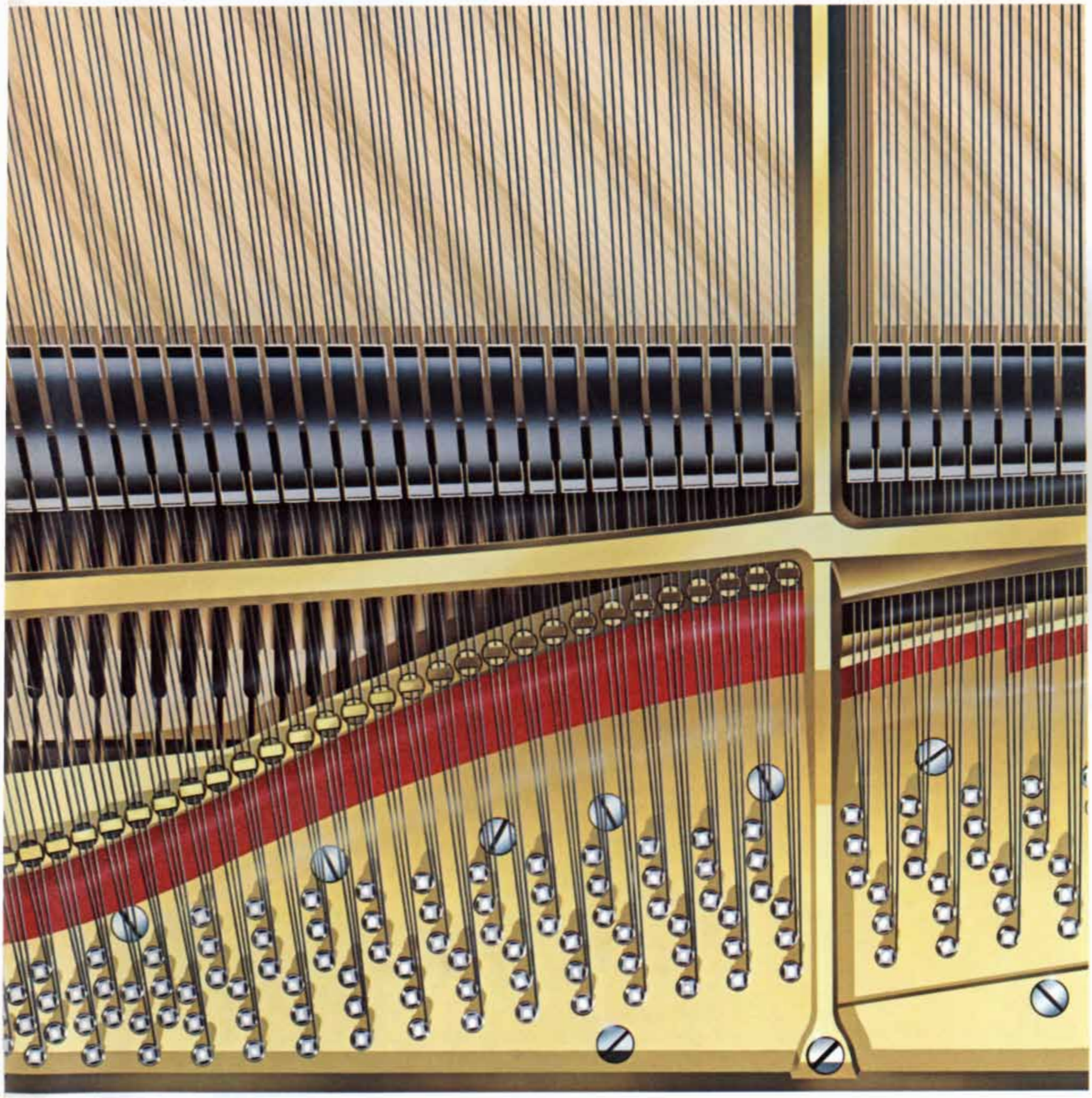
SCIENTIFIC AMERICAN



PHYSICS OF PIANO STRINGS

$1.50

January 1979

RABBIT THE #1 SELLING CAR IN SWITZERLAND.

The Swiss are no cuckoos.

They sit surrounded by Germany, Italy and France – all of the biggest car makers in Europe.

Their choice of cars is unlimited. Yet the car they buy most is the Volkswagen Rabbit.

It seems fair to ask why.

To begin with, mountain climbing isn't just a hobby in Switzerland; it's the way everyone drives. Good weather or bad (especially bad) there is nothing like Rabbit's front-wheel drive to get a car up an Alp. Or a Rocky, or even a steep driveway.

Also, the Swiss worship precision; it's what makes them tick. Car and Driver described the Rabbit this way: "Quality is exceptionally high throughout, with solid slamming doors and a structure that feels as substantial as a Mosler safe."

The Swiss also dote on technology. Fuel injection, for example.

You might be as interested as the Swiss to know that you can't get a Renault with fuel injection. Or a Fiat. Or a Lancia. Not to mention Toyota, Honda, or Mazda. But you can get a fuel-injected Rabbit.

Last, but hardly least, is the fact that the Swiss are – well – frugal. And so when they see a car that's built like a vault, climbs like a goat, is far ahead of its time and still sells for a reasonable price, the Swiss do what sensible people everywhere do.

They buy them in droves.

VOLKSWAGEN DOES IT AGAIN

On Photographing a Monster

It *hates* being photographed. It glares balefully at the camera. It sulks, fidgets, pouts.

It is that most fearsome of creatures, a perfectly normal four-year-old child.

The photographer, wise in the ways of distracting difficult subjects, abandons his camera. He wanders about the studio, talking and gesturing. The monster, no longer feeling threatened by the evil eye of the lens, relaxes and becomes once more the sweet child its mother knows.

From time to time, the eye of the camera blinks. There is a quiet whirring of an electric motor, and the camera blinks again. Yet the photographer is nowhere near it.

An Ingenious Little Dial.

The camera is a Hasselblad, the motor-drive 500EL/M.

A dial on the camera is set at "SR." This prereleases the entire reflex system. At the end of a 20-foot extension cord, the photographer holds a release mechanism. When triggered, it operates the leaf shutter without a moment's delay, thus reducing the risk of "eye blink" from the subject and delivering a higher number of useable photographs.

There are four other settings on this dial: Single-frame in the normal mode. Continuous, automatic firing for as long as the release is held down. A single frame in the prereleased, speeded-up mode for one frame only. And continuous, automatic firing in the speeded-up mode at the rate of one frame every 8/10th of a second.

Hasselblad Versus 35mm.

Why is this photographer not using one of the many excellent motor-drive 35's for this assignment?

One compelling reason might be the size of the Hasselblad negative. Each is 2¼ inches square, *almost four times the area of a 35mm frame.* (See box, below right, for actual size.)

To make an 11 x 14 print of the child, one would have to enlarge a 35mm negative 11.6 times. A comparable Hasselblad print requires only a 7-times enlargement. The ability to hold contrast is phenomenal.

Then there is the camera's leaf shutter. Unlike the shutter of a 35mm camera, it is fully synchronized for every type of flash at every speed up to 1/500th of a second.

Hasselblad Versus Others.

But why a Hasselblad when there are less-expensive alternatives? Why did eight out of ten top professional photographers surveyed name Hasselblad as the medium-format camera used in their work?

Consider this: The Hasselblad System makes available no less than 4 cameras, 21 lenses, 8 viewfinders, 9 film magazines (including Polaroid® backs) that give you a choice of anywhere from 1 to 500 exposures, and a staggering array of accessories.

A total of 332 pieces.

No other system that exists today in the medium-format field can muster more than a fraction of that number of pieces.

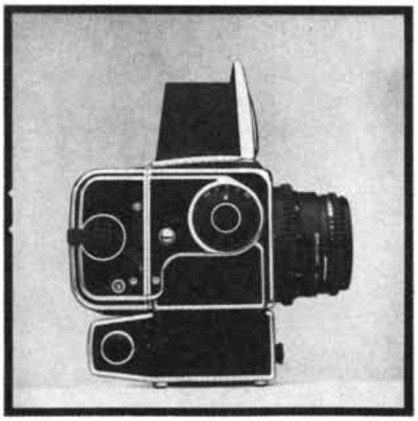
For that matter, no other system offers three different formats—6 x 6cm, 6 x 4.5cm, and a 4.5 x 4.5cm superslide format for showing in any 35mm projector.

A lavish brochure on the 500EL/M and other Hasselblad cameras is available free if you write: Braun North America, Dept. SA1H, 55 Cambridge Parkway, Cambridge, Mass. 02142—a division of The Gillette Company and exclusive marketer of Hasselblad in the U.S.

It is actually possible to custom-build for yourself over 100 different Hasselblad cameras for almost every imaginable photographic situation—from the most simple of family snapshots to the most sophisticated of professional work.

Is it any wonder that the space program came to Hasselblad for the equipment which was to record man's first voyage to the moon?

HASSELBLAD®



The Motor-Drive Hasselblad 500EL/M, made in Sweden. The standard lens is a multi-coated f2.8 80mm Zeiss Planar.

Established 1845 **SCIENTIFIC AMERICAN** January 1979 Volume 240 Number 1

ARTICLES

DEPARTMENTS

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

executive health

the report that briefs you on what to watch

Volume XV, Number 3 • December, 1978 • Pickfair Bldg., Rancho Santa Fe, Calif. 92067 • Area 714:756-2600

New research findings . . .

ON ZINC . . . the amazing metal so *essential* to your health

Deficiency of it can cause endless problems many of them just now being realized

Almost half of all men at some point can expect a bit or a lot of prostate trouble. But if a group of Chicago investigators are right, much of it might be avoided by attention to the role of a metal, zinc, in prostate health. With zinc, the investigators report, they have been able to overcome inflammation and even enlargement of the prostate in many men.

It may seem a far cry from prostate disorders to some queer problems of taste and smell — no taste or smell at all or disturbing perversions of one or both. Yet zinc of late has been used with considerable success in overcoming such problems.

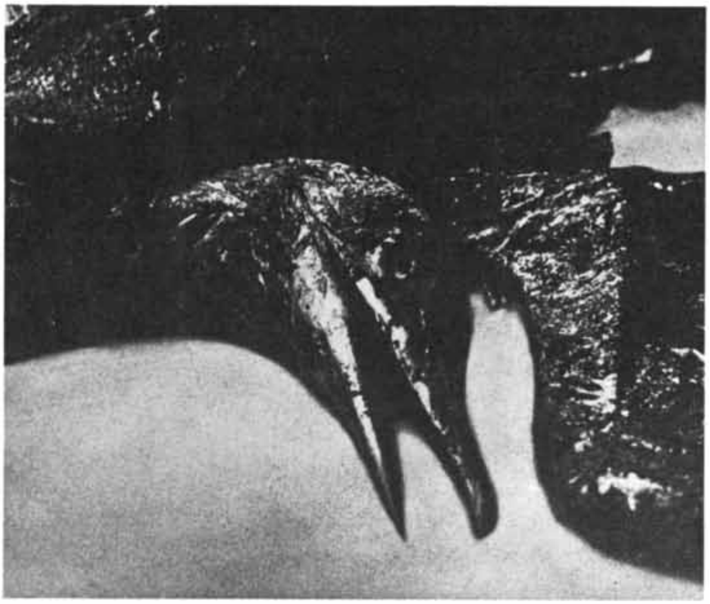
And these are just two of a considerable array of recent developments in the emerging story of the metal in human health and disease. It now appears that zinc may have notable potential for helping to overcome a wide variety of problems — ranging from slow wound healing to serious blood vessel disorders including atherosclerosis or "artery-choking," from proneness to infections to infertility, and from retarded growth and sexual maturation to a life-threatening skin rash.

Studies suggest that the metal, although common in many foods and long supposed to be in adequate supply, can be, for one reason or another, deficient in many people and, in fact, is to a surprising degree.

Zinc is one of a group of what are called trace metals, found in the human body in only tiny amounts, as little understood for the most part now as were vitamins 75 years ago, yet vital. Iodine is one: copper, another: chromium (See Executive Health, Vol. XI, No. 5) still another; and there are more.

Their combined weight in the body amounts to no more than an ounce. Yet, in their minute quantities, they play vital roles. The total of iodine in the body, for example, amounts to 30 milligrams (and a milligram is only one thirty-thousandths of an ounce). But

Visit the beautiful coast of France.



And see what the tide brought in.

NOVA goes to France to discover how one devastating oil spill has left its mark on the life of the Brittany coast. And what the chances are of the world's worst tanker accident happening again. Watch NOVA's **Black Tide.**

France to Utopia

NOVA's sixth season takes you from a beach in Brittany ... to the bottom of the sea where an international battle is being fought over mining rights ... to the high valleys of the Andes ... to the scientific breeding grounds of superathletes ... through the curved space of Albert Einstein ... to a utopian commune based on the behavioral theories of B. F. Skinner. NOVA takes you to worlds you've never seen.

Series of Surprises

Exciting journeys to unusual places. Rare encounters with fascinating people. Meetings with remarkable ideas.

Take NOVA to a new world every week.

NOVA

Season Premiere
Thursday, January 4
8:00 pm*

*Check local listings

PBS

A COMPANY CALLED
TRW

NOVA is made possible in part by a grant from TRW.

THE COVER

The design on the cover depicts the strings of the Steinway Model B piano that correspond to the notes in the middle range of the keyboard. Each key in this range sets in motion a triplet of strings that are not tuned precisely to one frequency. Such "mistuning" contributes in unexpected ways to the singing quality of the piano tone (see "The Coupled Motions of Piano Strings," by Gabriel Weinreich, page 118). When a piano key is depressed, a felt-covered hammer strikes the corresponding string triplet as the block of soft felt called a damper is lifted from the strings so that they can vibrate freely. The sound that emanates at first decays rapidly and then decays slowly. The mistuning couples the motions of the three strings in a way that affects the length of the slow decay.

THE ILLUSTRATIONS

Cover painting by Ted Lodigensky

Page	Source
20–28	Adolph E. Brotman
33	New York Public Library
34	Cedric S. Raine, Albert Einstein College of Medicine of Yeshiva University
39	Union Pacific Railroad Co.
40–47	Walken Graphics
49	David M. Knipe, Massachusetts Institute of Technology
50–63	George V. Kelvin
64	Floyd C. Clark, California Institute of Technology
65	Amnon Yariv and Pochi A. Yeh, California Institute of Technology, and Alfred Y. Cho, Bell Laboratories
66–72	Gabor Kiss
93–106	Allen Beechel
109	John Goeller, New York Eye and Ear Infirmary
110–117	Lorelle M. Raboni
119–127	Dan Todd
129–130	Tom Prentiss
131	Ilil Arbel
132–134	Tom Prentiss
136–138	Ilil Arbel
140	Malcolm R. Clarke, Marine Biological Association of the United Kingdom
142	Archives, California Institute of Technology
143	Archives, California Institute of Technology (*left*); Niels Bohr Library, American Institute of Physics (*middle*); Granger Collection (*right*)
144	Archives, California Institute of Technology
146–148	Jerome Kuhl
149–150	Archives, California Institute of Technology
151	Wide World Photos
153–154	James Bell
155–158	Michael Goodman

LETTERS

Sirs:

The astounding proposals of Philip Morrison and Paul F. Walker in "A New Strategy for Military Spending" in your October 1978 issue cannot go unanswered.

First and foremost, lest any of your readers missed the "punch line," there must be no misunderstanding about what the authors have proposed—truly massive cuts in U.S. defense expenditures: General Purpose Forces by more than 40 percent; Research and Development by more than 60 percent, and Strategic Forces by 90 percent. The cuts would come at a time when the Soviet Union is outspending the U.S. by as much as 40 percent and after an 18-year period in which annual Soviet expenditures have nearly doubled while U.S. expenditures have remained level in constant dollars.

Cuts on that scale go well beyond the often-debated line between "fat" and "muscle"; they would cut the heart out of the nation's defenses. And what rationale do the authors offer to justify such a radical change?

(1) "Most informed observers are satisfied there is a crude balance of forces along the frontiers of Europe."

The only attempt at substantiation of this assertion is a bar chart breaking down U.S. and Soviet manpower in such a way as to give the casual reader the impression that we alone can throw 1.9 million men into a "European confrontation"—almost as many as the Soviets' 2.1 million. Presumably that 10 percent disparity is what the authors mean by a "crude balance of forces" (although their proposal to cut U.S. manpower by a third would presumably increase the Soviet advantage to more than 60 percent, which some of us might crudely term an imbalance).

In any event the manpower comparison is hopelessly simplistic. It takes no account of allies on either side. It takes no account of our greater need for naval forces. It takes no account of differences in firepower or equipment on either side. It takes no account of how this balance is expected to change in the next few years. And particularly it takes no account of dynamics—how much force and equipment either side maintains as a function of distance to the inter-German border, what the logistic capacities of the road and rail nets are, how much warning we might have of a Warsaw Pact attack on NATO, etc., etc.

The primary problem facing NATO today, save in the most unlikely circumstances, has nothing whatever to do with the fully mobilized manpower of the U.S. and the Soviet Union. It has to do with making sure that NATO is not overwhelmed in the first week or so of a blitzkrieg war—an issue the authors simply do not address.

(2) "Military technology based on microelectronics . . . promises a dramatic advantage to the defense."

It is true that modern Precision Guided Munitions (PGM's) are improving defenses, but it is far from clear that PGM's will at some point in the future swing the balance in favor of the defense, thus making a Warsaw Pact attack in Europe less likely. The family of PGM's already includes weapons that greatly improve not the defense but the offense: antiradiation missiles that destroy defensive radars; cruise missiles (which the authors would presumably abandon along with our bomber force) of unprecedentedly high accuracy and low vulnerability; the Pershing II that brings terminal homing to ballistic missiles. The authors inadvertently make the case themselves. One of their illustrations shows an A-10 launching a laser-guided Maverick missile against an enemy bunker. The caption says: "New battlefield technology increasingly favors the *defense* over the offense" (emphasis added). Whether attacking fixed bunkers is more typical of the offense or the defense I shall leave as an exercise for the student.

Another factor the authors have failed to consider in connection with PGM's is that they are not the exclusive property of the West—and the proposed cut of more than 60 percent in the U.S. R&D budget that has supported the development of these weapons they find so intriguing is hardly likely to ensure future competitiveness in PGM's.

(3) "A couple of hundred missiles would suffice as a second-strike deterrent."

Again the authors argue against themselves. In two of the three indicators in which they claim a U.S. superiority (throw weight and megatonnage) our lead can clearly be seen from their bar chart as due solely to the U.S. bomber force. Yet referring to the bombers only two pages prior to that the authors flatly state that "in a missile age they are less than relevant."

The authors also have a mistaken notion that the triad (e.g., SLBM's, ICBM's and bombers) is based solely on a "hedge against some unforeseeable breakthrough in antimissile technology." It hedges, of course, not only against just that but also against unforeseen enemy breakthroughs in antisubmarine warfare (which the authors quite incorrectly allege "no one currently takes seriously"), in bomber defense, in defense against cruise missiles, hedges against failures in our own warning or communications systems and against unsuspected failures in any of the three components of our triad or ultimately errors in the ways we have planned to use those forces resulting from our (thank God!) inexperience in fighting a nuclear war. I might note, incidentally, that the authors' strategic force is completely vulnerable to an anti-ballistic-missile breakthrough, or breakout, or covert deployment. Is that prudent?

As to the authors' contention that "a couple of hundred missiles [meaning, as is clear from the text, warheads] would suffice as a second-strike deterrent," such insight into the Soviet mind, such assurance as to what would deter the Soviet leadership, particularly in times of deep crisis, is dazzling to contemplate. This is a classic example of what national-security scholars refer to as "mirror imaging"—assuming that whatever would deter us (in this case whatever would deter Messrs. Morrison and Walker) would certainly also deter the Soviet leadership. But of course we cannot know that, and neither Soviet writing on the matter nor U.S. intelligence offers much hope that the Soviet leadership views nuclear war in the same light that we do.

A glaring omission is the authors' failure even to mention the possibility of Soviet coercion of us or our allies, given the kind of strategic-force imbalance implied on the one hand by the continuing Soviet buildup and on the other by the authors' recommended cut of a

Scientific American, January, 1979; Vol. 240, No. 1. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, vice-president and treasurer; C. John Kirby, vice-president; Arlene Wright, assistant treasurer.

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

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

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 40 cents.

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

Name ______________________

New Address ______________________

____

____

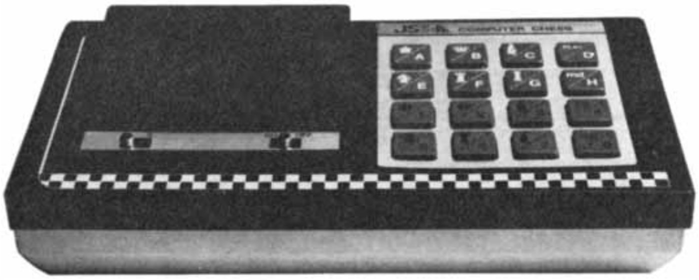
Old Address ______________________

____

____

CHESS DIPLOMACY

Soviet Challenge



This is the computer that may change the course of chess playing history.

Can an American chess computer beat the Soviet Chess Champion? A Confrontation between American space-age technology and a Soviet psychological weapon.

The Soviet Union regards chess as a psychological weapon, not just a game. It is a symbol of communism's cultural struggle with the West.

So when Russian Anatoli Karpov competed against the Russian Defector, Victor Korchnoi, he had the entire Soviet Union's resources at his disposal, including a hypnotist and neuro-psychologist.

Karpov won. And with it the world's undisputed chess championship. Karpov however, has never confronted American space-age technology and in particular JS&A's new Chess Computer.

So representatives of JS&A met with Karpov's representatives in Hong Kong in an effort to arrange a match between the Soviet Champion and the JS&A Chess Computer.

It wasn't easy negotiating with the Soviets. We offered them a $50,000 guarantee against royalties from the sales of our chess computers. But negotiations broke down.

Was the Soviet delegation afraid that American space-age technology would win? Were the Soviets fearful of negative publicity if Karpov lost to a $100 computer? Or were they fearful of a circus-type atmosphere that would degrade their prestige, even if he won?

Honestly, we don't know. We do know that our offer is still open, but we suspect Karpov will not accept.

Why did we challenge Karpov? Simple. We thought that having Karpov play against our computer would focus world-wide attention on our product. This attention would increase its sales and win or lose, we would sell more computers.

We had to sell more computers. We wanted to sell our unit for $100 even though it compares with units that sell for more than $300. But we had to do two things in order to sell our unit for $100. First, we had to manufacture it in Hong Kong where labor costs are very low. Secondly, we had to sell large quantities.

SOPHISTICATED DESIGN

The JS&A Chess Computer is designed to look several moves ahead to determine its next move. When we first designed it, it played five levels of chess. Level one was for beginners and as you played against the computer, you could increase its level of difficulty until the computer became more of a challenge. Level five was quite a challenge.

We thought we had the ultimate unit with five levels, until we developed our most sophisticated unit which has six levels. With six levels and all its previous features, the system is now a challenge for any Soviet Chess Champion.

The JS&A Chess Computer is a small unit that comes without a board or chess pieces. We felt that most players prefer their own board and pieces anyway.

LIKE PLAYING KARPOV

The system is the perfect way to sharpen your chess skills. It not only has six different skill levels, but if you are playing against the computer at level two and you are beating it, you can switch the unit to level six. It's like having Karpov as your new opponent–right during mid game.

To play against the computer, you enter your move on the unit's keyboard. You then wait until the computer examines all its options and selects its move You then move the computer's chess piece to correspond with its request as shown on the display. A board layout is provided to show you where each chess piece should be moved.

SHARPEN SKILLS

If you already play chess, the JS&A unit provides a new chess dimension. If you haven't played chess, the system is a good way to learn and sharpen your skills.

The JS&A Chess Computer measures only 2⅛" x 4⅞" x 8⅞" and weighs just a few ounces, so if service is ever required you can slip it in its handy mailer and send it back to our prompt service-by-mail center. Service should never be required, but it is reassuring to know that service is an important consideration in this program.

JS&A is America's largest single source of space-age products–further assurance that your modest investment is well protected.

We suggest you order a JS&A Chess Computer on our 30 day trial period. Play against it. Raise or lower the level as you play and watch how the computer's personality can change right in mid-game–from a tough competitor to a push over.

TEST LEVEL SIX

Test our level six and see if you'd have much of a chance against the Soviet Champion Karpov. Then, after you've really given it a workout, decide if you want to keep it. If not, you may return your unit for a prompt and courteous refund, including your $2.50 postage and handling charge. There is no risk. Each JS&A Chess Computer comes complete with instructions and an AC adapter (no batteries are required).

To order your JS&A Chess Computer, send your check for **$99.95** plus $2.50 for postage and handling (Illinois residents please add 5% sales tax) to the address below or credit card buyers may call our toll-free number below.

The Soviet Union may have the World's Chess Champion, but JS&A has a very powerful Chess Computer and something the Soviets don't have–a pretty good advertising department.

Why not order a JS&A Chess Computer at no obligation, today.

JS&A PRODUCTS THAT THINK®

Dept. SA One JS&A Plaza
Northbrook, Ill. 60062 (312) 564-7000
Call TOLL-FREE 800 323-6400
In Illinois Call .... (312) 564-7000
©JS&A Group, Inc., 1978

fourth (with more to follow) of our strategic submarine force, *90 percent* of our ICBM force and *all* of our bomber (and thus our cruise missile) force. The effect of such a unilateral renunciation of the will to remain strong, free and independent on the part of the leader of the free world is awesome to contemplate.

The authors' concern about nuclear war is commendable. But to place the blame for an "arms race" on our own country, to suspect evil influences in our leadership, to counsel for weakness in the face of growing Soviet strength, to doubt our ability to govern ourselves wisely, and to retreat into pious hope instead of forthright action is to serve this nation and the world poorly.

RUSSELL MURRAY, 2ND

Program Analysis and Evaluation
Office of the Secretary of Defense
Department of Defense
Washington, D.C.

Sirs:

Our article on a new strategy for U.S. military spending has drawn many responses from readers, whose characterizations of the proposals made in the article range from "tantamount to national suicide and the death of Western civilization" to broad concurrence with our arguments. We can report that many of the replies were sharply critical, not always but often from serving officers of the U.S. armed forces.

The text of the article began with the general principles for long-run stability we think are required in a world that grows no larger in geographic space but expands in firepower and military targets. We argued that "the U.S. buy as much force as it needs but not more." The article showed that "what matters is not input but output," for example not numbers of equipment and missiles but rather the potential amount of defense and destruction a military force can muster. In the strategic nuclear sphere the article strongly supported a policy of mutual assured destruction, and in the conventional sphere it forecast a *future* trend, a decade or two ahead, toward the ascendancy of the defense offered in land and sea battles by the rise of precision-guided munitions. The last columns of the article and the final tables showed our current proposals for a U.S. military budget of the 1980's; this included a sharp reduction in spending, based much more on our appraisal of current weapons and troop deployments than on the "smart" weapons of the future. Some of our critics unfortunately saw a direct connection between PGM's and major budgetary reductions that we did not put forward. What we were saying is that the future will see a profound change toward small mobile land, sea and air units. As one thoughtful critic, Air Force Major Duncan L. Dieterly, wrote: "The scenario of the future dictates mobility and compactness as the basic concept: a technologically sophisticated guerrilla force rather than fixed-base, high-density, low-mobility systems such as permanent bases, aircraft carriers, massive tanks or slow bombers. It will be a battle of shifting forces attempting to locate and destroy small, mobile firepower units. Only weapons systems operating in three dimensions, such as aircraft and submarines, will remain viable; all other systems will be high-risk, high-loss and short-duration systems."

A few careful students of history took our account of the surprise sinking of the Israeli destroyer the *Elath* by a Styx missile as an assertion that such "smart" missiles are invincible and called us to task, invoking the later Israeli successes in 1973 against such missiles as evidence. We need not take up the argument in detail, but the Israeli navy is well informed and may indicate the future. At the time of the loss of the *Elath* the Israeli navy had (considering only surface combat ships of more than 100 tons' displacement) three destroyers and a patrol gunboat. It now has in the same category 18 fast missile attack craft with Gabriel "smart" missiles; each of these vessels has less than a fourth the displacement of the oldest Israeli destroyer of a decade ago. That change is the essence of the point we were making, and it will continue.

We are pleased to see an official Department of Defense response to our arguments, here through Mr. Murray, the head of the Program Analysis and Evaluation office, and to note his emphasis on our proposed budgetary reductions. He is correct in stating that our proposal is not solely a critique of waste or goldplating but more a reexamination of the means and ends of U.S. military policy as a whole.

We regret that such a responsible official is willing to emphasize input above output; surely the point is not how much we or the Russians increase or decrease spending but rather what we each can accomplish. This is what Mr. Murray's predecessors in the early 1960's sought to do in instituting "systems analysis" techniques under Secretary of Defense McNamara, that is, to set "yardsticks of sufficiency" to overcome the past lack of quantitative and qualitative "standards of adequacy."

Mr. Murray will find in our book a full account of various indexes of balance in Europe, of allies, navies, dynamics and so on; clearly a subject so complex cannot be covered in one article. But we must express surprise at his "primary problem": "the first week of a blitzkrieg war" in Europe. We need not comment here on the true likelihood of this logically conceivable event in modern Europe. We should like to ask just what contribution is made to solving this problem by those particular systems of the general-purpose forces we recommend for deep cuts: the light Army divisions and the Marines, the amphibious forces, the underway-replenishment groups and the aircraft carriers? If Program Analysis and Evaluation has in fact recognized the primary problem, it has a major task to modify the U.S. posture quickly. Indeed, our own proposals improve the quick-reaction defense of Europe by pre-positioning more armor in Europe and rebasing most of the powerful combat air strength now held on the carriers to land bases in support of NATO; this would seem to lessen what Mr. Murray perceives as the primary danger.

The discussion of nuclear strategy we gave stands. Ringing phrases such as Mr. Murray's "will to remain strong, free and independent" do not justify the dangerous and costly plans for "fighting a nuclear war." The deterrence of nuclear war is the safest stance for our country and the world, not elaborate counterforce plans. The land-based missiles and the strategic bombers are quickly becoming obsolescent. The domain of reasonable discourse should be to debate the size of a deterrent force, not to discuss blame and suspicion. We stuck close to fact in our argument; the U.S. *is* largely responsible for setting the technical pace of the arms race. That is a matter of fact, not a question of faltering loyalties. While ours remains a government by the people, the citizens do well to question openly the wisdom of official policy.

Former Secretary of Defense McNamara stated in 1963 that "you cannot make decisions simply by asking yourself whether something might be nice to have. You have to make a judgment on how much is enough." We have sought to follow that dictum by expanding the spectrum of question and debate in national security matters while recognizing the increasing human and material burden of military spending.

PHILIP MORRISON

Professor
Department of Physics
Massachusetts Institute
of Technology
Cambridge, Mass.

PAUL F. WALKER

Program for Science
and International Affairs
Harvard University
Cambridge, Mass.


We reinvented the 35mm camera so others couldn't catch up. They haven't.
When Olympus examined the design concepts of traditional 35mm SLR's, they recognized the need for an engineering revolution: achieve quality without bulk, sophistication without complication. The result is called Olympus OM-1. The state-of-the-art SLR embodying the latest technology in electronics, metallurgy and optics. So advanced that others, while attempting to copy its compactness, have failed to duplicate its sophisticated internal design. ▪ OM-1 is lightweight, yet professionally-rugged. Compact, but with a viewfinder 70% brighter and 30% larger than conventional SLR's. And its special air dampers make shooting exceptionally quiet and vibration-free. All this, part of the most complete, compact SLR system, with almost 300 lenses and accessories to meet every photographic challenge. No wonder OM-1 is Number 1.
For the photographer demanding the last word in automatic exposure control, the one choice is the Olympus OM-2. A quantum leap ahead of traditional SLR design, incorporating the features of the OM-1 plus electronic sensor circuitry found in no other camera's light measurement system. It's called OTF ("Off-the-Film") light measurement, because it measures the light reflected off the film plane and the film itself, during actual exposure. The result is a range of capabilities found in no other camera in the world, including automatic electronic flash whose exposure is controlled by the OM-2's internal light sensors. And each frame is exposure-controlled with motor drive (5 fps!) and rapid winder (over 3 fps!). Learn more about the incomparable choice offered by the OM-1 and OM-2, detailed in our free brochure. Write OLYMPUS, Woodbury, New York 11797.
OLYMPUS
OLYMPUS OM-1
OLYMPUS OM-2

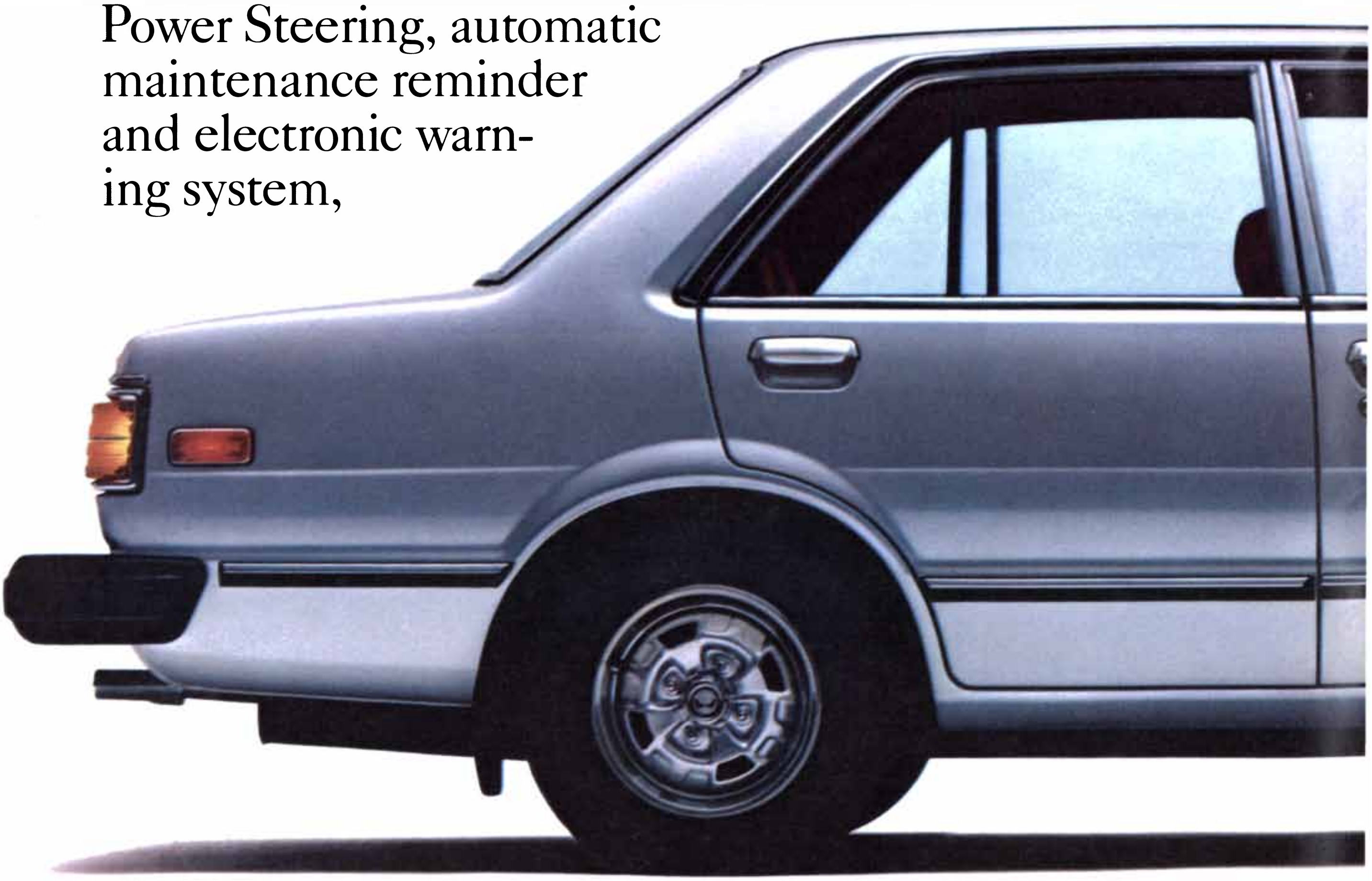
4-door simplicity.

Announcing the 1979 Honda Accord® 4-Door Sedan. This is, quite simply, the Accord that a lot of people have been waiting for.

For one thing, it's the first Accord to have four doors. And that's not all. This car also has a trunk.

The outstanding feature of the trunk is its roominess. But it also has an interior light, two stowage boxes for small items, and a lever that allows you to release the trunk lid from the driver's seat. So if you've been waiting to buy a Honda because you wanted a conventional trunk, it's been worth the wait.

Other standard features include Honda's Variable Assist Power Steering, automatic maintenance reminder and electronic warning system,

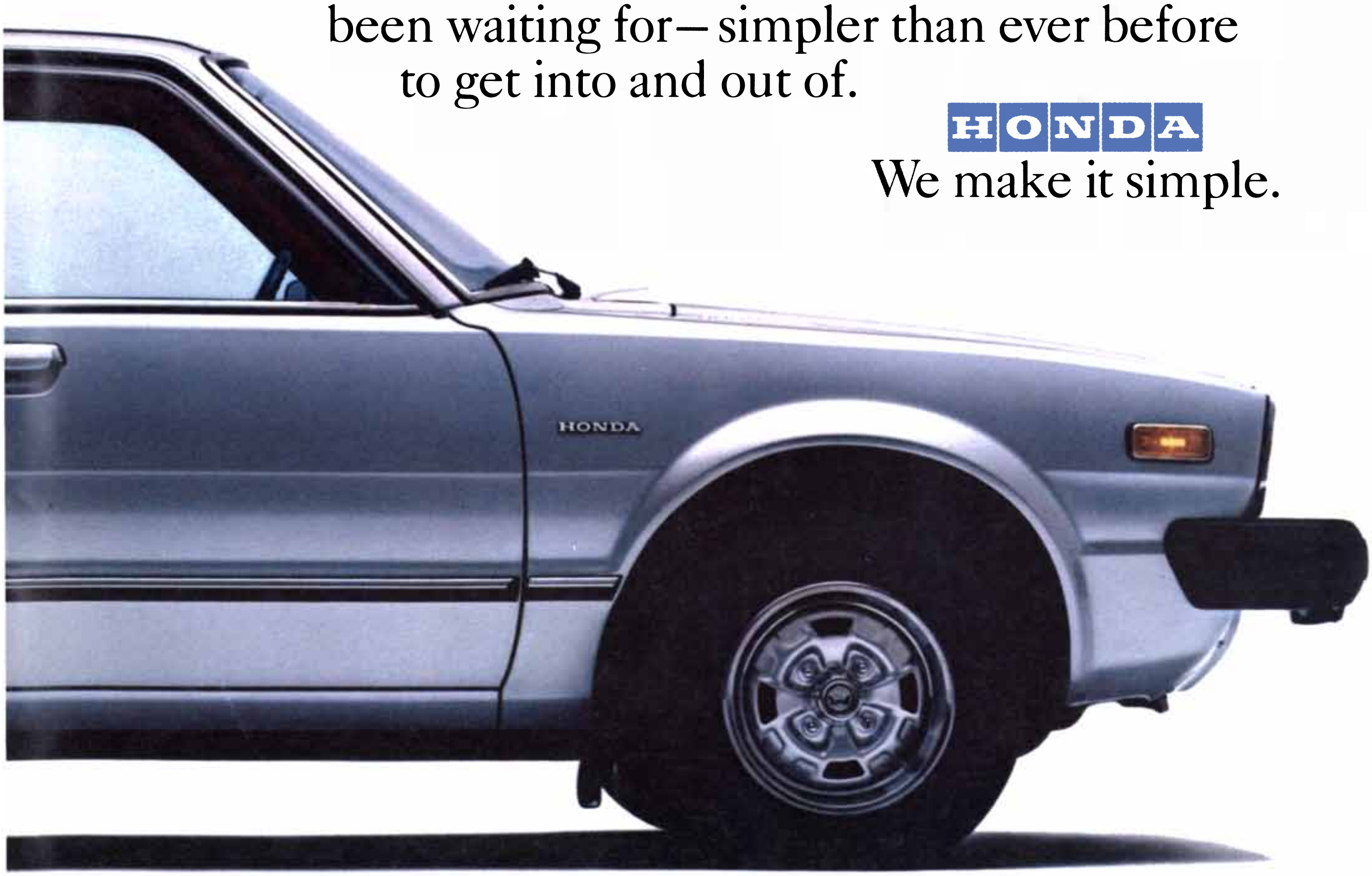


defrosters for front-door windows and rear window, tinted glass all around, tachometer, remote control rear-door locks, AM/FM radio, steel-belted radial tires, and 5-speed stick shift. Our 2-speed manually-selected Hondamatic transmission is optional.

Inevitably, the debut of such an automobile will cause some people to wonder if our devotion to simplicity is on the wane. Far from it. The Accord 4-Door Sedan carries on such Honda traditions as front-wheel drive, transverse-mounted CVCC® engine, and four-wheel independent MacPherson strut suspension.

So rest assured that even with all its doors and all its trunk space, the Accord 4-Door Sedan is all Honda. Simple to drive, simple to own, simple to enjoy. And—what you may have been waiting for—simpler than ever before to get into and out of.

HONDA
We make it simple.




THE SEAGRAM'S GIN CRYSTAL MARTINI.
Seagram's Extra Dry
THE PERFECT MARTINI GIN
Seagram's
Extra Dry
Gin
DISTILLED AND BOTTLED BY
LAWRENCEBURG, IND., RELAY, MD
750 ML (25.4 FL OZ) · 80 PROOF
DISTILLED DRY GIN
DISTILLED FROM AMERICAN GRAIN
Fill your glass with ice cubes made from natural spring water. Add Seagram's Extra Dry Gin, the real secret behind a perfect martini. And remember, enjoy our quality in moderation.
Seagram's Extra Dry. The Perfect Martini Gin. Perfect all ways.
SEAGRAM DISTILLERS COMPANY, N.Y.C. 80 PROOF. DISTILLED DRY GIN. DISTILLED FROM GRAIN.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JANUARY, 1929: "The main topic of astronomical interest is the great 200-inch reflecting telescope that is to be constructed for the California Institute of Technology. It is of vital importance that the new telescope should be located in a region of settled weather, out of the normal track of the great cyclonic storms that sweep over the northern and central parts of the United States. The southwestern part of the country is best situated in that respect. The observatory must also be located in a habitable place. Although the astronomers will doubtless come for observing 'spells' of a few days, returning to work up their photographs while others take their place at the telescope, a staff of mechanical and other assistants will establish their permanent homes close to the great dome. It has been tentatively decided, therefore, to place the telescope within a night's ride or thereabouts of Pasadena. This still leaves a very large extent of country open for investigation, including the best climatic zone (astronomically speaking) in the U.S., and a great variety of topography."

"Britain is of the opinion that we want to challenge her superiority on the seas. She adopted a suspicious attitude toward our proposals at Geneva and has consistently shown her distrust of our motives. Britain and many people in our country seem to have forgotten that when at the time of the Washington conference in 1923 we had unquestioned potential superiority in battle cruisers and battle-ships, built and building, we indicated our intention of avoiding competition and the resultant enormous expenditures. We could have gained superiority by simply continuing construction of ships already authorized or building and could have maintained it easily for a long time to come. Instead we agreed to scrap powerful ships that had cost us many millions of dollars. When Congress convenes, the second bill to be considered will be the revised Navy building program. This bill, if passed, will authorize the President to undertake construction of one aircraft carrier and 15 light cruisers. As further evidence of our peaceful intentions, the bill carries the proviso that 'in the event of international agreement for (further) limitation of armaments the President is authorized and empowered to suspend in whole or in part any of the construction authorized by this Act.'"

"According to a report emanating from the War Department, airplanes have found a new field of usefulness. The report states that in a demonstration at Brooks Field in Texas recently machine-gunners, together with their gun, were dropped from an airplane and on landing went into action in less than three minutes. In the demonstration six men jumped from a bombing plane, each with his individual parachute, picked up the machine gun, which had also been dropped by parachute, set it up and started firing. In this manner it is possible that a number of guns and their crews could be dropped and placed in action in commanding positions. The value of the maneuver in warfare would lie in the large element of surprise it contains. Its strategic value may prove to be very great."

SCIENTIFIC AMERICAN

JANUARY, 1879: "Mr. Edison's application for a patent on an electric lamp is before the Commissioner and is taking its regular course. According to the rules of the Patent Office, nothing concerning it can be divulged. It is understood, however, that it is progressing favorably. Mr. Edison has already received seven patents bearing on the electric light and has filed three caveats. Five more similar applications are now under way. He has had a man in the Astor Library searching the French and English patent records and scientific journals, from the earliest dates down to the past fortnight, and says nothing like his arrangement has been revealed. Mr. Edison is making elaborate preparations to introduce an experiment with the electric light. He purposes to commence at Menlo Park with 2,000 lights, using telegraph poles with 15 lights on each arm. This experiment, including the cost of the buildings, engine, generating machine and everything, is estimated at from $100,000 to $125,000."

"The Philadelphia Local Telegraph Company has perfected an arrangement putting their clients in the various parts of the city into immediate telephonic connection. This is done by means of an ingenious telephonic switch board recently devised. As described by a local paper, the front of the apparatus consists of a walnut frame and bright strips of brass, punctured with holes, into which wires are fitted to make the necessary connections. Behind all this all the wires converging in the office concentrate. The board just put into operation accommodates no fewer than 400 different lines, which have an aggregate length of 1,000 miles, thus placing each firm or individual having telephonic connection with the main office in direct communication with 399 other persons scattered over the city."

"Mr. John Muir has an interesting paper titled 'New Sequoia Forests of California.' He gives therein the details of a discovery by himself of a grand forest of *Sequoia* 70 miles long, lying considerably south of the isolated groups hitherto known and containing large numbers of saplings, which indicate that the species is still in a vigorous state of existence. It has heretofore been argued that the few groups of these trees known made it probable that the species was dying out from its last strongholds upon the earth, or that it has come down to us from preglacial times, when it existed in Europe also, as geology testifies. Mr. Muir's researches lead him to believe that the species has never been more extensively distributed on the Sierra in post-glacial times than it is now, and that to-day it is as full of life and vigor as it was 10,000 years ago."

"Loss of life by yellow fever in the South last year is estimated at about 15,000 persons, and of money and trade at from $175,000,000 to $200,000,000, as great as the loss from the Chicago fire. But some good is likely to come out of the calamity. It is thought that henceforth quarantine regulations will be more thoroughly established than they have ever been. Apart from death and human suffering, negligence is the worst kind of economy. Expenditure of a twentieth part of what the fever has cost might have prevented it altogether."

"Mr. Darwin, in his *Voyage of a Naturalist,* thus described a crab that makes its diet of cocoanuts and that he found on Kneeling Island in the South Seas: 'It is common on all parts of this dry land, and grows to a monstrous size. It has a front pair of legs terminated by a strong and heavy pincers and a last pair of legs terminated by other pincers that are narrow and weak. It would at first be thought quite impossible for a crab to open a strong cocoanut covered with a husk; but Mr. Liesk assures me that he has repeatedly seen the operation effected. The crab begins by tearing the husk, fiber by fiber. When this is completed, the crab commences hammering on the cocoanut till an opening is made; then, turning its body around with the aid of the narrow pair of pincers, it extracts the albuminous substance. I think this is as curious a case of instinct as I ever heard of, and likewise of adaptation in structure between two objects apparently so remote from each other in the scheme of nature as a crab and a cocoanut.'"

THE AUTHORS

EDWARD D. GRIFFITH and ALAN W. CLARKE ("World Coal Production") are economists specializing in energy consumption and demand. Griffith is senior consultant in policy analysis and forecasting in the Corporate Planning Division of the Atlantic Richfield Company. He studied economics at Cornell University and at the London School of Economics and received his M.B.A. degree in finance at Columbia University. At Atlantic Richfield he analyzes current and proposed Government energy policies and their impact on the economy and the energy industry. Clarke is head of energy information and analysis for the Shell International Petroleum Company in London. Since joining Shell in 1952 he has held a variety of positions concerned with energy-demand analysis and forecasting. From 1969 to 1971 he worked in the economic department of Shell Française in Paris, returning to London to take part in Shell's energy-demand studies. A statistician by training, he is currently developing a "worldwide data base on energy consumption," with emphasis on the developing countries.

HARVEY F. LODISH and JAMES E. ROTHMAN ("The Assembly of Cell Membranes") have collaborated on biochemical studies of cell-membrane structure. Lodish is professor of biology at the Massachusetts Institute of Technology and consulting scientist in hematology and oncology at the Children's Hospital Medical Center in Boston. He was graduated from Kenyon College with bachelor's degrees in chemistry and mathematics and obtained his Ph.D. from Rockefeller University in 1966. After two years as a postdoctoral fellow at the Medical Research Council Laboratory of Molecular Biology in Cambridge, England, he joined the M.I.T. faculty. Rothman is assistant professor of biochemistry at the Stanford University School of Medicine. He studied physics at Yale College, and in his senior year he did full-time independent research in molecular biophysics as a Scholar of the House, receiving the DeVane Award in 1971 for the best thesis by a student in the program. After being graduated from Yale summa cum laude he entered the Harvard Medical School, completing the requirements for both the M.D. and the Ph.D. degrees in 1976. Rothman spent two years in Lodish's laboratory as a postdoctoral fellow before moving to Stanford last fall.

