithm (for $n \geq 6$ barrels) next month and display a solution to the second desert-fox problem as well.

BOOKS

Celebrating the birth of a new physics, opening New Guinea, studying the spineless

by Philip Morrison

THE NEWTON HANDBOOK, by Derek Gjertsen. Routledge & Kegan Paul ($59.95). **THE BIRTH OF A NEW PHYSICS: REVISED AND UPDATED,** by I. Bernard Cohen. W. W. Norton & Company (paperbound, $5.95).

On July 5, 1687, Dr. Edmund Halley, secretary of the Royal Society, which was the publisher, wrote from London to author Isaac Newton, Lucasian Professor and Fellow of Trinity College at Cambridge, saying that he had "at last brought your book to an end." The first edition of the stunning *Philosophiae naturalis principia mathematica* was off the press. Halley had been that book's indispensable friend, critic and editor; he had even printed it "at his own charge," in an edition of 300 or 400 copies done by two printers. For its part the Royal Society was broke. Halley sent along to Cambridge "in the same parcell" 20 copies of the 511-page quarto volume for Newton "to bestow on your friends"; Newton also received 60 more copies, which he was to place for sale at a price of nine shillings with booksellers.

It is a very pleasant duty to mark the tricentenary of the most influential book in the youth of modern science through a notice of this handbook, the work of Derek Gjertsen, a historian of science who teaches at the Open University. Here is a one-volume encyclopedia of Newton: a book consisting of about 300 articles, arrayed in alphabetical order, that bear on the life, the works and the scholarship concerning him. The list of entries spans Aberration to Young, Thomas. It includes Blake, William, the Kit-Kat Club and Vellum Manuscript. The *Principia* is itself the topic of one of the longest articles. Nearly 50 pages long, the article includes a chronology of the book's complex embryology, a summary of its contents, an outline comparison of the three editions brought out during Newton's lifetime and much more. A fine copy of the first edition would be expected to sell for something over £20,000. Yet the work is easily available in English translation, even in paperback. The only complete translation is the one by Andrew Motte in 1729, often and variously revised and reissued.

The basis for all this matter is "the extraordinary advance which has taken place in Newtonian scholarship since 1945." A dozen illustrious scholars and many other workers have shared the harvest. Within the past two decades we have reaped seven volumes of Newton's collected correspondence and eight volumes of his mathematical papers, as well as a definitive two-volume version of the Latin text of the *Principia*, a meticulous presentation that treats all the changes among the three editions. That work is accompanied in turn by a fascinating detective tome that unfolds the entire history of the big book. In addition there are a variety of brilliant biographies and monographs.

All these scholars have been mining a mother lode: a pile of the tireless Newton's manuscripts that belonged to the family of the Earl of Portsmouth. They found their way to the family seat after an intricate squabble among Newton's "rapacious kin," and the bulk of them remained "unread, unexamined, undisturbed, and even unknown for another 200 years." They were never hidden, but their difficulty and sheer quantity were daunting, even after they had been stored and catalogued in the library at Cambridge 100 years ago. They run to some three and a half million diverse words, more than half of them treating theology, chronology and alchemy; after a public auction in 1936 that drew every learned eye, the somewhat dispersed papers have paradoxically been under intense study.

"*Principia* as a work rather than a set of ideas," Gjersten writes, "began some time in August 1684 with a visit by Halley to Newton in Cambridge." That was the famous occasion when Halley asked what Newton thought "the Curve would be that would be described by the Planets" under an inverse-square force. "'Why saith he I have calculated it.'" Newton soon sent a brief paper that did calculate it. Back Halley came in November, to encourage the wary Newton to prepare a fuller treatise that would be published by the Royal Society. By April, 1686, Newton indeed sent to London the first much elaborated part of the final version.

There was lots of action for Halley. Newton proposed a more discursive third part, on the system of the world, applying his austerely geometric theorems to the solar system and the earth. In June, 1686, Halley welcomed the idea, one that would make the work "acceptable to all Naturalists, as well as Mathematiciens," and "much advance the sale." Then Newton heard of Robert Hooke's claims for some share of the credit. His famous reply to Halley ran: "The third I now designe to suppress. Philosophy is such an impertinently litigious Lady that a man has as good be engaged in Law suits as have to do with her." Halley, "much troubled," replied with earnest and diplomatic arguments. Newton consented to retain the threatened less mathematical section. The balance of the manuscript reached Halley in April, 1687.

July was not quite the last act of the *Principia* drama. On September 29, 1687, Newton deposited with the university library, as he was obliged to do, the text of his Lucasian lectures for the year. It is possible that they were hardly ever delivered, since no students came. His amanuensis of those years recalled that "oftentimes he did...for want of Hearers, read to ye Walls." Still the text itself is of great interest; it was published 40 years later right after Newton's death, both in English translation and in the original Latin. Its content was the earliest version of the third part of the *Principia*. Under the title *The System of the World*, the text was much more readable and less mathematical than the version that in fact closes the *Principia*. Newton wrote: "I had indeed composed the third Book in a popular method, that it might be read by many; but afterward...I chose... the form of Propositions."