AMNON YARIV ("Guided-Wave Optics") is professor of engineering and applied physics at the California Institute of Technology. Born in Tel Aviv in 1930, he is a veteran of the Israeli war of independence. He came to the U.S. in 1951 to study engineering at the University of California at Berkeley, obtaining his B.S. in 1954 and his Ph.D. in 1958. The following year he joined the research staff of Bell Laboratories and worked on the development of a number of early solid-state laser systems. He became a member of the Cal Tech faculty in 1964. Yariv's current research interests include nonlinear optics, free-electron lasers, semiconductor lasers and integrated optics. He is the author of *Quantum Electronics* (1967) and *Introduction to Optical Electronics* (1971).

THOMAS H. JORDAN ("The Deep Structure of the Continents") is associate professor of geophysics at the Scripps Institution of Oceanography of the University of California at San Diego. He received his Ph.D. in geophysics and applied mathematics from the California Institute of Technology in 1972 and was on the faculty of Princeton University for three years before moving to Scripps in 1975. Jordan's primary research interests are in the fields of seismology and tectonics; his work has been aimed at "elucidating dynamical processes within the earth by the seismological study of earth structure."

A. TERRY BAHILL and LAWRENCE STARK ("The Trajectories of Saccadic Eye Movements") are bioengineers interested in the control of movement. Bahill is assistant professor of electrical engineering and bioengineering at Carnegie-Mellon University and assistant professor of neurology at the University of Pittsburgh School of Medicine. He was educated at the University of Arizona, San Jose State University and the University of California at Berkeley, where he obtained his Ph.D. in 1975. As director of the Neurological Control Systems Laboratory at Carnegie-Mellon, Bahill studies the control of eye movements and locomotion in man. He is also investigating the development of walking in infants, with his one-year-old son Alexander serving in the pilot study. Stark is professor of physiological optics and engineering science at the University of California at Berkeley and professor of neurology at the University of California at San Francisco Medical Center. He did his undergraduate work at Columbia University and received his M.D. from the Albany Medical College in 1948. After his internship and a research fellowship in England he was on the faculties of the New York Medical College, Yale University, the Massachusetts Institute of Technology and the University of Illinois. He moved to Berkeley in 1968. Stark's research involves computer analysis of the control of movement of the human eye and hand. He writes: "My hobby, bird-watching, brings together problems in the neurological control of flying and the cognitive pattern-recognition of bird species."

GABRIEL WEINREICH ("The Coupled Motions of Piano Strings") is professor of physics at the University of Michigan. He was educated at Columbia University, where he did his doctoral work in physics under I. I. Rabi. After seven years on the technical staff of Bell Laboratories he joined the Michigan faculty in 1960. His research has covered a broad range of topics, including the interaction of acoustic waves with electrons in semiconductors, which led eventually to the development of acoustic amplification. In 1975 he originated a course at Michigan on the physics of music. He writes: "Music has been an important part of my life since childhood. I am a self-taught pianist and took up the cello at age 40. My work on piano strings was done on my living-room piano with borrowed equipment in my spare time."

MALCOLM R. CLARKE ("The Head of the Sperm Whale") is a staff scientist in the laboratory of the Marine Biological Association of the United Kingdom in Plymouth. He studied zoology as an undergraduate at the University of Hull and went on to do graduate work on the gut of the sperm whale. His doctoral research involved working as a government whaling inspector for eight months aboard a British whaling-factory ship in the Antarctic. Of that experience he writes: "My early attempts to study the histology of the whale gut were foiled by the enormous size of the organ. A 50-foot sperm whale has a gut more than 700 feet long! Instead I collected parasites from the whales, including bile-duct tapeworms almost 70 feet long." After receiving his Ph.D. in parasitology in 1958, he did research on cephalopods at the National Institute of Oceanography at Wormley. He moved to Plymouth in 1971. Clarke's work on the buoyancy, migration, diving and parasites of sperm whales has involved more than 20 scientific cruises and visits to whaling stations in Norway, Madeira and South Africa.

DANIEL J. KEVLES ("Robert A. Millikan") is professor of history at the California Institute of Technology. He studied physics at Princeton University and went on to obtain his Ph.D. in history there in 1964. After working for a summer on the White House staff he joined the Cal Tech faculty. His historical research has focused on American science in the 19th and 20th centuries, and he is now writing a history of genetics and eugenics in the U.S. and Britain.

MATHEMATICAL GAMES

The diverse pleasures of circles that are tangent to one another

by Martin Gardner

When circles are tangent to one another, hundreds of beautiful problems arise, many of which have worked their way into the literature of recreational mathematics. Some historic examples have already been discussed in this department. This month I shall take up a few more. There is no room for the proofs, but if interested readers will take the theorems as challenges to find proofs, they cannot help but strengthen their understanding of elementary plane geometry.

We begin with a famous figure believed to have been first studied by Archimedes and known as the arbelos, from the Greek for "shoemaker's knife," because it resembles the blade of a knife used by ancient cobblers. It is the shaded region in the top illustration on page 20, bounded by the semicircles with diameters *AB, BC* and *AC. B* can be any point on *AC*. Here are a few of the most amazing properties of the arbelos:

1. The length of semicircular arc *AC* is, as is easily shown, equal to the sum of the arcs *AB* and *BC*.

2. Draw *BD* perpendicular to *AC*. The area of the arbelos equals the area of a circle with diameter *BD*.

3. *BD* divides the arbelos into two parts. Circles inscribed in each part are identical, each circle having a diameter of $(AB \times BC)/AC$. The smallest circle circumscribing these twin circles also has an area equal to that of the arbelos..

4. Draw a line tangent to arcs *AB* and *BC*. Tangent points *E* and *F* lie on lines *AD* and *CD*.

5. *EF* and *BD* are equal and bisect each other. This ensures that the circle with diameter *BD* passes through points *E* and *F*.

6. In the early 1950's Leon Bankoff, a mathematician by avocation, made a curious discovery. (Bankoff is a Los Angeles dentist who describes himself as "Russian by extraction." He edits the problems section of *The Pi Mu Epsilon Journal*, and he is known for his skill in composing and solving problems.) Bankoff improved on the discovery, credited to Archimedes, of the twin circles shown on each side of *BD* in the top illustration on page 20 as follows: Draw a circle (shown in broken outline) tangent to the three largest circles, and then draw a smaller circle (shown in color) that passes through *B* and the points where the broken circle touches arcs *AB* and *BC*. This circle also is identical with Archimedes' twin circles. Bankoff gives a proof in his article "Are the Twin Circles of Archimedes Really Twins?" (*Mathematics Magazine*, Vol. 47, No. 4, September, 1974, pages 214–218). No, they are not really twins, answers Bankoff; they are two circles in a set of triplets.

7. Construct inside the arbelos what is called a train of tangent circles. The broken circle in the top illustration on page 20 is the first in the train, which can be continued as far to the left as one wishes, in the manner shown in the bottom illustration on page 20. Label the circles C_1, C_2, C_3 and so on. The centers of all the circles in the train lie on an ellipse. The diameter of any circle C_n is $1/n$th the perpendicular distance from the center of that circle to the base line *ABC*. This remarkable result is in a fourth-century work by Pappus of Alexandria, who refers to it as an ancient theorem.

The proof of Pappus' theorem is simple if one uses inversion geometry, inverting the entire figure with *A* as the center of inversion. This converts semicircles *AB* and *AC* to parallel lines, and the train becomes a set of equal circles bounded by the two lines. You will find a good explanation of how this is done in Rodney T. Hood's article "A Chain of Circles" (*The Mathematics Teacher*, Vol. 54, No. 3, March, 1961, pages 134–137) and a briefer explanation in J. H. Cadwell's *Topics in Recreational Mathematics* (Cambridge University Press, 1966). Pappus did not know inversion methods (they were not developed until the 19th century) and so his proof is more cumbersome.

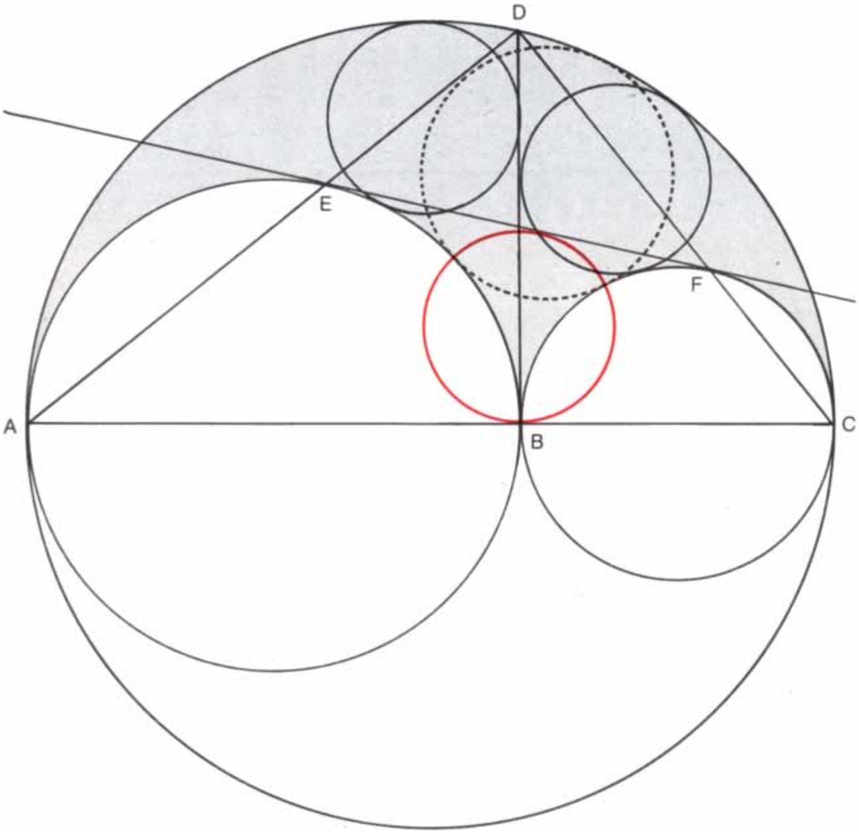
8. If *AB* equals 2 and *AC* equals 3, then the train of circles has many more surprises. The diameters of all circles in the train are rational fractions equal to $6/(n^2 + 6)$. Thus C_1 equals 6/7, C_2 equals 3/5 and so on. As Norman Pos recently pointed out to me, the center of C_2 lies on the diameter of the outside circle that is perpendicular to *AC*. Moreover, the centers of C_2 and C_3 are on a line parallel to *AC*, and that is also true of the centers of C_1 and C_6. The latter result is a special case of a more general theorem. If *AB* and *AC* are integral and *AC* equals $AB + 1$, the centers of every pair of circles whose subscripts have a product equal to $AB \times AC$ lie on a line parallel to *AC*. Hence if *AB* equals 3 and *AC* equals 4, circle pairs with subscripts 1 and 12, 2 and 6, and 3 and 4 all have centers on a line parallel with *AC*. (For a proof see M. G. Gaba's "On a Generalization of the Arbelos," *The American Mathematical Monthly*, Vol. 47, January, 1940, pages 19–24.)

9. If *B* divides *AC* in the golden ratio, many other striking properties result. These are discussed by Bankoff in his article "The Golden Arbelos" (*Scripta Mathematica*, Vol. 21, No. 1, March, 1955, pages 70–76). A note to publishers: Bankoff has an unpublished manuscript of 10 chapters on the arbelos, written in collaboration with the French mathematician Victor Thébault.

Closely related to the arbelos is a surprising theorem discovered by Jakob Steiner, a 19th-century Swiss mathematician, and depicted in the upper illustration on page 22. A small circle is drawn anywhere inside a larger one, and in the region between the circles a train is inscribed. In most cases the train will not exactly close to form a ring of tangent circles, that is, the end circles will overlap. In some cases, however, the train will form a perfect ring like the one shown with solid lines in the illustration. When this happens, the train is called a Steiner chain. What Steiner discovered was that if the two initial circles allow one Steiner chain, they allow an infinite number of chains. Put another way, no matter where you draw the first circle of the chain, if you add the other circles, the chain will always close exactly. An arbitrary second chain is shown in the illustration with broken lines.

As before, the easiest proof is by inversion geometry. One performs an inversion that transforms the two initial circles into concentric circles. The Steiner chain then becomes a chain of identical circles that fill the region between the concentric circles. Cadwell's book, cited above, gives the details.

Solomon W. Golomb, whose contributions to recreational mathematics often appear in this department, was on a trip through Europe last year and found himself carrying a variety of coins of different sizes. The following thought occurred to him. Suppose *n* coins of varying size form a closed chain that exactly surrounds a central coin, as is shown in the lower illustration on page 22. If the order of coins in the "wreath"



The arbelos, or "shoemaker's knife," of Archimedes

is permuted, will the coins still form a perfect wreath?

Most people guess yes, and there is even a "proof." Draw lines from the center of the interior coin that go between each pair of adjacent coins in the wreath, as is shown in the illustration. The sum of all these central angles must be 360 degrees, and this fact seems to be independent of the way the coins are arranged.

The proof is fallacious, because, as Golomb points out, it assumes that the radiating lines must be tangent to each pair of coins they pass between. That is not always the case, however, and when it is not, the order of coins in the wreath can make a difference. Of course, if there are three coins in the wreath, permutations will have no effect because they merely give rise to rotations or reflections of the original figure. When there are four or more coins, it is easy, Golomb discovered, to find examples where the wreath closes in certain permutations but not in others.

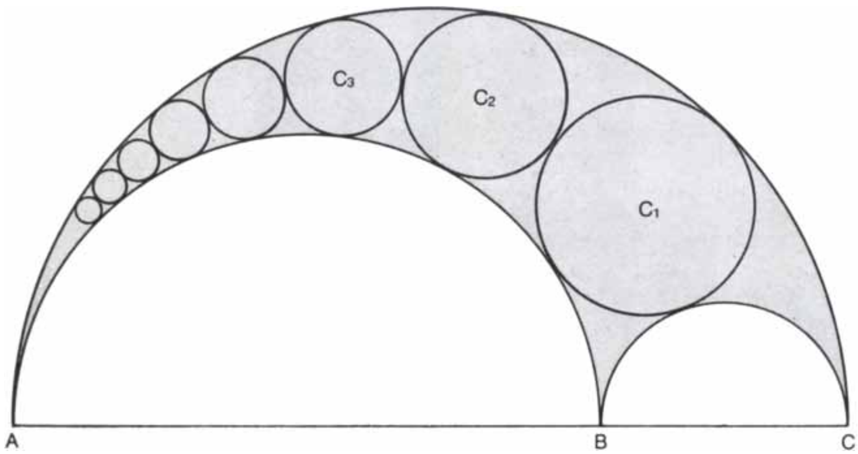
Differences between permutations are slight unless there are large discrepancies in coin sizes. Therefore if you arrange a half-dollar, a quarter, a nickel, a dime and a penny around a central circle, you will find that any permutation of the five coins seems to fit exactly. Nevertheless, the differences are there. I leave it to readers to find ways of proving this statement correct. You might also like to tackle another one of Golomb's discoveries. Given n coins no two of which are alike, what is the largest number of different interior circles they can exactly surround by permuting their order? The answer is $(n-1)!/2$. (The exclamation mark is a factorial sign.) Hence for four coins there are three permutations, for five there are 12, for six there are 60 and so on.

Golomb poses an interesting unsolved question. Given n coins of different sizes, what procedures will minimize and what procedures will maximize the size of the circle they can exactly surround? Golomb has some conjectures for both algorithms that he will soon be writing up in a paper on the subject.

A completely different kind of problem about touching circles, not well known, involves the packing of n identical circles, without overlap, into a specified boundary of the smallest area. This problem has practical aspects, because cylinders such as cans and bottles are often packed in containers with circular, square or other cross sections. What is the smallest area of the cross section that will make it possible to pack n cylinders? To formulate the problem another way, given the area of a region and n identical circles, what is the largest diameter of the circles that allows packing them into the region without overlap?

No general solution is known, even when the boundary of the region is as simple as a circle, a square or an equilateral triangle, and in each case optimal packings have been established only for very low values of n. When the boundary is a circle, proofs are known only for $n = 1$ through $n = 10$; they were first given in a 1969 paper by Udo Pirl. The cases $n = 2$ through $n = 10$ (taken from "Packing Cylinders into Cylindrical Containers," by Sidney Kravitz, *Mathematics Magazine*, Vol. 40, No. 2, March, 1967, pages 65–71) are shown in the illustration on page 24. The minimum diameter of the outside circle is given below each figure, assuming that the small circles are of unit diameter.

Kravitz supplies the best solutions he could find empirically for $n = 11$ through $n = 19$. The case $n = 12$ is of special interest. One would think that the close packing shown at the left in the top illustration on page 26 would be the densest, but Kravitz found the slightly better pattern shown at the right. Michael Goldberg, in "Packing of 14, 16, 17 and 20 Circles in a Circle" (*Mathematics Magazine*, Vol. 44, No. 3, May, 1971, pages 134–139) gives better packings for the four cases cited in his title. His packing for 17 circles was in turn improved by George E. Reis (*Mathe-*



The arbelos train

Discover

the most organized (and inexpensive) way to keep up with what's new in computer techniques and management —

The Library of Computer and Information Sciences

Take any 3 books for only $3.95

(values to $60.00)

if you will join now for a trial period and accept only 3 more books—at handsome discounts over the next 12 months.

(Publishers' Prices shown)

54455. IMS PROGRAMMING TECHNIQUES: A Guide to Using DL/1. *Kapp and Leben.* Spells out specific techniques to write application programs in ANS COBOL, PL/1, or Assembler Language in an IMS DL/1 data base environment. **$17.95**

41785. DEBUGGING SYSTEM 360/370 PROGRAMS USING OS AND VS STORAGE DUMPS. *D.H. Rindfleisch.* A superior guide to storage dump debugging. Methods for COBOL, FORTRAN, PL/1 and assembly language programs. Illustrations, examples, sample dumps. **$19.95**

52900. HIGH LEVEL COBOL PROGRAMMING. *Gerald M. Weinberg et al.* An advanced guide to proper and efficient COBOL usage. Introduces "sheltered programming" concepts, stresses modular design and maintainability of code. **$17.95**

45380. THE ENTREPRENEUR'S MANUAL: Business Start-Ups, Spin-Offs, and Innovative Management. *Richard M. White, Jr.* Techniques for uncovering lucrative markets and products; how to approach venture capital groups; more. **$15.00**

59920. MANAGEMENT: Tasks, Responsibilites, Practices. *Peter F. Drucker.* A brilliant task-focused handbook of the skills, techniques and tools a good executive needs to stay effective in modern management. **$17.50**

80150. STANDARDIZED DEVELOPMENT OF COMPUTER SOFTWARE. *Robert C. Tausworthe.* Shows you how to incorporate principles and techniques of structured programming in the development of large-scale systems. Over 100 well-chosen diagrams and flowcharts. **$19.95**

72690. THE PSYCHOLOGY OF COMPUTER PROGRAMMING. *Gerald M. Weinberg.* The classic study of computer programming as a human activity. Offers startling insights into the human factor of the programming environment. **$10.95**

59555. MAKING SYSTEMS WORK. *William C Ramsgard.* Examines specific, tested new tools for improved information processing. **$18.95**

62620. MICROCOMPUTER HANDBOOK. *Charles J. Sippl.* Detailed reference to the technology and applications of microprocessors. Expertly covers every phase, from software to the effect of micro technology on mainframe users. **$19.95**

60405. MANAGING A PROGRAMMING PROJECT. *Philip W. Metzger.* Sound advice on how to manage a project from the definition and design stage through testing to installation and operation. **$15.00**

81845. SYSTEM/370 JOB CONTROL LANGUAGE. *Gary DeWard Brown.* Introduction to JCL that also offers key information on the 370 operating system. Includes descriptions of hardware devices and access methods and reference to many useful JCL features. Softcover. **$12.50**

80845. STRUCTURED SYSTEMS DEVELOPMENT. *Kenneth T. Orr.* A practical handbook containing the basic steps for analyzing, developing, and installing systems with the emphasis on a structured "problem-solving" approach to systems building. **$17.00**

32485. ALGORITHMS + DATA STRUCTURES = PROGRAMS. *Niklaus Wirth.* A detailed explication of the logical steps necessary to program development and specification, employing minimum mathematics. A classic work on structured programming with practical examples. **$19.00**

40065-2. COMPUTER STORAGE SYSTEMS AND TECHNOLOGY. *Richard Matick.* A thorough examination of available storage devices. *Counts as 2 of your 3 books.* **$29.95**

48780. FORTRAN 77. *Harry Katzan, Jr.* Comprehensive coverage of the new FORTRAN standard language, from data types and constants to the latest I/O statements. **$16.95**

54465. I/O DESIGN: Data Management in Operating Systems. *Freeman and Perry.* Covers I/O devices, channel programming, device allocation, control statements, data organization and staging, and much more **$18.95**

62655-2. MICRO-ANALYSIS OF COMPUTER SYSTEM PERFORMANCE. *Boris Beizer.* A step-by-step approach to the construction, validation, and use of analytical models for determining the performance of a computer-controlled system. *Counts as 2 of your 3 books.* **$22.50**

41645-2. DATA COMMUNICATIONS: Facilities, Networks, and Systems Design. *Dixon R. Doll.* A straightfofward guide that covers everything from basic terminology to design trade-offs. Dozens of charts, diagrams and schematic drawings covering every topic. *Counts as 2 of your 3 books.* **$24.95**

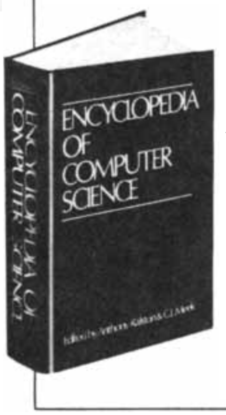
84406. TOP-DOWN STRUCTURED DESIGN TECHNIQUES. *Gloria Harrington Swann.* How to choose basic design tools, develop a work plan, write accurate and concise module specifications, and more. **$12.50**

85865. THE VAN NOSTRAND CONCISE ENCYCLOPEDIA OF MATHEMATICS. *Edited by W. Gellert et al.* 800-page survey of mathematical ideas and methods lists hundreds of techniques, worked examples, and practical applications. 950 illustrations. **$18.95**

78350. SIMULATION WITH GPSS AND GPSS V. *Bobillier, Kahan, and Probst* A detailed approach to the effective planning and execution of simulation experiments and projects. With concrete examples. **$18.50**

54450. THE IBM 5100 PORTABLE COMPUTER: A Comprehensive Guide for Users and Programmers. *Harry Katzan, Jr.* Includes BASIC and APL programs for payroll/inventory control applications. **$19.95**

EXTRAORDINARY VALUE!

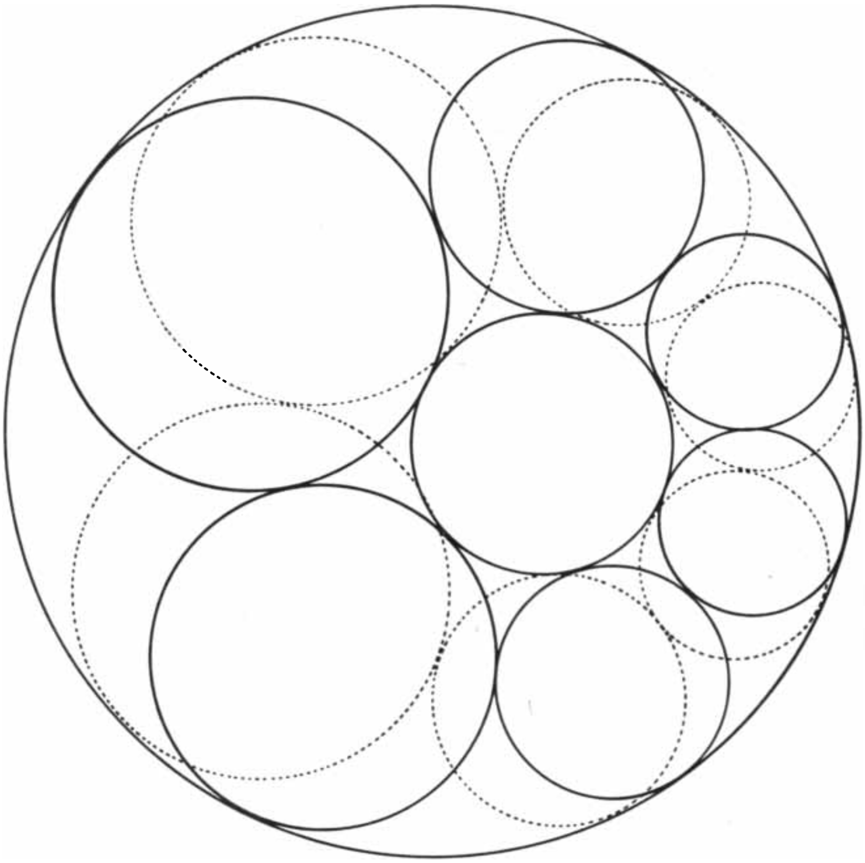


44900-3. ENCYCLOPEDIA OF COMPUTER SCIENCE. *Edited by Anthony Ralston and Chester L. Meek.* More than 1550 outsize pages of information on every aspect of computer science—from algebra to automata theory, from basic terminology to string-processing languages. 470 articles by 208 experts. Over 700 charts, tables, graphs, and diagrams. *Counts as 3 of your 3 books.* **$60.00**

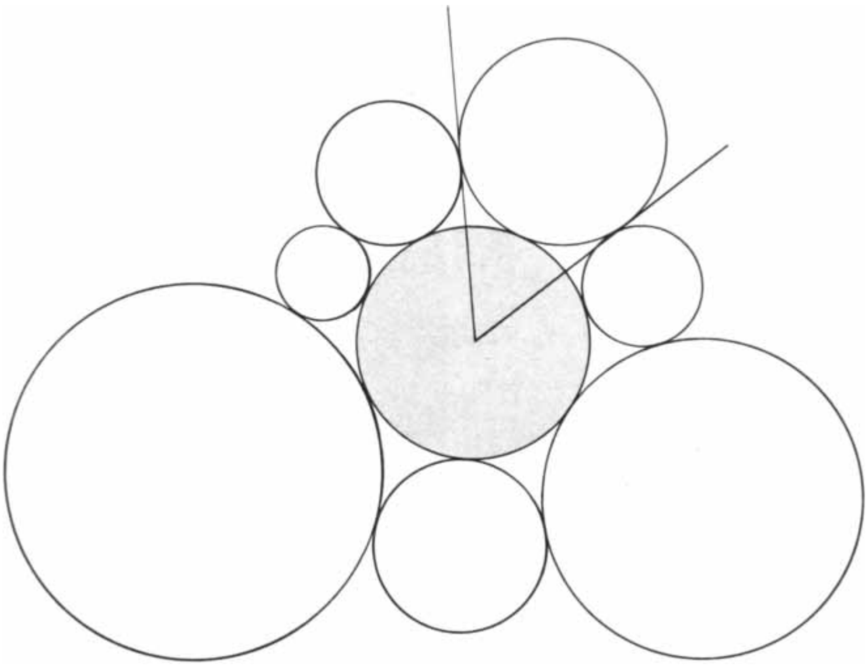
If the reply card has been removed, please write to
The Library of Computer and Information Sciences
Dept. 7-A6F, Riverside, N.J. 08370
to obtain membership information and application.

matics Magazine, Vol. 48, No. 1, January, 1975, pages 33–37), who also gives conjectured solutions for $n = 21$ through $n = 25$.

An inferior packing for $n = 12$ was the secret of "The Packer's Secret," a popular puzzle sold in France late in the 19th century. The puzzle consisted of a circular box containing 12 checkers. The task was to pack them into the box in a stable, rigid way, so that if the box were turned upside down without the lid, the checkers would not fall out. Shown in the bottom illustration on page 26 is a circle just the right size for working on the Packer's Secret with 12 U.S. pennies. Can you fit 12 pennies into this circle to form a rigid pattern in which no coin is movable? The answer will be given next month.



Jakob Steiner's chain



A false proof of Solomon W. Golomb's coin problem

Searches have also been made for the densest packings of n identical circles into squares. It has been shown that as n increases, the density approaches .9069+. That is the limit obtained by the familiar close packing of circles with their centers on a regular lattice of equilateral triangles. Proofs of the best packings are known, however, only for $n = 1$ through $n = 9$. Once again there is no known formula or algorithm that yields the densest packing. The upper illustration on page 28, reproduced from Goldberg's "The Packing of Equal Circles in a Square" (*Mathematics Magazine*, Vol. 43, No. 1, January, 1970, pages 24–30) shows solutions for $n = 1$ through $n = 9$. In the illustration, instead of displaying unit circles inside the smallest square, Goldberg shows the largest circles that fit into a unit square. His paper also gives his best conjectures for $n = 10$ through $n = 27$, and for selected higher values.

Goldberg proves that in each best solution there must be a structure of touching circles that connects all four sides of the square, and that within this structure each circle must make contact with at least three other circles or a side of the square. There may, however, be circles that are not part of this structure, as in the case of $n = 7$.

Below each square in the illustration is the diameter of the largest possible circle, assuming that the square's side is 1. In two cases, $n = 6$ and $n = 7$, proofs are not easy and have not been published. The case of $n = 6$ was first solved by Ronald L. Graham of Bell Laboratories. The case of $n = 7$ was announced by J. Schaer in 1965. In a note in *Mathematics Magazine* (Vol. 44, No. 3, May, 1971, pages 139–140) Schaer improved on Goldberg's conjecture for $n = 10$. When the problem is extended to packing spheres into spheres or into cubes, it becomes enormously more intractable, and almost nothing is known about this three-dimensional version of the problem. (For a summary of the known results on the packing of equal spheres in a cube see Goldberg's paper in *Mathematics Magazine*, Vol. 44, No. 4, September, 1971, pages 199–208.)

The packing of n equal circles into equilateral triangles also presents difficult questions. Little has been established except that when n is a triangular number (in the sequence 1, 3, 6, 10, 15 . . .), the densest packing is achieved by close packing in rows of 1, 2, 3, 4, 5 . . . circles. If the number of circles is a triangular number decreased by two or more, the remaining circles can always be shifted to fit into an equilateral triangle of smaller size. Hence if you remove two pool balls from inside the wood triangle used for closely packing 15 balls

The extraordinary new Sinclair Enterprise Programmable

A complete 79-step programmable calculator package for just $49.95.

One would expect to pay a higher price for such a programmable calculator.

Especially one with the capability of a 79-step keyboard program entry and full range of esoteric functions.

But incredibly, the new Sinclair Enterprise Programmable gives you that power for the remarkably low price of $49.95 — much less than the cost of any comparable machine. And it's perfect for everyone; students to scientists. As your first or only calculator. Or a handy pocketable second.

Yet the price does not in any way compromise performance you can expect: a look at the specifications of the calculator tells you that.

EVERY EXTRA INCLUDED!

When you buy your Enterprise Programmable, you buy a complete ready-to-go package:

You get a *comprehensive operating manual.* It shows you how to get the best from your calculator, how to enter and run, ready-made programs, how to write your own special programs.

You get a *3-volume program library,* the most comprehensive of its kind in the world. General calculations, games, finance, statistics, math, physics, engineering, electronics — 316 different programs make light work of everything from mortgages and business to vectors, matrices and circuit designs.

It's in special binders, so that you can detach selected pages, rebind them as a personal program collection, carry them in your pocket.

You get a powerful *manganese alkaline battery, an AC adaptor,* and a smart black *carrying case.*

And because writing your own programs is so easy, we've even included blank program tabs and a *Felt Tip pen!*

Still the Enterprise Programmable costs about half the price of comparable calculators (where the extras really *are* extra).

For $49.95, it's an incomparable value.

SPECIFICATIONS

Large, 8-digit, light-up angled red display, with fully-floating chemical point or fixed scientific notation.

Accurate to ± one digit in the eighth place on all functions.

+, −, ÷, +
1/×, ×, √x, π
× <—> y, +/−, y

- Common and natural logs and antilogs.
- Sin, cos, tan, arcsin, arccos, arctan.
- Radian/degree conversions.
- 2 levels of brackets.
- 7 independent 3-function memories.
- Full 79-step program facility with merged codes — equivalent, typically, of up to 120 keystrokes.
- Forward/backward stepping.
- Conditional/unconditional branching.
- Program edit facility.

TAKE THE ENTERPRISE PROGRAMMABLE ON A 10-DAY TRIAL

The only way to appreciate the capabilities of the Enterprise Programmable is to try it for yourself.

So we're offering a 10-day no obligation trial. Send your check or money order with the order form below. We'll send a complete Enterprise Programmable calculator package to you.

Put it through some of your more complex and challenging tests. If you don't feel it's the best $49.95 you ever invested, return the entire package.

We'll be surprised, but we'll refund your money in full without question.

Send your order today. You have nothing to lose — but a great deal to gain!

sinclair

To: Sinclair Radionics Inc., Galleria, 115 East 57th Street, New York, N.Y. 10022

Please send me ________ (qty) Sinclair Enterprise Programmable Package(s) at $49.95 each. Price includes Operating Manual, 3-Volume Program Library, Manganese Alkaline Battery, AC Adaptor, Pen and Carrying Case.

Add $2.50 to cover handling (plus 8% Sales Tax if NY resident)

I enclose my check or money order for $________ made payable to Sinclair Radionics Inc.

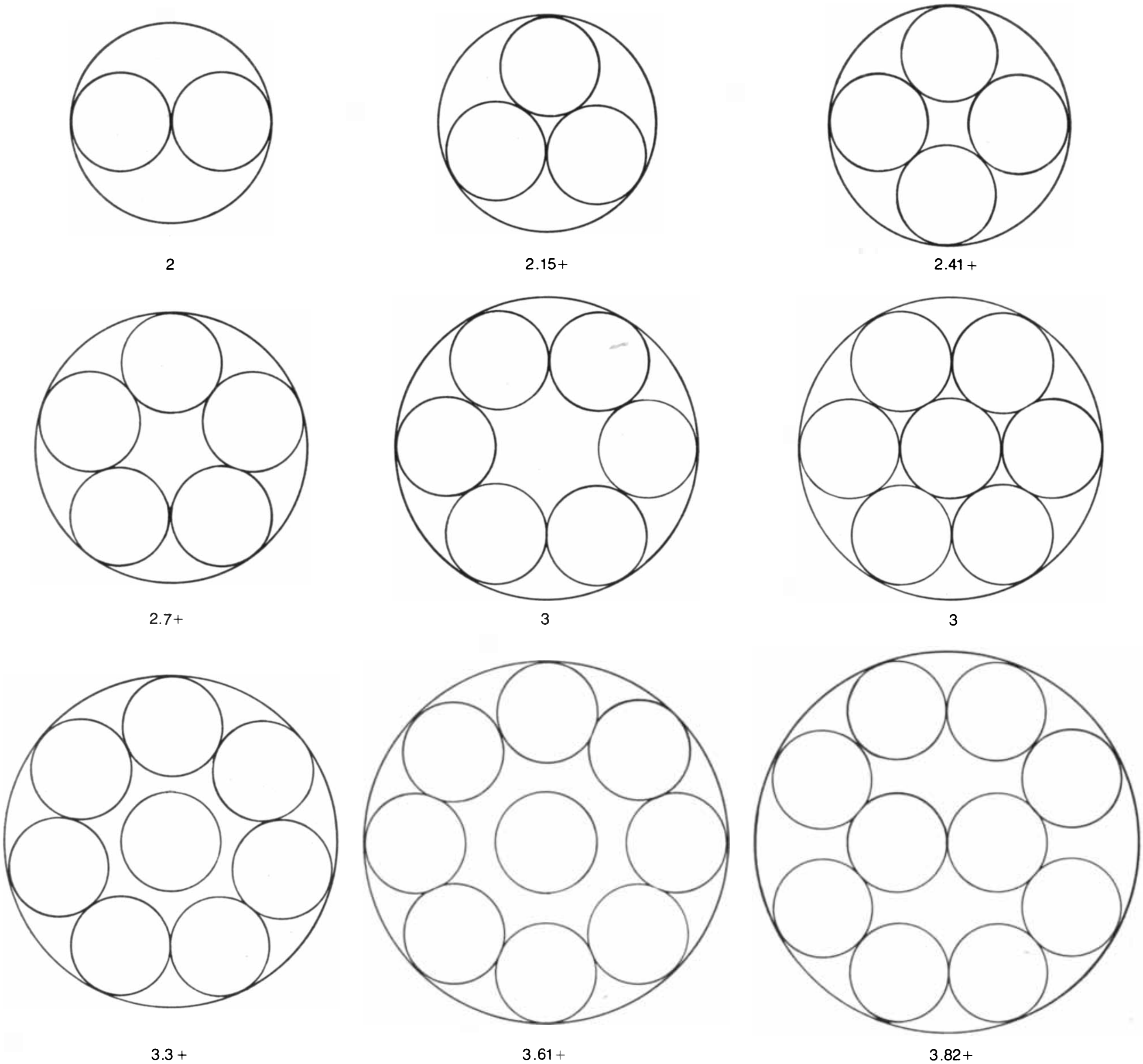
Name ____________________
(Please print)

Address ____________________

City ____________ State ____________ Zip ____________

Signature ____________________

I understand that you will refund purchase price in full if I return complete calculator package in saleable condition within 10 days of receipt



Densest packings of two through 10 unit circles into circles

at the start of a game, the remaining 13 balls can be rattled around so that none of them touches the boundary.

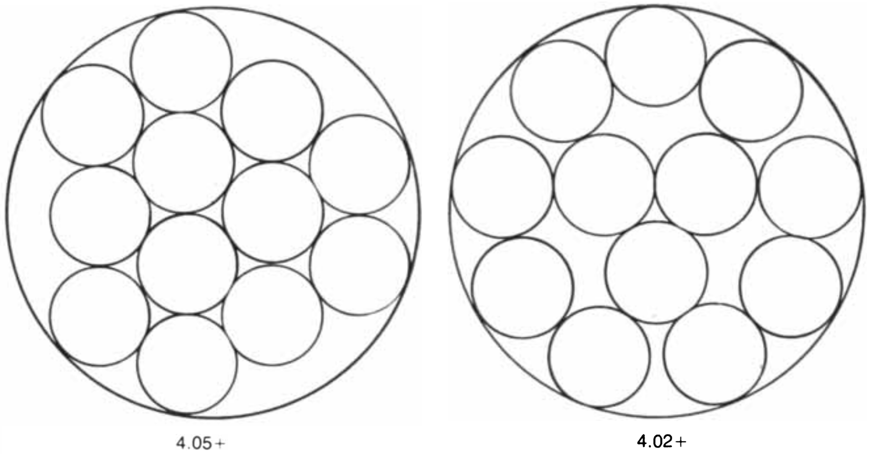
Suppose just one ball is removed. Can the remaining 14 balls be moved about until none touches the boundary? It seems unlikely, but no one knows for certain. Donald J. Newman has conjectured that in all cases where the number of balls (or circles) is one less than a triangular number there is no way to rearrange the balls to make possible a smaller enclosing triangle. The conjecture does not apply to the triangular number 1, and it is clearly true for 3. It seems to be true for the next number, 6, but I know of no formal proof that five balls cannot be squeezed into a smaller triangle than the one that holds six.

We can ask similar questions about the densest packing of equal circles into any defined region, including regions with holes. If the region has no holes and is bounded by a convex closed curve, the best result obtained so far is by J. H. Folkman and Graham in "A Packing Inequality for Compact Convex Subsets of the Plane" (*Canadian Mathematical Bulletin*, Vol. 12, No. 6, 1969, pages 745–752). Given the area and perimeter of a region, the authors establish an upper bound for the maximum number of unit circles that can be packed into it.

From hundreds of other theorems about touching circles I have space for just one more: an elegant result published in 1968 by the Canadian geometer H. S. M. Coxeter. It is shown in the lower illustration on page 28. An infinite sequence of circles is constructed so that every four consecutive circles are mutually tangent. It turns out that this sequence is unique. The radius of each circle is obtained by multiplying the radius of the next-smallest circle by the sum of the golden ratio and its square root, a number that is slightly more than 2.89. The contact points of the circles lie on an equiangular spiral shown by the broken curve.

The answers to last month's questions follow:

1. Inkstand.
2. Crankshaft.
3. If the positive integers are divided into even and odd numbers, any pair in either set will add to an even number greater than 2 and thus cannot add to a prime. (I am indebted to the late David L. Silverman for this problem.)
4. The social security number is 381–65–4729. Adding 0 at the end gives the unique solution to the same problem with the 10 digits from 0 through 9.
5. A 16-move solution to the knights-

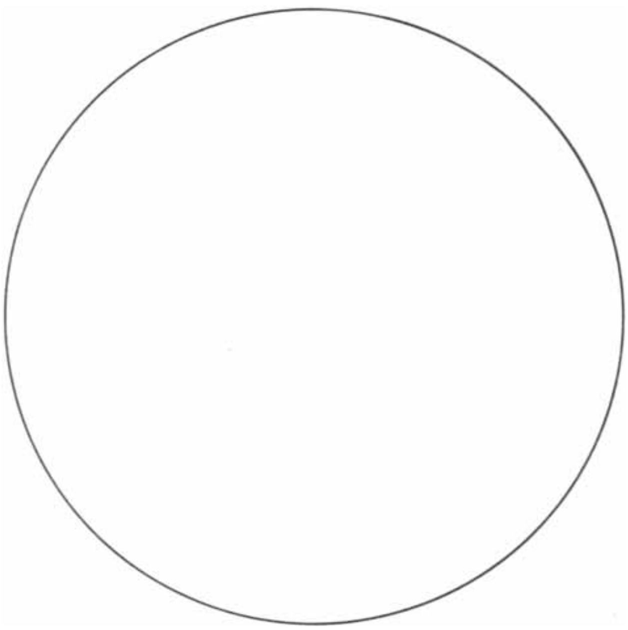


Inferior packing of 12 circles (**left**) *and conjectured best packing* (**right**)

switch problem is given below. As in last month's illustration, rows are numbered 1 through 4 from bottom to top and columns are labeled *A*, *B* and *C* from left to right.

1. *A*1–*C*2	9. *A*2–*C*1
2. *C*2–*A*3	10. *B*1–*C*3
3. *B*4–*C*2	11. *C*3–*A*4
4. *C*2–*A*1	12. *A*3–*C*2
5. *C*1–*A*2	13. *C*4–*A*3
6. *A*2–*B*4	14. *A*3–*B*1
7. *A*4–*C*3	15. *C*2–*A*3
8. *C*3–*A*2	16. *A*3–*C*4

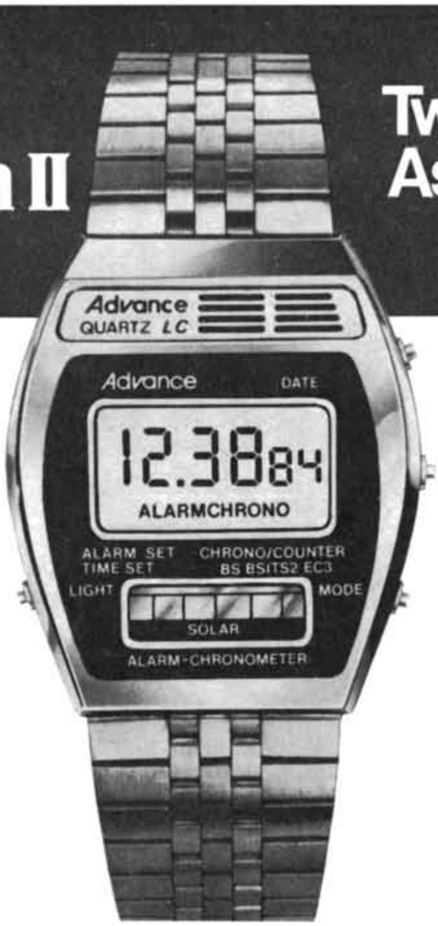
In my September column I said I knew of no marketed version of Percy MacMahon's 30 color cubes. The Educational Products Division of International Technical Associates (P.O. Box 7190, Menlo Park, Calif. 94025) tells me they have produced such a set that they can supply for $19.75 plus shipping and sales tax (where applicable) to any part of the world. In my October column on Egyptian fractions I gave the three-term expansion $5/121 = 1/25 + 1/759 + 1/208{,}725$. A three-term sequence is the smallest possible expansion for the fraction, but I wondered if this one could be improved by a three-term expression with a smaller largest denominator. Sin Hitotumatu, Alan F. Berndt and Pierre Tougne were the first three of the readers too numerous to list who found the "best" expansion: $5/121 = 1/33 + 1/121 + 1/363$. C. S. Metchette, Morton Raff and R. Bruce Welmers were the first of many readers to find the best four-term expansion for 8/11. It is $8/11 = 1/2 + 1/6 + 1/22 + 1/66$, which is a much better expression than the one given in the third answer at the end of my November column. In the fourth answer in that column *b* is equal to 25, not 5; both



Circle for working on "The Packer's Secret" with 12 U.S. pennies

LCD Alarm Chronograph II

Twice as Good As the Original



Watch War

1. We introduce the LCD Alarm Chronograph at a $200.00 savings.
2. They copy it (fair because we copied the Seiko).
3. We make improvements and lower our price.
4. They lower their price (no improvements).
5. Now, with five major improvements the Advance II is twice as good as the original—at the original low price.

This is no war of words. The differences you'll enjoy with our new LCD Alarm Chronograph are real and make it the absolutely best value on the market today.

Perfect Accuracy Imagine, at the touch of a button you can correct the slightest variation—to a tenth of a second. So, in addition to guaranteed quartz crystal accuracy of ± 5 seconds a month, you can instantly correct, or "trim the crystal" electronically, actually have to-the-second accuracy.

Now the Alarm's an Automatic Reminder. The alarm can be set in an instant for any minute of any hour. And now features a snooze alarm, personal reminder system you can repeat every seven minutes. Great, because you can't always leave for a meeting or your plane when the alarm rings, make that phone call at the time you promised. With the reminder you can be sure you don't forget.

When the alarm's set, an A appears. To check for the time, touch the alarm button.

Also, the new alarm has a warning or "executive tone." This sounds once and pauses—so only you hear it and can shut the alarm off if you want to before the regular 52 second beeping cycle begins.

Every one of these five new features is exclusive with the LCD Alarm Chronograph II. And make this watch the one you're sure to want.

Memory Bank for every mode. Which brings us to another major difference and advantage: every mode can be put into memory. For example, with the imitators you can't get the time when you're using the chronograph. It would wipe out timing of the event!

Exclusive touch button counting. (To 9,999)

Solar Rechargeable 3-year Battery. The battery in the Advance II is recharged—automatically by any kind of light. So, instead of the usual one year, you'll get guaranteed (by Mallory) 3-year battery life. And the bank of seven solar cells makes a handsome addition to the bold, impressive appearance.

The correct time all the time.

Digital Counter—for laps when you're running, golf score and more. Now, when someone asks, "What'd you shoot?" you can look at your watch instead of your score card. And it's terrific for keeping track of your laps, for counting cards, your ski runs, telephone calls, customers (daily, weekly, monthly cumulative comparisons), the people at a meeting, and in taking inventory. Touch a button to count, then store the tally in memory if you're interrupted or want to refer to the total any time after.

The Chronograph System. As to the chronograph, or split-second timer, its precision is so fine, it borders on the infinitesimal—to one-hundredth of a second for the first 20 minutes . . . to the second for a full 1 hour . . . and in a variety of ways unequaled by any other instrument.

Still Only $100.00

The best known LCD Alarm Chronograph, of course, is the Seiko, which regularly sells for $299.95. While Seiko pioneered in this area, its chronograph is limited and still times only to 1/10th of a second, its alarm has no reminder, its battery has to be replaced at least once a year.

Copies abound, of course. But Advance is a manufacturer in its own right. Forging ahead, building on existing technology, creating a reputation of its own for extraordinary quality, which we're sure the almost 30,000 people who ordered the first LCD Alarm Chronograph from us will verify.

(Actually, as one of the oldest and largest mail merchandisers, our only concern is to assure all these people that the refinements in the Advance II in no way minimizes the fact that the watch they already own does more and does it better than anyone else's.)

The LCD Alarm Chronograph II is still only $100.00 which includes shipping, handling, insurance and a handsome gift case. It's available in either chrome (white) or gold-plate (yellow) on a forged, hand-finished brass case. The back is stainless steel as is the band, which adjusts instantly to a perfect and extremely comfortable fit.

Then, so you can see when it's dim or you're in the dark, the face lights up.

The Watch and Electronic Calendar. Most important, the basic display gives you the exactly right time. The hour, minutes and running seconds, plus the day of the week. At a touch you can replace the time with the month and date, always right with no adjustment necessary.

Thirty day trial. Now we invite you to put everything we say and promise to the test—without risk or obligation for 30 days. And to send the LCD Alarm Chronograph II back to us for a complete refund if it isn't the equal of any watch at any price for accuracy, versatility and usefulness. A watch that wins the admiration of your friends as well as your own pleasure and lasting satisfaction.