One important argument he put forward "in a popular method" somehow never entered the *Principia* itself.

BREAKTHROUGH: HELPING COMPUTER PROGRAMMERS MAKE FAST AND STEADY PROGRESS.

Until now, manual skills were a big part of writing complicated computer programs. They required a major investment in time, effort and money—and days of boredom for the applications designer.

McDonnell Douglas information systems specialists developed a way to end the dull, turtle-speed tedium formerly necessary for good programs. Their software maps out the best program path and selects routine computer instructions.

The result is a working program that emerges with hare-like speed and is ready to use, allowing application systems designers to fully concentrate on the special characteristics of the project.

This breakthrough is called ALL—Application Language Liberator. Designers like it because it lets them try out ideas in prototype programs, quickly and easily, selecting only the best. Managements like it because it's faster, cheaper and yields better results.

We're making breakthroughs that add value and quality to information systems serving the people of the world.

We're McDonnell Douglas.

For more information, write:
ALL, McDonnell Douglas,
Box 14526, St. Louis, MO 63178

MCDONNELL DOUGLAS

INFORMATION SYSTEMS   MILITARY & COMMERCIAL AIRCRAFT   SPACECRAFT & MISSILES   HEALTHCARE SYSTEMS   HELICOPTERS   FINANCING

A thoughtful reader today, asked to accept the proposition of universal gravitation, is sure to be puzzled by the fact that the pull of one mass on another is not visible among the ordinary objects we see here on the earth. Apple and moon each fall to the earth's center, yet apples do not seem to pull on each other at all. Newton explains in the popular version that the attraction present between every pair of objects is simply too small to notice except on the cosmic scale. "Nay, even a mountain" is not large enough to attract noticeably, although he calculates its effect.

Within 20 years of Newton's death the pull of a nearby peak on the surveyor's plumb bob was measured roughly by Pierre Bouguer in the snows of the Peruvian Andes. Newton also reckoned that two one-foot balls of iron would attract each other across a gap of a quarter of an inch to meet in a month's time, if they were free to move. He got that wrong: it would take only a fraction of an hour, and yet even that much force is too small to outweigh friction.

The set of ideas Newton expressed in his forties had certainly been glimpsed in his undergraduate days. But they remained incomplete and uncertain for him, even if the moon-apple continuity came early to his mind, as he always maintained. This seems demonstrated by what happened five years before Halley's first visit. Hooke then wrote several times to Newton about planetary orbits and force laws; Newton worked out some results but held his counsel. All three men, Hooke, Halley and Newton, had grasped something of orbits under attraction, but the idea of universality was not there. In 1680 a bright comet streaked by; the astronomers mapped it as it fell near the sun and sped back outward again. Newton did not agree; he thought the two paths must be those of two comets, for comets must move in straight lines. John Flamsteed saw the comet as a single object and told Newton so. It was not until 1685 that Newton conceded. He must have realized by then that comets too were under universal attraction and curved under solar rule. At least a third of the last part of the *Principia* is devoted to the orbits and nature of comets.

Among the distinguished harvesters of the Newton material is I. Bernard Cohen. His book is more than an introductory history; it is an engaging account of the physics of simple orbits under gravity, presented in the context of the development of those ideas. It includes persuasive experimental material of today, including the wonderful time exposures by Berenice Abbott that show the path of a ball thrown up straight from a moving platform, a motion seen as curving from without but as vertical to the moving observer. It is not a new book but one revised and much improved by two decades of scholarship, particularly with respect to the experimental flavor of the work of Galileo. If any reader wants to study the grand if simple ideas of circular orbits and the persuasive similarity between apple and moon, the unity of all matter and motion, there is no better place than in this bargain paperback.

Nor is there a more fitting way to commemorate three centuries of the scientific Enlightenment than to delve for oneself among its roots.

FIRST CONTACT, by Bob Connolly and Robin Anderson. Viking Penguin, Inc. ($19.95).

The mountains of New Guinea stretch nearly east and west for about 1,000 miles, the roof ridge of that great equatorial island. Their highest peaks outtop the Alps or the Sierras. Monsoon-drenched and malarial, all dense forests and tall grass, wide coastal lowlands border the entire central ridge. Even empire came late and thin to New Guinea: it is reported that 100 years ago not a single European lived north of the ridge on mainland Papua, the island's eastern half, an area twice the size of Florida. Colonial rule extended a few tens of miles in from the sea, since the people lived near the coasts, perhaps a million of them after World War I in all Papua, village farmers and fishermen. To planters, labor recruiters, missionaries, traders and officials they were hands, souls and customers. A few hundred such colonial outsiders were resident when for vaguely strategic reasons Australia took over the old British and German rule over Papua and the Territory of New Guinea. (The western half of the island is now part of Indonesia and is not treated here.)