CALL 800-621-5554

(Illinois residents call 800-972-5858)

To order by credit card, call the toll-free number above. Send your check to Douglas Dunhill at the address below. Be sure to specify gold or chrome. (Illinois residents add the appropriate sales tax.)

The Advance LCD Alarm Chronograph II comes with a full ONE YEAR limited warranty and our assurance it will provide you with years of unmatched, trouble-free performance. Order now and enjoy the satisfaction that comes with wearing a watch that's second to none.

Four-way Chronograph System

No other instrument, at any price, gives you greater precision than the LCD Alarm Chronograph II, with its one 1/100th of a second accuracy. Or greater versatility and flexibility. Three separate chronographs that work in memory—combine in function.

Timed to 18 minutes, 14 seconds and 85/100ths of a second.

1 Add time . . . the stopwatch, which you can stop (for a time-out, say) and have pick up where you left off when the action resumes.

2 Split time . . . lets you stop to get a reading (of the 1/4, the 1/2, each contestant crossing the finish line) while overall timing continues in memory until displayed.

3 Lap time . . . gives you the time of individual segments of an event, automatically starting again from zero so you can time the following sequence. It literally splits elements that can't be separated any other way.

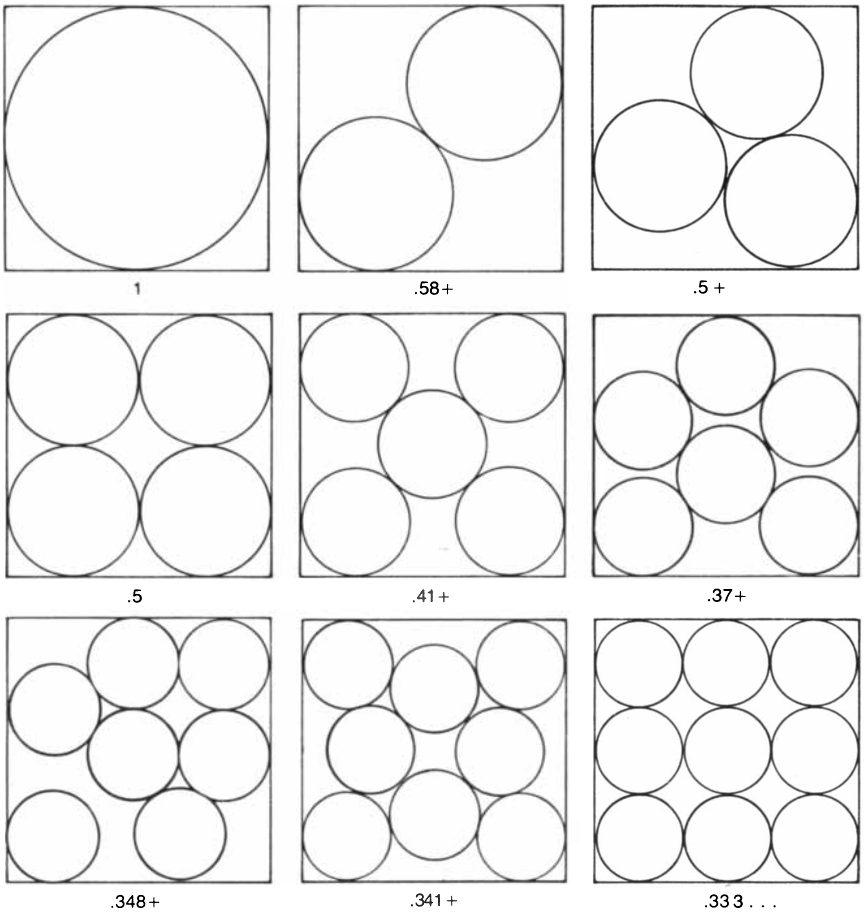
4 Twin timing . . . combines split time with add time so you can get the total time of an event and the times of individual portions within the event.

All this, plus a digital counter, in a sleek, superbly styled timepiece. Order now and take 30 days to prove how easy it all is to master and how useful when you have.

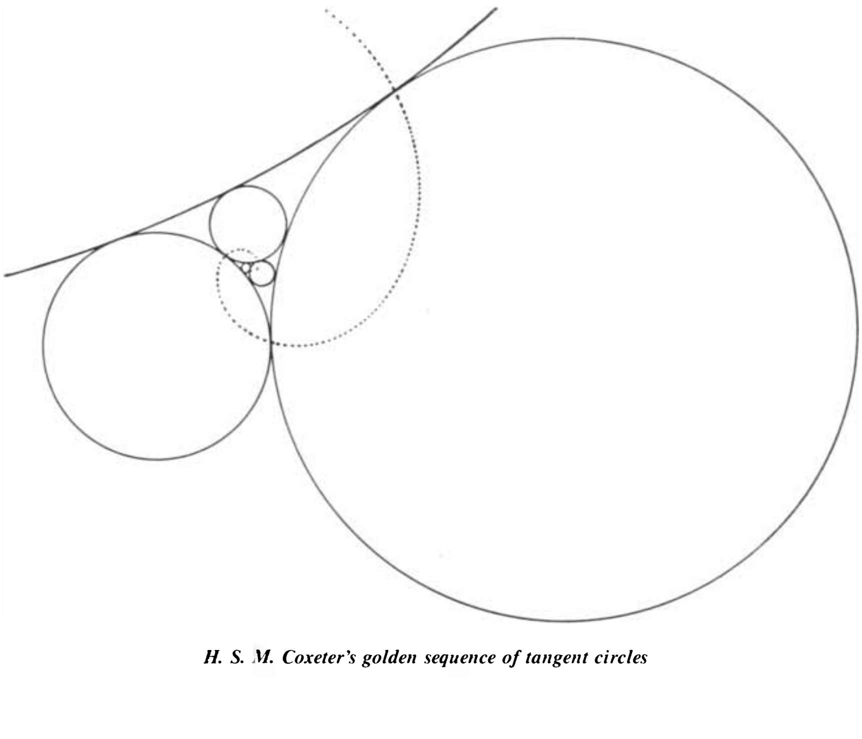
Douglas Dunhill
INC. AFFORDABLE QUALITY
Dept. 71-2502
4225 Frontage Road
Oak Forest, IL 60452

expansions in that answer are for 3/25.

I am pleased to report that a long-standing unsolved problem involving ticktacktoe on a three-dimensional 4-by-4-by-4 board has finally been solved. With the aid of a computer program Oren Patashnik of Bell Laboratories has established what had been suspected but was never completely proved, that the first player can always win if he plays correctly. Patashnik's work, which he completed in 1977 at Yale University, was finally verified last October by Ken Thompson, also of Bell Laboratories.



Densest packings of one through nine identical circles into squares



H. S. M. Coxeter's golden sequence of tangent circles

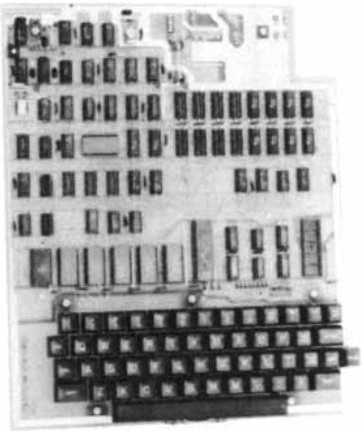
The Age of Affordable Personal Computing Has Finally Arrived.

Ohio Scientific has made a major breakthrough in small computer technology which dramatically reduces the cost of personal computers. By use of custom LSI micro circuits, we have managed to put a complete ultra high performance computer and all necessary interfaces, including the keyboard and power supply, on a single printed circuit board. This new computer actually has more features and higher performance than some home or personal computers that are selling today for up to $2000. It is more powerful than computer systems which cost over $20,000 in the early 1970's.

This new machine can entertain your whole family with spectacular video games and cartoons, made possible by its ultra high resolution graphics and super fast BASIC. It can help you with your personal finances and budget planning, made possible by its decimal arithmetic ability and cassette data storage capabilities. It can assist you in school or industry as an ultra powerful scientific calculator, made possible by its advanced scientific math functions and built-in "immediate" mode which allows complex problem solving without programming! This computer can actually entertain your children while it educates them in topics ranging from naming the Presidents of the United States to tutoring trigonometry all possible by its fast extended BASIC, graphics and data storage ability.

The machine can be economically expanded to assist in your business, remotely control your home, communicate with other computers and perform many other tasks via the broadest line of expansion accessories in the microcomputer industry.

This machine is super easy to use because it communicates naturally in BASIC, an English-like programming language. So you can easily instruct it or program it to do whatever you want, *but you don't have to.* You don't because it comes with a complete software library on cassette including programs for each application stated above. Ohio Scientific also offers you hundreds of inexpensive programs on ready-to-run cassettes. Program it yourself or just enjoy it; the choice is yours.



Ohio Scientific offers you this remarkable new computer two ways.

Challenger 1P $349
Fully packaged with power supply. Just plug in a video monitor or TV through an RF converter to be up and running.

Superboard II $279
For electronic buffs. Fully assembled and tested. Requires +5V. at 3 Amps and a video monitor or TV with RF converter to be up and running.

Standard Features

- Uses the ultra powerful 6502 microprocessor
- 8K Microsoft BASIC-in-ROM
  Full feature BASIC runs faster than currently available personal computers and all 8080-based business computers.
- 4K static RAM on board expandable to 8K
- Full 53-key keyboard with upper/lower case and user programmability
- Kansas City standard audio cassette interface for high reliability
- Full machine code monitor and I/O utilities in ROM
- Direct access video display has 1K of dedicated memory (besides 4K user memory), features upper case, lower case, graphics and gaming characters for an effective screen resolution of up to 256 by 256 points. Normal TV's with overscan display about 24 rows of 24 characters; without overscan up to 30 X 30 characters.

Extras

- Available expander board features 24K static RAM (additional), dual mini-floppy interface, port adapter for printer and modem and an OSI 48 line expansion interface.
- Assembler/editor and extended machine code monitor available.

ORDER FORM

Order direct or from your local Ohio Scientific dealer.

☐ I'm interested. Send me information on your:
☐ Personal Computers ☐ Business Systems
☐ Send me a Superboard II $279 enclosed
☐ Send me a Challenger 1P $349 enclosed
☐ Include 4 more K of RAM (8K Total) $69 more enclosed

Name ____________________

Address ____________________

City ____________ State ____________ Zip ________

Payment by: BAC (VISA) ____ Master Charge ____ Money Order ____

Credit Card Account # ____________________

Expires ________ Interbank #(Master Charge) ____________

Ohio Residents add 4% Sales Tax

TOTAL CHARGED OR ENCLOSED ____________

All orders shipped insured UPS unless otherwise requested. FOB Aurora, OH

Interested in a bigger system? Ohio Scientific offers 15 other models of microcomputer systems ranging from single board units to 74 million byte hard disk systems.

OHIO SCIENTIFIC

America's Largest Full Line Microcomputer Company

1333 S. Chillicothe Road • Aurora, Ohio 44202 (216) 562-3101

"CERTAINLY IT'S THE BEST EUROPE IN A LONG WHILE."

*The sedan that Road Test Magazine is talking about does not sell for $20,000. It does not sell for $15,000. It sells for a base price of $8995.**

This pleasantly surprising contradiction between the test results of the Audi 5000 and the price has been pointed out by several automotive publications.

Read these excerpts from the Road Test report on the Audi 5000 that appeared in the January, 1978 issue:

"At freeway speeds the ride is serene and smooth, there is no mechanical noise, and no wind noise. You get bored, turn on the radio, and just drive—until you glance down at the gauges and notice that you're making 85 mph. It's sneaky; it'll fool you if you don't pay attention. It'll also nip in and out of traffic like a much smaller car, thanks to the way it responds to its controls..."

"It may just be the ultimate sports sedan."

"...not only does the 5000 have a 5-cylindered mill, but it's got one that sips gas like a four,** makes horse-

*Suggested retail price $8995 P.O.E. Transportation, local taxes and dealer delivery charges additional. Fully equipped "S" model $10,575.

SEDAN TO COME FROM

—ROAD TEST MAGAZINE, ON THE AUDI 5000

power like a six and feels nearly as smooth as an eight."

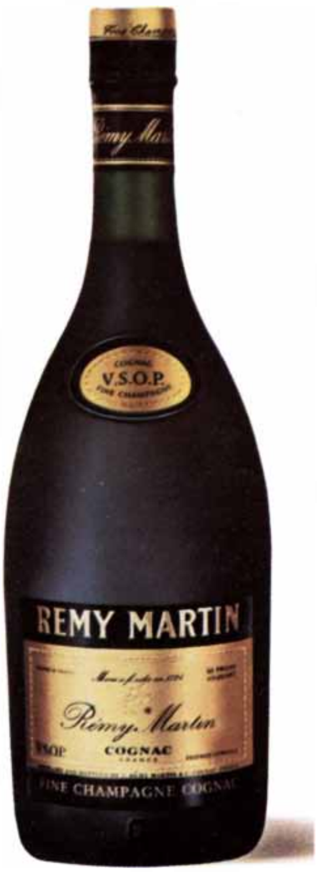
"...the price...isn't near as much as some of the 5000's closest competitors sell for. That, of course, is the name of the game, and Audi is playing it to the hilt with development of solid concept and good attention to detail. They're going to sell every one of these cars they can screw together."

"...the 5000 has an incredible heater/air conditioner that could heat half your house, or could freeze your—um, toes off, depending on which button is pushed."

Test drive the 1979 edition of the Audi 5000 at your Porsche + Audi dealer and see if it isn't so.

**26 mpg highway, 16 mpg city, std. transmission. (EPA estimates. Actual mileage may vary based on how and where you drive, car's condition, and optional equipment.)

PORSCHE+AUDI
Nothing Even Comes Close


V.S.O.P.
REMY MARTIN
Rémy Martin
COGNAC

BOOKS

The scientific exhibitions of 19th-century London, and myelin, the insulator of nerve

by Philip Morrison

THE SHOWS OF LONDON, by Richard D. Altick. The Belknap Press of Harvard University Press ($35). Chaucer's Pardoner was a traveling showman of the late Middle Ages; his glass case held dubious relics, including a miraculous "sholder-boon ... of an hooly Jewes sheep." Even lofty St. Paul's aspired little higher. In 1314 the chancellor of that cathedral displayed to the people "a piece of the true Cross, ... bones of martyred Virgins." The exhibitions of London thus begin with holy relics, concrete marvels for the pious; they break off, at least in this lively and learned account, a little after the famous Crystal Palace, in 1851.

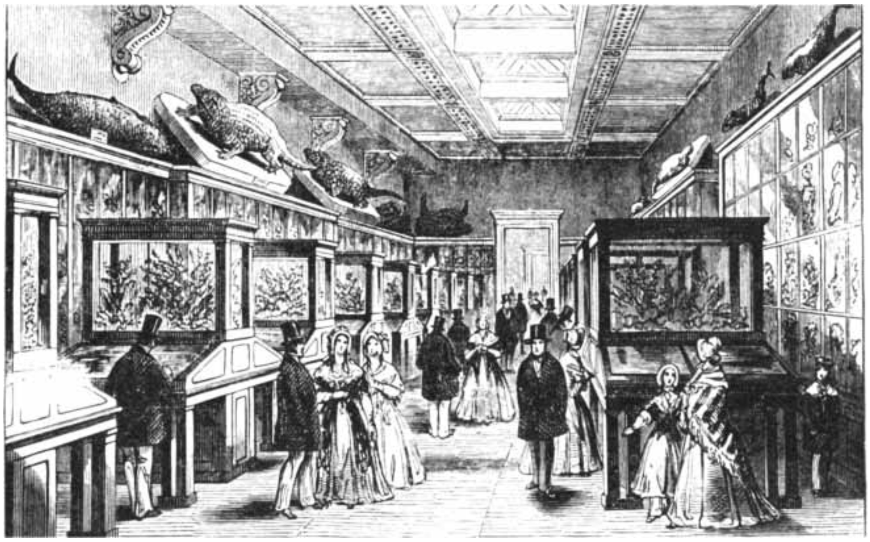
In between what a farrago of things there are to see and admire, real, false, ludicrous, sublime, petty, grand! This history presents a factual account of what London gaped at over 300 years, as recorded in clippings, prints, handbills, programs and guides preserved with "eccentric zeal" by the antiquarians of the past and ferreted out of many libraries by the author, a historian of popular culture at Ohio State University. His topic adjoins four others, the history of fairs, of the theater, of art and of scientific collections. It is actually none of them; it centers on the sights people paid to see, not so much for a performance as for the thing or creature itself. In this attention to the world it adumbrates science and technology.

No review can summarize the 500 big square pages and the impact of the prints, cartoons and posters that accompany them. Let us sample gingerly, the cornucopia of London held at arm's length, an item or two per century. In about 1650 a house and garden in South Lambeth became an obligatory sight for visitors to London, the finest natural history museum anywhere. The gardener-voyager (and his son, who had traveled to Virginia) kept the place, "Tradescant's Ark," expressly to display every rarity possible. His catalogue, the first such ever printed in England, listed hundreds of species of "strange Creatures," such as the "dodar from the Island Mauritius; it is not able to flie being so big." Ethnography was included, with Powhatan's "habit all embroidered with shells" and much more. The holy relics have vanished into the next world; now this world is much with us, across wide horizons. We are at the Renaissance.

A century later the Enlightenment had brought the thought of Europe to mechanism: Newton's orbits in heaven, equilibrium in trade, and law and clockwork for admiration. In 1742 the famous figures of Jacques de Vaucanson came to London. Voltaire wrote of him as "rival de Prométhée." The animated clockwork included a flute player (who could play 20 tunes on his three-holed flageolet), a drummer and a gilded-copper duck that ate and drank with lively motions, duly digested its grain and excreted, four times a day at the Opera House in Haymarket. A generation later the Jaquet-Drozes, *père et fils,* brought their subtler three automaton-humans, marvels still on display in Neuchâtel. Life itself seemed not beyond mechanism; children today still ask a bit anxiously as they watch those very animated figures play, write and sketch: "Are they alive?"

Even before Victorian times the optical arts joined the mechanical ones. What names they invented: Cosmorama, Eidophusikon, Phantasmagoria! A more romantic view of nature became popular, and the little magnified peep-shows and dioramas showed such wonders as "a Sea-Storm, Accompanied with all its awful effects ... distant Thunder ... the Skies darken, ... the Elements open with a Thunder Bolt darting through a SHIP." So went the proprietor's playbill. A cooler witness observed that the entire viewing area was no larger than a good-sized window, and that the "agitated sea is as large as half a blanket and is memorable for preserving a fine uniformity of wave even in its wildest agitation ... a squib explodes one poor ship's magazine." The magic lantern, first lighted by the new hollow-wick oil lamp and then by the oxyhydrogen "limelight," became a mainstay; dissolves and double-screen techniques became the stock-in-trade not only of professional exhibitors but also of comic lecturers and the very solemn Salvation Army. The Polytechnic showed microscopic life with the "gas microscope" projecting on a 20-foot screen in the 1840's.

The narrative ends with the Great Exhibition of 1851. Under a Royal Commission it was managed by private enterprise of unprecedented scale and seriousness. The Crystal Palace drew crowds in a euphoria of the age. It was genuinely instructive, in small doses at least, and broadened the "horizons of the millions of men, women and children who passed through the turnstiles." Here was the world brought close, not through the windowpane of a diorama but at full scale, in many contexts and with living people from every land. The fringe of exhibitors prospered too, and there rose in Leicester Square a Great Globe of the world, sponsored by the map seller James Wyld. It was a hollow sphere 60 feet in diameter. Built in a rush, at night under gas torches, it rose on schedule, and within a brick exterior



CORAL ROOM OF THE BRITISH MUSEUM was depicted in *The Illustrated London News* for April 3, 1847. The engraving is reproduced in *The Shows of London,* by Richard D. Altick.

structure visitors climbed stairs to four viewing platforms. There one could inspect portions of the earth, mapped in plaster relief, at a scale of 10 miles to the inch, in color and texture to please, cotton-wool tufts from every volcano, blue oceans, tawny deserts and green fertile lands. At once it became second only to the Crystal Palace in popularity, even though, as Philip Gosse wrote, "it was a poor affair; that was concave in it which should have been convex."

Then set in the decline of the London shows. People moved to the suburbs. Parks offered outdoor amusements and homes held the stereoscope, illustrated papers and books, which provided images aplenty. Out of the Great Exhibition surplus the Treasury got about 200,000 pounds. With that sum a process began that a decade after the exhibition left London with new national museums in South Kensington, an industrial exhibition (like the Crystal Palace itself) and above all a museum of ornamental art. (They were all forerunners of the present museums at the same general sites.) They were operated from the first with an eye to the mass public, the museum seen as "a powerful antidote to the gin palace." It took another decade or two, but the old shows dwindled; entertainment—drama and music hall—replaced the old exhibitions, in an era of plentiful printed pictures. The role of instruction and breadth of experience was increasingly assumed by the government, and "the age of exhibitions was succeeded by the age of public museums." The video screen was six or eight decades away.

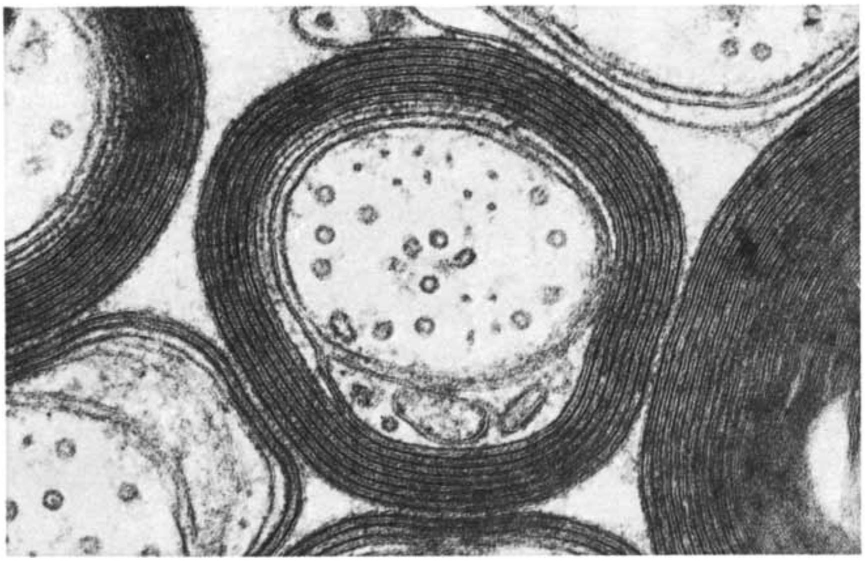
MYELIN, edited by Pierre Morell. Plenum Press ($39.50). The gray matter has a better press, but the white matter of the mature brain and spinal cord of mammals has equal claim to indispensability. Of the adult human brain as much as a third of the dry weight is this one substance: myelin. The cabled fibers everywhere in the nervous system and the bigger single motor nerves are richest in it, although it is found in all components of our internal wiring. It is just 100 years since the histologist L.-A. Ranvier saw in this quasi-liquid covering of the nerves both a mechanical shield against compression and an insulating envelope, serving to isolate the delicate nerve currents from their conducting saltwater environment: "It is on this principle that the construction of submarine cables rests," he wrote, a decade after the first successful telegraph cable was laid. "This extraordinarily astute observation" has been amply supported by the work of a century and is now confirmed by accurate electrochemical models for nerve conduction, in which the electrical parameters of the myelin sheath enter essentially. Myelin is not as good an insulator as paper, say, or any other such engineering insulator, but it is good enough and very thin. The typical thickness of some four micrometers serves to increase signal speed and reduce expenditure of energy to a remarkable degree.

We vertebrates all have the stuff (as do a few shrimps). Hence the 25-meters-per-second signal speed along a fine but myelinated frog nerve is about the same as that down the heavy bus bar of the squid, which is a giant axon, a bare wire 40 times bigger in diameter. That thin insulation saves packing volume by a factor of 1,000 and energy by a factor of 5,000. It seems pretty clear that we would be restricted to slow and simple electronics without this particular component of the nerve network. The myelin sheath is segmented; at the nodes (named after Ranvier) the covering is absent for a few microns, but the propagation of the action-potential pulse is well described by electrical modeling of the entire conductor, node and all. Indeed, the impulse spreads across the node just as the model suggests, with a safety factor for the transfer of five or so. The nodes are the seat of excitation and the unique seat for the passage of the ions responsible for the electrochemical process.

No one will be surprised to learn that this plentiful substance is no simple coating. It is a spiral wrapping of several hundred turns of a thin double layer of lipids, each layer some 50 or 60 angstroms thick, holding cholesterol and specific protein molecules, some as structural spacers and some free to diffuse through the ordered, liquidlike fatty bilayer. X-ray and neutron diffraction have told us much of the structure, and the electron micrographs of the sheath are striking.

We are not born so well insulated. The building of the sheaths for the peripheral nervous system starts in the fetus; by the end of the second year of life the brain has formed most of its sheaths, although the process continues into adulthood, perhaps for four or five decades in the cortex. The sheath is not unchanging during life: there is a maintenance turnover and even some replacement of damage. The newborn calf and lamb are better insulated, and they are capable of complex behavior almost at birth. Nesting creatures such as mice, however, are born most immature and need time (for myelinization, perhaps?) before they achieve physical independence. Special cells form the fully layered membrane; myelin has a remarkable "structural duality as a crystalline macromolecular assembly and a bit of differentiated tissue." They wrap it around the length of the axon, each cell wrapping a millimeter more or less, working very fast. At peak rate such a cell forms three times its weight in coating per day. Once wrapped in place, each coating segment automatically seals to the neighboring portions until a node is reached.

What price is paid for our fast signals and complex cabling? It is the risk of pathology in the coating system. A long and tragic list of diseases of the myelin system is given. Their effects are complex but manifest: both specific and more general failure of entire circuits and the loss of the synchronization of various actions, as signal rates come to differ between fibers. The most frequent such condition is multiple sclerosis, most prevalent in women between the ages of 20 and 40. So far there is no real therapy; the disease is generally very slow, so that three-fourths of the patients survive 25 years of the variable illness, with many ups and downs. Spontaneous remission is common, without



MYELIN SHEATH OF A NERVE FIBER in the spinal cord, seen in transverse section, is enlarged 145,000 times in this electron micrograph reproduced in *Myelin*, by Pierre Morell. Micrograph was made by Cedric S. Raine of Albert Einstein College of Medicine in New York.

the $988 Surprise . . .

If you haven't looked carefully at the Level-II 16K TRS-80, you're in for a big surprise! Level-II BASIC gives TRS-80 advanced features like comprehensive string handling, multi-dimension arrays, multi-letter variable names, named cassette files, full editing, integer arithmetic, single (6-digit) and double (16-digit) precision arithmetic, formatted printing, memory-mapped video (print directly at any of 1024 screen positions), 128x48 video graphics (may be intermixed with text), error trapping, auto line numbering, TRACE, PEEK and POKE . . . to name just a few. Because Level-II is in ROM, TRS-80 powers-up ready to go with the full 16K RAM available for your use.

This means TRS-80's memory is equivalent to a 28K RAM-based system.

New for 1979—TRS-80's numeric (calculator) keypad included on every 16K computer, and available as an add-on for present owners.

TRS-80's modular design allows easy expansion. Add up to 48K RAM, Expansion Interface, printers, 1 to 4 Mini-Disks, RS232C, telephone acoustic couplers, Voice Synthesizer, dual cassette recorders, our System Desk and Printer Stand. Surprisingly, these are not promises of things to come, but real products being delivered right now. Software from games to General Ledger are available, with more cassette and disk software being added monthly.

Radio Shack's 58 years of consumer electronics leadership, our 50 regional repair centers (growing to 100 this year), our new Radio Shack computer centers, and our NYSE-listed billion-dollar parent, Tandy Corporation, insure that customer support is always available right where it should be—locally.

So if you haven't seriously looked at TRS-80 yet, ask your local Radio Shack for our new 20-page fact-filled catalog and be prepared for a $988 surprise. Surprising power—features—price—support! Level-II 16K systems include everything pictured, plus the manual. Better to be surprised now . . . before you choose the wrong microcomputer system.

much reason to give for it. The Guillain-Barré syndrome, a rapidly progressing loss of ability to move and then even to breathe, is another common myelin disease (lately seen associated with influenza vaccination). Given external support of respiratory movements, the syndrome patients nearly all survive, and most of them avoid any severe lasting disability.

Virus infection just may induce these two diseases; less common failures of the myelin system in human beings and in laboratory animals may be hereditary, or nutritional, or induced by toxic substances (the antiseptic hexachlorophene is one of them). A very rare form was found among individuals who "drink large quantities of red wine"; it is fatal in half a dozen years, from a subtle onset that resembles other diseases of the brain. On rare occasions diphtheria can induce myelin loss. So can the genetic liver-enzyme deficiency, phenylketonuria. In that state a normal infant diet results in an excess of the amino acid phenylalanine, because it cannot be converted to tyrosine along the usual chain; one effect is decreased formation of myelin. The breakdown of the insulator may have many causes, from the death of the maintaining cells to the appearance of a detergentlike metabolite capable of soapily cleaning off the invaluable lipid layers. We do not yet know.

This comprehensive book, clear, technical, very knowing, has been nicely synthesized by the Chapel Hill editor from 14 chapters by U.S. and Canadian experts in a variety of disciplines, from electron microscopy and X-ray diffraction to clinical neurology. It is no mere collection of expert chapters but an organic text, a model of its weighty kind.

ASTRONOMY: THE COSMIC JOURNEY, by William K. Hartmann. Wadsworth Publishing Company, Inc. ($16.95). That the sun is a star is not obvious; "humans pondered the stars for many centuries" before they realized that. The conclusion requires an estimate of the variation of light intensity with distance, or now perhaps even requires the spectroscope. But to infer that our galaxy is one among many it is enough nowadays to compare two photographs, one a wide-angle shot of the Milky Way in our night sky, the other a selected photograph of an external galaxy seen edgewise in a large telescope. The similarity is striking in the pair of velvety halftones shown side by side at the same scale and orientation in one chapter of this fresh and visually outstanding elementary textbook. Dr. Hartmann is not only an expert in photographic techniques, a planetary astronomer and a writer of originality and wit but also a painter. His book, surely one of the most interesting among the dozens of introductory astronomy texts placed before us in the past few years, is wary of algebra, sequestering into typographical cages the five rather tame basic equations of his argument. Not one of these is any more taxing than the simple fourth-power dependence of thermal luminosity on temperature, but they are dealt with entirely qualitatively within the nonmathematical text; the option of a little algebra is left to those who will brave the boxes and try the "advanced problems" at the end of chapters. The vivid sense of reality—all objects are treated as actual places in the universe, not as mere points of light—and the excellent documentation both for the eye and the mind in the carefully listed inferences lift this well-written book above its modest algebraic ceiling. It should be good value for reading and reference to many outside of Astronomy 101.

Here is a little of what you can see, sights all worthwhile and by no means familiar: the Winona stone meteorite, honorably buried in its crypt by the Indians of northern Arizona long ago; the Nazca desert markings and the Medicine Wheel of Wyoming; the midwinter sky around Orion, with an adjoining key map pointing out the "panorama of star formation" there, nebulas, clusters and associations that form the strongest basis for the modern understanding of star birth; the Andromeda galaxy in half a dozen views, showing the effects of exposure, printing and image size on what one sees, including the famous high-resolution balloon-borne Stratoscope image of that brilliant starry nucleus and a reprojection of the spindle-shaped form onto a tilted screen, to "rectify" it by projection into its intrinsically circular pattern; and a dashing painting of a new-formed star seen within its cocoon nebula from an orbiting chunk of silicates and ice, as conceived by the artist-author. The graphs too are numerous, attractive and apt, for example the Hertzsprung-Russell diagrams shown here with the 100 nearest stars entered (all those commonplace, hard-to-see reddish dwarfs) and then with the 100 brightest stars entered as well, the less common but far more conspicuous familiar stars of our skies. One rich page bears eight sky maps in galactic coordinates, each one displaying the locations of some component system such as dust, nebulas, X-ray sources, galactic radio emission and globular clusters. The entire assemblage makes crystal clear the idea of the distinct distribution of the different galactic populations. More usual images and their explanation are here as well, from the Great Red Spot of Jupiter (in color) to portraits of Tycho and Newton.

The survey is broad. It proceeds from the dawn of human knowledge of the skies to the earth-moon system, the solar system, the stars, the stellar environment, groups of stars, galaxies and on to the frontiers of cosmology, cosmogony and the question of alien life. It should be clear that the volume is entirely elementary in techniques and purpose, but the arguments are made so engagingly and well, and the illustrations are so rich and useful, that the book handily transcends its textbook constraint of superficiality.

AN INTRODUCTION TO POPULATION ECOLOGY, by G. Evelyn Hutchinson. Yale University Press ($17.50). "Birth, and copulation, and death. That's all the facts when you come to brass tacks." Sweeney would be pleased, because that's what's here. Under the serious cover of a lucid introduction to the chief quantitative ideas of population ecology, this author, the grace of whose prose and the breadth of whose literary learning must make him unique among all those fluent in calculus, has put together a digressive volume, as entertaining as it is helpful. Would you read of birth control in the time of Thomas Malthus? (Two of his three children survived, not, as the canard has it, two of 14.) Would you wonder at *Pezophaps solitarius*, an extinct dodo-like bird of isolated Rodriguez Island of which only one tantalizing description has come to us? The Huguenot refugee François Leguat lived there for two years at the end of the 17th century and told of the bird, which performed, he maintained, a social ritual of carefully arranged "marriages" for its chicks. Would you care to examine the survivorship curves of the modern British peerage or of the European buttercup? All of them are here, with many reproductions, curves of data, diagrams and formulas, either caught up directly in the argument or held in delicious historical footnotes, some of which grew to occupy the bulk of the page.

The book offers six chapters and gives careful meaning to the fundamentals of its subject, theory and data alike. It begins with a thorough study of the logistic curve, from its history among ideas to its full derivation under plausible circumstances, without omitting the many variations and criticisms that now surround the simple form given it by Pierre-François Verhulst, L.-A.-J. Quételet's young mathematical colleague, in about 1840. Its two parameters, the undisturbed rate of increase and the limiting population, called r and K, are defined and considered at length; indeed, a later chapter describes the poles of evolutionary adaptation, r selection increasing the number of offspring (the prodigal mackerel roe with half a million eggs) and K selection ensuring survival (the proverbial royal litter, only one cub born at a time, but that one a lion). One chapter elaborates survival tables by theory and experiment, noting that the oldest actuarial data come from Roman funeral societies.

H. sapiens and the sheep of Mount

McKinley show survivor curves rather alike, both very different from those of lizards or of blackbirds. Two chapters discuss the heavier mathematics of species competition, with extension to the idea of a niche, a cell in some shadowy hyperspace of circumstance, where a species, or occasionally a few of them, can dwell. A reef overgrown by a single dominant form of coral is shown to us, and also what happens when a predator enters, here the crown of thorns starfish, whose voracious grazing cuts off the advantage of the chief coral form and whose predation increases the diversity of coral growth! All at once we see the shuttle of interdependence fly across the web of forms and we enter the last and most aloof chapter, one on the diversity of species. The text closes with a wise justification of the utility of oversimplified models, and an even wiser account of their limitations. The last illustration we see in these pages is the woodcut of *le solitaire,* old Leguat's vanished bird that had such a magnificent sense of family continuity.

Professor Hutchinson disdains to conceal his views. We ought, he believes, to preserve even the oddities of nature for the very sake of their oddness. Who knows to what use any knowledge might someday be put? Above all, the humdrum niches of simple survival are "totally inadequate" for our uniquely curious kind. "We also need all the wondrous things under heaven: 'Their leaues that differed both in shape and showe / (Though all were greene) yet difference such in greene / Like to the checkered bent of Iris bowe!'" (The verse is Christopher Marlowe's.)

MAYAN CALENDAR: 1979. Athanor, P.O. Box 65, Glen Ellen, Calif. ($3). Here is a modest calendar in a familiar format, one page per month, the days numbered and arranged in columns by weekdays. Above each month page there is a pictorial page. Each shows some Mayan work, such as the mustachioed ruler from a stela in Guatemala who so fiercely ornaments April. The typography is of the simplest kind, the text and illustrations are black on a colored stock, simply clear line drawings of the complex Mayan material originally carved in stone or painted in the colors of the codex. Nothing special, you think? But in every day square there are two carefully drawn sets of glyphs, each the correct notation for the date, one in the 260-day sacred-day count, the other in the 365-day civil calendar. Together they mark the day uniquely among 52 years. July 4, 1979, is 8 Etz'nab, 11 Tzec, although not many Americans will so write it.

The glyphs and their methods of use are derived from the work of Sir Eric Thompson; the explanatory material is brief but usable. The long count notation is not included.

Established 1845 SCIENTIFIC AMERICAN January 1979 Volume 240 Number 1

World Coal Production

As oil production peaks and declines, substitute energy sources will be needed on an enormous scale. Coal could fill a large part of the gap if its production and use were encouraged by national policies

by Edward D. Griffith and Alan W. Clarke

Before World War II coal was the world's dominant fuel. Since then it has steadily lost ground to other fuels, and today it supplies only 19 percent of the energy consumed in the non-Communist countries. During the past 20 years, as worldwide oil consumption roughly tripled, coal production remained nearly constant outside the U.S.S.R., China and Poland. In those countries production increased about 30 percent between 1960 and 1977 but still lagged behind total energy consumption. The virtual stagnation of coal consumption in the non-Communist countries reflects the preference of private and industrial consumers for the convenience offered by liquid fuels, natural gas and electricity, together with the growth of oil-specific markets such as transportation and petrochemicals. With the general recognition that world oil production is likely to peak and decline sometime during the next 25 years, nations will be obliged to seek alternative fuels. Once again coal may be called on to sustain industrial production and economic development throughout the world.

Since coal is not as easily extracted, transported and burned as oil and natural gas are, large increases in coal production and utilization will call both for strong incentives and for fundamental changes in attitudes toward coal. Making coal both available and acceptable will require a major expansion of facilities for coal mining, processing, transportation and utilization, together with new technologies for processing, converting and consuming coal. On the most modest projections the U.S. will have to mine about 60 percent more coal in the year 2000 than it does at present. Conceivably it will have to mine three times more. If such an expansion is to take place, serious economic, social and environmental issues related to coal mining and coal burning will have to be resolved.

We shall outline the implications of large-scale coal development as an alternative to liquid fossil fuels, the options that are open and the issues we believe policymakers will have to face over the next 25 years. Our estimates of coal demand and supply are based on the work of a two-year study, the Workshop on Alternative Energy Strategies, carried out under the direction of Carroll L. Wilson of the Massachusetts Institute of Technology. The study involved some 70 specialists recruited from industry, government and universities in 12 energy-consuming countries and three major oil-exporting countries (the U.S., Canada, Britain, France, West Germany, the Netherlands, Denmark, Sweden, Norway, Finland, Italy, Japan, Iran, Mexico and Venezuela).

The Workshop developed a series of energy projections to 1985 and 2000 based on various assumptions about economic growth, energy prices and government policies. The critical conclusion based on those projections was that sooner or later, and probably before the end of the century, oil demand would strain world supply and oil production would level off and decline. A detailed description of the Workshop's oil analysis has appeared in these pages [see "World Oil Production," by Andrew R. Flower; SCIENTIFIC AMERICAN, March, 1978]. In this article we shall consider the possible use of coal as a substitute for oil, the size of the world's coal resources, the outlook for coal demand, potential coal production in various regions of the world and the possible conversion of coal into synthetic fuels.

Although one lump of coal may look very much like another, there are many different grades and qualities, just as there are many different types of crude oil. High grades of coal have a heating value of some 7,000 kilocalories per kilogram (12,600 British thermal units per pound); lower grades have only half or a third as much. One system of classification divides coal into two broad categories: hard coal (anthracite and the various grades of bituminous coal) and soft coal (brown coal and lignite). The ability of different types of coal to form coke when they are heated in the absence of air also varies widely. Since coke for steelmaking is one of coal's major uses, a second important classification divides coals into metallurgical (or coking) coals and thermal (or steam) coals, which are burned to provide either heat or steam for the generation of electric power.

The total world production of all types of coal in 1977 was 3,400 million metric tons, of which hard coal represented 2,500 million metric tons, or about 70 percent. The leading producers of hard coal are the major developed countries. More than 75 percent of the world output is mined by six countries: the U.S., the U.S.S.R., China, Poland, Britain and West Germany. Among the developing countries the only large producer is India. The U.S. accounts for more than half of all the coal mined and consumed by the non-Communist countries. The U.S.S.R. and China each produce roughly 80 percent as much as the U.S.

World production of brown coal and lignite has been increasing slowly, reaching 950 million metric tons in 1977. East Germany is the largest producer, with 250 million tons, followed by the U.S.S.R. with 180 million tons.

Among the non-Communist countries West Germany is by far the largest producer of brown coal and lignite. Because of the lower energy content of such coals they contributed only 4 percent of the world's total primary energy in 1977, compared with 25 percent supplied by hard coal.

In the non-Communist countries the fraction of primary energy derived from coal (93 percent of it hard coal) declined from 33 percent in 1960 to 19 percent in 1976. The use of coal has been falling steadily in all markets except for the generation of electricity, which now accounts for more than 75 percent of the coal consumed in the U.S., 57 percent in western Europe, 18 percent in Japan and 36 percent in all other non-Communist countries.

Unlike crude oil, nearly two-thirds of which is shipped from its country of origin, most coal is consumed in the country where it is mined. In 1976 only 8 percent of the world's hard-coal production (190 million tons) moved in international trade, mostly as coking coal for the steel industry. Japan, with purchases of 50 million tons, half of it from Australia, was the largest importer.

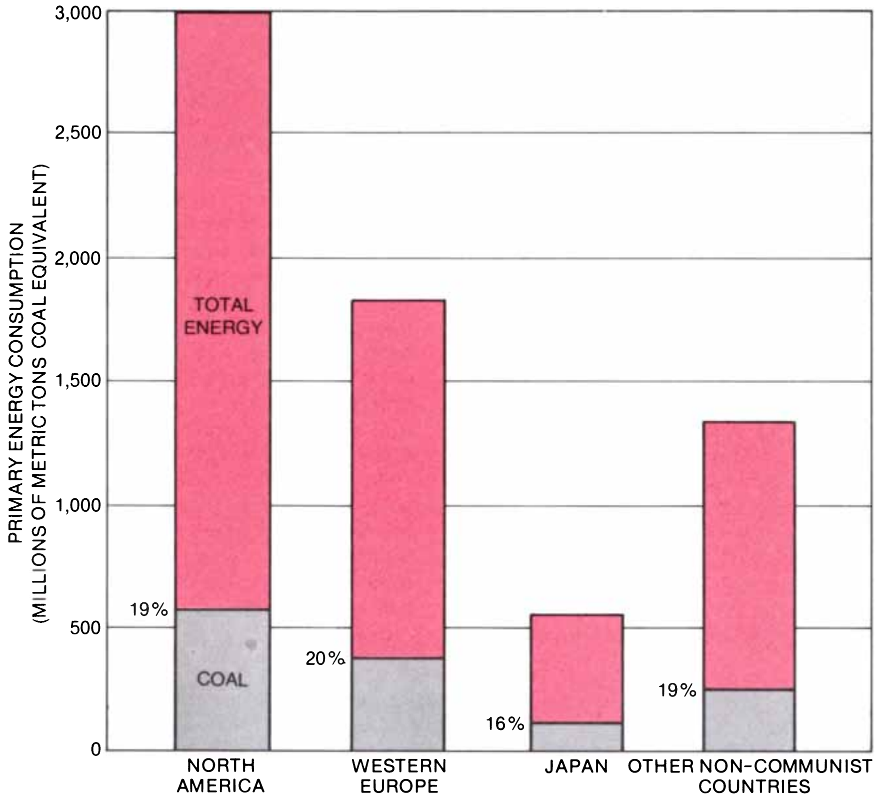
The world's reserves of coal far exceed those of any other fossil fuel and are sufficient to support a massive increase in consumption well into the next century. The World Energy Conference's *Survey of Energy Resources, 1976,* estimates the world's total resources of all ranks of coal to be 11,500 billion metric tons. Of this amount 1,300 billion tons count as known, or measured, reserves. Of this amount in turn 740 billion tons were deemed economically recoverable when the studies of the Workshop on Alternative Energy Strategies were made. Five regions hold 96 percent of the known reserves. North America has 31 percent, the U.S.S.R. and the other countries of eastern Europe 26 percent, western Europe 17 percent, China 15 percent and Australia 6 percent.

Estimates of ultimate recoverability vary from one coalfield to another depending on the accessibility of the coal, which is influenced by such variables as seam thickness and depth and type of terrain. Recovery rates range from 85 to 95 percent in surface mines and from 25 to 70 percent in underground mines.

The Workshop estimate of 740 billion tons of economically recoverable coal includes coal of all ranks. When adjustments are made for the inferior heating value of the lower ranks, the recoverable tonnage comes to about 600 billion tons of hard-coal equivalent, enough for more than 200 years' consumption at current rates. Expressed in terms of oil equivalent, 600 billion tons of hard coal has the energy content of about 3,000 billion barrels, which is four to five times the current estimates of proved reserves of crude oil.

The World Energy Conference esti-

COAL-CARRYING "UNIT" TRAINS have been developed to move coal expeditiously at low cost, usually between one mine and one customer, which is most often an electric utility. The trains shuttle back and forth without being uncoupled, acting much like a conveyor belt. The 110-car unit train that appears in this photograph is one of four trains that together deliver 2.6 million tons of coal per year 600 miles from a surface mine of Amax Carbon Products near Gillette, Wyo., to a power station of the Public Service Company of Colorado near Pueblo, Colo. In 1975 such unit trains carried 169 million tons of coal in the U.S., about half of all the coal that was moved by rail.



CONTRIBUTION OF COAL TO PRIMARY ENERGY SUPPLIES in non-Communist countries averaged about 19 percent in 1976, with little variation among the major regions. In 1960 coal supplied 33 percent of the total energy demand in non-Communist countries and roughly 45 percent in the world as a whole. By 1976 coal's contribution to world energy supplies, including those of the Communist countries, had fallen to about 30 percent. Most of the incremental demand since 1960 has been met by more convenient fossil fuels and nuclear power.

COMPARISON OF OIL AND COAL RESERVES shows that the world supply of potentially recoverable coal is equivalent to some 12,000 billion barrels of oil, or six times the estimated reserves of ultimately recoverable oil. The amount of coal that is economically recoverable with today's technology and prices is considerably smaller, equivalent to about 3,000 billion barrels of oil. Even this figure, however, is perhaps five times the proved reserves of oil.

mate of 11,500 billion tons of total world coal resources is equivalent to some 50,000 billion barrels of oil. On the conservative view that on the average 25 percent of the coal is recoverable, a total of 2,500 billion tons of coal, equivalent to 12,000 billion barrels of oil, could ultimately be produced. That is about six times the estimated quantity of the ultimately recoverable oil.

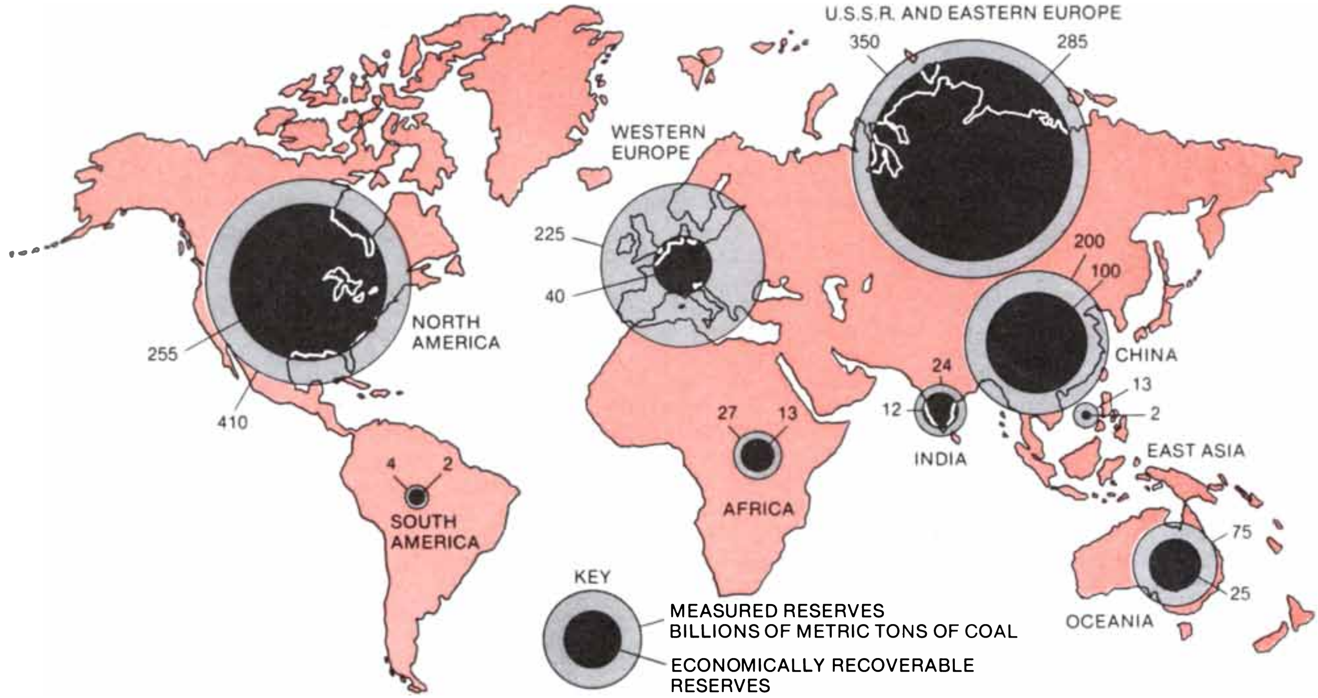
Assessments of world coal resources are less comprehensive than those for oil because exploration for coal has been less wide-ranging and generally less intensive. Many estimates of coal reserves were made when oil was both plentiful and cheap, so that there was little incentive to look for more coal. As interest in coal revives, more reserves are likely to be identified. For example, few of the developing countries have ever had to look for coal. Geology suggests, however, that large areas of the world, including much of the Southern Hemisphere, have a high coal potential.

Australia provides an example of recent coal development. Over the past 20 years geologists have carried out intensive coal explorations, largely inspired by the Japanese need for metallurgical coal. As a result substantial coal reserves have been found in New South Wales and Queensland, and Australia's annual output has been tripled. Measured Australian reserves of bituminous coal are now roughly comparable in energy content to the proved oil reserves of Saudi Arabia. By the same token exploration in southern Africa is now yielding favorable results, indicating that world coal resources may be far greater than has been estimated. Moreover, there are almost certainly major coal deposits offshore; for example, large coal deposits are thought to lie under the North Sea. Such potential coal resources are not included in even the most optimistic estimates of coal reserves.

The true size of the world coal reserve, however, is not a critical value. Known coal resources can support any likely level of exploitation for decades to come. What is in doubt is the willingness and ability of nations to accept large increases in coal production and utilization.

The key to estimating the future role of coal is determining the potential development of coal demand. Large increases in coal production will be realized only if the demand for coal develops and is perceived far enough in advance for the necessary investments in production and transportation to be made. The critical question, then, is whether the demand for coal will continue to stagnate or will rise sharply as the availability of oil declines.

A number of conditions will have to be met before coal can again fill a growing share of the world's energy needs.



TOTAL MEASURED RESERVES OF COAL are 1,300 billion metric tons, of which some 740 billion tons are estimated to be economically recoverable, a 200-year supply at current production rates. (Adjusted for heat content, the 740 billion tons are equivalent to 3,000 billion barrels of oil, the figure given in the bottom illustration on the opposite page.) "Measured reserves" are the quantities of coal in well-surveyed deposits whose extent and quality have been established by adequate sampling. Most of the currently measured coal reserves are located in the Northern Hemisphere, but the rest of the world is likely to increase its share through new exploration for coal deposits.

There must first be a clear recognition that alternatives to oil are needed and that coal is one of the cheapest and most abundant fuels. That implies an awareness of the long-term energy outlook, policy decisions by governments to encourage both the production and the utilization of coal, and public attitudes that make such policies feasible.

There must also be a clear recognition that widespread burning of coal without appropriate controls would pollute the air, with serious consequences for the environment and possibly for the climate. New technologies are being developed to control the emissions of sulfur dioxide and nitrogen oxide from burning coal. Moreover, because the hydrocarbons in coal have a high ratio of carbon to hydrogen the burning of coal yields more carbon dioxide per unit of heat released than does the burning of oil or natural gas. The carbon dioxide content of the atmosphere appears to be rising, and some scientists are concerned about possible effects on the climate. What contribution the combustion of fossil fuels will make to the rise in atmospheric carbon dioxide and what the ultimate consequences will be are not yet known, but the potential for harmful effects must be faced if the burning of coal is to increase greatly. There is need for additional research and development in flue-gas desulfurization, fluidized-bed combustion, chemical coal-cleaning and other techniques for the burning of coal with a minimum of pollution. Governments must also act to settle debates on clean-air standards and encourage investigation of the long-term effect of fossil-fuel combustion on the earth's atmosphere.

Extensive facilities for mining, transporting and burning coal will also have to be built. The construction of such facilities will call for a huge capital investment and for the development of improved methods for handling coal in clean and convenient ways. Better handling techniques are particularly important to industrial, commercial and domestic consumers, who would otherwise prefer to use oil or gas. Residential and small commercial consumers are not going to return to shoveling coal and ashes. Small-scale fluidized-bed units with automatic stoking and ash removal may someday, however, be practical for heating apartment houses and commercial buildings.

The cost of coal is currently competitive with the cost of oil and natural gas for large consumers, even after allowing for the added costs of coal facilities and pollution control. Energy pricing and tax policies could also be designed to supplement the economic incentive for the use of coal if governments want to encourage such use. Choices between coal and other energy sources will be based on a combination of price, environmental, convenience and other factors that bear on the consumer's perceived self-interest.

In the analysis carried out for the Workshop on Alternative Energy Strategies it was clear that in all cases coal production can be expanded to meet the projected demand for coal. Estimates were made of coal demand in the non-Communist countries consistent with various world economic growth rates (ranging from 3 to 6 percent per year) in combination with energy costs at various prices (ranging from $11.50 to $17.25 in 1975 dollars per barrel of oil or its energy equivalent), with various national energy-policy responses ("vigorous" or "restrained") and with policy emphasis placed either on coal or on nuclear power. The countries included now consume about 70 percent of the world's energy.

The lowest projected demand for coal in the year 2000 is 1,610 million metric tons of hard coal or its equivalent, an increase of only 27 percent over the demand in 1976. The low projection is based on an annual economic growth rate of 3 percent and an energy cost of $11.50 per barrel of oil or its equivalent (the 1975 price) in combination with government policies favoring nuclear development. The highest projected demand, 2,575 million metric tons (an increase of 103 percent over 1976), results from a model in which a 5 percent annual growth rate is combined with an energy cost of $17.25 per barrel of oil or its equivalent (a 50 percent increase over 1975) and policies emphasizing coal more than nuclear power. With the same rate of economic growth and cost of energy but with the emphasis changed in favor of nuclear power the demand for coal still reaches 2,225 million tons, an increase of 75 percent over 1976. On this model the North American demand alone would be 1,035 million tons, an 85 percent increase over 1976.

Since the production of coal can potentially exceed any projected demand, a major issue is the possibility of directly substituting coal for the declining oil-based fuels. Our analyses indicate that an all-out program to substitute abundant coal for limited oil could add from 200 million to 600 million tons to the estimates given above for the demand for coal in the year 2000. For example, in an extreme case the industrial demand for coal in 2000 could be twice the Workshop estimates if coal were substituted for an equivalent amount of oil. Although the degree to which energy consumers will be willing and able to use coal instead of oil is uncertain, these analyses emphasize the potential usefulness of coal as a major substitute for oil. At the same time the magnitude of the changeover that would be required indicates the scale of the difficulties that would be faced in making a major transition from oil to coal.

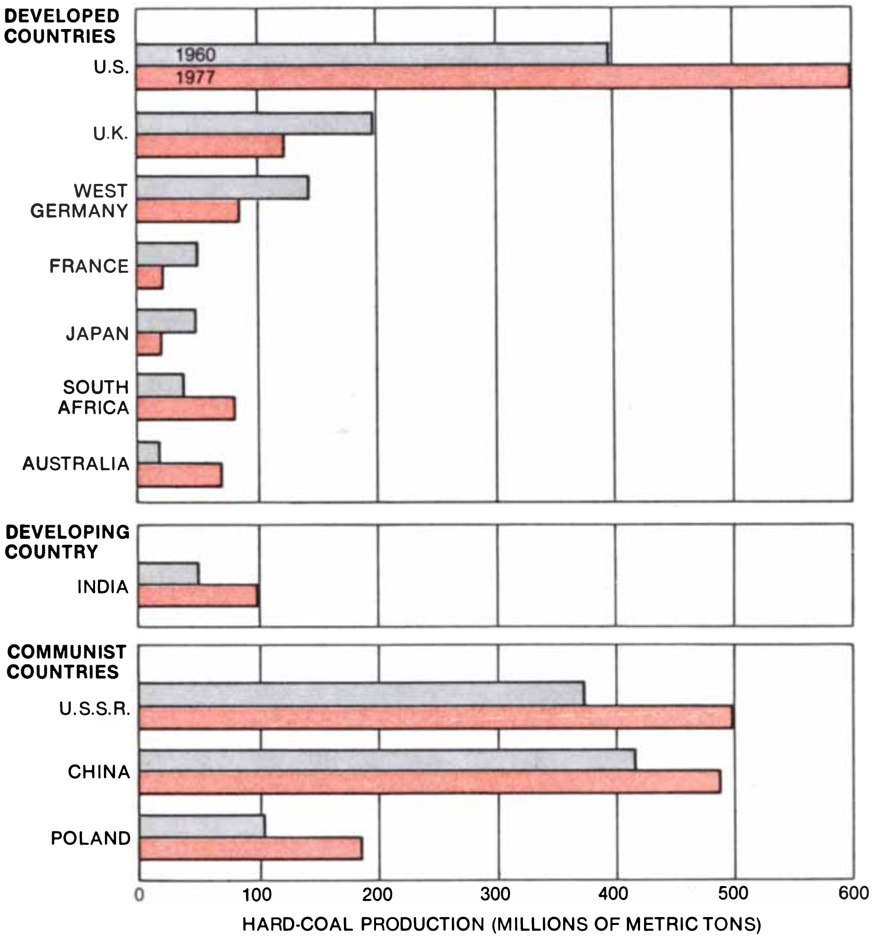
Individual national governments will have to decide the extent to which they will encourage or discourage the expansion of coal-consuming systems. Since long lead times are needed in shifting away from an energy system dominated by oil, decisions must be made soon if such a shift is to be accomplished before oil becomes scarcer and more expensive than it is today. The choices made in the next few years may set the course for the rest of the century.

On the other side of the coin from demand is the potential supply of coal, realistically estimated on the basis of the resources available for opening new mines and building the needed ancillary facilities. Careful estimates of potential coal production for various models of economic growth, energy price and fuel-policy emphasis were prepared by each of the national groups in the Workshop study. Also undertaken was a special analysis of potential coal production in non-Communist countries not represented in the study group and exports from those countries.

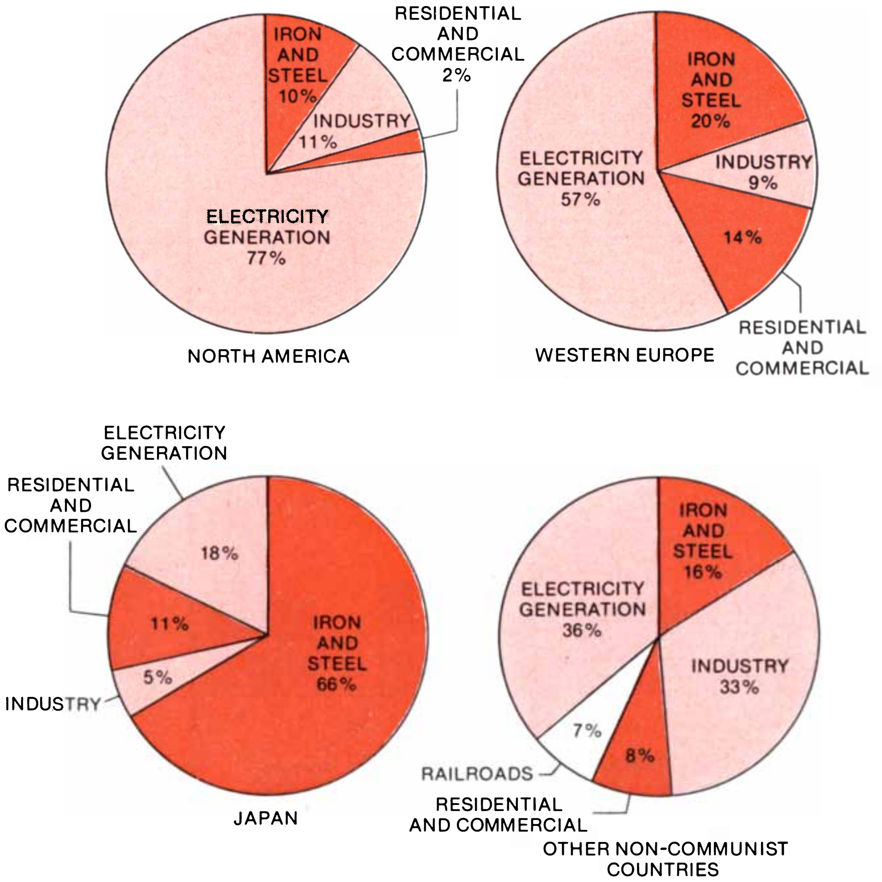
As in the case of the demand estimates, the largest values of potential coal production at the end of the century were elicited by postulating a 5 percent annual economic growth rate, fuel priced at $17.25 per barrel of oil or its equivalent and an energy policy that emphasized coal rather than nuclear power. The projected volume of coal available under these conditions was 3,170 million metric tons, an increase of nearly 150 percent over the 1977 production of 1,280 million tons. The total was arrived at by projecting large increases for North America, the developing countries, Australia and southern Africa. North America was deemed capable of reaching an output of 1,835 million tons, or nearly three times the 1977 tonnage. The lowest projection for coal production from all non-Communist countries in the year 2000 was 2,070 million tons, representing a 62 percent increase over the amount produced in 1977. As we have noted, all the projections of potential production for the year 2000 exceeded by comfortable margins the projected demands for coal made under comparable assumptions of growth rate, energy price and fuel-policy emphasis.

Through the period between 1985 and 2000 western Europe and Japan will be net importers of coal, potentially on a large scale, since their demand will exceed local production. Areas with a substantial potential for exporting coal will be North America, parts of Latin America, southern Africa, South Asia, Australia and certain of the Communist countries. Among the oil-poor developing countries India has the largest potential for producing coal, possibly in the range of 200 million to 250 million tons per year by the end of the century. In several other developing countries with indigenous coal reserves coal could provide a domestic alternative to costly imported oil and a potential source of export revenue. The large coal reserves in China and the U.S.S.R. also make those countries potentially important exporters of coal.

The eventual volume of coal made available for export to Europe and Japan is uncertain because it depends on estimates of indigenous supply and demand for each of the exporting nations. It does appear, however, that international shipments of coal could become a major source of energy for the energy-



WORLD PRODUCTION OF HARD COAL IN 1977 approached 2,500 million tons, an increase of 3.2 percent over the year before and a figure 23 percent higher than the 1960 one. The term hard coal excludes brown coal and lignite, fuels of lower heating value whose production came to about 910 million metric tons in 1977, nearly half of it mined in East Germany and the U.S.S.R. Much of the increase in hard-coal production since 1960 has come in the Communist countries. Total output in non-Communist countries has been between 1.1 billion and 1.2 billion tons throughout the 18-year period. Within this virtually static total there has been a gradual rise in the production of coal in North America, a decline in production in western Europe and Japan and substantial increases in production in southern Africa, Australia and India.



MAJOR MARKETS FOR COAL around the world in 1976 were electric-generating plants and the iron and steel industry. More than three-fourths of the coal consumed in North America was used to generate electricity. In Japan the dominant user of coal was the iron and steel industry. Industrial uses of coal were also significant in developing countries, notably India.

poor industrialized nations and could range up to between 400 million and 900 million metric tons by the year 2000.

The projections of supply and demand made by the Workshop on Alternative Energy Strategies include both steam coal and metallurgical coal. The largest increases available for export will be steam coal. In the long run, however, the distinction between the types of coal may become blurred as new technology makes more types of coal usable for metallurgical purposes.

Coal powered the Industrial Revolution in Europe and North America and was the main industrial and domestic fuel for more than a century. This long history has given rise to a sophisticated coal-production technology. Underground mining, however, is still labor-intensive, particularly in western Europe, and underground-mine coal-production costs are therefore sensitive to wage inflation. Underground mining is also a hazardous occupation, a fact that has led to social and political pressures to enact and enforce strict coal-mining safety and health regulations. One result has been declining labor productivity in underground mines, which is only partly offset by new mining equipment and techniques.

Although underground mining of favorable seams of steam coal and comparatively thin seams of coking coal will remain economically attractive in many areas, significant increases in total coal production will call for a sizable expansion of surface mining. Productivity, or tons of coal per man-day, is typically two or three times higher in surface mines than it is in underground ones. Surface mining, however, has its own problems, most notably the environmental objections to the disturbance of large tracts of land in the course of mining operations and before reclamation of the land.

The environmental, economic and social issues related to coal are potential causes of political conflict in areas of the world that have not experienced large-scale coal mining or coal use. Whether environmental and social opposition will be a constraining factor in the expansion of coal production outside the U.S. remains an open question. In any event the success or failure of U.S. efforts to deal with the issues involved and to meet President Carter's goals for expanded coal utilization (1.2 million short tons per year by 1985) could have an important effect on attitudes toward the development of coal in other countries.

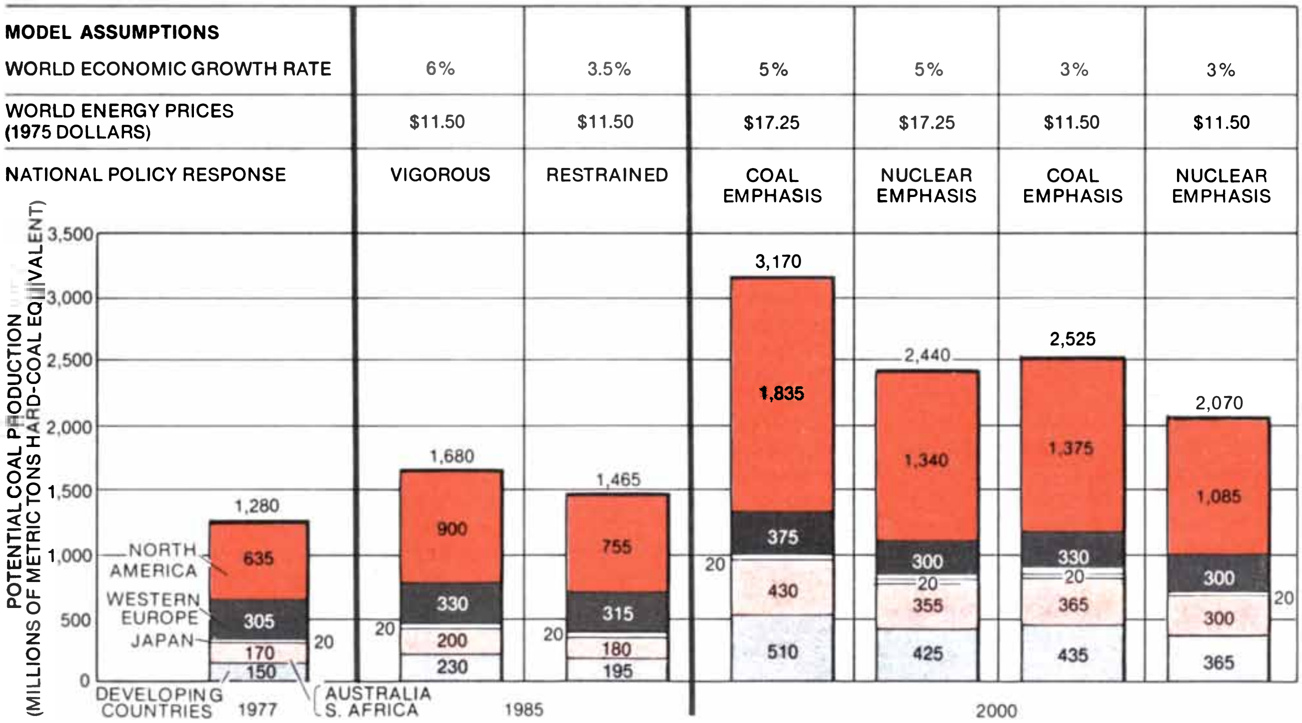
Since the U.S. has half of the world's coal reserves outside the Communist countries and is the largest producer and consumer of coal, its potential for expanded production is important both to itself and to its prospective overseas customers for coal. Projections made for the Workshop indicate that by the end of the century the U.S. could reach an annual production of between 1.1 and two billion metric tons (between 1.2 and 2.2 billion short tons). The low estimate assumes constant real energy prices ($11.50 per barrel of oil or its equivalent) and a policy choice that emphasizes nuclear power as the principal replacement for oil. The high estimate assumes that the real price of energy will rise by 50 percent (to $17.25 per barrel of oil or its equivalent) and that Government policy will encourage the adoption of coal as the principal replacement fuel. In all cases the projections reflect major increases over the 1977 coal production of 695 million short tons.

Although 70 percent of U.S. coal reserves can be reached only by underground mining, it is expected that most new coal production over the next 20 years will come from surface mining in the Western states, particularly on the northern Great Plains. There coal seams are thick, near the surface and under fairly flat land. Western coal typically has a lower heat content per ton than Eastern coal, but in compensation much of it has a low sulfur content and therefore emits less sulfur dioxide per ton of coal burned.

Widespread surface mining in previously undisturbed areas is opposed by environmental groups and local agricultural interests. There are two principal objections to the extension of surface mining in the Western states: the potential despoliation of large tracts of land and the potentially undesirable social and economic effects in a primarily rural, sparsely populated, agriculturally oriented society of small towns and isolated ranches. Agriculturists are concerned that some of the land currently being grazed by livestock may be ruined for future agricultural purposes and that the water requirements of mining and reclamation may conflict with the needs of farmers. Local residents fear that a large influx of construction workers and miners will tear the social fabric of the small towns. They are worried about noise, dust and accidents resulting from large numbers of coal trains passing through their communities. Such concerns cannot be dismissed.

Large new surface mines are nonethe-

POTENTIAL COAL DEMAND TO THE YEAR 2000 was estimated by the Workshop on Alternative Energy Strategies, in which the authors participated. The potential demand for the non-Communist countries in 1985 and 2000 is compared here with the actual 1976 demand when various assumptions are made about the world economic growth rate, the world energy price in constant (1975) dollars and the vigor and direction of national policies. In 1975 the world oil price was $11.50 per barrel for Saudi Arabian light crude oil shipped from the Persian Gulf. The largest potential demand in 2000 is projected for a model that assumes an economic growth rate of 5 percent per year, a 50 percent increase in real energy price (to $17.25 per barrel of oil equivalent) and policies that favor coal over nuclear power.



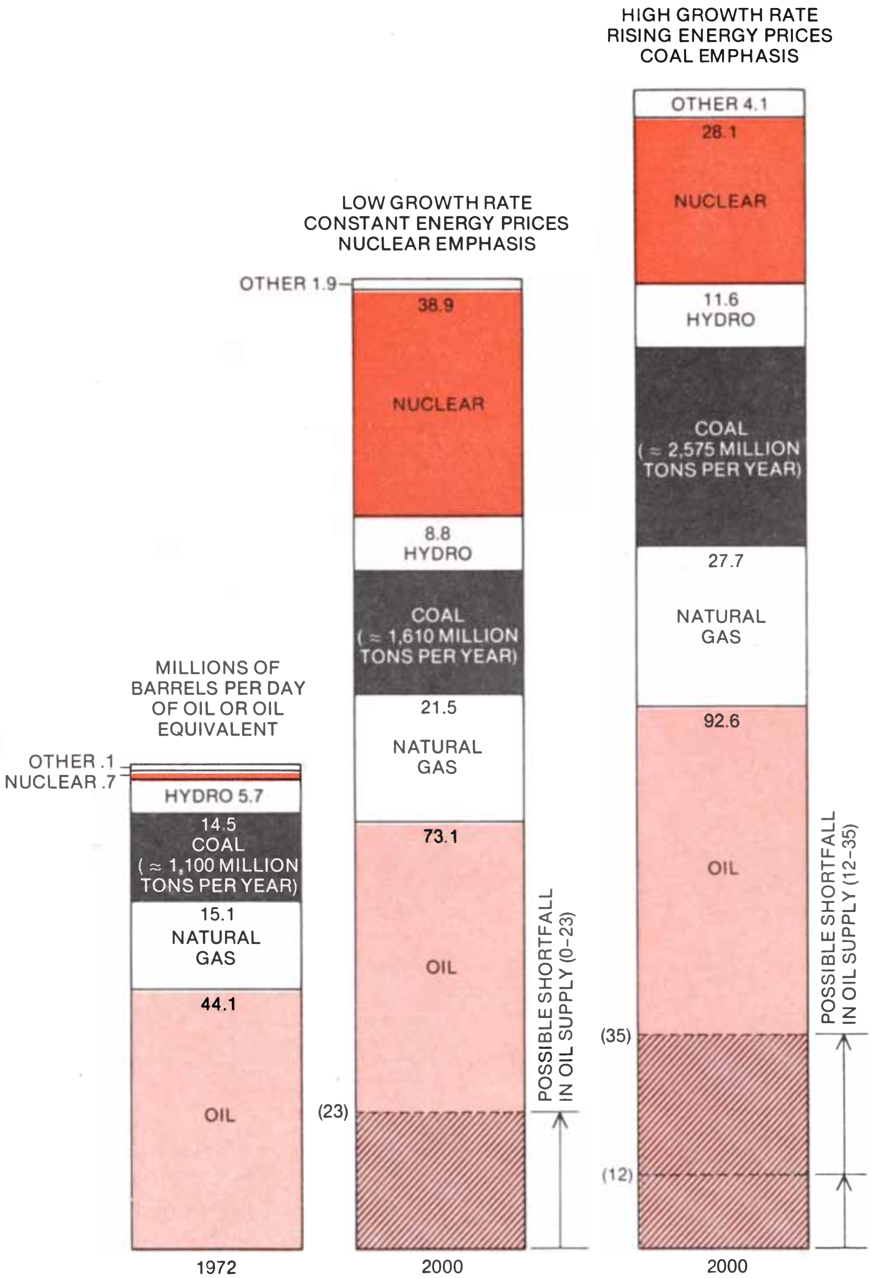
POTENTIAL COAL PRODUCTION TO THE YEAR 2000 was similarly estimated by the Workshop participants and compared with actual 1977 production. In all models but one (1985, low economic growth rate, constant energy prices, restrained policies) the potential production of coal exceeds the potential demand. With the exception of short-term bottlenecks due to restrained energy policies, the consumption of coal in non-Communist countries should not be constrained by shortages of supply up to the year 2000. Assuming high economic growth, a 50 percent increase in real energy prices and a national emphasis on coal, potential production could reach 3,170 million metric tons by 2000, an increase of almost 150 percent over 1977 production. For North America alone potential production of 1,835 million tons would be an increase of 190 percent over 1977. Increases of 200 percent might be achieved in developing countries.

less being developed in the Powder River basin of Wyoming and Montana. The two states have adopted strict surface-mining regulations stipulating that new mining development must proceed hand in hand with reclamation in order to avoid the kind of despoliation seen in the mountainous coal-mining areas of the eastern U.S. In addition Congress has passed comprehensive legislation to regulate surface-mining activities, to protect water resources and to limit mining in certain areas. Over the next several years the mine operators will be in a position to demonstrate whether they can conduct mining operations acceptably and can rehabilitate disturbed lands effectively. The mining companies are confident it can be done at a cost that is reasonable with respect to the value of the coal produced. If the acceptability of rehabilitation programs is conclusively demonstrated, as it has already been in Europe, environmental issues may recede as a constraint on the future development of coal mines.

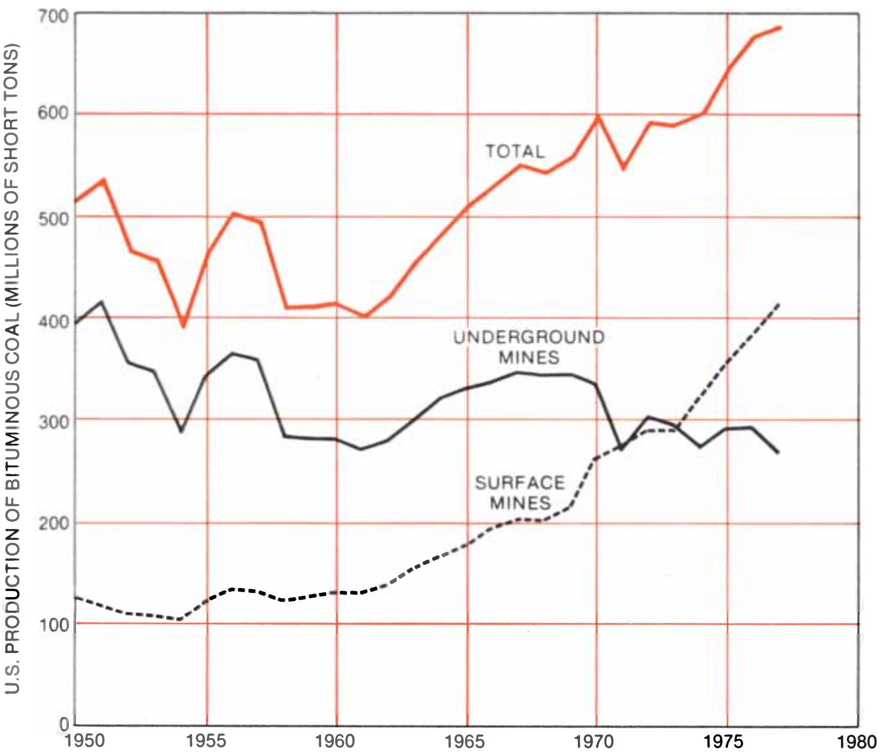
Programs are also being evolved at both the Federal and the state levels to help local governments in coal-mining areas plan for (and meet the cost of) expanded social and community services. Apprehension about long rumbling coal trains may eventually be allayed by policies that favor coal-slurry pipelines for moving coal to market or call for the rerouting of some rail lines. Nevertheless, continued opposition to new mines and conflicts over detailed mining regulations, regulatory jurisdiction, royalty sharing, land-use planning, water resources and local government assistance may slow the expansion of Western surface-mining activities. All in all, the uncertainties are so great that the Workshop projections of potential coal production in the year 2000 differ by as much as 500 million metric tons annually, depending on whether or not national policies effectively encourage the development of U.S. coal resources.

The year-2000 projections that embody high estimates of coal production also assume substantial increases in underground mining in the eastern part of the U.S., consistent with the assumed increases in energy prices. For some years the cost of coal mined underground has been rising because of increasingly stringent mine-safety regulations, an associated decline in the productivity of underground mining and a gradual depletion of the most accessible underground reserves. It is impossible to predict whether this trend will continue. The development of Eastern underground coal would, however, stimulate employment and economic activity in parts of the country that have a history of unemployment and underdevelopment. The continued importance of coal mining in the Eastern states seems to be reflected recently in certain Federal policies favoring Eastern coal. In any case most projections of U.S. coal production predict significant increases not only in the West but also in the East and the Middle West.

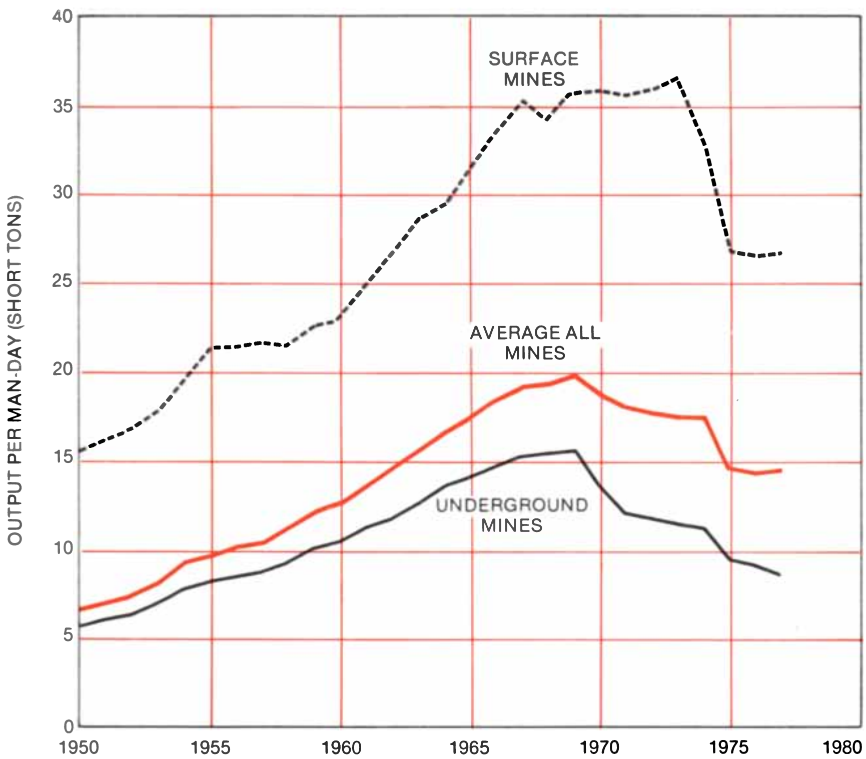
The forecast of a rising international trade in coal made by the Workshop on Alternative Energy Strategies implies large increases in coal exports from the U.S. Will the U.S. be willing to accept environmental and social consequences and the depletion of a nonrenewable national resource in order to provide energy for other countries? Good arguments can be advanced for its doing so. For the U.S. coal is not a scarce resource: there-



PROJECTION OF TOTAL ENERGY DEMAND IN THE YEAR 2000, as estimated by the Workshop, suggests how high the consumption of oil might climb in non-Communist countries if supplies were unlimited. The low-economic-growth, low-energy-demand projection for the year 2000 shows that the appetite for oil might barely be satisfied if new oil discoveries remained high, if the Organization of Petroleum Exporting Countries (OPEC) placed no limits on production, if coal consumption were to increase 50 percent and if a strong emphasis were placed on nuclear energy. A shortfall of up to 23 million barrels per day could arise, however, if discoveries were to falter and OPEC were to impose production limits. In the high-economic-growth, high-energy-demand projection oil shortfalls of between 12 million and 35 million barrels per day are possible. If coal were developed to its full potential of 3,170 million metric tons per year, compared with the expected demand of 2,575 million tons, 600 million tons (equivalent to eight million barrels of oil per day) would be available to relieve shortfall in oil.



TOTAL OUTPUT OF U.S. COAL MINES has increased by less than 3 percent per year since 1954, when production fell to the lowest point since the depression years of the 1930's. Output of underground mines remains below the levels of the early 1950's. In the meantime, however, output of surface mines has climbed by more than 7 percent per year for the past 10 years.



PRODUCTIVITY OF U.S. COAL MINES, measured in tons per man-day, rose rapidly between 1950 and 1969. The sharp drop since then has had a number of causes, including new mine-safety and health regulations and, in the case of surface mining, legislation requiring the restoration of the lands disturbed by mining operations. Productivity in the hard-coal mines of Europe, virtually all underground, is much lower than productivity in the U.S.: it is about 2.5 tons per man-day in Britain, about 1.9 tons in France and about 3.8 tons in West Germany.

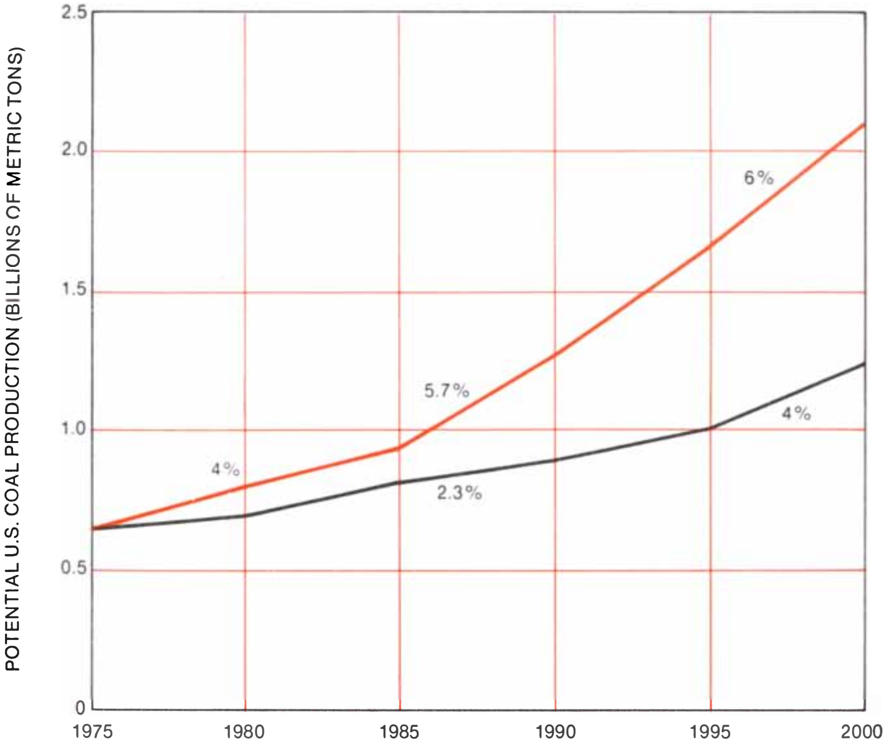
fore exporting it to help other countries meet their energy needs, while simultaneously offsetting the high cost of importing oil to the U.S., would seem to be in the broad national interest. Exporting coal would have a positive effect on economic development, employment and the balance of foreign-trade payments. These benefits will not accrue, however, unless three conditions are met. First, the demand for coal exports must be ensured by firm, long-term contracts, based on the buyers' expectation that the coal will be available for export. Second, the environmental and social consequences of expanded coal exports must be made acceptable to the American public. Third, there must be sufficient advance planning to provide for all the facilities that an increase in coal exports will require, particularly the expansion of deep-water ports. Such matters are not yet prominent in the U.S. energy debate. We expect they will be in the 1980's.

Major increases in coal production will call for the construction and operation of a large mining and transportation network to win coal and move it to its markets. Special analyses of mining equipment, transportation facilities, labor and the other resources needed for the rapid expansion of U.S. coal production were prepared for the Workshop. One study conducted by the Bechtel Power Corporation showed that in order to reach an annual production of 2,000 million metric tons of coal at the end of the century the U.S. would need to open 377 new Eastern underground mines, each mine yielding two million tons per year, 75 Eastern surface mines yielding four million tons per year, 232 Western surface mines yielding six million tons per year and eight Western underground mines yielding two million tons per year. The total cost, in 1975 dollars, would be $32 billion. The investment in transportation facilities to carry the coal to consumers would be much higher: $86 billion. That investment would provide 1,400 "unit" railroad trains (trains of about 100 hopper cars each, dedicated solely to carrying coal between the minehead and the consumer), 3,200 conventional railroad trains, 500 large river barges, 9,400 trucks and nine coal-slurry pipelines, each pipeline with a capacity of 25 million tons per year.

The necessary transportation capacity for the bulk of the coal could be provided by either railroad or slurry pipeline or some combination of the two. Railroads are the most suitable where maximum flexibility is required for local transport and small-to-moderate volumes. Slurry pipelines would be the most economic for large and constant volumes over long distances and fixed routes. An extensive railroad network is already in place. The development of slurry pipelines in the U.S., on the other hand, faces formidable political obstacles that would have to be overcome. Federal legislation to encourage such development has been introduced, but its passage is not yet assured.

The bottleneck most likely to develop in expanding coal exports is seaport facilities. Although the U.S. is already the world's leading exporter of coal (50 million metric tons in 1977), its port facilities are generally inadequate for any substantial expansion of steam-coal exports and have a limited capacity for handling large bulk carriers.

The Workshop concluded that the infrastructure and transportation requirements for a major coal expansion are large but not unattainable in view of the fact that the maximum annual growth rate projected for U.S. coal production is about 6 percent. Nevertheless, even more modest goals will become increasingly hard to achieve if major investments are deferred into the mid-1980's. The industries that provide mining and transportation equipment and resources are capable of doing the job. Clear signals of increased demand are needed, however, to get the necessary expansions under way.



U.S. COAL PRODUCTION BY THE YEAR 2000 could climb to more than two billion metric tons per year if economic growth rates are high, if world energy prices rise 50 percent above current levels and if Government policies encourage coal production and use. The lower curve shows that potential coal production could be substantially less if growth rates are low and if policies favor nuclear power rather than coal. Percents adjacent to different curve segments indicate average annual growth rates in coal production corresponding to the two models.

Even with an assumed 50 percent increase in the real price of energy ($17.25 per barrel of oil or its equivalent in 1975) and policies to support the development of synthetic fuels the Workshop studies indicate only modest potentials by the year 2000 for the conversion of coal into synthetic gas and oil. The maximum volume of coal-based synthetic fuels in our studies was slightly more than four million barrels per day of oil or its equivalent, with two-thirds of the synthetic gas and most of the synthetic oil being produced in North America. At a typical conversion efficiency of 60 to 70 percent this volume of synthetic fuel would require about 500 million metric tons of coal.

For the U.S., with its large coal reserves, the production of synthetic fuels from coal could be an attractive alternative to rising imports of oil and liquefied natural gas. It would allow consumers to keep their existing oil- and gas-fired equipment, thereby putting the burden of new investment on the energy producers and processors rather than on the distributors and consumers. It would also simplify the control of air pollution by concentrating the utilization of coal in fewer locations. On the other hand, converting coal into other forms of fuel adds to the consumer's cost of energy because of the relatively low efficiency of the conversion process and the high capital cost of the conversion facilities. Synthetic-fuel production requires the mining of larger volumes of coal than direct coal utilization, and if it were pursued in the Western U.S., it would place an additional load on the available resources of water for cooling and for land reclamation after surface mining.

Unless there are sharp increases in the price of petroleum, synthetic fuels will continue to cost more than natural oil and gas. Because of the high cost of synthetic fuels they will not be developed without vigorous policy support by governments. Eventually the decline of world oil production and the rise of oil prices may cause some major oil-importing regions (particularly North America and possibly western Europe) to begin exploiting their coal resources to produce synthetic oil and gas. Having a synthetic-fuel capability ready in time calls for early policy action to support the demonstration of commercial synthetic-fuel technologies. The lead times are long, and governments would be prudent not to wait for rising oil prices before they undertake to encourage investments in synthetic fuel.

Several major synthetic-fuel projects have already been proposed in the U.S. but will not be viable without Government support in the form of loan guarantees, special pricing arrangements or favorable regulatory treatment. Although some of the projects have preliminary approval from regulatory agencies, the appropriate level of Federal support for synthetic fuels is still being debated in the Department of Energy and in Congress. If favorable Government decisions are made soon, the production of synthetic fuels could begin before 1985 and could reach significant levels by 2000.

Coal is the one fossil fuel likely to remain abundant at relatively low cost for the remainder of this century and well into the next. Therefore it is one of the major replacement fuels available to bridge the gap from the waning era of abundant oil and gas to a future era for which the world must develop alternative energy sources. Coal will be even more important if nations are reluctant, as some are today, to turn strongly to nuclear power. Conversely, decisions to avoid or limit the consumption of coal will increase pressure for the further development of nuclear power and for restrictive energy-conservation measures as alternatives to an unsatisfied demand for oil.

If coal is to serve as such a bridge to a new energy era, governments will soon have to adopt the policies necessary to allow and encourage its expanded utilization. Both public and private investment decisions will be needed to provide funds for opening new mines, building transportation systems including deep-water port facilities and developing the infrastructure for the handling and burning of coal by consumers. Finally, the expanded exploitation of coal will call for satisfactory solutions to the environmental problems associated with the mining and burning of coal, wherever it is done.

The Assembly of Cell Membranes

The two sides of a biological membrane differ in structure and function. Studies of animal viruses and bacteria have helped to reveal how this asymmetry is preserved as the membrane grows

by Harvey F. Lodish and James E. Rothman

The membrane that surrounds a living cell is much more than a simple container or boundary; it not only defines the extent of the cell but also acts to maintain a distinction between inside and outside. Certain ions, for example, are pumped into the cell by large molecules embedded in the membrane and other ions are pumped out. Nutrients are taken up by the cell and concentrated inside by other components of the membrane.

For a membrane to preserve such concentration gradients one absolute requirement is that it form a closed vessel; otherwise a substance pumped in through the membrane would immediately leak back out. All known biological membranes form closed compartments. Another essential property is sidedness: the inner surface and the outer surface of a membrane must function differently. If they did not, an ion or a molecule pumped in at one point might be pumped out at another, with an expenditure of energy in each direction. Sidedness is also expressed in other properties of membranes. In animal cells, for example, receptors for hormones and for other chemical signals are associated with the membrane, and so are markers that identify a cell to its neighbors. These facilities for intercellular communication should be accessible at the outer surface; they would be useless inside.

In the past several years it has been established that the functional asymmetry of membranes reflects an underlying structural asymmetry. The protein molecules embedded in membranes or attached to them have a fixed orientation. Some of them are exposed only on the inner surface and some only on the outer surface. Others span the thickness of the membrane, but even they can invariably be regarded as having a fixed, asymmetric orientation. Every protein of the same type points in the same direction. The smaller phospholipid molecules that make up the structural matrix of the membrane have also been found to have an asymmetrical distribution, although it is a partial rather than an absolute asymmetry.

A major challenge to biologists is to understand how cells replicate and divide, and an important part of this problem is how cells form membranes. When a cell divides, new membrane constituents must be synthesized and then properly assembled. The new membrane must be laid down with great precision, and in particular the new components must have the same topographical arrangement as the old ones to ensure that the new membrane functions correctly.

In the past few years the outlines of the schemes cells use to solve these problems have been greatly clarified, although many biochemical details remain to be elucidated. The progress has resulted from the study of a few highly specialized kinds of membranes, which incorporate only a handful of molecular components and so are well suited for experimentation. It is these studies that we should like to discuss here.

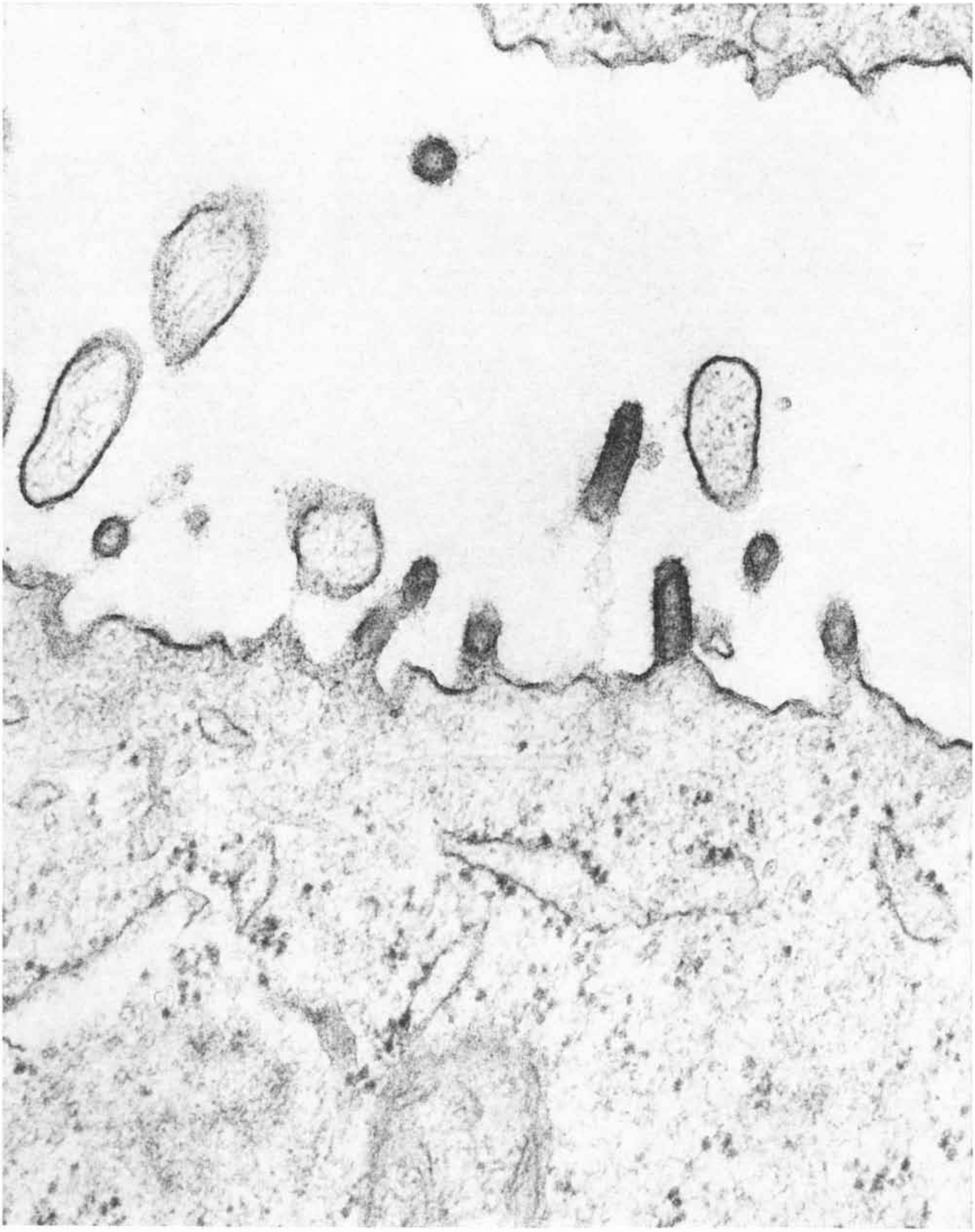
The simplest cells, those of bacteria, have only one membrane, the plasma membrane that envelops the cell itself. Such organisms are prokaryotes, or cells without well-formed nuclei. Higher plants and animals are made up of eukaryotic cells, which do have nuclei. In addition to a plasma membrane eukaryotes have a great diversity of specialized internal membranes, which surround and define the subcellular structures called organelles. The nucleus itself is bounded by a system of membranes. In mitochondria there are two membranes, and the inner one has a vital role in the manufacture of adenosine triphosphate (ATP), the "energy currency" spent elsewhere in the cell. The chloroplasts of green plant cells also have two membranes, which may operate in a similar way. Another organelle of particular importance in the study of membranes is the endoplasmic reticulum, which is specialized for the fabrication of secreted proteins and new membrane material.