Two large river valleys drain the interior. The Fly River flows a couple of hundred miles south to the sea and the Sepik flows north. Each rises far inland on wild, fog-shrouded mountain slopes. Along the wide valleys there is a diverse set of cultures, the population thinning out toward the steep foothills. There was an easy interpolation: out there on the high, narrow ridge between the slopes could be found only sparse nomadic mountain bands. They would be the highland people who traditionally had taken, through indirect trade from hand to hand, the cowries and the showy gilt-edged pearl shell gathered on the coasts, paying with plumage and bird skins. The best maps left a blank space as big as Scotland across the Papuan highlands, which were thought surely to be wild, rugged and empty. But then, in 1926, stream-borne gold was found near the northern coast of the island.

In May of 1930 two bold young Australian prospectors, Michael Leahy and Michael Dwyer, with 15 coastal men hired as carriers, six of them armed and licensed *gunbois,* climbed into the Bismarck Range, bound for the first of a set of short trips upriver seeking new placers. They started at a tiny foothills settlement the Lutheran missionaries had set up at the northern edge of the blank space. In one day's hard climb they had reached the lip of the mountain; before them spread a startling vista, a grassy plain, one of a few long plateau valleys that part the twin mountain ridges of the interior. By nightfall their wonder had turned to alarm: as far as they could see the stream-laced valley was filled with the pinpoint lights of fires.

At sunrise the villagers came to meet them—men armed with bows and arrows and women offering sugarcane. Followed by crowds, touching and hugging, sharing the sweet cane, the travelers walked past endless hedged gardens with their long rows of beans, cane and sweet potatoes, while fat pigs wandered freely everywhere.

From time to time the smiling, excited people would drop away with tense warning gestures. Soon there would assemble a new emotional crowd: the strangers had entered a distinct but similar neighboring territory. After a week they reached a high vista and looked out to the horizon. An early photograph shows the wider scene: a grove of trees surrounds a ceremonial clearing amidst a dozen or two log huts adjoining large, neat gardens; within a mile in any direction there is another grove, another cluster... Thousands of enclaves, splintered into bands of friends and enemies, politically divided by language and by ancient ritual wars, spread over the fertile lands. There were no metals (a few steel knives and axes from trade), no malaria, no woven textiles, no beasts of burden, no chiefdoms; instead the competitive feasting displays of the hardworking and charismatic "big men." The populous highlands had at last been united with the rest of our species.

Mike Leahy was robust, inspired, brilliant, even though he never finished secondary school. His extraordinary story is not a new one; he published his account after six years of contact with the highlanders. That contact was as intimate as the begetting of his three sons by Mount Hagen women, as ruthless as his Mauser rifle (which, along with the shorter-range muskets of his *gunbois*, killed perhaps 40 of the people in swift encounters that met tragic misunderstanding with an old colonial ruthlessness), as fascinating as the 5,000 candid shots he took with his Leica, as novel as the light aircraft he talked the mining company into dispatching to link forever the million highlanders with the outside world they had never known.

It is a new story that is the core of this absorbing book by two Sydney filmmakers turned ethnographers. They present the testimony from the other side. In the early 1980's the authors followed Leahy's carefully recorded itineraries to seek out men and women who recalled the times of first contact, when tens of thousands of people first saw the outsiders. Some 60 interviews in eight highland languages are the ore they worked. They also present images made by Leahy and others, including a picture of Mike's influential and wealthy son Clem, shown with his mother in 1983, and two Leica shots of that same woman when, five decades earlier, she had both smiled and frowned at Mike Leahy in all her youthful beauty.

Concede that recollections told 50 years late are suspect; that cannot be helped. Moreover, these informants recall the most astonishing perception of their lives with utter clarity. Remembering just what they were doing at the time, these old people look around carefully to spot and point out "the boy or girl whose age they were when news first arrived."

The ancestors of the million people of the highlands have lived there for a time so long that it passes for always; archaeological dating puts some of the gardens at 9,000 years old. Thoughtful people knew there were other places, for they prized greatly the shells they gained by trade, exotic objects of enigmatic beauty. (The ocean itself was unguessed at; salt was a costly stuff.) Some, in places where the view was bounded by ridges, knew only that on the next ridge there were enemies: "We couldn't go past them." Others with a wider view of eye and mind "used to see smoke in the distance...I wonder who that is. I'd like to meet those people one day."

When the new men came, they were neither allies nor enemies. The cosmology had room for nothing besides, except spirits. The white men were just the color expected of those from the place of the dead, "our dead people, come back!" Even an Australian could be seen as a dead father by a bereaved young person and be taken before the uncles to be looked at straight in the eye for confirmation. The coastal men who were serving as carriers looked familiar enough, even if they could speak only gibberish: those men were often tearfully made out to be the familiar dead. "I saw half his finger missing, and I recognized him as my dead cousin...the very same man. His facial expression, the way he talked, laughed—exactly the same."

Perhaps they all turned into skeletons in sleep, befitting their spirit nature. In one place two heroic warriors are recalled by name as those who dared to enter the sleeping camp to peep into the tents. Elsewhere that story faded away slowly. Did all that wrapping mean they did not need to rid themselves of wastes? It did not: one man hid to watch them at the latrine pit. The smell was unifying. Was that loosely folded skin or a costume? Did those clothes-covered bodies mean they had something to hide, perhaps the giant penises of spirit myth? That was not true either, as watching the bathers soon proved. Were their women hidden, packed in the baggage? The prospectors spent much time panning stream gravel; the people long before had themselves regularly burned the bones of the dead and disposed of the ashes in the rivers. Were the strangers perhaps seeking their own old bones?