In our work we have been concerned mainly with the plasma membrane. (This should not be confused with the cell wall, which lies outside the plasma membrane in plant and bacterial cells and is not a membrane at all but a porous skeletal organ.) All cells have a plasma membrane, and in animals it is the only barrier separating the cell from its environment. There are intriguing differences between the plasma membrane and the various internal membranes of eukaryotic cells, but all biological membranes are remarkably similar in their basic structural features.

If a membrane is broken down into its molecular constituents, it is found to consist of three kinds of substance: lipids, proteins and carbohydrates. Lipid molecules are by far the most abundant; proteins, however, are much larger molecules, and in mass these two components are roughly equal. In any given membrane there are only a few species of lipid, each present in many copies; there is a vast assortment of proteins, on the other hand, some of them represented by no more than a few molecules per membrane. Carbohydrate is a minor constituent (in mass) and is mainly associated with the plasma membrane rather than with internal membranes. The carbohydrate in the membrane is always combined chemically with other membrane components; it binds to lipids to form glycolipids and to proteins to form glycoproteins.

It is the lipids that are responsible for the structural integrity of the membrane. Each lipid molecule is said to be amphipathic, which means it has one part that is hydrophobic and another part that is hydrophilic. If the hydrophobic part were separated from the rest of the molecule, it would be insoluble in water, just as salad oil is. The hydrophilic part, if it were isolated, would be soluble in water and insoluble in oil.

Two kinds of lipids are included in membranes, cholesterol and phospholipids. Cholesterol is found almost exclusively in the plasma membrane of mammalian cells and is absent entirely in



REGION OF CELL MEMBRANE envelops a virus particle as the virus buds off from an infected cell. The membrane appropriated by the virus incorporates only a single kind of surface protein, which makes the viral coat membrane a convenient experimental system for studies of how membranes are assembled. Each molecule of the surface protein is oriented in the membrane so as to form a spike on the surface of the virus. The cell is infected with vesicular stomatitis virus (VSV), which causes an influenzalike disease in farm animals. The virus particles are the small, dark, cigar-shaped objects emerging into the intercellular space at the top; a few are seen end on and appear round. The larger, irregular vesicles are cellular material isolated by the sectioning of the cell. This transmission electron micrograph, which magnifies the image some 60,000 times, was made by David M. Knipe of Massachusetts Institute of Technology.

bacteria. The hydrophilic head in cholesterol is a hydroxyl group (OH). Phospholipids are found in all membranes. The hydrophilic region includes a negatively charged phosphate group (PO_4^-); in many phospholipids a compensating positive charge is contributed by some other chemical entity, such as an amino group (NH_3^+).

Because the two parts of a membrane lipid have incompatible solubilities the molecules spontaneously organize themselves in the form of a bilayer, or double layer of molecules. In this way the hydrophobic portion of each molecule is shielded from water, whereas the hydrophilic heads are immersed in it. Such a bilayer sheet is a minimum-energy configuration for a suspension of lipids. The only place where the nonpolar tails must interact with water is at the edge of the sheet, and even that unfavorable contact can be eliminated: the growing bilayer simply folds over to form a closed vesicle, which has no edges.

The lipids of all membranes have the same organization, with two hydrophilic surfaces separated by a hydrophobic core. It is this lipid bilayer that determines the overall morphology of membranes and accounts for the fact that they form large sheets or vesicles. The structure of the bilayer also provides for one of the essential functions of the membrane: it is impermeable to most water-soluble molecules because they are insoluble in the oily core region.

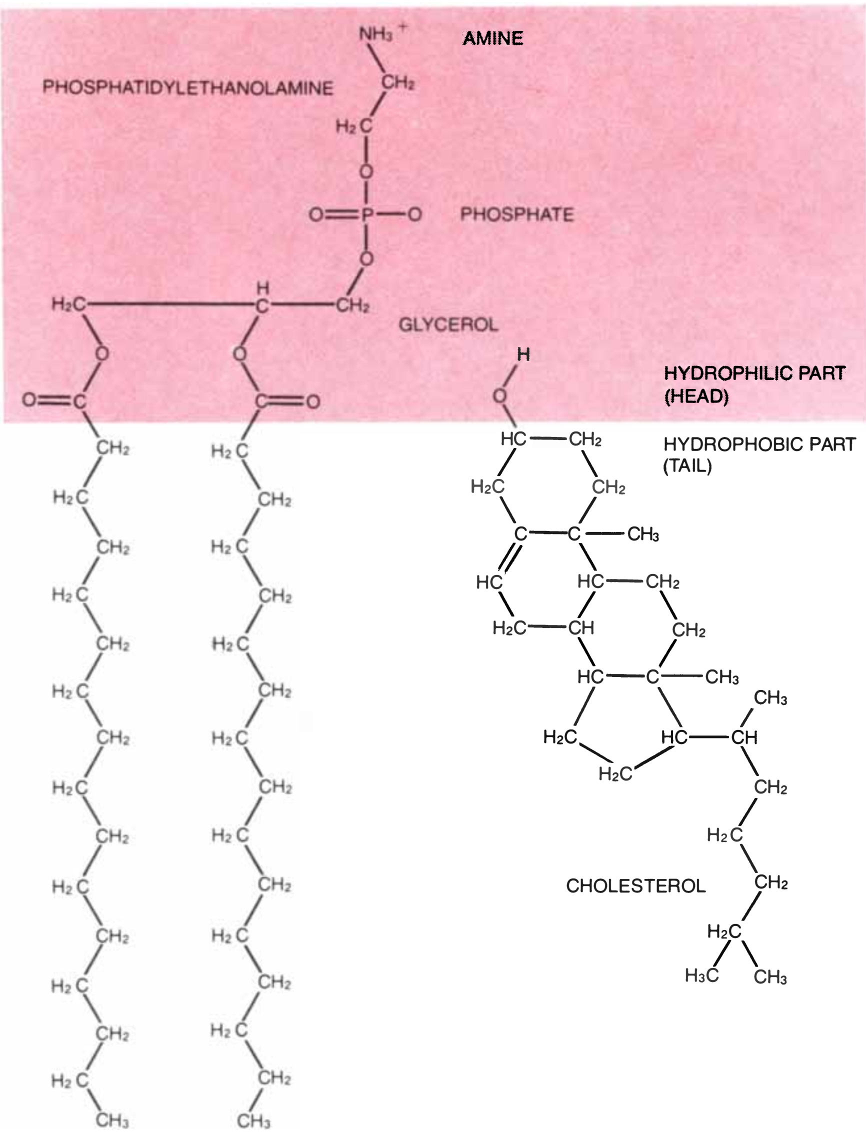
The bilayer is by no means static; on the contrary, each monolayer is a two-dimensional fluid. The lipid molecules are free to diffuse laterally, like the molecules in a thin film of liquid. They exchange positions as often as a million times per second. In the third dimension, however, the mobility of the lipids is severely restricted. For a lipid molecule to jump from one monolayer to the other, a motion called a flip-flop, the polar head must pass through the hydrophobic core of the membrane, where it is insoluble. Recent measurements indicate that the flip-flop rate is so low that a given lipid molecule makes the passage no oftener than once a month.

Because the lipid bilayer is a two-dimensional fluid any protein molecules embedded in it can also diffuse laterally. The fluidity simplifies the task of membrane assembly, since both lipid molecules and protein molecules can be inserted anywhere in a monolayer with confidence that they will eventually reach any other point. At the same time the low rate of flip-flop could allow the two opposing monolayers to preserve different lipid and protein compositions. Indeed, the lipid bilayer has been found to be asymmetric in composition in every biological membrane that has been examined for this property. The function of the asymmetry is unclear.

The interactions of hydrophilic and hydrophobic chemical groups are important in proteins as well as in lipids. Proteins are polymers (polypeptides) made up of amino acids strung together in a linear sequence. Of the 20 amino acids specified by the genetic code six are strongly hydrophobic and a few others are weakly hydrophobic; the rest are hydrophilic. If a protein is considered as being merely a straight chain of amino acids, then a pattern can seldom be discerned in the sequence of hydrophilic and hydrophobic units. The native conformation of protein molecules, however, is not a straight chain but a densely folded one. In that form the soluble proteins of the cell cytoplasm generally have an excess of hydrophilic units at the surface of the molecule and an excess of hydrophobic ones inside.

Proteins associated with membranes fall into two classes. Integral membrane proteins are those with a portion of each molecule embedded in the lipid bilayer. All integral proteins that have been studied in detail have been found to span the full width of the bilayer and so to have regions exposed on both sides of the membrane. The other class is made up of peripheral membrane proteins, which are not inserted into the bilayer at all but reside at one surface or the other. Each peripheral protein molecule is bound to an integral protein.

In peripheral proteins the balance of hydrophilic and hydrophobic amino acids is much like that in cytoplasmic proteins. The peripheral proteins can be removed from the membrane by treatments that do not destroy the integrity of the lipid bilayer, and when they have

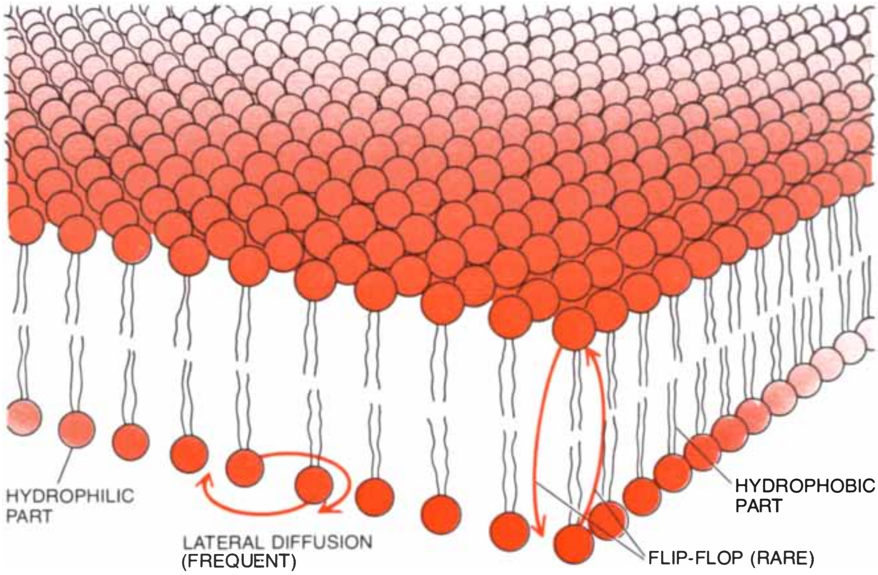


LIPID MOLECULES in the cell membrane include phospholipids and cholesterol. Both kinds of lipid are amphipathic: one end of each molecule, called the head group, is polar and hydrophilic; if it were isolated, it would be water-soluble. The other end, the tail, is nonpolar and hydrophobic, or insoluble in water. Cholesterol is found mainly in the outer, plasma membrane of mammalian cells; phospholipids of various kinds are components of all biological membranes.

been removed, they are fully soluble in water. Integral proteins, in contrast, are generally insoluble in water because they include substantial regions where most of the exposed amino acid units are hydrophobic. It is these regions that are embedded in the hydrophobic core of the membrane. They cannot be removed without disrupting the lipid bilayer itself.

The regions of an integral protein that extend into the external fluid or into the cytoplasm have the hydrophilic character common to soluble proteins. These hydrophilic regions essentially preclude the possibility of a flip-flop transition for membrane proteins. If the small polar head of a lipid molecule can rarely be forced through the core of the bilayer, then the much larger hydrophilic volume of a protein molecule is permanently excluded.

A cell that has proved useful for studies of membrane topography (although not for studies of membrane assembly) is the red blood cell. The plasma membrane of the red cell has only two main species of integral protein and several peripheral ones, and their orientation can readily be determined. From studies of the red cell several principles of membrane construction have emerged. All integral proteins are inserted asymmetrically, so that every molecule of a given protein species has the same orientation in the lipid bilayer. The lipid bilayer itself is asymmetric. Carbohydrates can be attached either to proteins or to lipids, but in each case they are invariably on the exterior surface rather than on the cytoplasmic surface. Peripheral proteins are usually found on the cytoplasmic side.



BILAYER OF PHOSPHOLIPIDS forms the structural matrix of a membrane. The lipids are arranged tail to tail so that only the hydrophilic heads are exposed to the aqueous medium on both sides of the membrane. This is the minimum-energy configuration for a suspension of lipids in water. Each monolayer is a two-dimensional fluid: lipid molecules can diffuse laterally as in a liquid film but can rarely execute a "flip-flop" transition from one layer to the other.

If suspended lipids can spontaneously form a closed vesicle, it is not unreasonable to suppose a complete and functional biological membrane might be assembled the same way. Certain viruses, muscle fibers and ribosomes spontaneously assemble themselves, but these structures are highly ordered and in some cases even crystalline, whereas membranes are not.

The self-assembly hypothesis can easily be tested. A membrane can be dispersed into individual protein and lipid molecules by disrupting the lipid bilayer with a high concentration of detergent. The detergent is a synthetic lipid that does not form a bilayer but instead forms droplets called micelles. The lipids and proteins of the membrane no longer need to stay together to avoid water; they can achieve the same thing inside the detergent micelles. With a sufficiently high concentration of detergent the components of the membrane can be completely separated, with no more than one membrane protein in each detergent micelle.

If the detergent is removed by any of several procedures, a membrane vesicle re-forms spontaneously, incorporating both lipids and proteins. The individual protein molecules are often functional, with the capability, for example, of transporting ions across the permeability barrier. There is just one flaw in such reconstituted membranes: they are almost always symmetrical. The proteins are embedded in the lipid bilayer, but no longer do all the molecules of a given species of protein point in the same direction. There are exceptions, which are interesting in themselves, but in most cases the absolute asymmetry of the original membrane is lost.

Why is the asymmetry not re-created in this experiment along with the rest of the membrane structure? In the self-assembly of a membrane the formation of the lipid bilayer and the insertion of the membrane proteins take place concurrently. Because there is no permeability barrier present during the assembly, proteins can be inserted from both sides, resulting in an approximately symmetric structure.

The alternative to the self-assembly hypothesis is that integral membrane proteins are given their orientation at the time of insertion. That mechanism requires that the protein always be inserted from the same side. For the sidedness to be defined and preserved at all times, the membrane must always have the form of a closed vessel. From these arguments a fundamental conclusion follows: Membranes can grow only by expansion, by the insertion of new material into a membrane that is already a closed vessel.

This mechanism also has a difficulty. If a protein (or a lipid) is to be inserted from one side of the membrane, then in many cases a hydrophilic region will have to be pushed through the core of the bilayer. It is the very inability of these regions to cross the membrane that preserves the asymmetry of the proteins once they are inserted. Our experiments have helped to identify the methods by which this intracellular rope trick can be performed.

It would be a formidable task to study the synthesis of membrane proteins in a typical eukaryotic cell or even in a prokaryotic one. There is too great a variety of proteins, each one present in too small a quantity. We and our colleagues have therefore turned to a simpler membrane system, that of a virus whose outer coat is a phospholipid membrane appropriated from the plasma membrane of an infected animal cell. The membrane segment is densely coated with integral membrane protein molecules, but they are all of one species, determined by the viral genome (the complete set of viral genes).

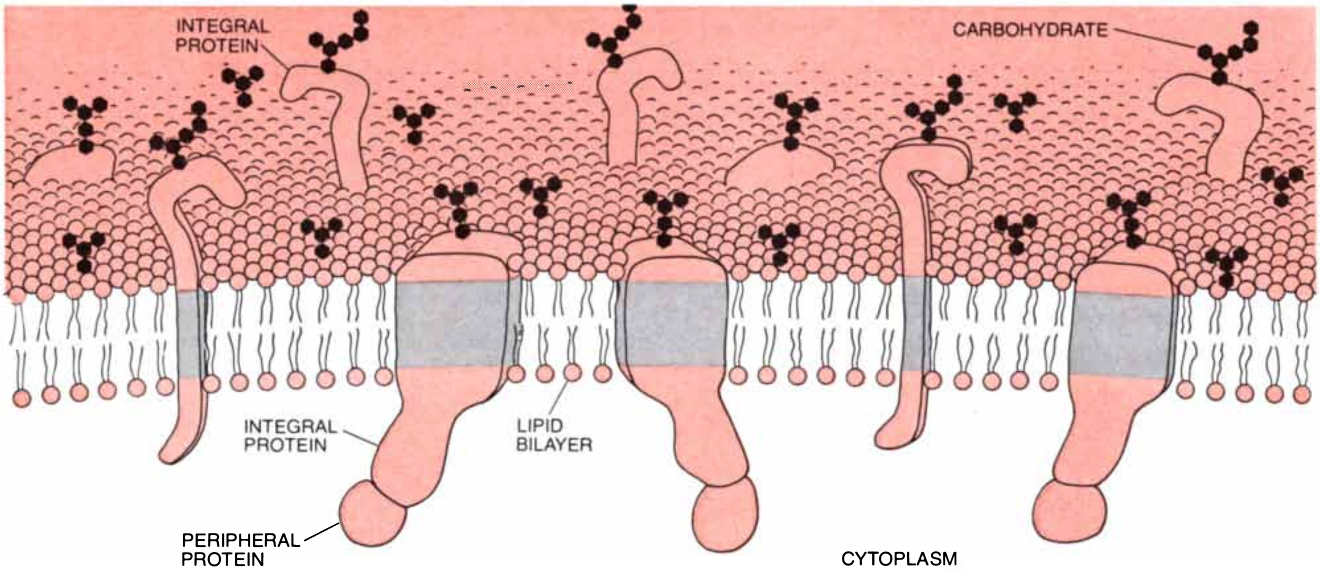
The virus is vesicular stomatitis virus (VSV), which infects mainly farm animals and elicits symptoms similar to those of influenza. When a VSV particle enters a cell, it halts the normal synthesis of the host cell's protein and commandeers enzymes and other components of the synthetic apparatus for its own purposes. In the subsequent period only viral proteins are assembled. At the same time the genome of the virus, which is a strand of RNA, is replicated many times by means of enzymes coded by the viral genome and carried in the virus particle. The life cycle is completed when the replicated RNA and the viral proteins come together at the plasma membrane. Each copy of the RNA, along with a set of proteins, is enveloped in a section of membrane and buds from the cell to form a new generation of virions, or virus particles. The viral membrane has the lipid composition of the host-cell membrane, but it incorporates only proteins specified by the virus.

The VSV genome encodes just five polypeptides, all of which are found in the mature virus particle. Three are closely associated with the genome RNA; they are called nucleoproteins, by analogy with certain proteins in the eukaryotic cell nucleus. Two of the nucleoproteins are the enzymes responsible for the replication and transcription of the viral genome; the third nucleoprotein, designated *N*, probably has a structural function and is the most abundant of the three. A fourth protein is labeled *M* for matrix; it is synthesized as a soluble cytoplasmic protein but subsequently is incorporated into the virion as a peripheral protein on the inner surface of the plasma membrane. The last protein, called *G* for glycoprotein, has been the major focus of our investigations. *G* is a membrane protein that spans the lipid bilayer; in electron micrographs it is visible as a pattern of spikes on the surface of the virion. It can also be detected in the plasma membrane of infected cells during budding. The amino acid sequence and three-dimensional structure of *G* have not been determined, but the protein is known to include about 550 amino acid units; there are also two carbohydrate side chains attached to it. The orientation of *G* in the bilayer is unquestionably asymmetrical: most of the polypeptide and all the carbohydrate is exposed on the external side of the membrane and only a stub of about 30 amino acids protrudes on the cytoplasmic side.

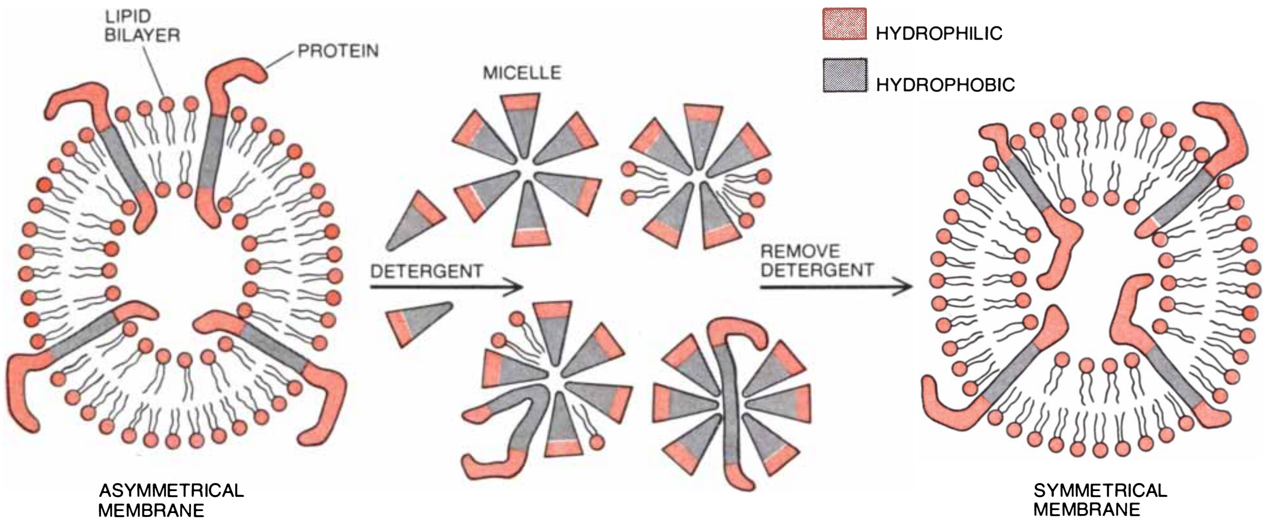
The steps involved in the budding of the virion are not known with certainty, but a likely sequence of events can be proposed. The genome RNA and the nucleoproteins evidently come together in the cytoplasm, forming a structure called the nucleocapsid. At the same time *G* appears in the plasma membrane. It is the only membrane glycoprotein being synthesized by the cell and it is made in large quantities, so that at least some parts of the membrane probably acquire a dense coat of it. The matrix protein, *M*, is thought to serve as a bridge linking the nucleocapsid to the cytoplasmic stub of *G*. If that is so, a plausible mechanism of budding is easily imagined. Whenever a nucleocapsid happens to approach the membrane, *M* proteins form cross-links at the point of first contact between the nucleoproteins and the cytoplasmic stub of *G*. As a result of the attractive interaction between the components, more of the membrane is then wrapped around the nucleocapsid and is stitched in place by the matrix protein. Any membrane proteins of the host cell that happen to be in the vicinity are shouldered aside. In the end the entire nucleocapsid is enveloped and the virion pinches off from the cell. The nucleocapsid has not been pushed through the membrane but rather is drawn into it by the cross-links of the matrix protein.

The viral proteins are made by the same process as normal host-cell proteins. The code specifying the amino acid sequence of each of the five viral proteins is carried by a length of messenger RNA, which is synthesized as a copy of a region of the viral genome RNA. The messenger RNA is translated into a polypeptide by the subcellular structure called a ribosome. The ribosome is made up of several dozen proteins and three or more kinds of RNA, organized into a large subunit and a small one. Messenger RNA is threaded through the ribosome, and the growing polypeptide is assembled in the large subunit. Each amino acid added to the polypeptide is carried by a molecule of transfer RNA, which recognizes a particular sequence of three nucleotide bases in the messenger RNA. Translation begins with the amino terminus and proceeds to the carboxyl terminus of the protein. The point where amino acids are joined to the growing chain is deeply embedded in the large subunit, and throughout translation about 40 amino acid units—the 40 most recently added—are buried inside the ribosome. As the chain grows, these units are successively pushed out by the new ones added behind them, and on completion of the polypeptide the entire protein molecule is released. In general a molecule of messenger RNA is translated by many ribosomes simultaneously, each ribosome generating a polypeptide chain.

It has often been pointed out that there are two kinds of ribosomes in the eukaryotic cell; some are free in the cytoplasm, whereas others are bound to membranes of the endoplasmic reticulum. It is now apparent that the ribosomes in these two classes are structurally identical and indeed are interchangeable. A membrane-bound ribosome can be released into the solution



MODEL OF THE PLASMA MEMBRANE includes proteins and carbohydrates as well as lipids. Integral proteins are embedded in the lipid bilayer; peripheral proteins are merely associated with the membrane surface. The carbohydrate consists of monosaccharides, or simple sugars, strung together in chains that are attached to proteins (forming glycoproteins) or to lipids (forming glycolipids). The asymmetry of the membrane is manifested in several ways. Carbohydrates are always on the exterior surface and peripheral proteins are almost always on the cytoplasmic, or inner, surface. The two lipid monolayers include different proportions of the various kinds of lipid molecule. Most important, each species of integral protein has a definite orientation, which is the same for every molecule of that species.



SELF-ASSEMBLY OF A MEMBRANE preserves its basic structure but not its asymmetry. A membrane can be disrupted by a high concentration of a detergent, which is an amphipathic molecule that forms the small droplets called micelles. The detergent dissolves the components of the membrane by enveloping the hydrophobic portions of both lipids and proteins in micelles, where they are protected from contact with water. If the detergent is then removed, the lipids spontaneously form a new bilayer, incorporating the integral proteins in it. The proteins, however, generally assume random orientations. Experiments such as this one have shown that membranes in the cell cannot be self-assembled; instead the integral proteins must be inserted in a membrane that already exists and has a defined sidedness.

and can later take up another place on the membrane. The two kinds of ribosomes do differ in function, however. Free ribosomes synthesize mainly soluble proteins, whereas membrane-bound ribosomes manufacture integral membrane proteins and proteins destined for secretion by the cell.

The VSV matrix protein is assembled on free ribosomes. That is not surprising, since immediately after synthesis the protein itself is free in the cytoplasm; it binds to the membrane (on the cytoplasmic side) only when budding begins. Other peripheral proteins have also been shown to originate as soluble proteins, and it is likely that all proteins attached to the cytoplasmic surface but not embedded in it are synthesized the same way.

The glycoprotein *G* follows a more complicated trajectory. It is synthesized by ribosomes on the endoplasmic reticulum and then transported to the plasma membrane, with at least one stop along the way. The endoplasmic reticulum is a network of membranes that is found near the cell nucleus and often partly surrounds it. The membranes are arranged in layers, and in electron micrographs each layer is seen to be a flattened, elongated vesicle, a cell within the cell.

The VSV glycoprotein first appears in the cell as a component of the rough endoplasmic reticulum, which is called rough because it is studded with ribosomes. *G* is embedded in the membrane of a reticulum vesicle, with the large spike portion, bearing the two carbohydrate chains, protruding into the lumen, or internal cavity, of the vesicle. The short stub at the carboxyl terminus of the protein faces the cytoplasm. After 20 to 30 minutes *G* can be found in the Golgi apparatus, another organelle made up of many stacked membrane vesicles. The Golgi apparatus is often called a smooth membranous organelle, because it lacks bound ribosomes. Only after the protein has passed through the Golgi apparatus does it reach the plasma membrane.

The means of transport between these organelles is obscure. The most likely possibility is that small vesicles bud off from one closed membrane and later fuse with another membrane, carrying the protein with them and preserving its orientation at all times. In this regard it is important to keep track of the orientation of the protein throughout its journey. At first it may seem that the orientation is reversed at some point in transit. In the endoplasmic reticulum the protein points inward, toward the lumen of the vesicle, but in the plasma membrane it points outward, toward the exterior of the cell. Actually this does not represent a reversal but the conservation of the same orientation. Note that the same portion of the protein (the stub) is always exposed to the cytoplasm and the same portion (the spike) is always outside the cytoplasm. Indeed, the lumen of an organelle such as the endoplasmic reticulum or the Golgi apparatus is topologically equivalent to the exterior of the cell, a fact that George E. Palade of the Yale University School of Medicine was the first to point out. In a topological sense the spike end of the protein is already outside the cell from the time it enters the endoplasmic reticulum.

The carbohydrate side chains of the glycoprotein are attached before it reaches the plasma membrane. In the virion the protein bears two identical carbohydrate structures, each made up of a dozen or so linked monosaccharides, or sugar units. The function of the sugar side chains in the VSV glycoprotein is not known, but in structure they resemble the carbohydrates of many normal glycoproteins found in the blood serum.

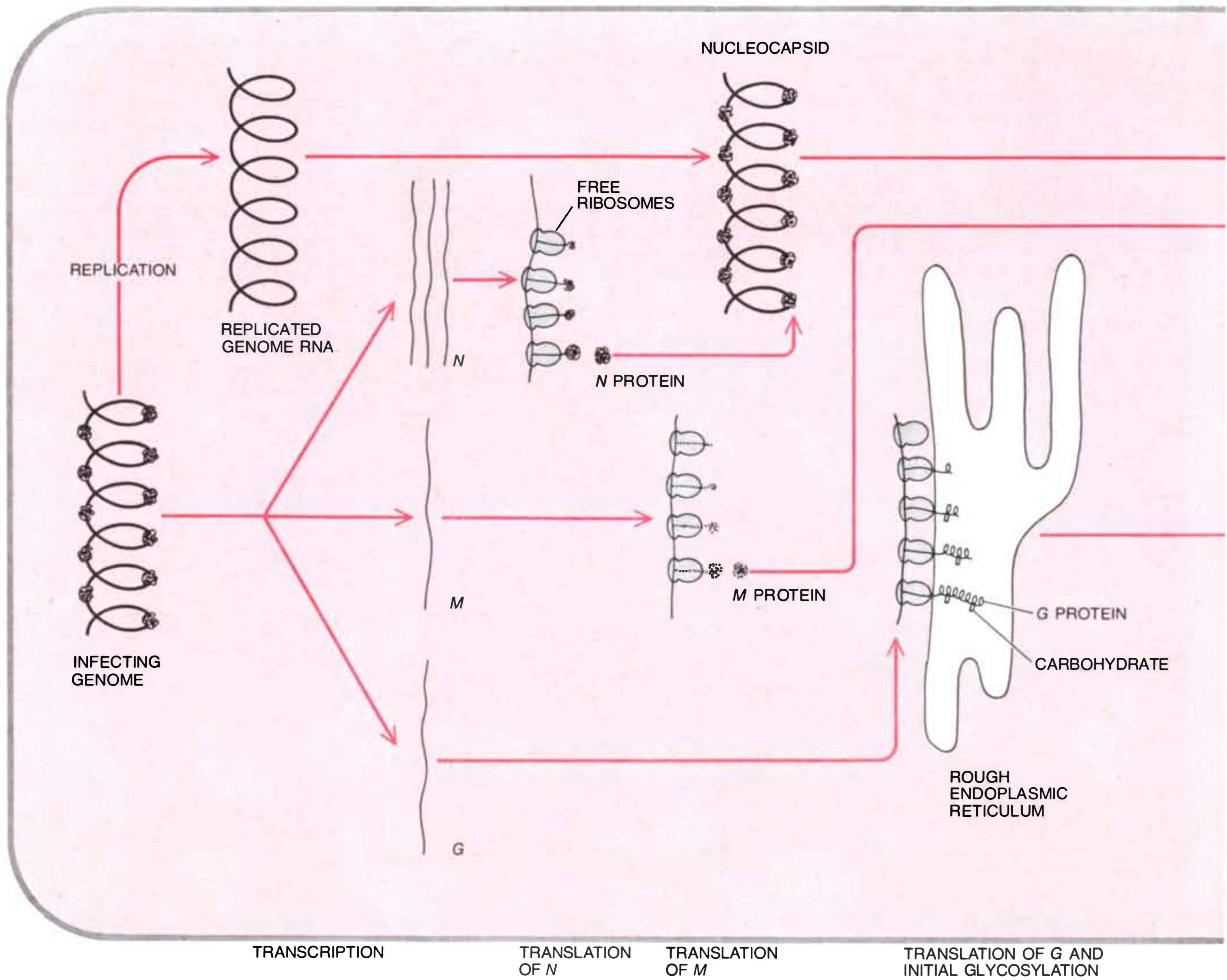
Donald F. Summers and his co-workers at the University of Utah have determined the sequence of monosaccharides in the carbohydrate chains and have shown that each chain can be considered to have two parts. The region nearer the protein and attached directly to it they designate the core, and they note that it consists exclusively of the monosaccharides mannose and *N*-acetylglucosamine. The outlying part of the carbohydrate constitutes the terminal region, and it is built up from three monosaccharides, *N*-acetylglucosamine, galactose and sialic acid. As we shall explain below, the core region is attached to the protein while it is still in the rough endoplasmic reticulum. In the Golgi apparatus a few of the core sugars are removed and the several new terminal monosaccharides are added. When the glycoprotein leaves the Golgi apparatus, it is complete.

We have studied the synthesis, glycosylation and insertion of the *G* protein in a cell-free system, a liquid medium that includes ribosomes, transfer RNA's, amino acids and various enzymes and other ingredients necessary for protein synthesis. To study the insertion and glycosylation of the protein, membranes from the endoplasmic reticulum that have been stripped of their endogenous ribosomes must also be present. Protein synthesis is initiated in this system simply by adding purified messenger RNA for the *G* protein, which can be obtained from infected cells. In order for newly made polypeptides to be identified later, one of the amino acids in the medium (methionine) includes an atom of a radioactive isotope of sulfur.

These experiments have been carried out at the Massachusetts Institute of Technology. Much of the work was done in collaboration with Flora N. Katz of M.I.T. and Vishu Lingappa and Gunter Blobel of Rockefeller University. Similar experiments have been conducted independently by Hara Ghosh and Frances Toneguzzo of McMaster University. We shall first present some of the conclusions derived from our findings and then discuss the experiments themselves.

The central thesis of our model is that an integral membrane protein is inserted through the lipid bilayer as the protein is synthesized and before it has assumed its ultimate, folded configuration. This mechanism accounts for two important observations. First, it helps to explain how the bulk of the protein can cross the hydrophobic core of the membrane: it passes through before the folding of the polypeptide makes the insertion even more difficult. Second, the model explains the asymmetry of the protein's orientation: the protein can be inserted in only one direction (the direction of polypeptide elongation) and from only one side of the membrane (the cytoplasmic side, where ribosomes are found).

The translation of *G*, like that of all proteins, begins with the binding of messenger RNA to the small subunit of a ribosome; then the large subunit is added. At this stage the ribosome is not associated with a membrane but is free in the cytoplasm. Some 40 amino acid units must be assembled before the polypeptide begins to emerge from the large subunit of the ribosome; these are the 40 units at the amino terminus of the molecule. After another 30 amino acids have been added to the growing chain the first 30 units are exposed. Within this section of the polypeptide is a group of amino acids, called the signal sequence, that is essential to the distinction be-

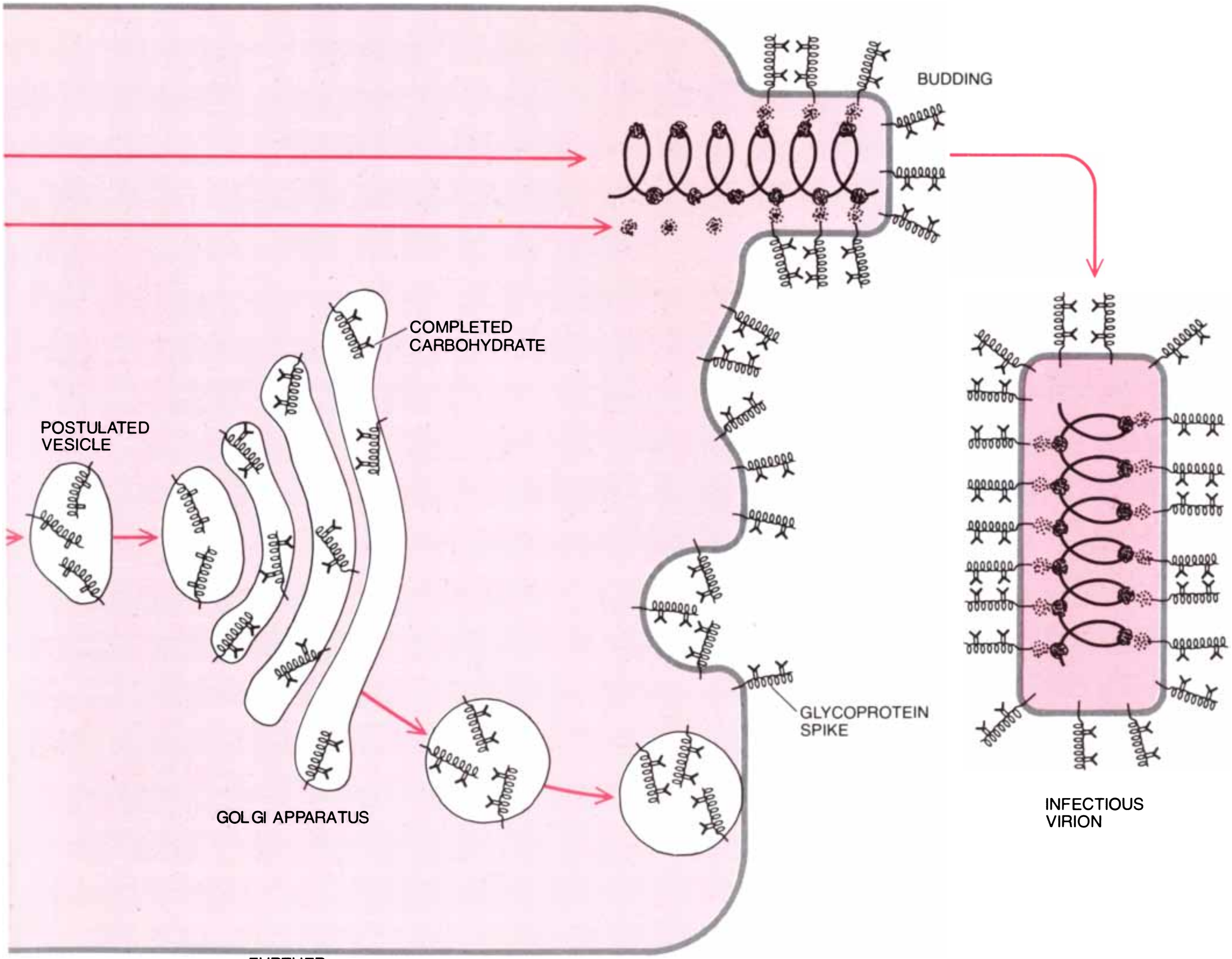


VIRAL PROTEINS are manufactured by the genetic machinery of the cell, but their structure is specified by the genome of the virus. The infectious cycle of VSV begins when a virus inserts its genetic material, a strand of RNA, into a cell. The genome RNA is transcribed by two viral enzymes into five molecules of messenger RNA. Later the genome RNA is replicated by similar viral enzymes. Each strand of messenger RNA is translated by the organelles called ribosomes into many identical protein molecules; since there are five messenger-RNA molecules, the viral genome encodes five proteins. Three are nucleoproteins; these include the two enzymes needed for transcription and replication and a third protein, designated *N*, whose main role is probably structural. The three nucleoproteins bind to the genome RNA to form a complex called the nucleocapsid. A fourth protein, designated *M* for matrix, is initially a soluble component of the cytoplasm. All four of these proteins are synthesized on ribosomes that are free in the cytoplasm. The last protein is designated *G* be-

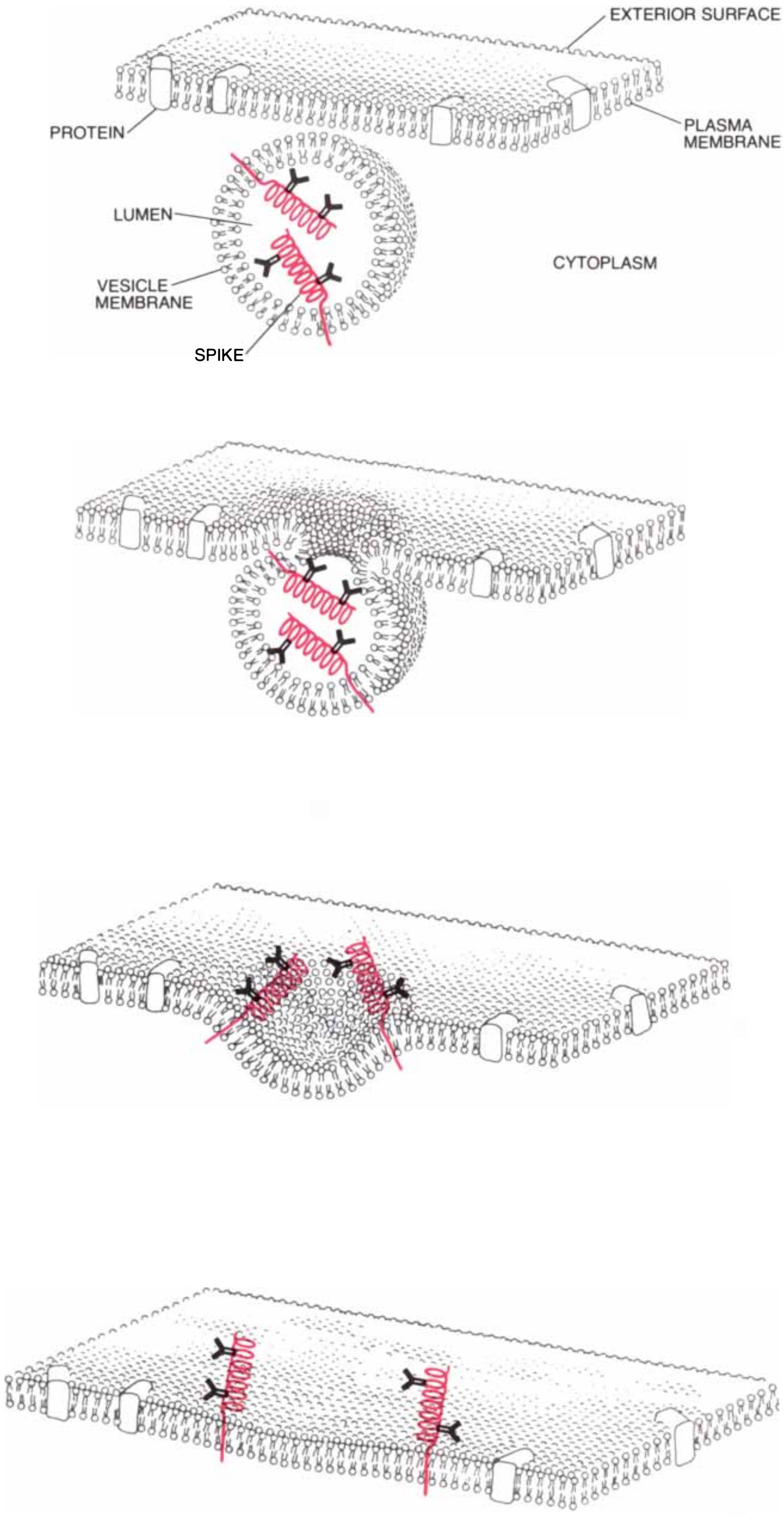
tween membrane proteins and secreted proteins on the one hand and soluble proteins on the other. The signal sequence is evidently recognized by some component of the endoplasmic-reticulum membrane, presumably a membrane protein. This membrane transport site binds to the nascent polypeptide and helps to initiate its passage through the bilayer.

The existence of a signal sequence was first proposed by Cesar Milstein and George G. Brownlee and their colleagues at the Medical Research Council Laboratory of Molecular Biology in Cambridge, England, based on their studies of the synthesis of antibody proteins. Much additional evidence for the sequence has been provided by the experiments of Blobel and his co-workers, who have examined primarily proteins destined to be secreted by the cell. It appears that the cellular mechanisms for handling the two classes of proteins are similar in many ways. Both are extruded into the endoplasmic reticulum; the principal difference is that secreted proteins pass all the way through and are released into the lumen, whereas membrane proteins become embedded in the phospholipid bilayer before the polypeptide is complete. When a vesicle bearing such proteins fuses with the plasma membrane, the secreted proteins are released outside the cell, whereas the integral proteins remain bound to the membrane. In secreted proteins, once the signal sequence has piloted the polypeptide across the lipid bilayer it is clipped off by enzymes in the reticulum; it has recently been found that a signal sequence of the *G* protein, made up of 16 amino acids, is also removed. The signal sequences of the various secreted proteins and of *G* are not identical, but they share certain properties; in particular, they include a preponderance of hydrophobic amino acids.

It is only when the signal sequence draws the polypeptide to the endoplasmic reticulum that the ribosome becomes associated with a membrane. Initially the ribosome itself is probably not bound directly to the membrane but is merely tethered by the polypeptide chain. Later a weak electrostatic inter-



cause it is a glycoprotein, one with carbohydrate side chains (whose structure is not specified by the viral genome). *G* is assembled by ribosomes associated with membranes of the rough endoplasmic reticulum, and from the time the protein first appears it is bound to these membranes. Core monosaccharides are added to growing *G* molecules in the rough endoplasmic reticulum, and after 20 to 30 minutes the carbohydrate side chains are completed in the Golgi apparatus. The mechanism of transport between these organelles is not known, but it may depend on vesicles that bud off from one membrane and fuse with another. The glycoprotein could reach the plasma membrane, where it forms spikes on the surface of the cell, by the same means. The life cycle of the virus is completed when the nucleocapsid is wrapped in a piece of plasma membrane bearing the *G* protein; this newly assembled virus particle then buds from the cell. Budding is probably mediated by the matrix protein, which could stitch together nucleoproteins and a stub of *G* that protrudes into the cytoplasm.



FUSION OF A VESICLE with the plasma membrane preserves the orientation of any integral proteins embedded in the vesicle bilayer. Initially the large spike end of the *G* protein faces the lumen, or inner cavity, of such a vesicle. After fusion the spike is on the exterior surface of the plasma membrane. That the orientation of the protein has not been reversed can be perceived by noting that the other end of the molecule, the small stub, is always immersed in the cytoplasm. The lumen of a vesicle and the outside of the cell are topologically equivalent.

action may develop between the ribosome and the membrane.

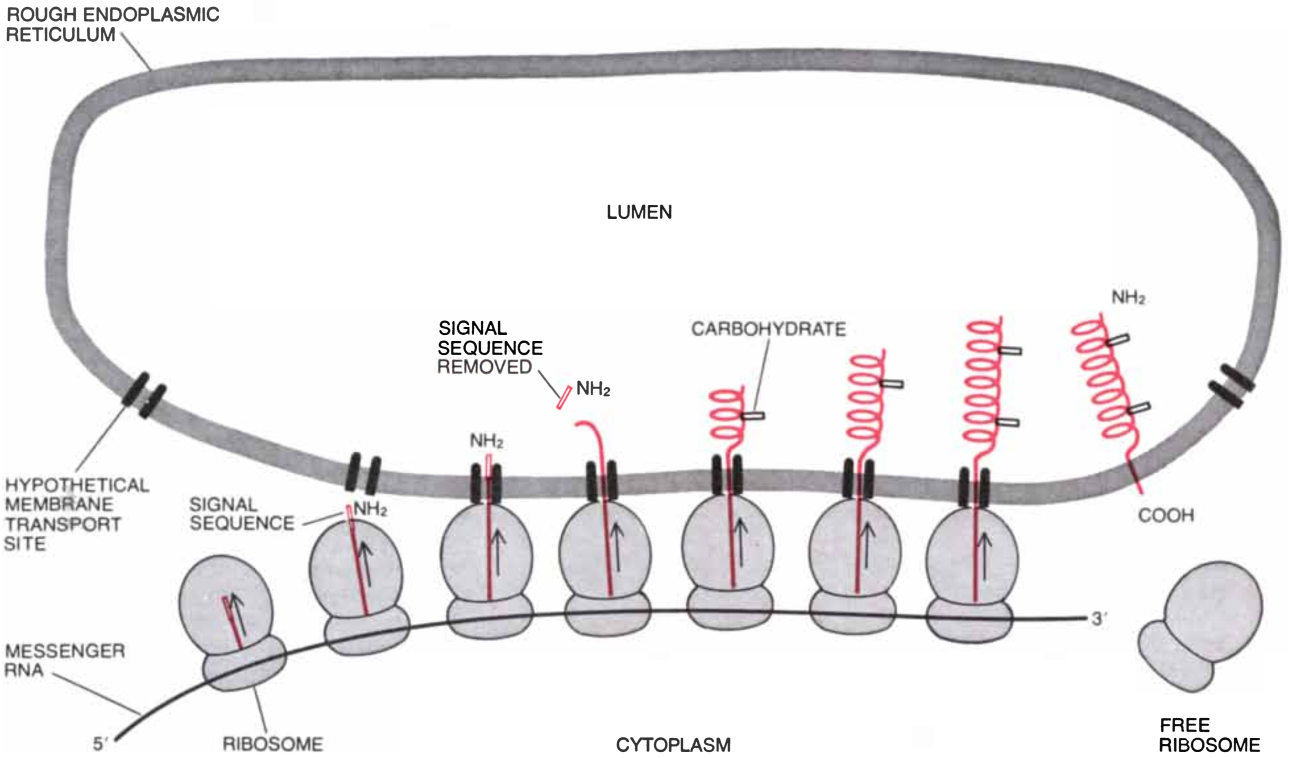
In all likelihood the protein does not penetrate the lipid bilayer directly but is aided by some other integral protein already in the membrane, which diffuses away when the insertion is complete. As more of the chain enters the lumen of the reticulum the polypeptide folds spontaneously, like a soluble protein, to achieve a configuration of minimum energy in an aqueous environment. Once in that state the protein cannot slide back through the bilayer; it is anchored like a rivet. What remains to be explained is why the small stub at the carboxyl terminus of the polypeptide does not slide through to the lumen side, as it does in secreted proteins. It has also not been established whether energy in addition to that normally expended in protein synthesis is required to transport the nascent protein across the membrane.