Above all, the newcomers had with them a partly decipherable but rich inventory. At first their goods were mainly magic: the lid of a tin can might become a precious talisman, or a discarded matchstick might be found and eaten for its unknown powers. Soon the loaded airplanes arrived (the first airstrip was prepared by late 1932), coming down among throngs swept by sudden terror as the daunting noise approached. Yet the thundering craft seemed always to be friendly; they "came with things—trade goods, axes, shells, to name just a few. Heaps of them! Then

Michael Leahy in 1934

we said, 'These men must be men-of-all-things.'"

Fifty years have gone by. Many highlanders have themselves gone to view the salt sea. Papua New Guinea is a nation; the highlands grow coffee at profit, and now they know colleges and police barracks, movies and beer gardens, hospitals that admit both for arrow wounds and for heart attacks. The roar of heavy diesels rolling down dirt roads has also become familiar. The big men still aspire, but their lovely pearl shells, steadily inflated by air cargo from dazzling rarity into cloying excess, have lost their prestigious appeal; the profligate displays the big men now organize center around tusked pigs, motorcars and cold cash.

LIVING INVERTEBRATES, by Vicki Pearse, John Pearse, Mildred Buchsbaum and Ralph Buchsbaum. Blackwell Scientific Publications, Inc., P.O. Box 50009, Palo Alto, Calif. 94303 ($45).

This thick, image-crowded, up-to-date bestiary admits neither fish nor fowl, neither frog nor fox nor fer-de-lance. Its contents are defined by their otherness; none of the five classes of our true vertebrate kin are here (although some chordates are included) solely because it is we who make the rules. The topic is invertebrates, "the great spineless majority," representatives of groups that include about 97 percent of the 10 or 20 million living species. Those species are included within the 10 principal phyla, or distinct body plans, that together make up most living animal life. The remaining 20 minor phyla include no more than hundreds of species each. This is a highly readable catalogue of the products of a slow, ingenious natural engineering, still at the margins of our understanding. Quite a few pages at the end list additional tempting readings, some popular, some erudite, categorized mostly by animal group.

Many hundreds of photographs and diagrams spill over these 800 pages, which are organized into some 30 readable chapters, each one a much illustrated review of the natural history of one or another phylum. The treatments include descriptions of materials, structures, habitat and behavior for each form; they are given a degree of unity by the rich comparisons, explicit and implied. The book is a fond celebration of living diversity, visually and in closely integrated clear, lively prose, meant both for students and for general readers.

A few examples will have to stand for the hefty whole. The most familiar ciliate, the eyelashed slipper-shaped paramecium, is freshly seen over 10 pages, photographed and diagrammed both whole and (in one detail of its cortex) at electron-microscope scale. Its complex genetics and behavior, its size (large compared with most animal cells), and its statistical life course (with clear analogues to maturity, aging and death) all lead to the comment that multicellular animals are here foreshadowed both by the individual paramecium and by a clone of these protists. The aggregating ciliates, like the aggregating amoebas with their flagellas, are shown also, in lovely circus imitations of permanently multicellular animals.

Look at the roundworms, the nematodes. They are the animal phylum next in size to arthropods. Highly adaptable pressurized tubes of tissue, they frequent an astonishing variety of habitats both as free-living organisms and as internal parasites. "If all the matter in the universe except the nematodes were swept away...and if, as disembodied spirits, we could then investigate it, we should find its mountains, hills, vales, rivers, lakes, and oceans represented by a film of nematodes. The location of towns would be decipherable, since for every massing of human beings, there would be a corresponding massing of certain nematodes. Trees would still stand in ghostly rows representing our streets and highways. The location of the various plants and animals would still be decipherable,...even their species." (The telling quotation is credited to N. A. Cobb; unfortunately no more detailed source is given.)

Even though most readers will never see a nematode, people remain anxious about the trichina, a nematode parasite of both pigs and human beings. It is argued here that the two Aesculapian serpents entwined on a staff, still the symbol for the medical professional, arose from a therapy long practiced in earnest against the parasitic roundworm *Dracunculus,* common in the Middle East. Its serpentine ridge, a foot long, bulged visibly under the skin of the patient's leg or trunk; the professional cure was to gingerly wind the creature out alive and unbroken around a stick, leaving behind no source of infection. That is clearly no task for a nervous tyro.

So we read on, past the colonial man-of-war, the pelagic violet snail and the blue crab—the crab unhappily bearing on its back an outsize barnacle and probably also long inhibited from molting by an internal parasite, a rhizocephalan, of the same subclass as the barnacle outside. (It is able to take over hormonal control of crab growth on entry to form a "rootlike system of nutrient-absorbing tubules...that branch throughout the host's blood spaces, even to the tips of the legs.") Millipedes earn a couple of pages. (In truth the leg count among their 8,000 species does not exceed 752.) A really big tropical millipede may be two feet long. Herbivores, they do not bite, but they often secrete toxic substances in defense, including hydrogen cyanide. "It is best not to hold any millipede too close to one's eyes."