In a cell-free system where the only membranes present are those of the endoplasmic reticulum a complete glycoprotein cannot be constructed, because the modification of the carbohydrate portion is completed only in the Golgi apparatus. *G* protein grown in the laboratory does acquire the two chains of core sugars, however, and what is equally important, they appear to be identical with those formed in the intact cell.

Each of the two carbohydrate structures is attached to a side chain of the amino acid asparagine. One asparagine unit is thought to be roughly 150 amino acid units from the amino terminus of the protein and the other is about 400 units from the same end. The core carbohydrates are probably attached to the asparagine units very soon after they emerge on the lumen side. They are not built up on the protein sugar by sugar but rather are added all at once. The core regions are assembled beforehand on a membrane lipid and are transferred to *G* as a unit. Since the reticulum membranes employed in these experiments come from cells not infected with VSV, the carbohydrate chains must be components of the normal cell rather than structures assembled for the virus.

The glycoprotein created by this series of steps has its amino terminus and most of its mass, including about 500 amino acid units, on the lumen side of the reticulum membrane. The two carbohydrate chains are on the same side; it is this end of the molecule that is eventually observed as a spike on the exterior surface of the plasma membrane and on the virion. At least a few amino acid units must remain in the bilayer, and although their number is not known exactly, it is probably about 20 or 30. Only a stub made up of the last 30 units or so, including the carboxyl terminus, protrudes on the cytoplasmic side of the membrane.

How was this model of protein inser-



SYNTHESIS, INSERTION AND GLYCOSYLATION of the *G* protein are closely coupled. The protein, or polypeptide, is built up from amino acids linked together in a linear sequence specified by the messenger RNA. Synthesis begins with the amino terminus (NH_2) of the polypeptide; amino acids are added one at a time behind that terminus. Among the first 30 amino acid units of *G* is a "signal sequence" that identifies the protein as one destined to be inserted into the membrane of the rough endoplasmic reticulum. Because some 40 amino acids remain buried in the ribosome, the signal sequence does not emerge until the polypeptide is about 70 amino acid units long. At that time the signal sequence is recognized by some molecule, presumably a protein, in the membrane of the endoplasmic reticulum. This hypothetical protein is thought to facilitate the passage of the polypeptide through the lipid bilayer. Once in the lumen of the reticulum the signal sequence is removed. The protein continues to elongate, and as it grows it is extruded through the membrane and folds up in the lumen. As it enters, two identical, preformed carbohydrate side chains are fastened to it. Proteins secreted by the cell pass all the way through the membrane in this manner, but for reasons that are not fully understood *G* becomes stuck at about the time translation is completed, with some 30 amino acids remaining in the cytoplasm. Thus the completed glycoprotein has its amino terminus, most of its bulk and all its carbohydrate in the lumen of the reticulum, and a short stub that includes the carboxyl terminus (COOH) on the cytoplasmic side. Once the protein has folded it cannot be pulled out of the membrane, nor can it execute a flip-flop; it is anchored in an asymmetric orientation. Components of the cell are not shown to scale: the ribosome is some 50 times larger than *G* protein.

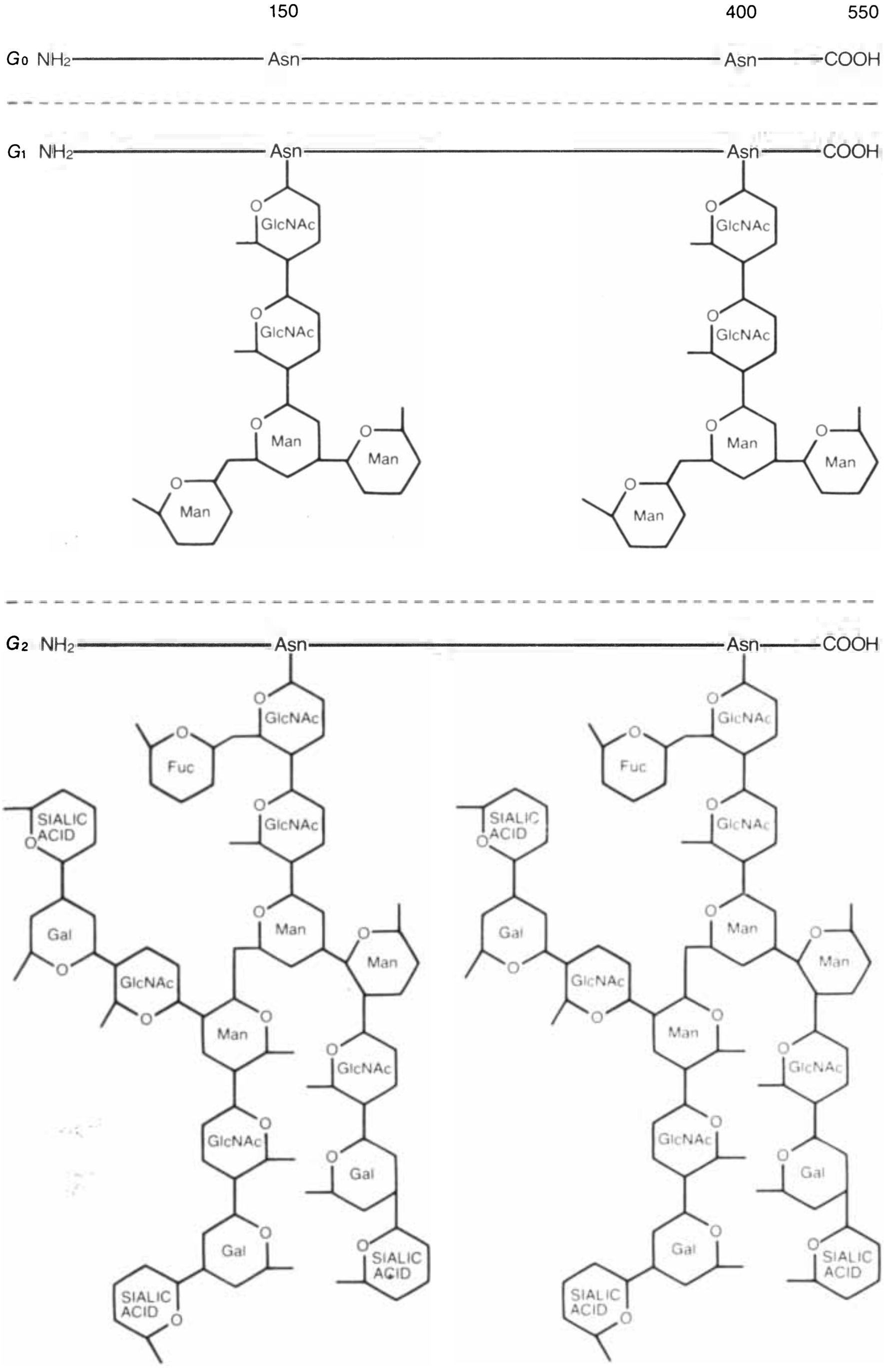
tion constructed and tested? A technique that had a part in almost all our experiments was gel electrophoresis, which separates proteins according to their size. A typical experiment yields a mixture of proteins, some in their native form, others reduced to fragments by enzymatic digestion, others with only a partial complement of carbohydrate. The mixture is placed on a porous gel and an electric field draws the proteins through the gel. Small molecules migrate through the pores of the gel faster than large ones, and so the various fractions are separated and can be identified.

In one experiment all the enzymes and other molecules required for the normal synthesis of *G* were provided except for reticulum membranes. Under these conditions the polypeptide can be assembled by the ribosome, but of course it cannot be inserted into a membrane. The resulting protein appeared to be normal in all respects except two: it had no carbohydrate, since that can be added only by enzymes of the reticulum, and it retained the signal sequence of 16 amino acid units at the amino terminus. Such nude protein is designated G_0. Protein with only core carbohydrates (and lacking the signal sequence) is G_1, and the finished protein with both core and terminal carbohydrate is G_2. Although G_1 and G_2 are normal intermediates found in infected cells, G_0 is not formed during infection.

If endoplasmic-reticulum membranes are added to the medium before translation begins, the protein is fabricated in the G_1 form, with both core chains of carbohydrate but without the signal sequence. A telling result was obtained, however, when the membranes were added while the synthesis was under way. The model presented above predicts that unless the membranes are added very early in the synthesis the emerging protein will not be inserted into the membrane. Part of the polypeptide will already have folded in the cytoplasm and will be unable to interact with receptors on the membrane or to cross the permeability barrier of the lipid bilayer. In fact, adding membranes well after synthesis had begun accomplished nothing; when it was subsequently completed, the protein remained in the G_0 form.

In order to check the orientation of the molecules, *G* protein was synthesized in the presence of membranes and after completion of protein synthesis the enzyme trypsin was added to the solution. Trypsin is a digestive enzyme that can cut a polypeptide at many points, but because it is a soluble protein it cannot cross a lipid membrane. Hence it can digest only those portions of the protein molecule that lie on the cytoplasmic side of the membrane. All molecules of *G* subjected to this procedure yielded a large fragment of *G* about 30 amino acids shorter than the normal G_1. We were able to show that only the carboxyl terminus of the polypeptide had been digested and that the large protected fragment carried both carbohydrate chains. These findings indicated that only the 30-unit carboxyl stub projects into the cytoplasm and that the carbohydrate is in the lumen of the vesicle. Moreover, the fact that all copies of *G* were reduced to the fragment indicated that all the G molecules in the membrane had the same orientation.

In order to find out exactly when insertion into the bilayer must begin, we carried out a series of experiments in which progressively longer segments of protein were synthesized before membranes were added to the medium. We found that the polypeptide was correctly inserted and went on to form a normal G_1 if the membranes were added before the first 70 amino acids had been polymerized. If the membranes were added later, only the sugarless G_0 was formed, and this polypeptide was not properly embedded in the membrane but was probably loose in the cytoplasm. Since about 40 of the first 70 amino acids are buried in the ribosome, the first 30 units must bear the crucial signal sequence.



CARBOHYDRATE SIDE CHAINS are constructed on the *G* protein in two steps. In the rough endoplasmic reticulum the growing protein receives the core regions of the two carbohydrate chains, which are made up exclusively of the monosaccharides mannose (Man) and *N*-acetylglucosamine (GlcNAc). In the Golgi apparatus the core regions are modified (additional core mannose units not shown here are removed) and the terminal sugar units are attached; these include additional *N*-acetylglucosamine as well as the monosaccharides galactose (Gal) and sialic acid. Fucose (Fuc) is also added. The completed glycoprotein has two identical carbohydrate chains, each bonded to the polypeptide through a side chain of the amino acid asparagine (Asn). The two asparagine units are thought to be about 150 and about 400 amino acid units from the amino terminus of the polypeptide. There is evidence that the core carbohydrates are attached to the protein as a unit, having been assembled on a glycolipid.

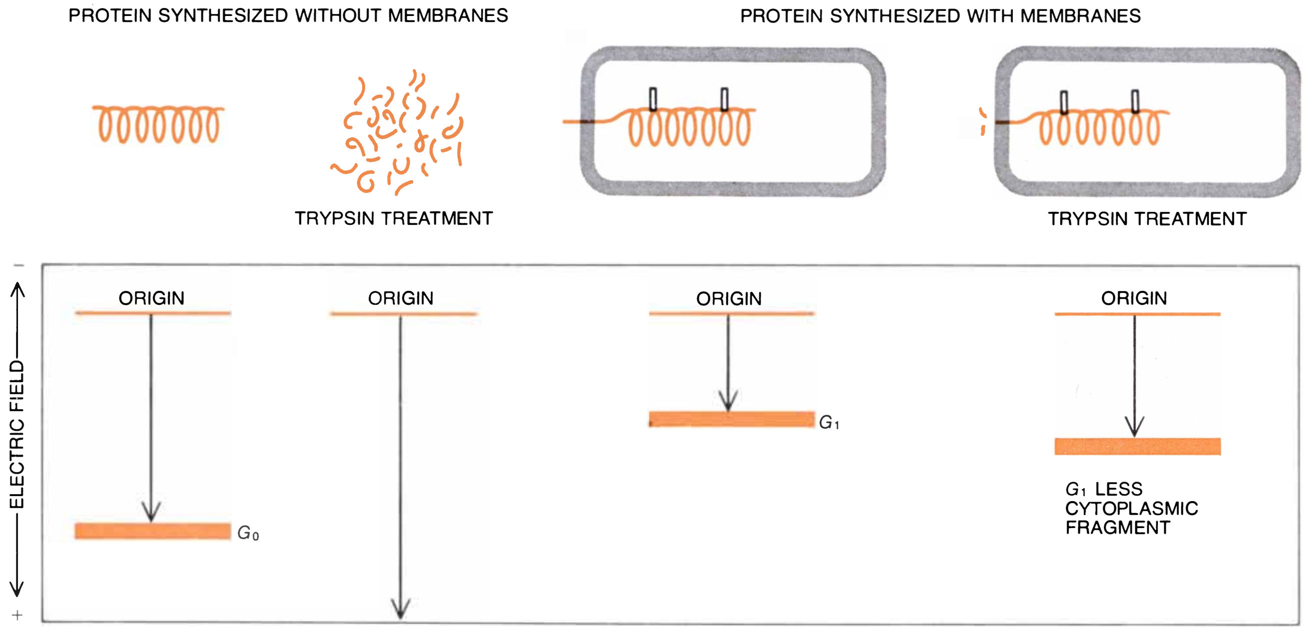
An experiment that was in some ways the obverse of this one tested the assertion that the protein is extruded through the membrane as it elongates. The interpretation of the experiment depended on the assumption that carbohydrate can be added only in the lumen of the endoplasmic reticulum.

Various lengths of polypeptide were allowed to grow in the presence of membranes; then the membrane was destroyed by dissolving it with a detergent. The detergent has no effect on protein synthesis, but it does halt the attachment of carbohydrates, probably because the enzymes that transfer the sugar side chains are bound to the lumen side of the endoplasmic-reticulum membrane and are separated from the ribosomes by detergent treatment. We found that when detergent was added before the polypeptide was about 150 amino acids long, and the protein was then completed, only the G_0 protein was observed. Thus no carbohydrates were attached before the polypeptide was 150 amino acids long. If the membrane was maintained intact until after some 400 amino acids had been incorporated in the chain, then the completed glycoprotein was the G_1 form, with two core carbohydrates.

This meant that all core sugars are added during protein synthesis, while the growing chain of G is still attached to the ribosome. It was a crucial finding. If the carbohydrate is added to G on the lumen side while the protein is still growing on a ribosome on the cytoplasmic side, then G must be extruded through the membrane during protein synthesis.

An intriguing result was obtained when we disrupted the membrane after 150 amino acids had been linked to the polypeptide but before 400 amino acids had been. The resulting protein was an intermediate form, never observed in nature, with just one carbohydrate chain. The creation of this novel intermediate provides strong evidence that the two carbohydrate structures are added in sequence as the polypeptide passes through the membrane.

The asymmetry of membrane lipids differs in detail from that of the proteins, but the same fundamental principle applies: The membrane can grow only by expansion; it cannot be broken open for the insertion of new material. In the case of the phospholipids this requirement has been met in a simple way:



CELL-FREE SYSTEM for synthesizing *G* was employed to test the hypothesis that the protein is inserted into the membrane of the endoplasmic reticulum as the polypeptide elongates. The system is a liquid medium containing ribosomes, amino acids, various enzymes and other molecules necessary for protein growth. Synthesis is initiated by adding messenger RNA for *G* to the medium. In this system the protein can be grown under different conditions and the results assayed by gel electrophoresis, which separates molecules according to their size; small molecules move faster through the gel than large ones. If *G* is made in the absence of membranes, the product is a protein labeled G_0, which is normal except that it has no carbohydrate and retains the amino-terminal signal sequence. Repeating the procedure and adding the digestive enzyme trypsin after completion of protein synthesis reduces the protein to a multitude of small fragments, which migrate off the gel. When the membranes of the endoplasmic reticulum are added to the medium before protein synthesis begins, a different product is obtained: the protein G_1, which has the core regions of both carbohydrate chains and lacks the signal sequence. When trypsin is added in this procedure, only the cytoplasmic stub of the protein is degraded because trypsin cannot cross the lipid bilayer. The results of these experiments demonstrate that *G* acquires carbohydrate chains only if it enters endoplasmic reticulum.

the lipids are synthesized within the membrane itself.

An example is the lipid phosphatidylethanolamine, or PE, whose synthesis in the bacterium *Escherichia coli* follows a biochemical pathway that has been traced mainly by Eugene P. Kennedy and his colleagues at the Harvard Medical School. The synthesis begins with two molecules of a fatty acid, which are transferred from a donor to a molecule of glycerol phosphate, forming the most primitive phospholipid, phosphatidic acid. The two fatty acids are the hydrophobic tails of the molecule and the glycerol phosphate is the hydrophilic head. In subsequent steps a phosphorylated nucleotide, CMP, is grafted onto the head of phosphatidic acid, then the CMP is replaced by the amino acid serine. Finally, a molecule of carbon dioxide is extracted from the serine, yielding PE. Each step in the synthesis is catalyzed by an enzyme, and it has been found that all the enzymes except one are integral proteins of the bacterial plasma membrane. The lipid substrates and products of the reactions must therefore also be components of the bilayer. The membrane grows by fabricating lipids in situ.

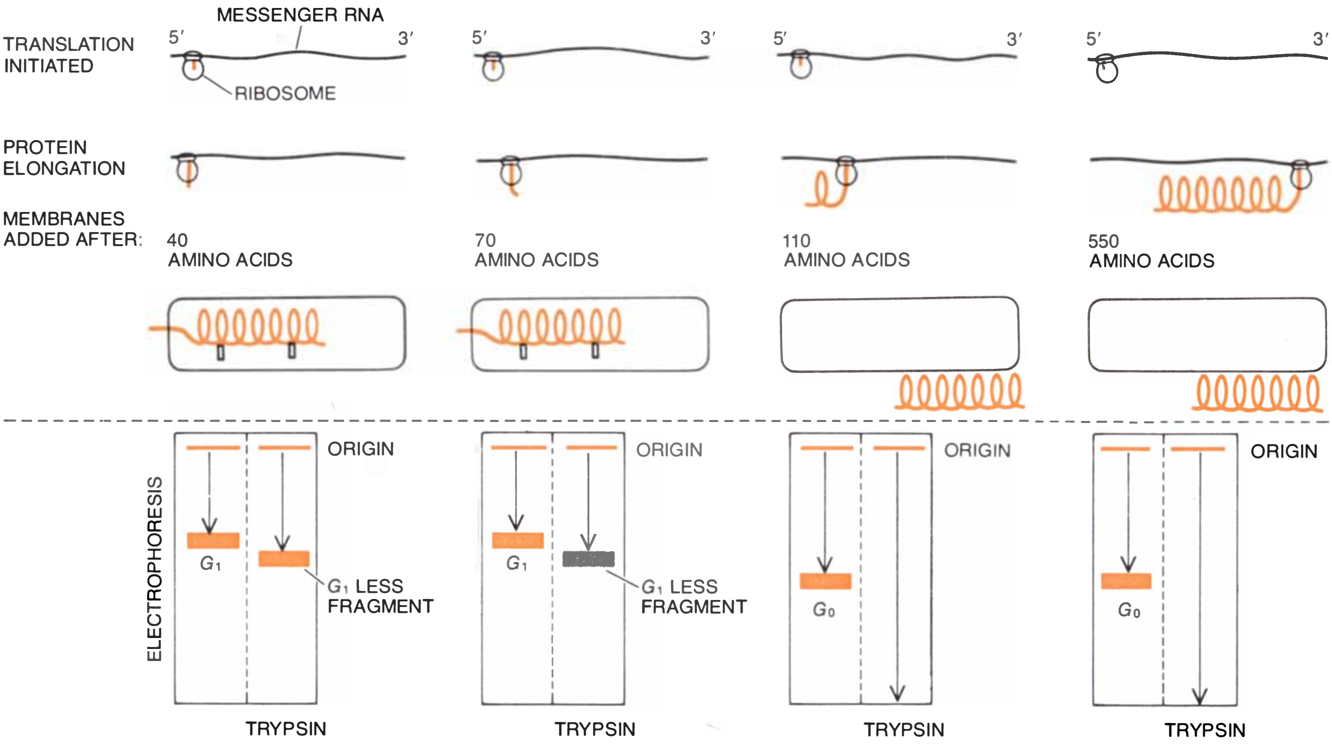
Phospholipid synthesis is fundamentally similar in eukaryotic cells, the chief difference being that the events take place not in the plasma membrane but in the endoplasmic reticulum. Again the enzymes concerned are themselves membrane proteins, indicating that newly made lipids and intermediates are part of the bilayer. Hence the endoplasmic reticulum is a true membrane factory: it is the site where both the lipid and most of the protein components of membranes are assembled before they are dispatched to other parts of the cell.

It is easy to understand in a general way how lipid membranes can grow by expansion, but the analysis presented above overlooks the one feature of membrane structure we set out to explain, namely that bilayers have two sides, which are not identical. Are new lipids formed on only one side of the bilayer or on both sides? Many of the precursors in lipid synthesis and one of the enzymes are soluble molecules in the cytoplasm; the fact that these substances cannot cross the hydrophobic core of the membrane argues that lipid synthesis should take place only at the cytoplasmic surface. If that is the case, however, some means must be provided for ferrying lipid molecules to the external surface.

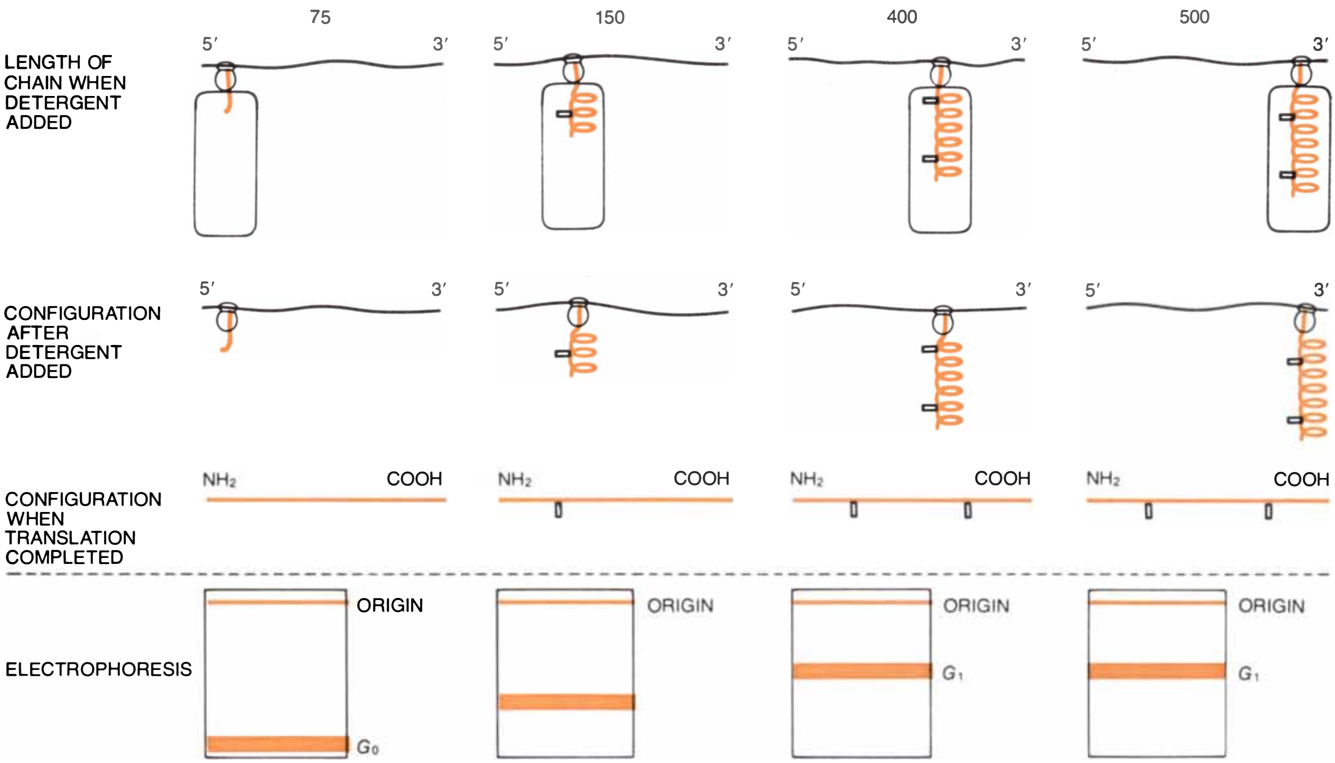
The mechanism of lipid synthesis must also be consistent with the compositional asymmetry of the bilayer. This asymmetry is not an absolute one, as it is for proteins, where each species of protein has a designated orientation observed by every molecule of that species. The asymmetry of lipid composition is better described as a statistical bias. There are no lipids found exclusively in one monolayer or the other, but most lipids do seem to favor one surface. In certain bacteria, for example, PE is roughly twice as abundant on the cytoplasmic side of the membrane as it is on the external side. Other phospholipids are found predominantly in the external monolayer.

Asymmetries of lipid composition have been investigated at the Harvard Medical School by one of us (Rothman) in collaboration with Kennedy. The membrane employed in several of these experiments was that of *Bacillus megaterium*. A bacterium of this type, called gram-positive, was chosen because the presence of only one membrane per cell simplifies the analysis of lipid composition. This particular organism was selected in part because its membrane includes large quantities of PE, which makes up some 70 percent of the total membrane lipid. Most of the remainder is phosphatidylglycerol, in which the ethanolamine group of PE is replaced by a second glycerol. In other gram-positive bacteria phosphatidylglycerol is the major lipid constituent of the membrane. The abundance of PE in *B. megaterium* is an experimental convenience because PE is easily labeled.

The distribution of PE in the membrane of *B. megaterium* was measured by



ADDITION OF MEMBRANES after protein synthesis has already begun demonstrates that insertion must be simultaneous with synthesis if it is to succeed at all. If the polypeptide is allowed to reach a length of 70 amino acid units before reticulum membranes are introduced, the protein can still penetrate the endoplasmic reticulum membrane, and it subsequently develops normally. Electrophoresis shows that it acquires the two core carbohydrates and that only a small fragment (the stub) is degraded by trypsin. With any longer delay between the initiation of synthesis and the introduction of membranes, however, the folding of the protein has progressed too far for it to enter the bilayer. Because the folding exposes many hydrophilic amino acids the polypeptide is permanently excluded from the reticulum. When proteins are analyzed by electrophoresis, they are found to be without carbohydrate and to be completely digested by trypsin.



DESTRUCTION OF MEMBRANES after protein synthesis has begun proves that core carbohydrates can be joined to the polypeptide only within membranes of the endoplasmic reticulum. The translation of many copies of *G* is begun simultaneously in a medium with abundant reticulum membranes and is allowed to continue for various periods before a detergent is added to the medium, destroying the membranes. Elongation of the polypeptide is then allowed to proceed to completion in the absence of membranes. When the membranes are disrupted before the chain is about 150 amino acid units long, the only protein observed is G_0, which lacks all carbohydrate. If detergent is not added until some 400 or more amino acids are completed, then normal G_1, with two carbohydrate chains, is produced. Applying detergent at intermediate intervals yields a protein that is never observed in nature, with only the first carbohydrate chain.

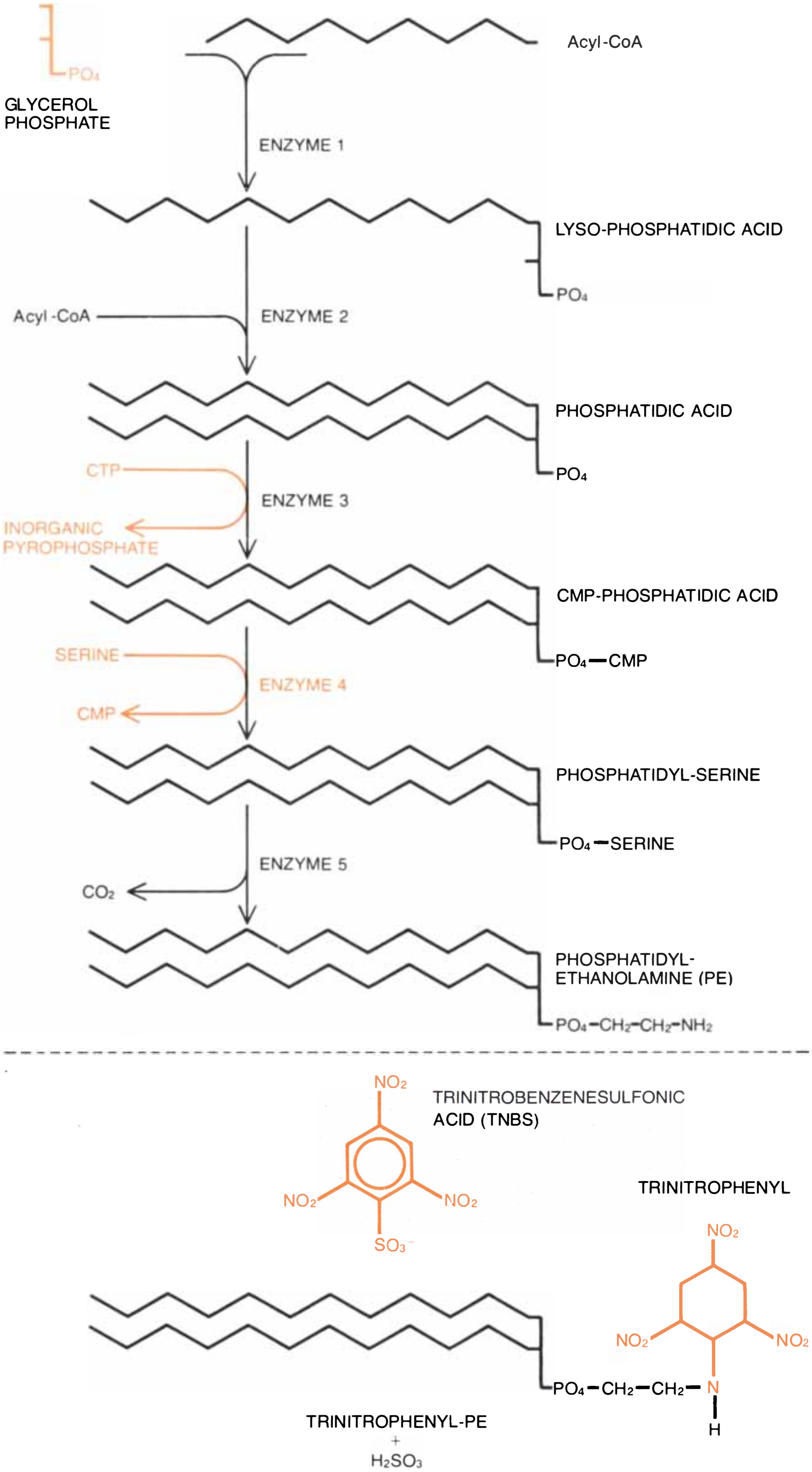
labeling all the PE molecules exposed at the exterior surface but none of those at the cytoplasmic surface. The proportion of the PE on each side could then be determined by comparing the number of labeled and unlabeled molecules.

The labeling reagent was trinitrobenzenesulfonic acid, or TNBS, which reacts with the amino group of the PE molecule but cannot combine with phosphatidylglycerol or other lipids that lack an amino group. If TNBS is present in sufficient concentrations, it will tag all molecules of PE to which it can gain access; the accessible molecules are precisely those of the exterior membrane surface. TNBS cannot reach the cytoplasmic side of the membrane because it is water-soluble and cannot cross the permeability barrier. This simple but important principle of labeling membranes with nonpenetrating reagents was devised by Mark S. Bretscher of the Medical Research Council laboratories in Cambridge, who applied it in pioneering studies of lipids and proteins in the membrane of the red blood cell.

When the labeling was completed, the bacterial membranes were broken down by dissolving the lipids in an organic solvent. It was then necessary to separate the labeled and unlabeled PE, which was done by the technique of thin-layer chromatography. A sample of the extracted lipids was applied to the end of a thin sheet of a silica gel and a mixture of organic solvents was allowed to rise through the gel by capillary action, sweeping past the lipid extracts. PE labeled with TNBS is more soluble in these solvents than unlabeled PE, and so the labeled lipids were carried farther through the gel. In a sense the two halves of the phospholipid bilayer were separated by this method. When the amount of each material was determined, it was found that only about 30 percent of the PE had been labeled. It follows that about 30 percent of the PE is on the outer surface of *B. megaterium* and the other 70 percent is inside. Since the total amount of lipid in the two monolayers is roughly equal, the other major lipid, phosphatidylglycerol, must have the opposite distribution.

Through a somewhat more elaborate procedure we were able to approach the question not only of where the lipids reside in the membrane but also of where they are made. The idea of the experiment was to employ two independent labels: one would mark only the newly synthesized lipids and the other would distinguish inner-surface lipids from outer-surface ones. In this way the distribution of the new lipid material would be revealed.

In order to label freshly synthesized PE we incubated bacterial cells with inorganic phosphate prepared with the radioactive isotope phosphorus 32. The radioactive phosphate was incorporated



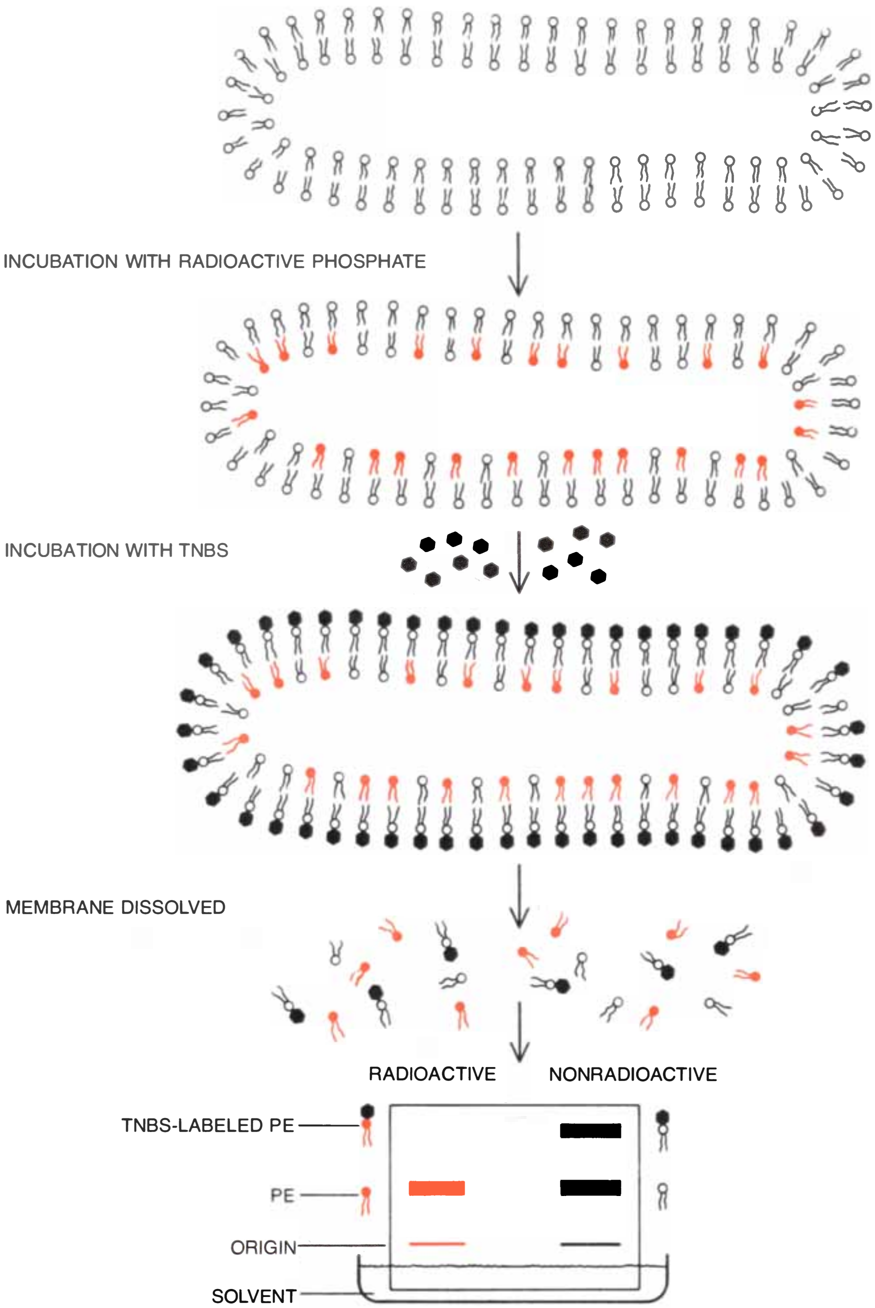
SYNTHESIS OF A LIPID relies on several enzymes that are integral membrane proteins. The lipid is phosphatidylethanolamine, or PE. Enzymes, substrates and precursors that are soluble in the cytoplasm are shown in color; all other molecules are oil-soluble and are confined to the lipid bilayer. The synthesis begins with the transfer of two fatty acids from donor molecules (Acyl-CoA) to glycerol phosphate. A molecule of cytosine monophosphate (CMP) is attached to the glycerol phosphate and is then replaced by the amino acid serine. In the last step of the synthesis PE is made by extracting a molecule of carbon dioxide (CO_2) from the serine. Because four of the five enzymes required for this procedure can be found only in the lipid bilayer, it appears that lipids are formed within the membrane itself. Because the one remaining enzyme and some substrates are present only in the cytoplasm, the lipid synthesis must take place on the cytoplasmic side of the bilayer. PE can be selectively labeled with trinitrobenzenesulfonic acid (TNBS), which bonds to the amino group in the hydrophilic head of the lipid.

into glycerol phosphate and hence into membrane lipids. The incubation lasted for only about a minute, and so only the lipids made during that period would exhibit radioactivity. TNBS was then added to the suspension of cells and allowed to combine with PE molecules on the exterior surface of the membrane. The membrane was next dissolved and the lipids were separated by chromatography and tested for radioactivity. If both the TNBS-labeled PE and the unlabeled PE were radioactive, the lipid must be synthesized in both monolayers, since these fractions come from opposite sides of the membrane. We found, however, that radioactivity could be detected only in the PE that had not combined with TNBS, indicating that synthesis of this lipid is confined to the cytoplasmic surface, which TNBS cannot reach.

This finding is consistent with the cytoplasmic origin of the lipid precursors, but it also raises a troubling question. If lipids are made at the cytoplasmic surface and immediately become part of the cytoplasmic monolayer, how do they ever get to the other side of the membrane? As we pointed out above, the rate of spontaneous flip-flops across the bilayer is exceedingly low, so that a given lipid could be expected to make the transition only about once a month. Yet a bacterium can double in mass and divide in less than an hour, thereby doubling the surface area of the plasma membrane. If lipids could not cross from the inner surface to the outer one, the bilayer would soon become a lipid monolayer.

Another experiment has shown that lipids do in fact cross the membrane, although it has not revealed how they do so. Bacteria were again briefly incubated with radioactive phosphate, but instead of their being labeled immediately with TNBS another reagent was added that inhibits further lipid synthesis. The inhibitor, hydroxylamine (NH_2OH), inhibits the last enzyme in the sequence leading to PE, the enzyme that extracts a molecule of carbon dioxide from serine. TNBS could then be added after various intervals and would show how much lipid had migrated across the membrane in the period after PE synthesis was halted.

When TNBS was added immediately after the inhibitor, essentially all the radioactivity was found in the inner monolayer, again showing it to be the site of lipid synthesis. If the labeling with TNBS was delayed by 30 minutes, however, we found that the distribution of new, radioactive PE was the same as the distribution of the old, nonradioactive lipid, with about 30 percent of both on the exterior surface. Evidently the new PE molecules were able to cross the membrane rapidly enough to reach their natural asymmetric distribution.

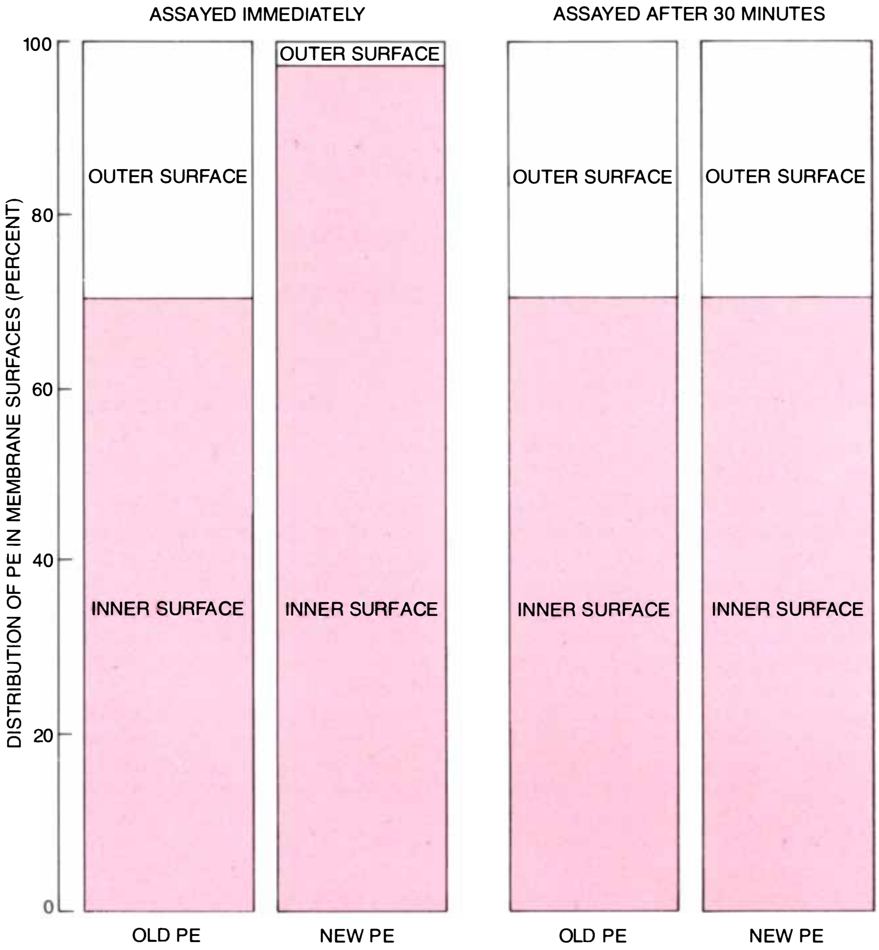


SITE OF LIPID SYNTHESIS was determined by independently labeling newly made PE and the fraction of the PE that is exposed on the outer surface of a bacterial cell. The newly formed PE is labeled by briefly incubating the cell with radioactive phosphate, which is incorporated in all phospholipids. The external PE is marked with TNBS, which combines with every PE molecule it can reach but which cannot cross the bilayer. When the membrane is dissolved, the lipids can be segregated according to whether or not they are radioactive and whether or not they are labeled with TNBS. (The latter determination is made by thin-layer chromatography.) Almost none of the radioactive PE molecules—the recently synthesized ones—are labeled with TNBS, indicating that the lipid is made only on the cytoplasmic side of the membrane.

From a series of such measurements we calculated that at physiological temperature the time required for a batch of new lipids to return halfway to their final distribution is about five minutes. That is almost 100,000 times faster than the rate of flip-flop observed in other membrane systems, and it cannot plausibly be attributed to spontaneous flip-flops. We have therefore proposed that growing membranes include proteins that facilitate the equilibration of lipids across the membrane, a notion first put forward by Bretscher. Such a protein need not influence the direction of lipid movement; it might merely provide a channel. The proteins would be needed only where membranes are actively growing, such as the plasma membrane of bacteria and the endoplasmic reticulum of eukaryotes; their absence elsewhere might explain the much lower rate of interchange between the layers of other membranes. It should be emphasized, however, that no such proteins have yet been shown to exist.

In this account of the lipid bilayer one further perplexity arises: If lipids can pass freely from one side of the membrane to the other in a growing membrane such as that of *B. megaterium,* how can the asymmetry of lipid composition be maintained? The explanation now favored by many investigators is that the asymmetry is not actively maintained at all but reflects a thermodynamic equilibrium between the molecules at the two surfaces. This hypothesis assumes that some species of lipids have a lower free energy on one side of the membrane than they do on the other. The difference in energy need not be great to explain the observed asymmetries, which are almost always modest compared with those of proteins. It is not hard to imagine a mechanism that might give rise to such a small energy difference. For example, if proteins found only at the cytoplasmic surface or ions in the cytoplasm bound PE with greater affinity than the proteins outside the cell, then each PE molecule might spend more of its time on the cytoplasmic side. In this way an asymmetric distribution would arise even though the lipid molecules were not physically restrained.

This hypothesis makes the asymmetries of proteins and lipids fundamentally different. Protein asymmetry is a nonequilibrium state enforced at the time of synthesis; if the proteins are allowed to come to equilibrium (by disrupting and reconstituting the membrane), they assume a more random configuration. Lipid asymmetry, according to the hypothesis, results from a somewhat biased equilibrium and does not depend on how or where the lipids are synthesized; if they were inserted in the outer monolayer, their ultimate distribution would be the same.



TRANSPORT OF LIPIDS across the bilayer is required if new lipid is formed at only one surface. In the bacterium *Bacillus megaterium* the distribution of PE in the two layers of the membrane is asymmetric: about 70 percent of it is found at the cytoplasmic surface. When the distribution of newly synthesized PE is measured, however, as in the experiment diagrammed on the opposite page, almost all of it is found at the inner surface. In order to maintain the integrity of the bilayer some of the new material must be ferried to the outer surface. In fact, within just 30 minutes the distribution of newly made PE matches that of the older material. It has been proposed that proteins in the growing bacterial membrane facilitate the movement of lipids across the bilayer, perhaps allowing them to reach thermodynamic equilibrium.

Our conclusions about the synthesis of lipids in bacterial membranes have recently been extended by other investigators who have found similar mechanisms at work in the endoplasmic reticulum of eukaryotic cells. Rosalind Coleman and Robert M. Bell of Duke University have shown that the enzymes responsible for the synthesis of several phospholipids all have their active sites on the cytoplasmic side of the reticulum membrane. Donald B. Zilversmit and his colleagues at Cornell University have measured transmembrane movement in the reticulum. They find that the lipids reach an equilibrium across the membrane of the endoplasmic reticulum in minutes, just as they do in bacteria. Therefore our model of how the lipid bilayer is assembled appears to be of general validity.

The studies described here, although they are confined to a few simple experimental systems, have brought to light several principles of membrane synthesis that seem to be widely applicable. It is only fair to point out, however, that other aspects of membrane assembly remain obscure.

One aspect is the transport of assembled membrane material from the endoplasmic reticulum to other sites in the cell. Each of the various organelle membranes, and the plasma membrane, has its own characteristic suite of lipids and proteins; to preserve these distinctions newly assembled membrane materials must somehow be sorted out as they leave the endoplasmic reticulum or the Golgi apparatus to ensure that they reach the proper destinations. There are also proteins peculiar to the endoplasmic reticulum itself, and it is not clear what prevents them from being exported along with all the rest. The supposed mechanism of intracellular transport—vesicles that bud off from the reticulum and fuse with another membrane—may well serve to ferry material to the plasma membrane, but it meets with difficulties elsewhere. Certain membrane proteins of mitochondria, for example, must cross both of the membranes surrounding that organelle.

Another deficiency of our models is that they cannot yet describe in satisfactory detail the one event that is crucial to the emplacement of both proteins and lipids: transport across the bilayer. For proteins this happens only once, at the time of synthesis and insertion. For lipids there is probably a rapid but undirected interchange for as long as new material is being added to the membrane. In both cases a membrane-bound protein probably serves as a gatekeeper, but the identity of these proteins and how they work are not known. It is in these areas, where present understanding is most vague and tentative, that the results of future investigations should prove most illuminating.