The next-to-last chapter is a brief evolutionary account. The main phyla here are half a billion years old, a brief part of life's long and mainly biochemical evolution. The multicellular phyla can be divided into two classes by their early embryology. Often the distinction can be seen as early as the third generation of cell division. The eight embryonic cells may be arranged radially with four cells above four, or the two layers of cells may be twisted 45 degrees out of alignment: the root of a spiral. Every species within a given phylum shares the same beginning; no phylum is mixed. Radial phyla include worms and mollusks; we chordates share the spiral cleavage pattern with the phyla of corals, starfishes and other such creatures, our cousins many times removed.

The final chapter is exceptional: a cross-classifying look at color among animals of all groups. There are a few dozen plates in color, each with four or five photographs—an attractive gallery to finish the long, bizarre tale. Color comes from pigments, from tanning, from structural interference patterns. An animal may gain almost any color from those causes (observe live lobsters bright blue or greenish brown, and boiled ones red); it may look sky blue from scattering by tiny granules strewn within a clear layer, or it may borrow hue from colorful endosymbionts. Quick change is achieved too, either photochemically or by the physical flattening of many rounded sacs of pigment in the skin. The long body of one Tasmanian damselfly scatters a glorious sky blue in the midday sun but remains enclosed within a warming sleeve of black pigment during the cool of the morning.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

THE NEXT COMPUTER REVOLUTION

LOGIC FOR PROBLEM SOLVING. R. Kowalski. Elsevier Science Publishing Co., Inc., 1979.

SMALLTALK-80: THE LANGUAGE AND ITS IMPLEMENTATION. Adele Goldberg and David Robson. Addison-Wesley Publishing Co., 1983.

NO SILVER BULLET: ESSENCE AND ACCIDENTS OF SOFTWARE ENGINEERING. Frederick P. Brooks, Jr., in *Computer,* Vol. 20, No. 4, pages 10-19; April, 1987.

HERE COME THE BILLION-TRANSISTOR IC. Bernard C. Cole in *Electronics,* Vol. 60, No. 7, pages 81-85; April 2, 1987.

ADVANCED COMPUTER ARCHITECTURES

INTERCONNECTION NETWORKS FOR SIMD MACHINES. Howard Jay Sussman in *Computer,* Vol. 12, No. 6, pages 57-65; June, 1979.

PARALLEL COMPUTERS. R. W. Hockney and C. R. Jesshope. Adam Hilger, Ltd., 1981.

A SURVEY OF HIGHLY PARALLEL COMPUTING. Leonard S. Haynes, Richard L. Lau, Daniel P. Siewiorek and David W. Mizell in *Computer,* Vol. 15, No. 1, pages 9-24; January, 1982.

CONCURRENT VLSI ARCHITECTURES. Charles L. Seitz in *IEEE Transactions on Computers,* Vol. C-33, No. 12, pages 1247-1265; December, 1984.

A SURVEY OF PROPOSED ARCHITECTURES FOR THE EXECUTION OF FUNCTIONAL LANGUAGES. Steven R. Vegdahl in *IEEE Transactions on Computers,* Vol. C-33, No. 12, pages 1050-1071; December, 1984.

EXPERIMENTAL PARALLEL COMPUTING ARCHITECTURES. Edited by Jack C. Dongarra. Elsevier Science Publishing Co., Inc., 1987.

PROCEEDINGS OF THE THIRD CONFERENCE ON MULTIPROCESSORS AND ARRAY PROCESSORS. Edited by Walter J. Karplus. The Society for Computer Simulation, 1987.

CHIPS FOR ADVANCED COMPUTING

THE EVOLUTION OF DIGITAL ELECTRONICS TOWARDS VLSI. Robert W. Keyes in *IEEE Transactions on Electron Devices,* Vol. ED-26, No. 4, pages 271-279; April, 1979.

ULTRA-LARGE SCALE INTEGRATION. James D. Meindl in *IEEE Transactions on Electron Devices,* Vol. ED-31, No. 11, pages 1555-1561; November, 1984.

THE FUNDAMENTAL PHYSICAL LIMITS OF COMPUTATION. Charles H. Bennett and Rolf Landauer in *Scientific American,* Vol. 253, No. 1, pages 48-56; July, 1985.

PROGRAMMING FOR ADVANCED COMPUTING

THE COSMIC CUBE. Charles L. Seitz in *Communications of the ACM,* Vol. 28, No. 1, pages 22-33; January, 1985.

DOMESTICATING PARALLELISM. David Gelernter in *Computer,* Vol. 19, No. 8, pages 12-19; August, 1986.

PARA-FUNCTIONAL PROGRAMMING. Paul Hudak in *Computer,* Vol. 19, No. 8, pages 60-71; August, 1986.

ADVANCED COMPILER OPTIMIZATIONS FOR SUPERCOMPUTERS. David A. Padua and Michael J. Wolfe in *Communications of the ACM,* Vol. 29, No. 12, pages 1184-1201; December, 1986.

DATA PARALLEL ALGORITHMS. W. Daniel Hillis and Guy L. Steele, Jr., in *Communications of the ACM,* Vol. 29, No. 12, pages 1170-1183; December, 1986.

PARALLELISM, PERSISTENCE AND META-CLEANLINESS IN THE SYMMETRIC LISP INTERPRETER. David Gelernter, Suresh Jaganathan and Thomas London in *Proceedings of the SIGPLAN '87 Symposium on Interpreters and Interpretive Techniques.* ACM Press, 1987.

DATA-STORAGE TECHNOLOGIES FOR ADVANCED COMPUTING

DISK-STORAGE TECHNOLOGY. Robert M. White in *Scientific American,* Vol. 243, No. 2, pages 138-148; August, 1980.

PERPENDICULAR MAGNETIC RECORDING—EVOLUTION AND FUTURE. Shunichi Iwasaki in *IEEE Transactions on Magnetics,* Vol. MAG-20, No. 5, pages 657-662; September, 1984.

MAGNETO-OPTIC RECORDING TECHNOLOGY. Mark H. Kryder in *Journal of Applied Physics,* Vol. 57, No. 8, Part IIB, pages 3913-3918; April 15, 1985.

INTERFACES FOR ADVANCED COMPUTING

FUNDAMENTALS OF INTERACTIVE COMPUTER GRAPHICS. James D. Foley and

Andries van Dam. Addison-Wesley Publishing Co., 1982.

DIRECT MANIPULATION: A STEP BEYOND PROGRAMMING LANGUAGES. Ben Shneiderman in *Computer*, Vol. 16, No. 8, pages 57–69; August, 1983.

THE HUMAN INTERFACE: WHERE PEOPLE AND COMPUTERS MEET. Richard A. Bolt. Lifetime Learning Publications, 1984.

DIRECT MANIPULATION INTERFACES. E. Hutchins, J. Hollan and D. Norman in *User Centered System Design*, edited by Donald A. Norman and Stephen Draper. Lawrence Erlbaum & Associates, 1986.

NETWORKS FOR ADVANCED COMPUTING

COMPUTER NETWORK DEVELOPMENT TO ACHIEVE RESOURCE SHARING. L. G. Roberts and B. D. Wessler in *Proceedings of the AFIPS Spring Joint Computer Conference*. American Federation of Information Processing Societies, 1970.

A PROTOCOL FOR PACKET NETWORK INTERCOMMUNICATION. Vinton G. Cerf and Robert E. Kahn in *IEEE Transactions on Communications*, Vol. COM-22, No. 5, pages 637–648; May, 1974.

SPECIAL ISSUE ON PACKET COMMUNICATION NETWORKS. *Proceedings of the IEEE*, Vol. 66, No. 11; November, 1978.

EVOLUTION OF THE ETHERNET LOCAL COMPUTER NETWORK. John F. Shoch, Yogen K. Dalal, David D. Redell and Ronald C. Crane in *Computer*, Vol. 15, No. 8, pages 10–27; August, 1982.

INTEGRATED SERVICES DIGITAL NETWORK. *AT&T Technical Journal*, Vol. 65, No. 1; January–February, 1986.

ADVANCED COMPUTING FOR SCIENCE

GALACTIC BRIDGES AND TAILS. Alar Toomre and Juri Toomre in *The Astrophysical Journal*, Vol. 178, No. 3, Part 1, pages 623–666; December 15, 1972.

INTRODUCTION TO VLSI SYSTEMS. Carver Mead and Lynn Conway. Addison-Wesley Publishing Co., 1980.

THE PHYSICAL UNIVERSE: AN INTRODUCTION TO ASTRONOMY. Frank H. Shu. University Science Books, 1982.

BINARY FORMATION AND INTERACTIONS WITH FIELD STARS. Piet Hut in *Dynamics of Star Clusters: Proceedings of the 113th Symposium of the International Astronomical Union*, edited by Jeremy Goodman and Piet Hut. D. Reidel Publishing Co., 1985.

A DIGITAL ORRERY. James H. Applegate, Michael R. Douglas, Yetka Gürsel, Gerald J. Sussman and Jack Wisdom in *The Astronomical Journal*, Vol. 92, No. 1, pages 176–194; July, 1986.

ADVANCED COMPUTING FOR MEDICINE

READINGS IN MEDICAL ARTIFICIAL INTELLIGENCE: THE FIRST DECADE. Edited by W. J. Clancey and E. H. Shortliffe. Addison-Wesley Publishing Co., 1984.

COMPUTER-ASSISTED MEDICAL DECISION MAKING. Edited by J. A. Reggia and S. Tuhrim. Springer-Verlag, 1985.

A COMPUTATIONAL MODEL OF REASONING FROM THE CLINICAL LITERATURE. Glenn D. Rennels. Springer-Verlag, 1985.