Guided-Wave Optics

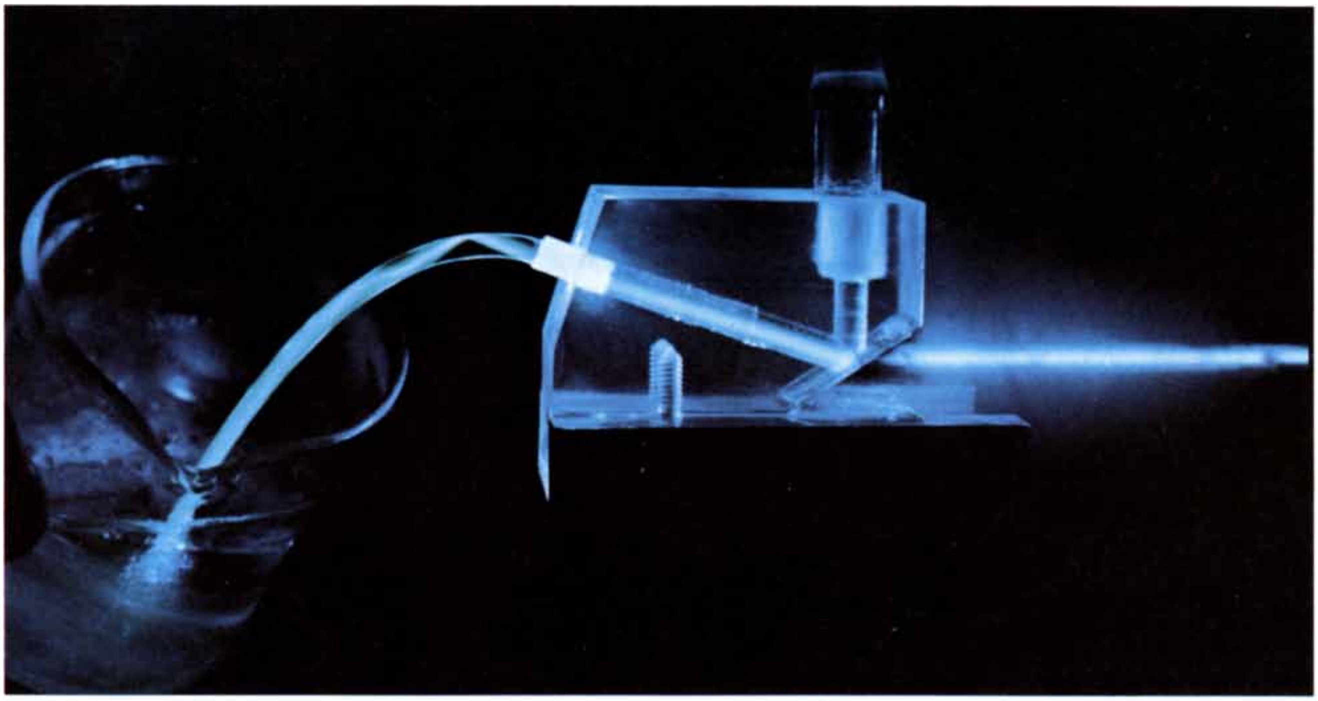
Progress is reported in the design of devices for manipulating laser beams in thin transparent films. The goal is to provide tiny, efficient components for communication by light waves

by Amnon Yariv

A conventional microwave communication system has a number of basic constituents. First there must be an oscillator of some kind to generate the "carrier" waves in a highly coherent form (that is, with the waves all in step at the same frequency). Next the waves must be modulated to carry information by varying either their frequency or their amplitude. The modulated microwave beam is then transmitted, either by channeling it through a metallic waveguide or by propagating it through the atmosphere. Since the waves become attenuated in transit, a long-distance communication system of this type requires in addition a series of intermediate relay stations, where the microwave signals are detected, amplified, reshaped to their original modulated waveform and sent on to their ultimate destination. There the beam is again detected, amplified and reshaped before being processed to recover the information impressed on it at the point of origin.

Until fairly recently it would have been unthinkable to try to put together a comparable communication system based on light waves. Yet the motivation for doing so had long been recognized: the signal-carrying capacity of any electromagnetic wave increases with its frequency, and light waves are typically between 20,000 and 200,000 times higher in frequency than microwaves.

The advent of the laser in the early 1960's set in motion a train of events that is steadily changing the outlook for optical communications. The past few years have seen three important advances. One was the invention of the semiconductor-diode laser, which satisfied the need for a cheap, efficient and long-lived source of coherent light waves to serve as the carrier waves for encoded signals. Another noteworthy advance was the development of ultrathin, low-loss optical fibers, which are well suited for the long-range transmission of light waves. Finally, an entire new generation of compact optical devices based on the precise guiding of laser light is being devised, making the



LIGHT-GUIDING PRINCIPLE employed in the development of miniature thin-film waveguides is demonstrated in this "tabletop" experiment conducted by Richard Moyer and the author at the California Institute of Technology. The original demonstration of the principle was given by John Tyndall more than a century ago. In this modern reenactment a coherent beam of blue-green light from an argon-gas laser is directed into a flowing jet of water. (Tyndall of course worked with an ordinary incoherent light source.) The difference between the refractive index of water and that of air makes it possible for beam to be guided by total internal reflection around bend in jet.

design of optical microcircuits a realistic prospect. It appears to be only a matter of time before efficient high-capacity optical-communication systems begin to replace microwave systems in a variety of applications. Here I shall review the current status of research on an important aspect of any future optical-communication system: the development of components based on the guiding of laser light in thin transparent films.

A beam of light propagating in the atmosphere always tends to spread out. This spreading behavior—diffraction—is a basic property not only of light waves but also of all other kinds of propagating waves. In practice diffraction means that if any light beam (even a highly collimated laser beam) is directed from one point to another, the beam will become greatly attenuated, provided the distance between the two points is great enough. Attempting to increase the intensity of the beam by reducing its "waist" with the aid of a focusing lens will actually cause the beam to fan out even more. As early as 1910 Peter J. W. Debye proposed that the problem could be avoided by propagating the light through a waveguide consisting of a multilayered transparent material.

To understand why a light beam is confined and transmitted efficiently through such a waveguide one must take a closer look at what happens when light is reflected at the interface separating two transparent materials. Light is transmitted through any transparent material at a velocity that is less than the velocity of light in a vacuum. The degree of retardation with respect to the velocity of light in air (300,000 kilometers per second) is a characteristic property of each material and is referred to as that material's index of refraction. For example, in a glass with a refractive index of 1.5 the velocity of light is 300,000 divided by 1.5, or 200,000 kilometers per second.

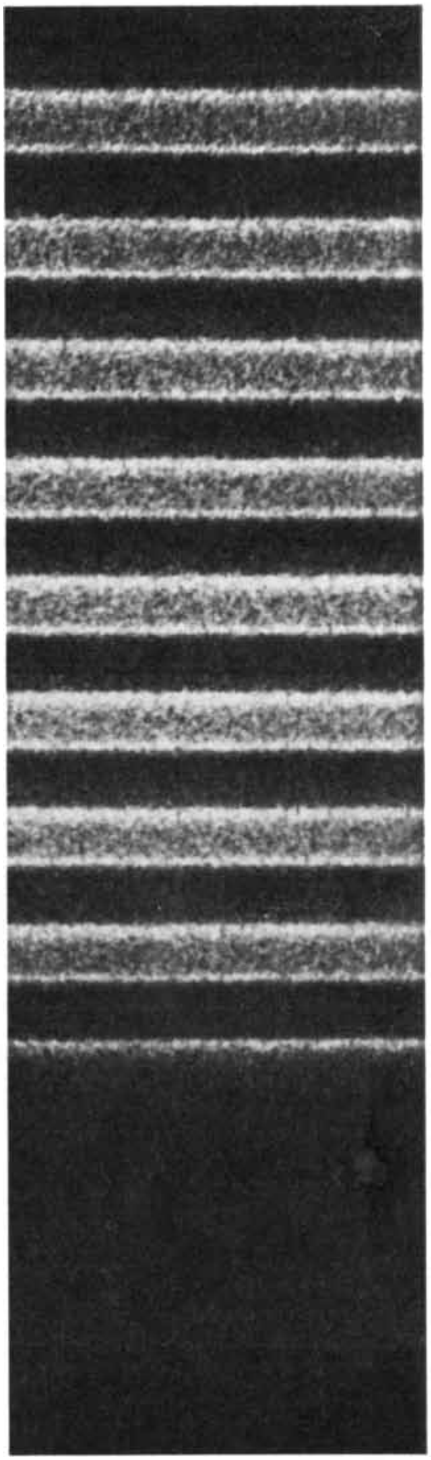
Whenever a light wave traveling through one material strikes a flat interface separating it from another material that has a lower refractive index, the discontinuity in the refractive index causes part of the wave to be reflected back into the first material, while the rest of the beam is refracted (transmitted at an angle) into the second material. At some critical angle of incidence (measured from a line perpendicular to the interface) the refracted wave will emerge parallel to the interface, and at greater angles of incidence (that is, shallower grazing angles) the refracted wave will disappear. In that case no light will be transmitted into the second material, and the incident wave is said to undergo total internal reflection. The phenomenon of total internal reflection takes place only if the refractive index of the light-conducting material is greater than that of the adjacent transparent material. For example, for a light beam propagating in a glass slab with a refractive index of 1.5 the critical angle of incidence for total internal reflection from a glass-air interface is 41.8 degrees.

In general the simplest kind of transparent waveguide is obtained when a material with a higher refractive index is sandwiched between materials with a lower refractive index. In such a waveguide it is possible for a single light ray to zigzag back and forth through the inner layer, undergoing successive total reflections at both the upper and the lower interface. The only condition that must be met is that the angle of incidence of the light ray at both interfaces has to exceed the critical angle for total internal reflection. By embedding the light-conducting material on all sides in a material with a lower refractive index one can get an efficient channel waveguide. The cross section of such a channel can be rectangular, as in the case of an actual thin-film waveguide, or circular, as in the case of an optical fiber.

Even though the light is totally reflected at the interfaces of such a waveguide, a small fraction of the propagating electromagnetic energy is always transmitted outside the guiding channel. Typically the intensity profile of the electromagnetic field is strongest at the center of the guiding channel and falls off gradually away from the center. As it happens, the fact that the transmitted energy is not entirely confined to the channel plays an important role in many applications.

How are transparent thin-film waveguides made? One important type of thin-film waveguide is fabricated by growing successive crystalline layers of semiconducting materials from their liquid phase. In this technique one usually starts with a highly polished substrate of some transparent crystalline semiconductor such as gallium arsenide (GaAs). One then proceeds to grow on it a thin crystalline layer of gallium-aluminum arsenide ($Ga_{1-x}Al_xAs$), a closely related semiconductor in which some fraction (designated x) of the gallium atoms are replaced by aluminum atoms. This layer is surmounted by a thin layer of pure gallium arsenide, which is in turn surmounted by another layer of gallium-aluminum arsenide. The index of refraction of gallium arsenide is greater than that of gallium-aluminum arsenide, and so light can be guided in the intermediate gallium arsenide layer.

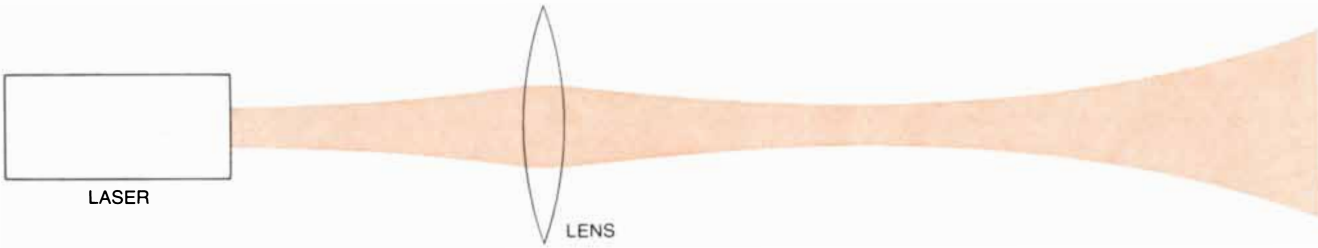
The various layers of such a structure are grown by bringing the gallium arsenide substrate in contact with pools of molten gallium that have been saturated with one or the other compound and then lowering the temperature to initiate crystal growth. Gallium arsenide and



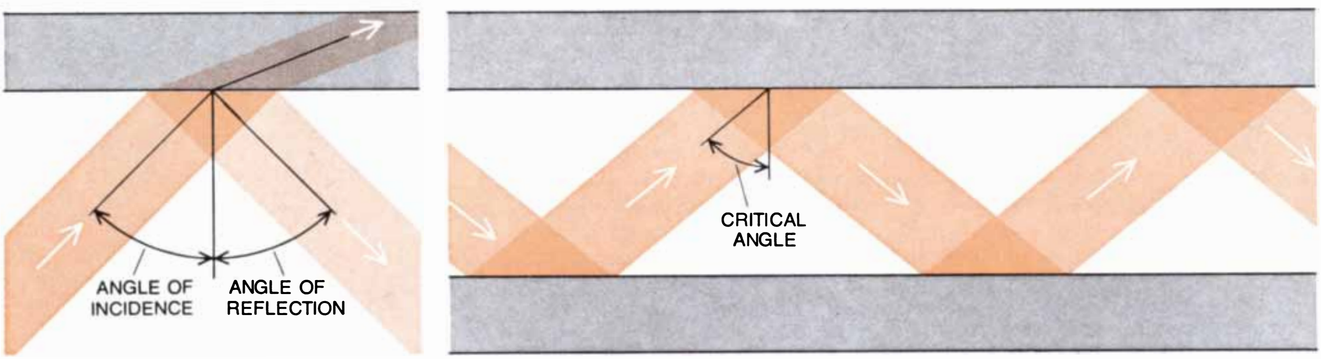
ALTERNATING LAYERS of two transparent semiconducting materials, gallium arsenide and gallium-aluminum arsenide, are seen in this scanning electron micrograph. The layers, which are only a few microns thick, were fabricated by the extremely precise new technique of molecular-beam epitaxy. The resulting single-crystal structure can be used to trap light waves entering the uppermost layer from either side, confining them to the first few layers of the device as they travel along parallel to the interface separating the surface layer from the ambient air. This unusual mode of light propagation was observed for the first time by the author and Pochi A. Yeh of Cal Tech and Alfred Y. Cho of Bell Laboratories.

gallium-aluminum arsenide have the same crystal structure, so that the resulting multilayer structure consists in effect of a single uninterrupted crystal.

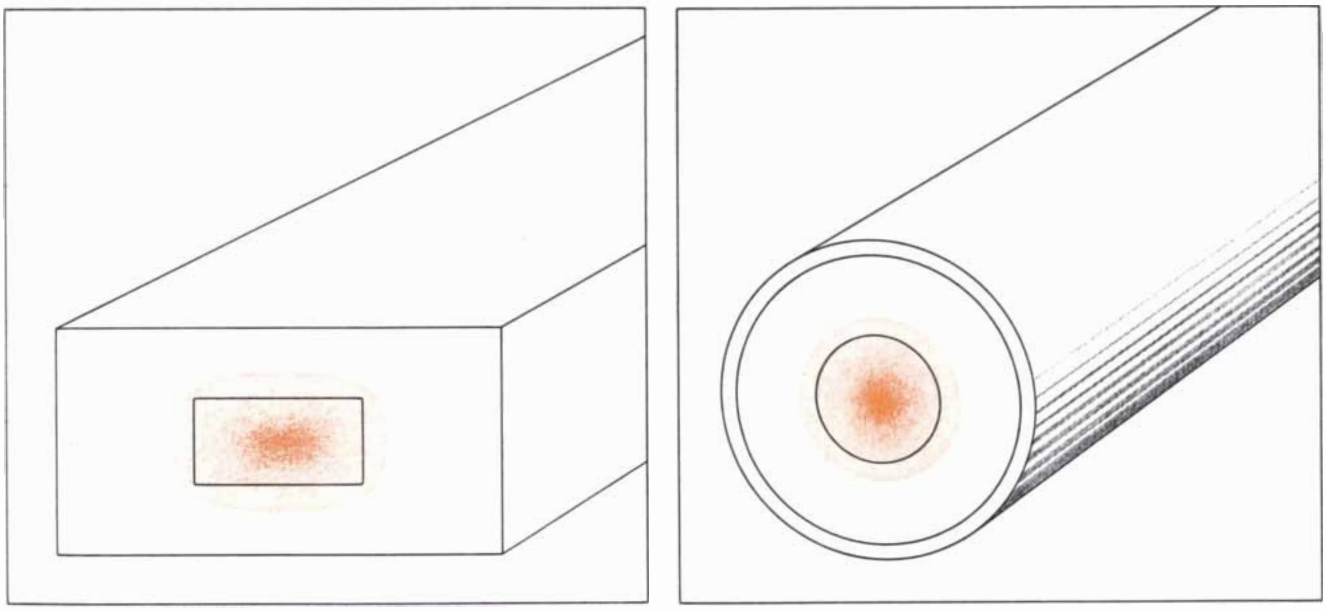
Another important property of such a structure is that the electrical properties of the semiconducting layers can be controlled individually by adding impurity atoms to the melt pool. The addition of electron-donating atoms, such as those of tin or tellurium, results in an *n*-type material, in which the current is conducted by negative charge carriers (the extra donated electrons). The addition of electron-accepting atoms, such as those of zinc, results in a *p*-type material, in which the current is conducted by positive charge carriers (the mobile electron vacancies, or "holes"). The mutual annihilation of electrons and holes near the interface of a semiconducting



DIFFRACTION of a laser beam propagating through the atmosphere, shown greatly exaggerated here, means that beyond a certain critical distance the intensity of the beam will be attenuated in proportion to the square of the distance. For example, a laser beam with an initial diameter of one centimeter will spread in a distance of one kilometer to a diameter of approximately 20 centimeters; the intensity of the beam will accordingly be reduced by a factor of 400. Focusing the beam with a lens will cause it to fan out even more.



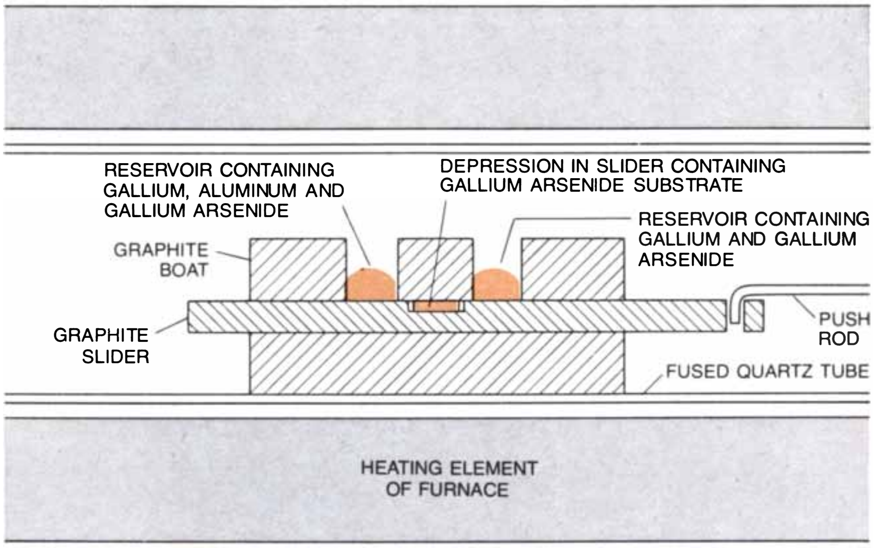
REFLECTION of a laser beam traveling in a transparent material takes place at the flat interface with another transparent material that has a lower refractive index (*left*). Part of the beam is reflected back into the first material, while the rest of the beam is refracted into the second material. As the angle of incidence increases beyond some critical value the beam undergoes total internal reflection (*right*).



CHANNELING of a laser beam can be accomplished by embedding the light-conducting material on all sides in a material with a lower refractive index. The cross section of such a channel waveguide can be rectangular (*left*) or circular (*right*). In either case a small fraction of the propagating electromagnetic energy (*color*) is always transmitted outside the guiding channel. Structure at right is typical of optical fibers, which consist of a glass core surrounded by a glass cladding; core diameters range from a few micrometers to 100 micrometers.

p-n junction leads to the emission of light waves. In 1963 Rogerio C. Leite and I, working at Bell Laboratories, found that the light emitted by such a device was channeled by dielectric waveguiding in the vicinity of the *p-n* junction. This observation helped to stimulate interest in the launching of research on guided-wave optics.

Transparent waveguides can also be made by changing the chemical composition of a material through the introduction of a trace constituent that alters the material's index of refraction. A common waveguide of this type, developed by Jacob M. Hammer and his colleagues at RCA Laboratories, is obtained by diffusing titanium atoms into crystalline lithium niobate ($LiNbO_3$). The thin surface layer of the lithium niobate in which the titanium atoms are present has a higher refractive index than that of the surrounding bulk material, and hence it can serve as an optical waveguide. The titanium can be diffused into the substrate material in patterns by means of a photolithographic masking technique, creating not only straight-channel waveguides but also more complex networks.



FABRICATION of a semiconducting thin-film waveguide for use in research on light-wave communication is carried out in the specially designed apparatus shown here. The various layers of such a structure are grown epitaxially (that is, in successive identical-crystal stages) by bringing a crystalline substrate of gallium arsenide in contact with pools of molten gallium that have been saturated with either gallium arsenide (GaAs) or gallium-aluminum arsenide ($Ga_{1-x}Al_xAs$), a compound in which some fraction (x) of the gallium atoms are replaced by aluminum atoms, endowing the latter material with a lower refractive index than that of the former material. The substrate is moved by means of a push rod under the appropriate reservoir, and the temperature is lowered to initiate crystal growth. The thickness of a layer is controlled by adjusting the cooling rate of the furnace and the period of contact with the molten pool. The procedure is repeated at the second reservoir to grow a second layer. Some recent devices of this type call for as many as eight layers, and hence they require eight reservoirs.

The semiconductor-diode laser is the only type of laser that seems likely to have a strong impact on the consumer market in the near future. This unique potential arises primarily from its simplicity and small size. In a laser of this type the energy of a small electric current passing through a multilayer semiconducting crystal is converted directly into laser light. The extremely small size of the device (less than a millimeter on a side) means that it can be used in conjunction with integrated electronic microcircuits, which can drive the laser directly. That factor, combined with the semiconductor-diode laser's convenient range of output wavelengths in the near-infrared region of the spectrum, make it the best available choice for generating the carrier waves for a light-wave communication system employing optical fibers as the transmission link.

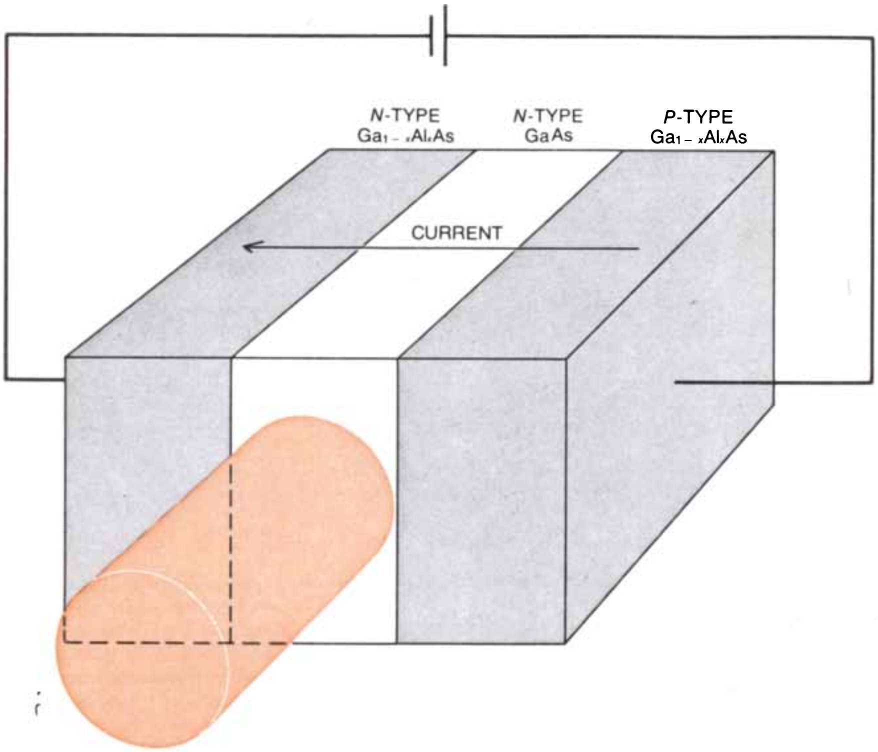
From the point of view of the waveguide designer a semiconductor-diode laser is nothing more than a thin-film waveguide composed of a gallium arsenide guiding layer sandwiched between two layers of gallium-aluminum arsenide. To understand how light guided in the inner gallium arsenide layer can be amplified by an electric current flowing across the layers, one needs to first examine how the controlled "doping" of the layers with impurity atoms affects the spatial distribution of the electrons and their energies.

In the absence of an applied voltage the conduction electrons in the *n*-type gallium-aluminum arsenide on one side of the gallium arsenide layer and the electron vacancies in the *p*-type gallium-aluminum arsenide on the other side have approximately the same energy. Both the electrons and the holes are prevented from penetrating the inner gallium arsenide layer by the high potential-voltage barriers at the interfaces. The application of a large "forward bias" voltage to the structure raises the energy of the electrons on the *n*-type side with respect to the energy of the holes on the *p*-type side. The applied voltage has the effect of lowering the potential barriers, thereby allowing the electrons from the *n*-type region and the holes from the *p*-type region to flood into the central gallium arsenide layer.

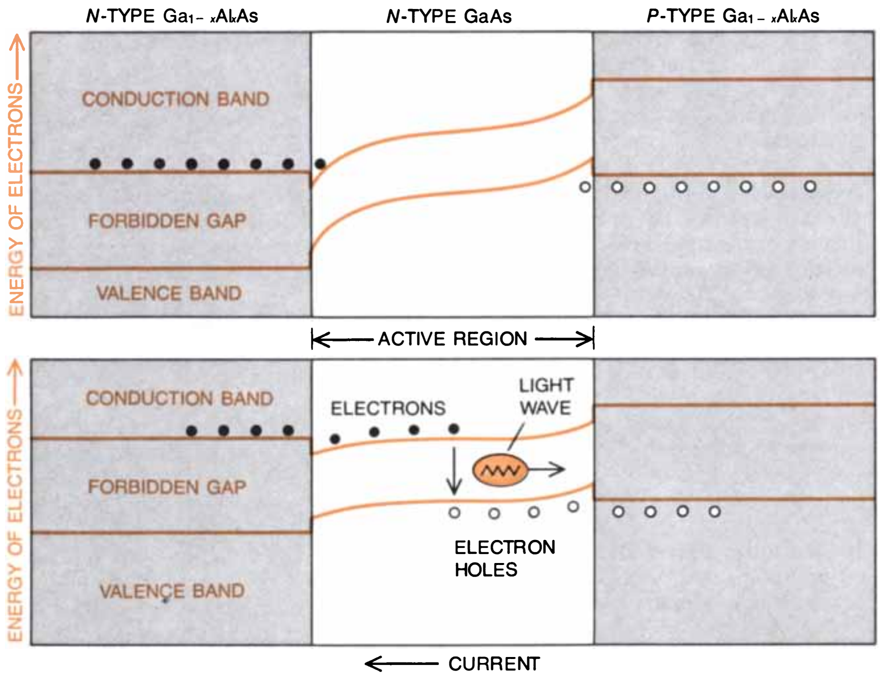
The result is a situation in which electrons at a certain energy share the same volume of space with holes at a lower energy. In that situation the electrons can be stimulated by the presence of light of the appropriate wavelength to undergo a transition across the "band gap" from the higher-energy conduction band into the empty energy states (the holes) in the lower-energy valence band, emitting a photon, or light quantum, for each transition. Since the energy of the stimulated light is added coherently to that of the stimulating light, the process qualifies as a laser amplification mechanism. The other main criterion for laser action—the provision of a feedback mechanism—is satisfied by the reflection of the emitted light waves at the cleaved end surfaces of the crystal. The wavelength of the emitted radiation is determined by the energy gap separating the electrons from the holes; in gallium arsenide, for example, the resulting wavelength is approximately .85 micron.

The entire process can be summarized by following the path of a single electron. The electron is first extracted from the *p*-type layer and enters the external wire, leaving a hole in the material's valence band. The circulating electron is next elevated in energy as it passes the voltage source (a battery, say) by an amount that is proportional to the applied voltage. The electron then diffuses without losing energy through the *n*-type layer and into the active gallium arsenide layer. Simultaneously the hole that was left behind drifts across from the *p*-type layer into the active region. The laser radiation, which is trapped by total internal reflection within the same gallium arsenide layer, stimulates the electron to recombine with the hole in the valence band, imparting the difference in energy to the propagating wave, which in the process is amplified.

The original version of this laser was invented in 1962 by groups of investigators working independently at the General Electric Company research laboratories, the Thomas J. Watson Research Center of the International Business Machines Corporation and the Lincoln Laboratory of the Massachusetts Institute of Technology. The first semiconductor-diode lasers were made entirely of gallium arsenide. The lack of effec-



SEMICONDUCTING-DIODE LASER consists essentially of a thin film of gallium arsenide sandwiched between two layers of gallium-aluminum arsenide. The application of an external voltage across this single-crystal stack causes an electric current to flow at right angles to the layers (*from right to left*). The current is carried by extra electrons flowing inward from the negatively charged (*n*-doped) layer and by electron vacancies (or "holes") flowing inward from the positively charged (*p*-doped) layer. The electrons and the holes recombine in the central gallium arsenide layer, where they emit their excess energy in the form of visible light.



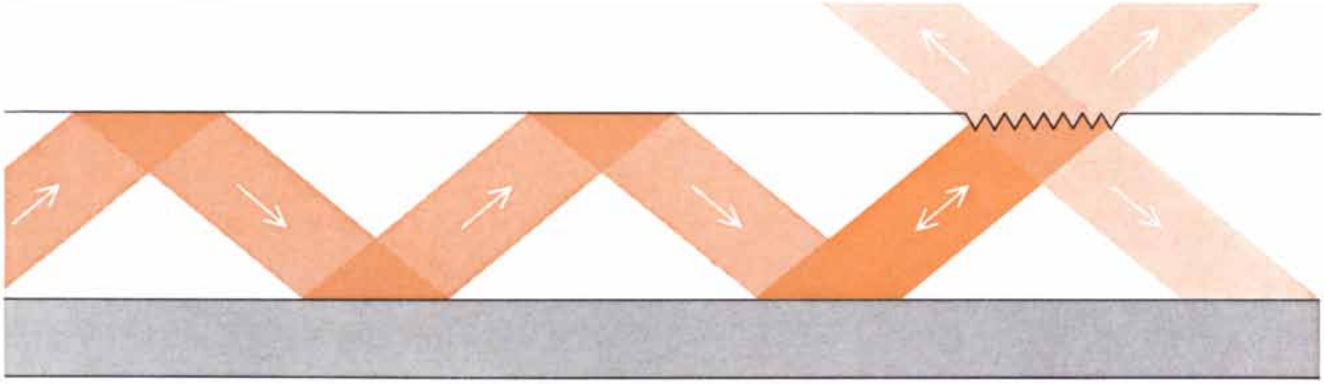
ENERGY-LEVEL DIAGRAMS show the relation between the energy of the electrons and the holes on the one hand and their location in the layered structure of a semiconducting-diode laser on the other. In the absence of an applied voltage (*diagram at top*) the electrons (*black dots*) and the holes (*open circles*) have the same energy but are spatially separated. When a "forward bias" voltage is applied (*diagram at bottom*), the electrons and the holes become separated in energy but are now able to cross over into the central gallium arsenide layer. There the electrons are free to jump across the "forbidden gap" of energies between the conduction band and the valence band, recombining with holes in the lower-energy band. Each such transition results in the emission of a photon of radiation with a wavelength of .85 micrometer.

tive waveguiding in these early devices meant that the threshold current for laser action had to be quite high, necessitating the maintenance of very low temperatures for continuous operation. The idea of combining electron injection with a layered-waveguide structure by sandwiching the gallium arsenide between layers of gallium-aluminum arsenide was put forward in 1969 by two Russian workers, Zhores J. Alferov and Rudolph Kazarinov. Their suggestion has led to the low-threshold-voltage semiconductor-diode lasers available for widespread use today.

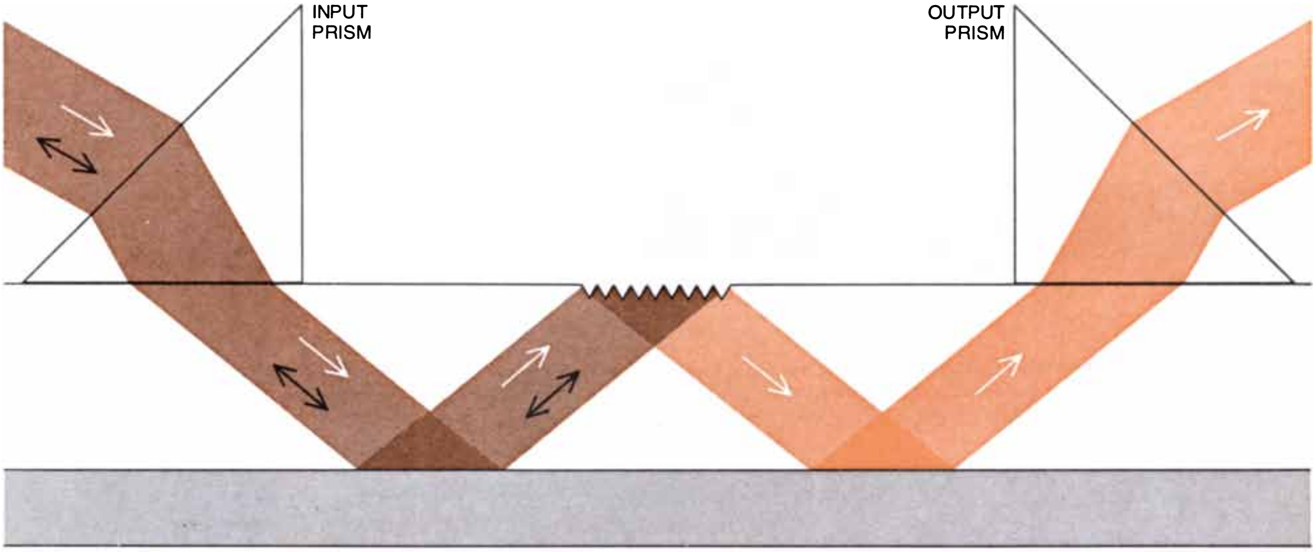
A number of applications in the area of guided-wave optics call for the incorporation of a periodic perturbation into the structure of the waveguide. The most common perturbation takes the form of a corrugated section in one or more of the interfaces of the waveguide. Consider what happens when a light wave propagating in a waveguide encounters such a corrugation. Each ridge of the corrugated section scatters reflected and refracted light in all directions. There can exist, however, a set of discrete directions along which the radiation scattered by all the ridges together adds up in step. In any particular case these directions depend on the wavelength of the light and the period of the corrugation (that is, the distance between two adjacent ridges).

If the period of the corrugation is equal to half the wavelength of the light propagating in the waveguide or to some multiple thereof (a situation known as the Bragg condition), then the light will be reflected exactly backward, propagating in a direction opposite to that of the incident wave. In other words, the corrugation can act as a kind of mirror. A short section of such a corrugated waveguide will strongly reflect light waves that satisfy the Bragg condition and will let other waves pass through with little loss. This property can be exploited as the basis of a new class of optical filters (that is, devices that select or reject certain bands of wavelengths). In one such filter the input wave propagates in the waveguide at right angles to the corrugation. Since the filter in this case rejects the narrow band of wavelengths that satisfy the Bragg condition, it is referred to as a band-rejection filter. The first filter of this type was operated successfully at Bell Laboratories by Dale C. Flanders, H. W. Kogelnik and Charles V. Shank in 1974.

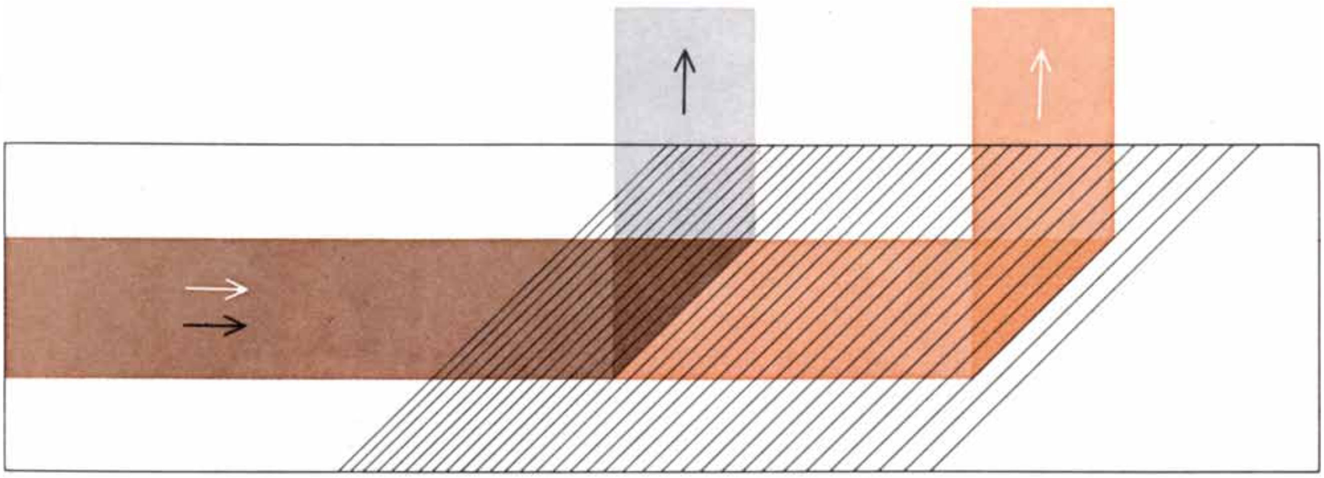
Another application of the Bragg-reflection principle in corrugated waveguides was developed recently in my laboratory at the California Institute of Technology by my colleagues Chi-Shain Hong, Jeffrey B. Shellan, Abraham Katzir and Alexander C. Livanos. In their design a corrugation with a variable period is first made in a transparent waveguide. The input beam is then fed



CORRUGATED SECTION in one of the interfaces of a thin-film waveguide reflects and refracts light from a propagating laser beam in a set of discrete directions, depending on the wavelength of the light and the period of the corrugations. Four directions are shown.



COINCIDENT LASER BEAMS with different wavelengths can be made to enter a corrugated thin-film waveguide with the aid of a prism (*left*). If the period of the corrugation is equal to half the wavelength of one of the beams or to some multiple thereof (a situation known as the Bragg condition), then the corrugation will act as a kind of mirror, reflecting that beam (*gray*) exactly backward. The other beam (*color*) will be unaffected by the corrugation and will emerge from the output prism with little loss. The device acts as a band-rejection filter.

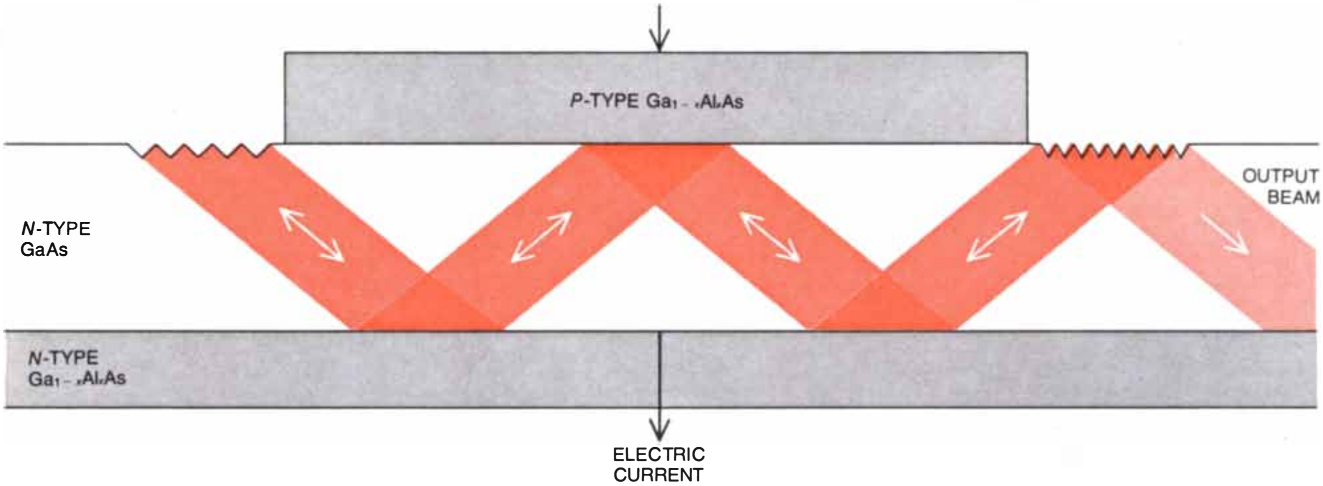


ANOTHER SOLUTION to the problem of separating coincident laser beams of different wavelengths in a corrugated waveguide was devised recently by members of the author's research group at Cal Tech. In this approach the corrugation has a variable period, and the coincident input beams are fed in at an angle of 45 degrees to the corrugations (as seen here from directly overhead). Light with wavelengths satisfying the Bragg condition will be reflected at a right angle with respect to the incident beams. Since the period of the corrugation varies along the direction of the propagating light, different wavelengths (*gray and color*) will be reflected at different locations.

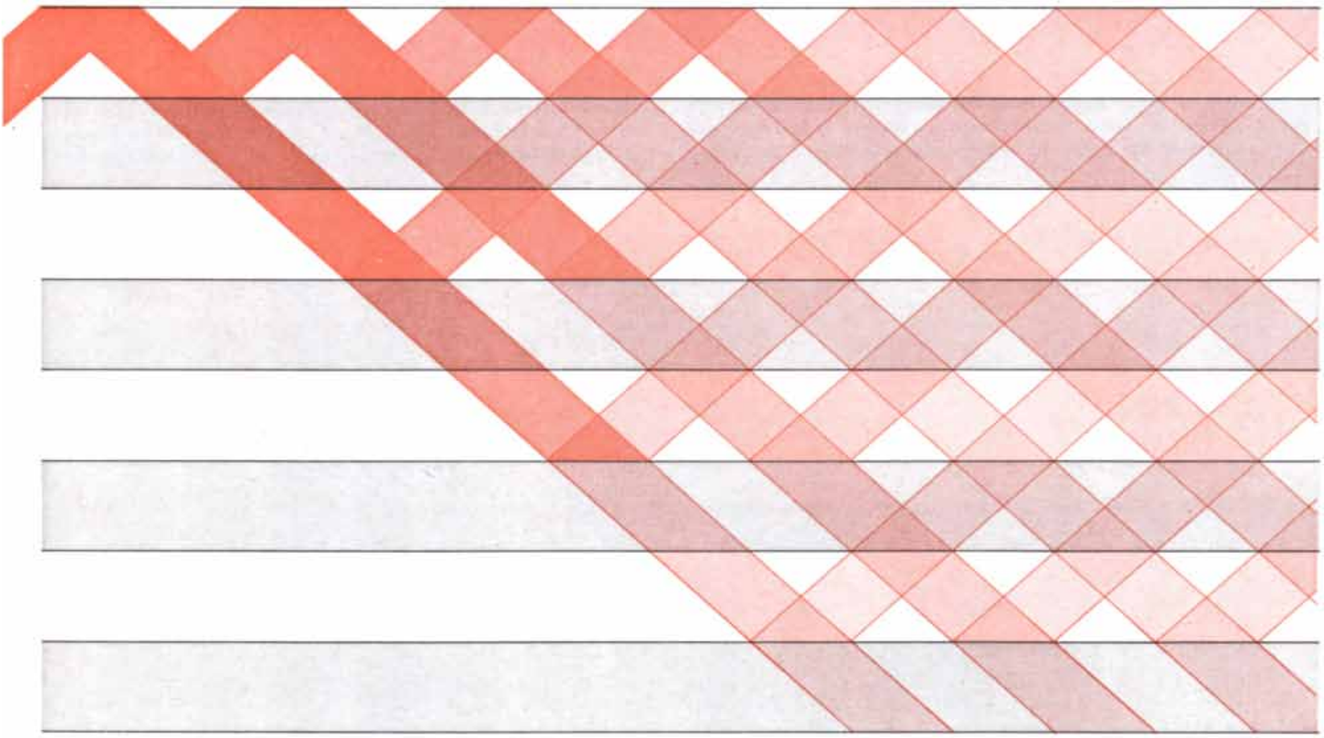
in at an angle of 45 degrees with respect to the corrugation. Wavelengths satisfying the Bragg condition (which now requires that the period of the corrugation be equal to a multiple of the wavelength divided by the square root of 2) are reflected at a right angle to the direction of the incident beam.

Since the period of the corrugation varies along the direction of the propagating light, different wavelengths are reflected at different locations. This scheme should find applications in cases where a number of light beams, each having a different wavelength and each carrying different information, are transmitted in a single channel (say an optical fiber). At the receiving station the beams would be separated spatially by means of another variable-period Bragg filter and sent to a set of detectors for recovery of the information carried by each beam.

When two corrugated sections are incorporated in a single semiconductor-diode laser, one on each side of the amplifying section, they can replace the usual cleaved end-face mirrors as the laser reflectors. The high degree of wavelength selectivity of these reflectors forces the laser to emit an extremely monochromatic (that is, single-wavelength) light. Accordingly it is possible to choose, within a broad range, the characteristic wavelength of a semiconductor-diode laser by selecting the corresponding period of the corrugation fabricated into the laser's structure. Lasers operating on this principle were first



NOVEL LASER can be built by incorporating two corrugated sections into a single thin-film semiconducting-diode structure, one on each side of the central amplifying section, where they can replace the usual cleaved end-face mirrors as the laser reflectors. The reflectivity of the corrugated sections can be adjusted to let a predetermined fraction of the light (*light color*) emerge as the output beam.



PROPAGATION OF LIGHT WAVES near the interface of a stratified semiconducting material and a homogeneous surrounding medium (in this case air) is explained with the aid of a schematic diagram. The layered structure is the same as the one shown in the electron micrograph on page 55. The laser beam enters the uppermost layer of gallium arsenide from the left and is confined to the first few layers of the device by total internal reflection at the interface with the air and by repeated in-phase Bragg reflections at the interfaces of the gallium arsenide and gallium-aluminum arsenide layers. In this case the intensity falls virtually to zero within seven layers.

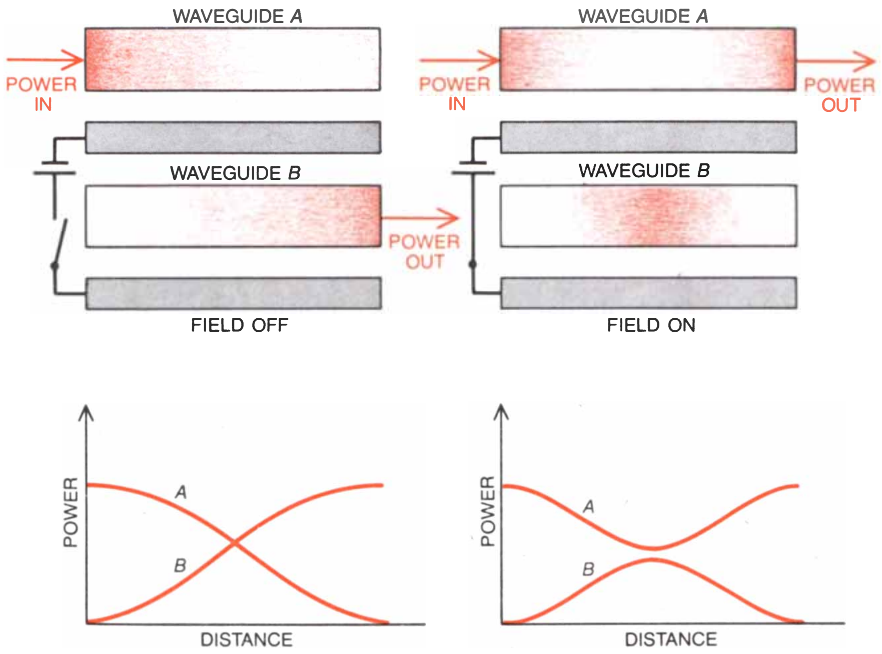
demonstrated by Huan-Chun Yen, Michiharu Nakamura and me at Cal Tech in 1975.

The period of the corrugated gratings employed in the applications described above is on the order of a fraction of a micron, which corresponds to a few thousand grating lines per millimeter. These ultrafine gratings are fabricated by means of a photolithographic technique utilizing a holographic exposure method. The technique is based on the interference fringes created by the interaction of two laser beams within a thin photosensitive film coating the surface to be corrugated. The period of the gratings is determined by the wavelength of the laser beams and their angle of incidence.