SPECIAL ISSUE ON MEDICAL INFORMATICS. *The Western Journal of Medicine*, Vol. 145, No. 6; June, 1986.

A COMPUTER DATA BASE FOR INFORMATION ON CANCER TREATMENT. Susan Molloy Hubbard, Jane E. Henney and Vincent T. DeVita in *The New England Journal of Medicine*, Vol. 316, No. 6, pages 315–318; February 5, 1987.

COMPUTER PROGRAMS TO SUPPORT CLINICAL DECISION MAKING. Edward H. Shortliffe in *The Journal of the American Medical Association*, Vol. 258, No. 1, pages 61–66; July 3, 1987.

ADVANCED COMPUTING FOR MANUFACTURING

IMPACT OF COMPUTATIONAL METHODS ON AIRCRAFT DESIGN. P. E. Rubbert and E. N. Tinoco. AIAA Atmosphere Flight Conference, 1983.

HIGH-SPEED COMPUTATION. Edited by Janusz S. Kowalik. Springer-Verlag, 1984.

VECTOR AND PARALLEL PROCESSORS IN COMPUTATIONAL SCIENCE. Edited by I. S. Duff and J. K. Reid. Elsevier Science Publishing Co., Inc., 1985.

DIRECT METHODS FOR SPACE MATRICES. I. S. Duff, A. M. Erisman and J. K. Reid. Oxford University Press, 1986.

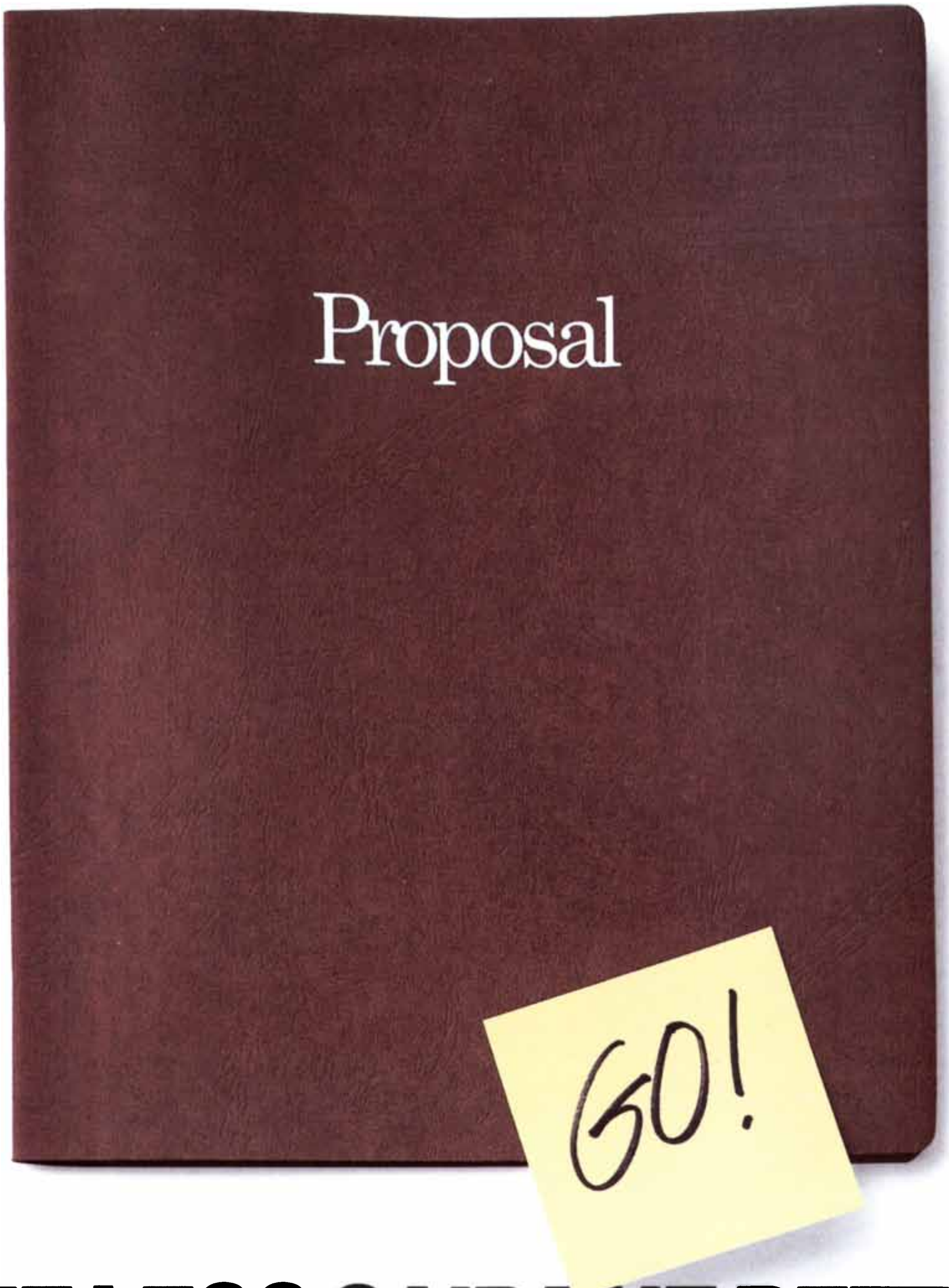
THE AMATEUR SCIENTIST

SIMPLE SOLUTIONS TO RUBIK'S MAGIC. James G. Nourse. Bantam Books, 1986.

COMPUTER RECREATIONS

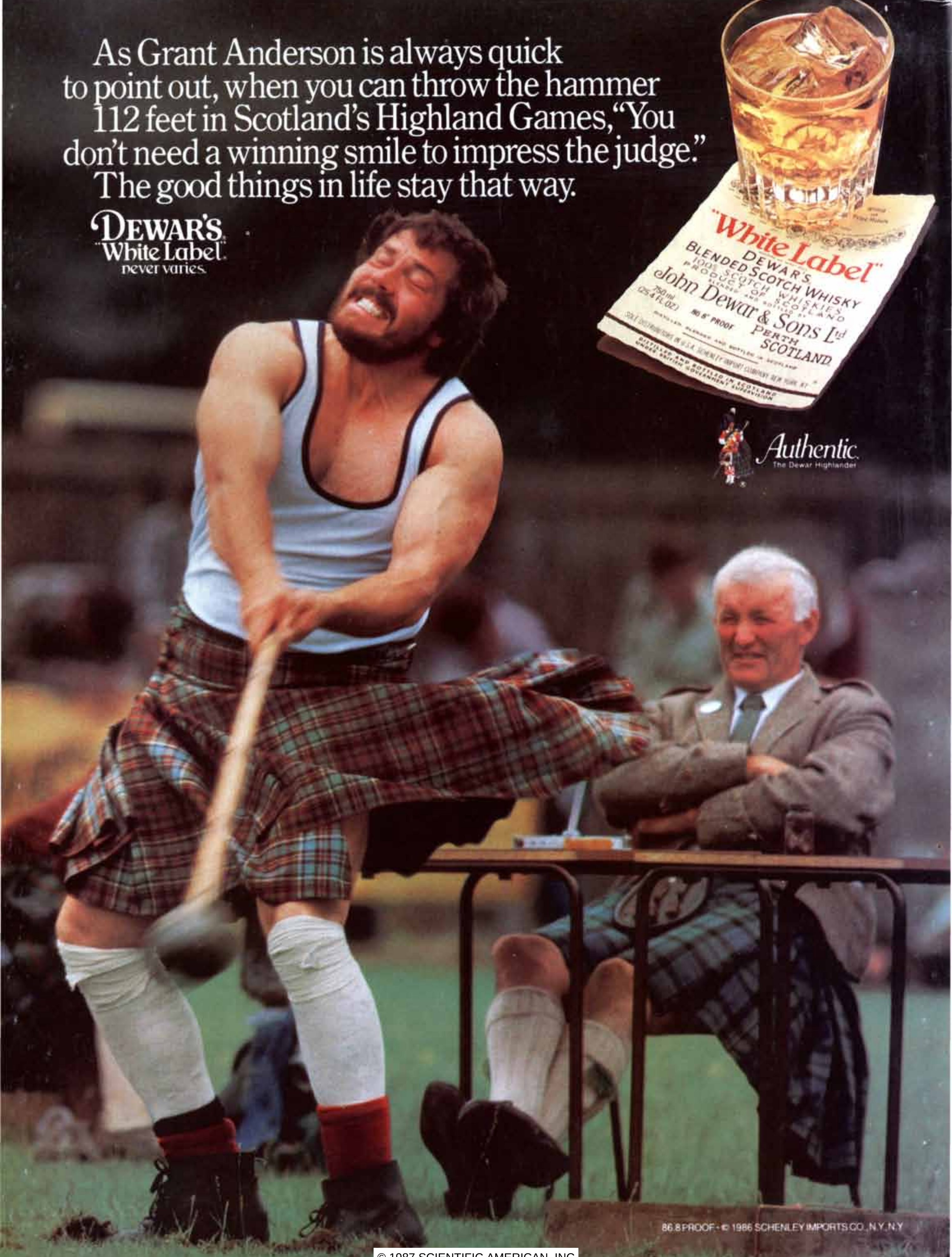
FIGHTS, GAMES, AND DEBATES. Anatol Rapoport. University of Michigan Press, 1960.

THE EVOLUTION OF COOPERATION. Robert Axelrod. Basic Books, 1984.



THE LESS SAID THE BETTER.

When People Count On You,
Count On Post-it™ Notes.
BRAND



As Grant Anderson is always quick to point out, when you can throw the hammer 112 feet in Scotland's Highland Games, "You don't need a winning smile to impress the judge."
The good things in life stay that way.
DEWAR'S
"White Label"
never varies.
"White Label"
DEWAR'S
BLENDED SCOTCH WHISKY
John Dewar & Sons Ltd
PERTH
SCOTLAND
Authentic.
The Dewar Highlander
86.8 PROOF · © 1986 SCHENLEY IMPORTS CO., N.Y., N.Y.