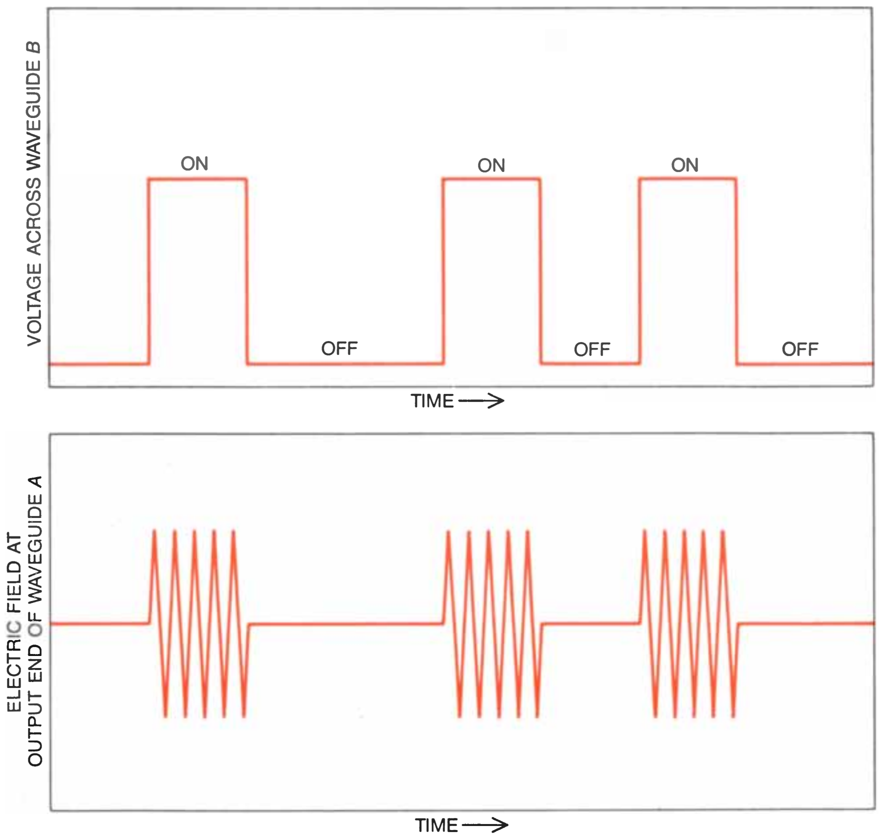
The study of light reflection under the Bragg condition led recently to the discovery by me and my colleagues Pochi A. Yeh of Cal Tech and Alfred Y. Cho of Bell Laboratories of a new mode of propagation of light waves near the interface of a stratified semiconducting material and a homogeneous medium. If the period (in this case the repeat distance) of the layered material is approximately half the wavelength of the incident light, then the light waves will be prevented from entering the bulk of the layered material, since successive reflections from neighboring interfaces will be exactly in phase with one another and so will be reinforced. (This situation is analogous to the Bragg reflection of X rays from crystal planes.) The incident wave, unable to enter the bulk of the layered material, would normally escape to the homogeneous surrounding medium (air), but in this case it will be turned back by total internal reflection. The net result is a propagation mode in which most of the energy of the light is confined to the first few layers and is carried along parallel to the planes defined by the interfaces separating the layers.

The fabrication of such precisely periodic layered materials has become feasible only with recent improvements in a method of crystal growth called molecular-beam epitaxy, which make it possible to control the thickness of the layers to within a few angstrom units. Optical waveguides based on the Bragg-reflection principle are free from some of the fundamental constraints imposed on conventional waveguides and are expected to play an important role in applications where a high degree of wavelength selectivity is important or in applications where conventional waveguiding cannot be used.

The fact that most optical waveguides are made of crystalline materials opens up a range of possibilities not available to the microwave-circuit designer. For example, the index of refraction of many crystalline materials can be controlled by the application of an electric



DIRECTIONAL WAVE-COUPLING DEVICE consists of two adjacent transparent waveguides in which the outlying portions of the propagating light waves overlap. With the control field off a light wave introduced into waveguide *A* will be coupled coherently and completely into waveguide *B* (*diagram at upper left*). The application of an electrical control field across waveguide *B* will change the velocity of the waves in that guide and so spoil the perfect power exchange (*diagram at upper right*). By the midpoint of the effective interaction length of the device half of the power in waveguide *A* will have been transferred to waveguide *B*, but at the output end the full power will have returned to waveguide *A*. The graphs below the diagrams represent the intensity of the light wave in each waveguide as a function of distance.



DIRECTIONAL WAVE-COUPLING CAN BE USED to impress pulse-coded binary information on the emerging light waves. By applying a train of coded voltage pulses across waveguide *B* (*curve at top*) one can modulate the amplitude of the optical output beam (*curve at bottom*) in such a way that its envelope becomes a replica of the electric-pulse train.

field, a phenomenon known as the electro-optic effect. Since the waveguiding properties of a material depend strongly on its index of refraction, the application of an electric field to a transparent waveguide can drastically modify the nature of the propagated light. This electro-optic control method can be harnessed for rapidly modulating, switching and coupling light beams.

One example of the electro-optic control of waveguiding is the process called directional coupling. The basic effect here involves two parallel waveguides that are close enough to each other for the outlying portions of the light propagating in one waveguide to reach across to the other waveguide and vice versa [*see top illustration on preceding page*]. Under these conditions a wave propagating in one waveguide (designated *A*) will gradually leak into and excite the second waveguide (*B*), provided the velocity of the waves in both waveguides is the same. The equality of the wave velocities ensures that the excitation of waveguide *B* by the "leaky" field of waveguide *A* is in exact phase with the wave propagating in waveguide *B*. Since the total amount of propagating power is a constant, the excitation of the light wave in waveguide *B* must be accompanied by a diminution of the power in waveguide *A*.

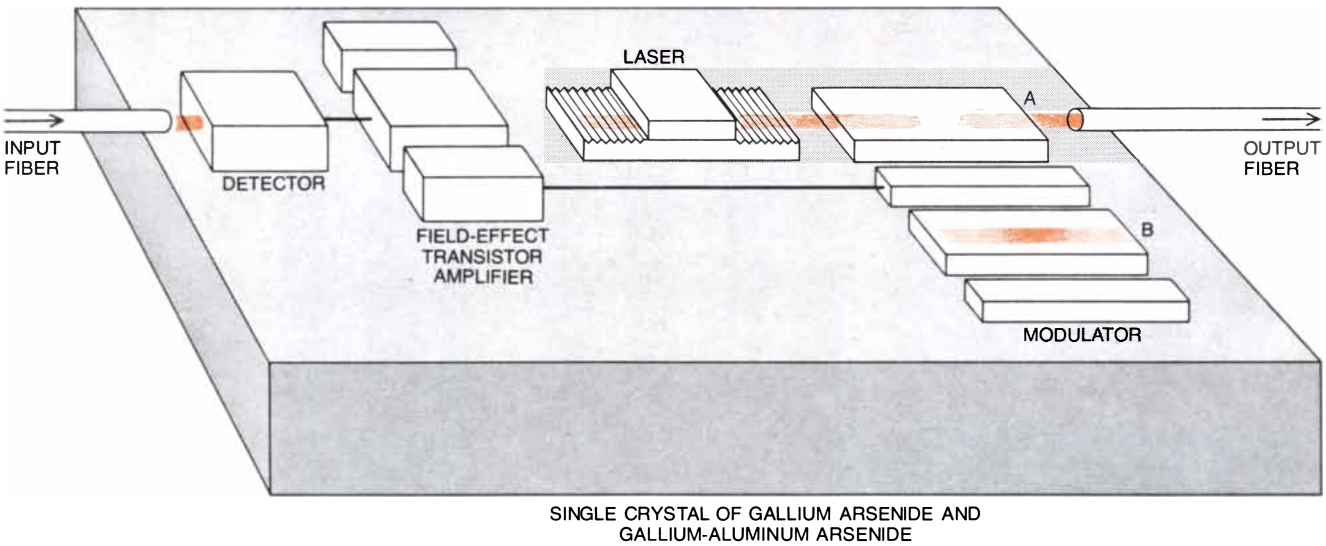
The result is a complete transfer of power from waveguide *A* to waveguide *B* within a certain distance. When an electric field is applied across waveguide *B* by supplying a voltage to a pair of nearby electrodes, the electro-optically induced change in the refractive index causes the wave velocity in that waveguide to change slightly. The resulting mismatch in velocity prevents the cumulative buildup of power in that waveguide, and only a fraction of the total power of the light wave in waveguide *A* is transferred to waveguide *B*. Farther along the waveguides the coupled power returns to waveguide *A*. In effect what one has is a device in which power entering one guide can emerge either from the output end of the other guide (when the voltage is "off") or from the first guide (when the voltage is "on").

This electrical switching of guided light, first demonstrated in 1975 by James Campbell and Fred A. Blum of Texas Instruments Incorporated can be used to construct optical distribution networks in which a number of incoming optical fibers are connected in any desired combination with a number of output guides. Equally important, it can be used to modulate an optical beam with an electrical signal. To understand how this modulation is effected consider a stream of data encoded onto a train of binary ("on" or "off") voltage pulses applied across waveguide *B* [*see bottom illustration on preceding page*]. A continuous laser beam is fed into the input end of waveguide *A*. During the "on" period of a pulse the power leaves by way of the output end of waveguide *A;* during the "off" period it emerges from the output end of waveguide *B*.

The electric-pulse train is thus reproduced in the form of a modulation of the envelope of the optical carrier beam; in other words, a plot of the spatial distribution of the intensity of the propagating light wave is a replica of the waveform of the voltage. In this way the information can now be carried by the light beam propagating inside a fiber or through the atmosphere. An optical detector such as a photomultiplier or a semiconductor diode can later reconvert the light beam to an almost perfect replica of the original electrical signal.

The phenomena and the devices described in this article, which are based on the waveguiding of light in thin transparent films and its manipulation by impressed fields and currents, are the result of research conducted almost entirely during the past eight years or so. Now that most of the main building blocks for optical communication have been demonstrated, the shift in research and development is toward the integration of thin-film devices on a single chip to perform more complex functions.

One example of a "next generation" circuit that could be built with existing technology is a repeater station in a fiber-optics communication link. Information carried as an envelope modulation of the light beam entering the repeater in a fiber is recovered (converted to an electrical signal) by a *p-n*-junction detector. The detected electrical signal (which consists of a string of 1's and 0's) is reshaped, amplified in a transistor amplifier and used to modulate the output of a laser beam. All the devices included in this repeater have been demonstrated separately in gallium arsenide crystals. It is therefore possible in principle to build the device (and even more complex optical circuits) on a single-crystal chip. Challenges of this kind provide a continuing stimulus to those of us active in the field of guided-wave optics.



ENTIRE OPTICAL CIRCUIT can be manufactured with currently available technology on a single-crystal chip of semiconducting material. This particular circuit, for example, could serve as a relay station for an optical communication link in which the information-bearing signal, degraded and attenuated by long-distance transmission over an optical fiber, is detected, amplified and reshaped electronically. The resulting voltage signal would in turn be used to modulate the output of a corrugated semiconducting-diode laser in a twin-waveguide directional coupling device of the type shown on preceding page. More complex optical circuits are also becoming feasible.

No grants needed, thank you

There is science for its own precious sake, and there is the use of scientific instruments, techniques, and personnel to make a buck and stay out of trouble. What P. Mark Henrichs does certainly sounds like basic science. Nevertheless, if the management were convinced that his contributions had nothing whatsoever to do with making some bucks for Kodak, *he'd* be in trouble. Not in finding other suitable employment, of course, but in troubling to write grant proposals, the way scientists have to do when less directly tied into the general economy.

Indeed, a scant five years after joining us from post-docs in Basel and South Carolina Dr. Henrichs got the management to buy him a nuclear magnetic resonance (NMR) spectrometer with a 63-kilogauss field supplied by a magnet with windings kept superconducting by liquid helium. He uses it for Fourier transform spectroscopy to explore structures of compounds tagged with such nuclides as ^{13}C, ^{15}N, and ^{109}Ag, along with ^{1}H.

Lately Henrichs has focused on developing a method for separating relaxations of degenerate transitions. He has also come up with interesting information about the molecular structure of a variety of silver complexes in solution, and about charge distribution, aggregation, and conformation in various dyes. We are furthermore much impressed by his skill in detailing composition and microstructure of polymers.

Conformational Analysis of Carbocyanine Dyes with Variable-Temperature Proton Fourier Transform Nuclear Magnetic Resonance Spectroscopy

P. M. Henrichs* and S. Gross

Contribution from the Research Laboratories, Eastman Kodak Company, Rochester, New York 14650.

"Selective excitation transfer and separated relaxation measurements of degenerate transitions in NMR spectroscopy," by P. M. Henrichs and L. J. Schwartz

(to be published in J. Chemical Physics)

Application of ^{13}C Chemical Shifts and Relaxation Times to the Study of Charge Distribution, Aggregation, and Conformation in Carbocyanine Dyes. Part I

By P. Mark Henrichs, Research Laboratories, Eastman Kodak Company, Rochester, New York 14650

"Diad, Triad, and Pentad Sequence-Distribution Analysis of Acrylonitrile-Vinylidene Chloride Model Copolymers by ^{13}C-NMR," by D. B. Bailey and P. M. Henrichs

(to be published in J. Polymer Science)

Complexation of Silver(I) with Thiourea and Tetramethylthiourea in Dimethyl Sulfoxide Solution as Studied by ^{13}C and ^{109}Ag Nuclear Magnetic Resonance Spectroscopy

P. M. Henrichs,* J. J. H. Ackerman, and G. E. Maciel*

Contribution from the Research Laboratories, Eastman Kodak Company, Rochester, New York 14650, and the Department of Chemistry, Colorado State University, Fort Collins, Colorado 80521.

Selective Excitation Transfer in Nuclear Magnetic Resonance

P. M. HENRICHS   *Research Laboratories*
L. J. SCHWARTZ   *Eastman Kodak Company*
Rochester, New York 14650

"Selective Excitation Transfer in Spin-Spin Multiplets: A Method of Obtaining Separated Relaxation Measurements on Degenerate Transitions," by P. M. Henrichs and L. J. Schwartz

(paper presented at Experimental NMR Conference)

If information presented in any of the above offers hope of benefitting others than Kodak, so much the better. For a reprint: note Henrichs' address above.

SCIENCE AND THE CITIZEN

The Strategy of Fusion

The competition for funds within the U.S. fusion-research establishment has sharpened in recent years as different groups pursuing different approaches to the problem have intensified their efforts to achieve the "break even" conditions needed to demonstrate the ultimate utility of this alternative long-range energy source. From the point of view of the Department of Energy the situation has reached the point where a broad restatement of the objectives and priorities of the national fusion program has become advisable. Accordingly in a paper prepared for a recent International Nuclear Fuel Cycle Evaluation (INFCE) conference in Chicago, John M. Deutch, the department's director of energy research, spells out for the benefit of both the general public and workers in the field the current policy of the Carter Administration regarding the support of research on controlled fusion.

Of the three major "inexhaustible" energy sources thought to be potentially available to meet the nation's long-range energy needs (solar-radiation collectors, fission reactors of the breeder type and fusion reactors) fusion is regarded by the Administration's energy planners as "furthest ... from practical economic utility." Nevertheless, Deutch emphasizes, "its potential rewards are great." For the distant promise of cheap, safe and unlimited fusion power to be realized, however, "there is no lead time to spare: if fusion energy is to be available when it is needed—thousands of megawatts economically produced about 2030—the research and development program, rigorously directed toward the goal of commercial utility, must be undertaken in earnest now."

The high priority the department places on fusion research, Deutch points out, "has recently been demonstrated tangibly" by the fact that the total budget for fusion research (comprising both the magnetic-confinement and the inertial-confinement approaches) is expected to remain "essentially at its present level [about $500 million per year] through fiscal year 1980." The decision to exempt the fusion program "from even a straight-percentage share" of the President's current round of Federal budget cuts is characterized as "a strong affirmation of support."

The overall funding decision was made following a six-month study by the department's newly established Fusion Review Committee, which found the existing program "to be technically sound and to possess promise of achieving economical fusion power eventually." In spite of the confidence expressed in the program's ultimate success, the paper submitted to the INFCE conference states, "we expect that success will take many years of work, more than 30 years at best," and that "very substantial research and development costs ... will be incurred as the fusion program passes through the stages of technical development, from demonstration of scientific feasibility to commercialization."

According to Deutch, he and his colleagues are now confident that scientific feasibility (defined as the attainment of a net energy gain from fusion reactions) will be demonstrated in about five years, "based on the program's steady progress over the past few years, including such gratifying results as the recent achievement of high ion temperatures in the PLT," a tokamak-type magnetic-confinement device now in operation at Princeton University's Plasma Physics Laboratory. (The abbreviation stands for Princeton Large Torus.) Presumably the first demonstration of scientific feasibility will be accomplished with the Tokamak Fusion Test Reactor (TFTR), a much larger magnetic-confinement machine currently under construction at the Princeton facility. "Although one cannot yet be quite as confident in predicting particular dates for success in inertial-confinement experiments," Deutch adds, "it is possible that by 1986 scientific feasibility will have been demonstrated by one or more of three systems: the Nova laser at Livermore, the Antares laser at Los Alamos and the EBFA electron beam at Sandia Laboratories."

Following the demonstration of scientific feasibility, the paper goes on, "the program will move from applied research into a development phase," the first step of which will be the construction of "Engineering Test Facilities (ETF) ... for the most promising prospects in both magnetic and inertial confinement." Before deciding which of the competing alternatives in each category will serve as the basis for its Engineering Test Facility, Deutch stresses, "we must be prepared to suspend judgment to guarantee that we have selected the system which will satisfy commercial requirements."

The present expectation is that about eight years will elapse in each program between the ETF decision and initial operation of these facilities. "Thus reasonable dates for operation of the two ETF's are 1992–5 (for magnetic) and 1995–8 (for inertial), assuming funding follows the current projection. That projection keeps the entire program's operating budget roughly constant in 1979 dollars and estimates each ETF at about $600 million. Between now and the time both ETF's are operational in the late '90's, we will have spent about $11 billion.... Experience with the ETF's will teach us what sort of system—magnetic or inertial-confinement—we want to make the basis of the Engineering Power Reactor, and how to build it. That construction will take another eight years and will cost in the neighborhood of a billion dollars. The EPR should be fully operational by 2005, with the decision on the design and funding of the commercial demonstration reactor in progress. At that point the entire fusion program will have expended about $18 billion. The Initial Operating Capability would accordingly be estimated as 2020–2030, a time frame consistent with current predictions of the necessity for large contributions from alternative energy sources."

Coal to Oil to KWH

A major objective of recent energy legislation is to induce electric utilities that burn oil to burn coal instead. Nevertheless, industry projections suggest that oil consumption by the nation's utilities, which amounted to 1.5 million barrels per day in 1976, could well reach 2.4 million barrels per day in 1986, an increase of 60 percent. The difficulty is that most oil-fired generating plants cannot be converted to coal except at prohibitive cost. As a result the utility industry is vitally interested in technologies for converting coal to liquids by hydrogenation. Coal hydrogenation was developed by Friedrich Bergius before World War I and helped to fuel the German air force in World War II.

In a lump of coal the carbon atoms outnumber the hydrogen atoms in the ratio of about 10 to eight. By the addition of hydrogen under pressure to a heated slurry of coal, usually with the aid of a catalyst, the ratio can be shifted so that the hydrogen atoms outnumber the carbon atoms by about 18 to 10, a value typical of hydrocarbons.

Three coal-hydrogenation processes have now been carried through small pilot-scale development by U.S. companies. All yield between 2.5 and three barrels of liquid fuel per ton of coal under pressures of between 2,000 and 3,000 pounds per square inch and at temperatures of between 400 and 460 degrees Celsius. Two of the processes involve a catalyst; one does not.

The noncatalytic process, known as SRC-II, most closely follows the original German process and is the most advanced in physical scale. SRC-II development has been pursued largely by Gulf Mineral Resources Company (a subsidiary of the Gulf Oil Corporation) with funding of about $100 million, pri-

BERLINETTA

A NEW WAY TO TAKE YOUR PULSE. The New Camaro Berlinetta offers a superb balance between the qualities of a sporty road car and the comfort of a fine touring car.

Beneath its classic Camaro styling, Berlinetta also offers a wealth of engineering advances. For example, special insulation surrounds you to help reduce engine, wind, and road noise. A new suspension system, unique to Berlinetta, gives you a smooth, responsive ride with a road-hugging pleasure you would expect from a car that wears the Camaro name.

And if its roadability and ride are special, Berlinetta's interior is equally so. There you'll find custom-tailored cloths in fine cuts and colors. Freshly designed front bucket seats carry you in comfort. Berlinetta's instrument panel—including tachometer, voltmeter, temperature gage and electric clock—is newly designed for quick, easy reading.

Outside, there's a special Berlinetta grille treatment, dual pin striping, blackened rocker panels, color-keyed custom-styled wheels, sport mirrors, and distinctive Berlinetta nameplates, front, side, and rear.

So, for a new way to take your pulse, buy or lease a new Camaro Berlinetta. Then go out and hug a road —in comfort and style.

Aluminum wheels shown are available at extra cost.
Available white-lettered tires are supplied by various manufacturers.

CAMARO. THE HUGGER.

Chevrolet

marily from the Department of Energy. A Department of Energy pilot plant at Tacoma, Wash., which began operation in July, 1977, produces 75 barrels of liquid fuel from 30 tons of coal per day. The department has recently asked Gulf for preliminary designs of a plant to produce 20,000 barrels per day of coal liquids. The plant may be running by 1983.

The second process, H-Coal, has been developed by Hydrocarbon Research, Inc., with recent funding by Ashland Oil, Inc., the Department of Energy, the Electric Power Research Institute (EPRI) and a group of other oil and coal companies. A $100-million H-Coal pilot plant with an output between 750 and 1,800 barrels of liquid fuels per day is being built at Catlettsburg, Ky. It is scheduled to begin operating this summer. Hydrocarbon Research and Ashland Oil believe a 50,000-barrel-per-day plant could be built for $1.1 billion and could be operating by 1985.

The third hydrogenation process, EDS, was developed by the Exxon Research and Engineering Company and has recently drawn financial backing from the Department of Energy, EPRI, several oil companies and Japan Coal Liquefaction Development Co., Ltd. A 700-barrel-per-day pilot plant adjacent to Exxon's refinery in Baytown, Tex., is scheduled for completion early next year. The next step visualized by Exxon is a 60,000-barrel-per-day "pioneer" plant, estimated to cost $1.4 billion, that might be in operation as early as 1985.

EPRI engineers believe that if no serious problems are encountered, a commercial industry capable of producing 450,000 barrels of coal liquids per day could be achieved by 1990 and that output could be more than doubled to 950,000 barrels by 1995. The coal requirement would be about 45 million tons in 1990 and nearly 100 million tons in 1995. For purposes of comparison, in 1976 the nation's electric utilities used 450 million short tons of bituminous coal and lignite.

The future of coal hydrogenation turns critically on costs. EPRI estimates that the potential utility market for coal liquids is 2.5 million barrels per day in the year 2000 if the price can be held to $18 per barrel (in 1975 dollars). (The current price of fuel oil is $13.50 per barrel.) On the other hand, if the price of coal liquids were to turn out to be as much as $24 per barrel, the potential utility market would shrink to 60,000 barrels per day, a volume far too low to support a commercial coal-hydrogenation industry.

Metrication at the Crossroads

In spite of its name the Metric Conversion Act of 1975 does not provide for the adoption of the metric system as the predominant system of measurement in the U.S. The act merely establishes a Federal agency to coordinate and support the efforts of those in business, Government and education who choose to convert to the metric system. According to a recent study conducted by the General Accounting Office, the fact that conversion is voluntary rather than mandatory means that many of the presumed benefits of metrication may never be realized.

Proponents of metrication have always maintained that the metric system is easier to work with because the measuring units come in multiples of 10. In a familiar example the liter, the metric unit of volume, consists of 1,000 milliliters, whereas a quart consists of a clumsy 32 fluid ounces. It would therefore appear that the metric system would make it easier to compare the unit prices of variously packaged items in a supermarket. Such is the case, however, only when the actual dimensions of products are changed to metric dimensions ("hard conversion"), not merely when the customary measurement units are replaced by their metric equivalents ("soft conversion"). In other words, little is accomplished if one quart of a beverage is simply relabeled .95 liter; rather, the size of the container must be changed to one liter, which happens to equal 1.06 quarts. Only in the latter case will the ease of working with multiples of 10 come into play. The goal of hard conversion, the General Accounting Office maintains, may well be unattainable without the imposition of laws and regulations.

Even in industries that have undertaken hard conversion, price comparisons are not necessarily easier to make. The distilled-spirits industry is a case in point. Before the industry began to adopt the metric system 94 percent of its sales were in five customary sizes: 1/2 gallon, quart, 4/5 quart, pint and 1/2 pint. Except for the 4/5 quart these sizes are all integral multiples of one another, enabling consumers to make price comparisons easily. The 200-, 500- and 750-milliliter and the one- and 1.75-liter containers that have come to replace the customary sizes are more difficult to compare in price.

Curiously the average unit prices of the 200-milliliter and 1.75-liter sizes, the ones whose prices are the most difficult to compare, showed the greatest increase over their nonmetric equivalents. The unit price of the 200-milliliter size is 11.4 percent higher than its 1/2-pint equivalent and the unit price of the 1.75-liter size is 6.1 percent higher than its 1/2-gallon equivalent. The General Accounting Office does not predict whether or not these price hikes will persist once the industry is completely metricated. The report does indicate, however, that the metrication of the entire U.S. would cost billions of dollars.

The General Accounting Office study calls on Congress to commit itself either to the traditional U.S. system or to the metric system but not to both. The present course of voluntary metrication will lead to a dual system that is "impractical, inefficient, uneconomical and confusing." The General Accounting Office does not say, however, which of the two systems it prefers, although it belittles many of the presumed advantages of metrication. The U.S. is now the only major nation not using the metric system, and it may still be years before Americans join the rest of the world in realizing that 28.3 grams of prevention are worth .453 kilogram of cure.

The Neutrino Telescope

Neutrino astronomy is a problematical science. As an elementary particle with no electric charge and no apparent mass a neutrino is capable of crossing the universe without being absorbed or even deflected. Hence neutrinos should be visible coming from regions of space where light and other electromagnetic radiation are obscured. For the same reason, however, the neutrino is extraordinarily difficult to observe. The same reluctance to interact with matter that gives the particle its long range makes it almost impossible to capture one of them in the laboratory. Indeed, most of the neutrinos that encounter the earth zip all the way through it unperturbed, leaving no evidence of their passage.

In spite of this difficulty a proposal to build a neutrino telescope has been gaining support for the past three years. The project is called DUMAND, for Deep Underwater Muon and Neutrino Detection, and the telescope would comprise a vast array of detectors near the ocean floor. Such an instrument could serve both astronomers, whose primary interest is the source of the neutrinos, and physicists, who would be able to study the properties of neutrinos at energies substantially higher than those available from particle accelerators.

An essential feature of any neutrino detector is a large mass. Because a neutrino has a vanishingly small probability of colliding with any given atom, many atoms must be put in its path. An experiment begun in the 1960's that may recently have measured the flux of low-energy neutrinos emitted by the sun employed as an antenna some 675 tons of the cleaning fluid tetrachloroethylene. The high-energy neutrinos of astrophysical interest in the DUMAND project come from more distant sources and fewer of them reach the earth (just as starlight is dimmer than sunlight). Measurements of this smaller flux are estimated to require a detector with a mass of about a billion tons. The only suitable material available in that quantity is seawater. The proposal calls for distributing sensors throughout a cubic kilo-

America's hospitals may be stricken by a dangerous epidemic.

meter of the ocean at a depth of about six kilometers.

The telescope would be sensitive only to high-energy neutrinos, and the great majority of even those would pass through it unnoticed. The only neutrinos with a chance of being detected would be those few that happened to interact with an atom in the water, yielding electrically charged particles that recoil with high energy. The charged particles themselves would not be captured, but the apparatus would record the thunder and lightning that mark their passage. The "thunder" is a high-pitched noise generated as the charged particle ionizes atoms along its trajectory. The "lightning" is Čerenkov radiation, the blue light given off by a particle moving through a transparent medium faster than the speed of light in that medium. (These speeds are always less than the speed of light in a vacuum, which no particle can exceed.)

For these two signals two kinds of sensor are proposed. Acoustic signals would be detected by a few large arrays of hydrophones, each array being made up of as many as 10,000 elements. Only a few arrays would be needed because sound has a long range underwater, being limited ultimately only by the finite depth and homogeneity of the ocean. Čerenkov radiation would be collected and the signal amplified by photomultiplier tubes. The range of the light is limited by absorption and scattering in the water and so depends on the water's clarity. It has been found that deep ocean water in some areas is clearer than singly distilled water; even so, light detectors would probably have to be spaced uniformly throughout the volume every few tens of meters. At least 15,000 photomultipliers and perhaps as many as 125,000 might be required.

The acoustic and optical sensors would report by cable to a central computer on shore. From the information they provide a crude image of a particle track could be constructed. Moreover, the energy and the original trajectory of the incoming neutrino could be deduced. Two sites for the DUMAND array are under consideration; both are in deep subsidence basins near the Hawaiian Islands.

There are several possible astrophysical sources of high-energy neutrinos. Some neutrinos must be created in the upper atmosphere when cosmic rays, which are mainly protons, collide with atoms in the air. Among the products of such collisions are short-lived particles whose decay products include neutrinos. Of greater astronomical interest are possible "point" sources, where the presence of energetic particles could give rise to high-energy neutrinos by a similar collisional mechanism. These sites include supernovas, pulsars, quasars and the nuclei of various active galaxies. The telescope might resolve these sources with a precision of better than half a degree. What is more, since neutrinos penetrate both the earth and the atmosphere, the telescope would scan the entire celestial sphere continuously.

The core of our own galaxy might also emit energetic neutrinos. It is a region hidden from view for light astronomy by dense clouds of gas and dust. Another possibility is that neutrinos of very high energy might have been created in the early years of the universe, shortly after the "big bang." By now their effective energy would have been reduced by the expansion of the universe, just as the wavelength of light is shifted toward the red, or lower-energy, end of the spectrum. Such cosmic-background neutrinos would come from those regions of the universe that are most remote from the solar system in time and space.

The energy spectrum of the natural neutrino flux is thought to extend well beyond the highest energies available from present particle accelerators, or from any accelerators expected to be built in the next several years. Accelerator energies are now approaching 10^{12} electron volts, or one teravolt. DUMAND would be most sensitive to neutrinos whose energy extends upward from one teravolt. The rate of neutrino-induced events presumably drops off with increasing energy, but estimates suggest that a cubic kilometer of seawater should yield at least a few interactions per day with energies greater than 10 teravolts and a few per year with energies approaching 100 teravolts. These rates are based on the minimum estimated flux of neutrinos produced in the earth's atmosphere. Neutrinos from other sources could increase the flux substantially.

Events at these energies might give rise to several new phenomena in the physics of elementary particles. One that has been eagerly awaited is the creation of intermediate vector bosons, the particles labeled W^+, W^- and Z^0 that serve as carriers for one of the four basic forces in nature, the weak force. It is because neutrinos respond only to the weak force that they are so little influenced by matter. Thresholds for the production of certain other exotic particles, such as the particles called Higgs bosons, should also be crossed. If there are additional kinds of neutrinos beyond the two known today, they might be observed in the DUMAND experiments, and a change in the pattern of the weak force itself might be discovered. At the energies accessible now the weak force grows stronger as the energy of the interaction increases, but this trend is expected to level off.

The idea of employing the ocean as a neutrino detector was conceived independently by several workers. The DUMAND project itself was established by a small group of investigators, the principal organizers being Frederick Reines of the University of California at Irvine, John G. Learned, who is at Irvine on leave from the University of Wisconsin, and Arthur Roberts of the Fermi National Accelerator Laboratory. Reines, who is chairman of the committee that directs the project, collaborated with Clyde L. Cowan, Jr., in the first experimental detection of neutrinos in 1956. About 100 workers are now affiliated with the DUMAND organization.

A few preliminary studies of the proposed sites and of the physics of the acoustic and optical detectors have been completed. A pilot project intended to investigate the technology of the detectors and to demonstrate the feasibility of the overall project is expected to begin early this year. An important aim of the pilot program is to establish the cost of the full-scale installation. Reines estimates that the cost could not be less than a few tens of millions of dollars, but it depends strongly on factors such as the number of elements in the optical-sensor array. If the project is approved and the money is provided, construction and installation would probably require from five to 10 years.

The Swine Flu Affair

In February, 1976, the Federal Government's Center for Disease Control identified the agent responsible for a small outbreak of respiratory disease among recruits at Fort Dix, N.J., as a particular subtype of the influenza virus: *Hsw*1 *N*1, the swine flu virus. There was fear of a major pandemic. On March 24 President Ford announced an unprecedented nationwide campaign to inoculate every American against swine flu. Congress appropriated $135 million to finance the effort, which was coordinated by health agencies in the Department of Health, Education, and Welfare and conducted by state health departments. After successive delays caused by uncertainty about the proper dose for children, by production difficulties and by a serious controversy about liability insurance, vaccination finally got under way on October 1. In November several cases of a serious neurological side effect of vaccination were reported. The mass-immunization program was suspended on December 16 and was not resumed. The epidemic never came. In February, 1977, the new Secretary of Health, Education, and Welfare, Joseph Califano, commissioned a review of what had become known as "the swine flu fiasco" by Richard E. Neustadt of Harvard University's John F. Kennedy School of Government and Harvey V. Fineberg, M.D., of the Harvard School of Public Health. Their report to Califano, now published by the Department of Health, Education, and Welfare, is a meticulous reconstruction of "decision making on a slippery disease" and a thoughtful examination of the admin-

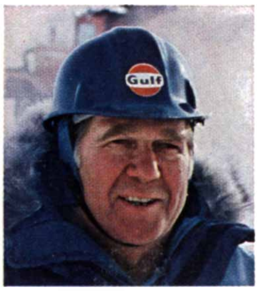
"It's 52° below zero up here, but, crazy as it sounds, we need this machine to keep the ground frozen."

Left to right: Bob Toole, Jacob Kuhoktak, roustabout, and the freezer.

"When you drill for gas north of the Arctic Circle, the way we're doing here at Canada's Mackenzie Delta, you expect a lot of problems," says Bob Toole, drilling superintendent for Gulf.

"But this one's a real fooler. It's a freezing unit. Even at 52° below zero, we need it to keep the ground around the wellhead frozen.

"At temperatures as low as this, you have to invent new ways to do almost everything."



"The problem is that when the drilling mud comes up from the bottom of the hole, maybe a mile down, it's hot enough to melt the permafrost that's holding up the whole rig.

"If the permafrost melted, the hole would get bigger and bigger, and the operation would have to come to a halt.

"Our freezer keeps the top thirty feet of the casing around the drill pipe at temperatures below freezing, so that doesn't happen.

"This country is probably one of the toughest spots on earth to drill for natural gas. But we're drilling the wells. We're meeting the challenge."

Gulf

Gulf people: meeting the challenge.

Gulf Oil Corporation

The same Polavision instant movie system that families are enjoying all across the country could turn out to be one of the most useful business tools ever invented. Consider these pertinent facts: 1. Polavision gives you an *instant*, full color eyewitness report on anything from a complex manufacturing process to a documentary tour of insured property. ☐ 2. The camera is so simple that any employee can use it. (A movie on a technical subject, for instance, can be filmed by the person who understands it best: your technical expert.) 3. The system is so portable, it can be easily taken to any location for filming. (You can have a distant situation brought in for review by your office staff instead of sending your staff to the situation.) ☐ 4. Since the film doesn't have to be sent to a lab for processing, it can be viewed and analyzed immediately. ☐ 5. The optional instant replay feature lets you review important scenes in seconds for in-depth study or to emphasize a sales point. ☐ We'll send you a free booklet that lists 152 ways

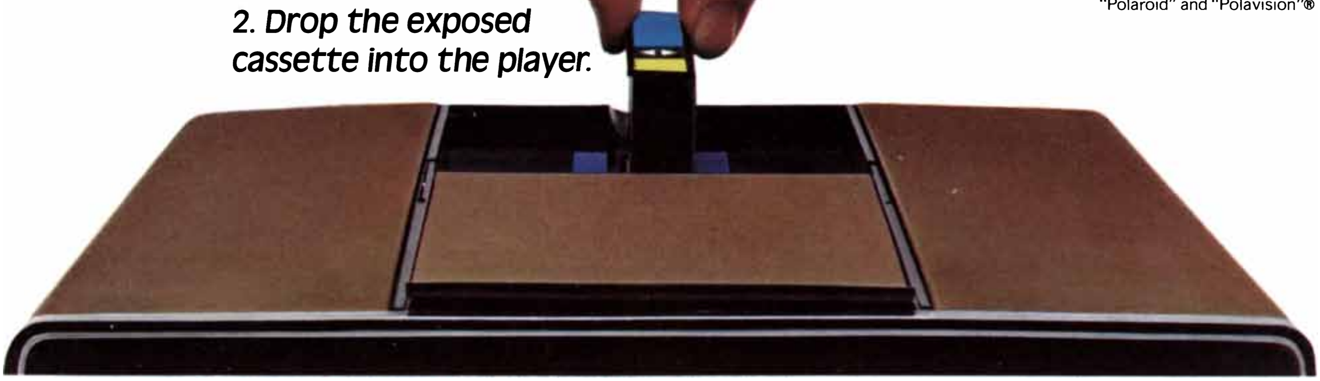
1. You film your sub
with our easy-to-
camera.

Put a little show in your business.

POLAVISION INSTANT MOVIES FROM POLAROID

"Polaroid" and "Polavision"®

2. Drop the exposed cassette into the player.



to sharpen your competitive edge, improve efficiency, and increase profits with instant movies. Then, your local Polavision business applications dealer will call to set up a demonstration and answer your questions. All *you* have to do is write to Polaroid Corporation, Dept. A410, Cambridge, MA 02139. Or call us collect (anywhere in the U.S.): 617-864-1534. ☐ One last fact: The same Polavision System that can make moving documentaries for your business, will make *very* moving documentaries of your whole family.

3. In seconds, your movie appears on the screen. You can even get instant replay at the touch of a button.

Polavision lets you demonstrate your 80-ton earthmover in your prospect's office.

Polavision can help teach your workers the most efficient production techniques.

It can make a documentary tour of work in progress.

Simulated pictures.

And, on weekends, it'll entertain your whole family.


THE FIRST ZX.
DATSUN'S LEGENDARY Z CAR REBORN AS THE NEW 280-ZX.
Behold! Our new ZX—this is not transportation...this is a love affair. This is a car you never let near the car wash. This is a car you tip the parking attendant so you can park it yourself... a car that turns heads as quickly as corners. This is a car you write a new license plate for. This is the sensuous new ZX...so intimate, so thrilling...you'll take the long way home.
The new ZX is a sports car with moves you've never seen. They come from a furious fuel-injected six-cylinder engine; fully independent suspension; power disc brakes all around; a sensitive power-steering system that keeps you in touch with the road (standard on the 2+2 Coupe); and burly radials at all four corners. The performance runs torrid...the quality runs deep. We've fitted doors to frame, buckets to body, carpets to floor with nary a tolerance for error. A superb example of perfection from the worldwide company whose name stands for quality: Nissan Motor Company, Ltd. Buy or lease one at your Datsun dealer.
At its price, the performance, quality and luxury of the new ZX make it a fuel-injected value. It is...AWESOME!
280-ZX
A CREATURE OF COMFORT.
280-ZX STANDARD LUXURY FEATURES. 4-Wheel Power Disc Brakes—Power Steering*—Remote Hatch Opener—Remote Control Electric Outside Mirrors*—Power Windows*—Quartz Sweephand Clock—Ceiling-mounted Swivel Light—Cut Pile Carpeting—Reclining Bucket Seats with Adjustable Lumbar Support (Driver's only)—Aircraft-style Backlit Instrument Panel Gauges—Air Conditioning (Optional on 2 Seater)—AM/FM 4-Speaker Stereo*—Automatic Electric Radio Antenna—Rear Window Wiper*—Computerized Check-out System*—5-Speed with Cruise Control*—
*Standard on Luxury Package.
NOBODY DEMANDS MORE
DATSUN
WE ARE DRIVEN
NISSAN

istrative and technical difficulties that arise wherever science and public policy interact.

The report makes it clear that there was ample reason for concern when swine flu was first identified at Fort Dix. That particular subtype of the influenza virus had been the agent of the pandemic of 1918, which killed 20 million people worldwide and 500,000 in the U.S. Since the late 1920's, however, the strain had confined itself to pigs; no human being under 50 could have built up antibodies to it. The "antigenic shift" now observed meant that what might (or might not) be a virulent human flu virus had acquired a new outer coat of antigenic proteins that might (or might not) make it very contagious to humans. The potentiality for epidemic was clear to the first line of decision makers within the Center for Disease Control. They recommended a major effort not only to produce a vaccine against the new strain but also to administer it. That original decision, generally accepted and reinforced by advisory bodies and at successive levels within the Department of Health, Education, and Welfare, led inexorably to President Ford's March decision. The decision, Neustadt and Fineberg conclude, was not a matter of "party politics" but a valid public-health judgment. Acknowledging the benefits of hindsight, however, they indicate that a better option was insufficiently weighed: stockpiling a vaccine and waiting for evidence of significant spread of the virus before proceeding to mass immunization.

The authors go on to subject each step of decision making and implementation to close analysis. They examine in particular detail the complex issue of liability, which arose when insurance companies refused to cover the vaccine manufacturers against the enormous liability potential of a mass-immunization campaign. That contingency had been foreseen by some Government health officials; it should have been dealt with at the outset. As it was, the program was delayed while Congress passed new legislation specifying that any swine flu claims would be filed against the Government, not the drug companies.

Immunization began on October 1. There had been plenty of publicity; people knew about the program and most were willing to participate; in the first 10 days more than a million were vaccinated. Then, first in Pittsburgh and later elsewhere, there came reports of deaths following vaccination. The deaths were attributed to heart attacks and other coincidental causes, but some states suspended immunization and the public backed off. President Ford and his family got their shots on television and the rate of vaccination picked up for a while, but in December the rate fell off again. The reasons included the death scare, the low level of participation by private physicians and the lack of enough of the specially formulated doses approved for children under 18. Most important, perhaps, "there was no swine flu, or almost none." Aside from one case in Missouri that could not be traced directly to pigs, there was no flu that fall that "the swine flu virus could have caused, or [swine flu] vaccine cured."

The final blow came in December. A rare paralytic disease called Guillain-Barré syndrome was reported in a few people who had been vaccinated. It is still not clear whether the syndrome is a side effect of swine flu vaccination, of any flu vaccination or possibly of any vaccination at all. It was immediately clear, however, that the program had to be suspended, if only to determine the statistical association of the syndrome with vaccination, so that the consent form signed by people receiving the shot could include some assessment of the risk. The health officials put the proposal for suspension to Ford. "He heard them out, sighed and agreed. For most intents and purposes the swine flu program was over."

Neustadt and Fineberg do not allocate blame or even list the "lessons," many of which they feel "leap out of the narrative." They confine themselves to "reflections" on five phenomena bearing on decision making at the level of the Secretary of Health, Education, and Welfare: "program reviews, implementation analyses, media reactions, agency reputations and slippery diseases."

Complicated decisions based on limited data require periodic reevaluation. The necessary step-by-step review requires in turn that at each stage there be "a tracing out of the relationships between deadlines and each decision" and "an explicit statement of assumptions underlying each decision." It may have been necessary, for example, to decide immediately in March, 1976, whether or not to undertake vaccine production; it was not necessary to decide as quickly whether to establish a mass-immunization program. Moreover, assumptions need to be defined and aired if they are ever to be reviewed. One would like to see each issue "posed according to its component parts and argued in probabilistic terms." That would surely be difficult; an alternative would be to ask repeatedly, as one medical consultant put it, "what evidence on which things, when and why, would make us change the course we now propose, and to what?"

As for implementation, it "is not only something to be done after decision, it is as much or more a thing to think about before decision, right along with substance. Of this there was but little in the swine flu case." The authors believe many of the problems in implementing the program could have been anticipated. "More attention to the do-able would almost certainly have altered emphasis and scope." They urge the value of weighing, "*in* the decision, estimates of some sort about difficulties, likelihoods and costs of going wrong."

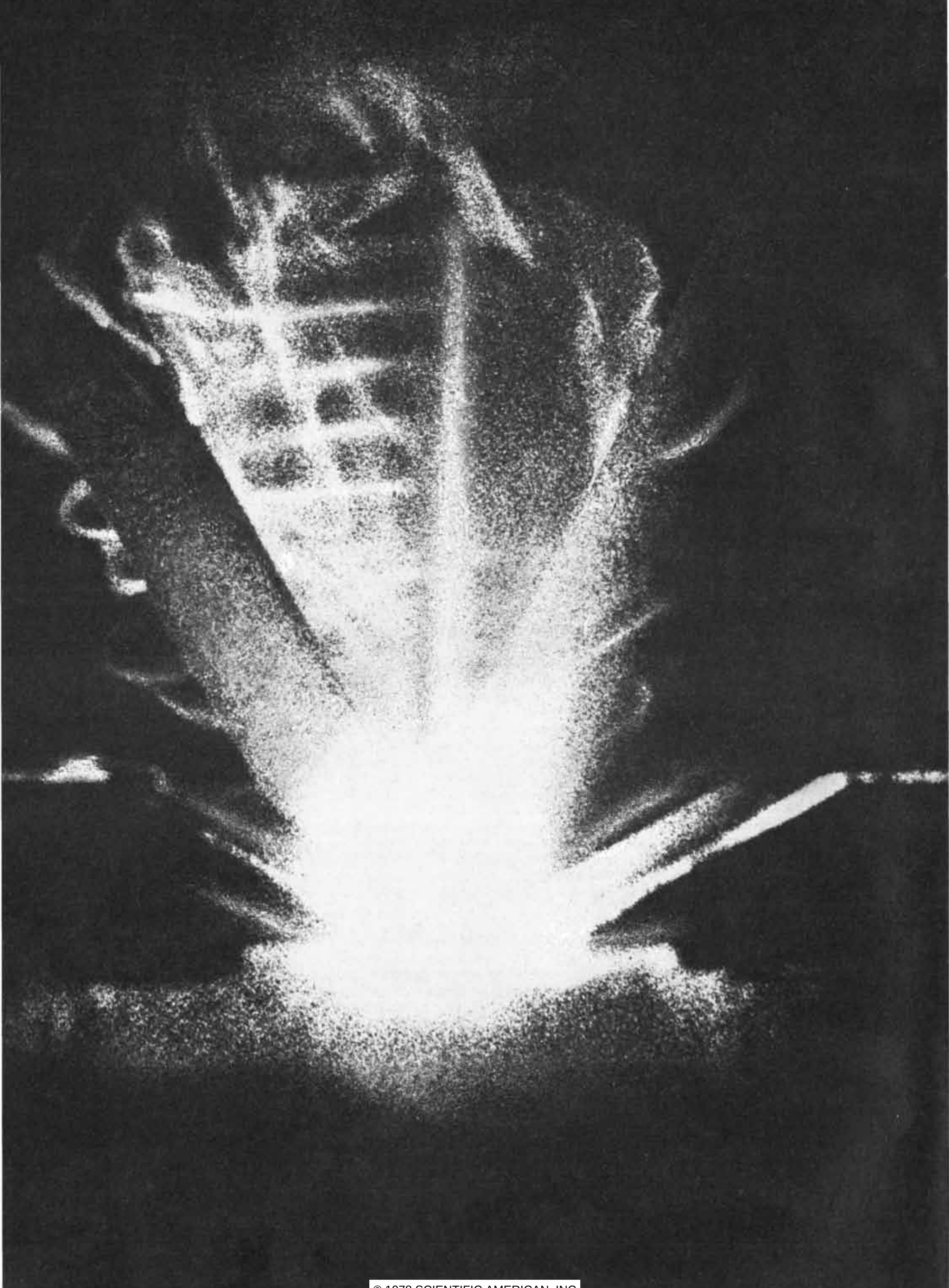
Not enough attention was paid to the realities of news coverage—by which the authors mean not "influencing coverage" but "anticipating it, preparing for it, weighing in the balance of decision both prospective benefits and costs. In a mass program this is crucial to the thinking about doing." It was apparently not anticipated that experts who disagreed with the mass-immunization plan would surface to provide the controversy that "spices life on television news." There was no preparation of the public for such foreseeable incidents as the Pittsburgh deaths. There was no plan for reacting to the appearance of a side effect such as the Guillain-Barré syndrome.

The credibility of the Center for Disease Control was unnecessarily compromised, the authors believe, by the decision of its director, David J. Sencer, to put himself in a "supersalesman's role" as the advocate of mass immunization. A technical expert should confine himself to stating the probabilities, outlining the options to his superiors for their judgment; instead Sencer "pushed his bosses without stint."

Influenza is a "slippery disease," first because its spread and timing are "mortgaged to the process of antigenic change" [see "The Epidemiology of Influenza," by Martin M. Kaplan and Robert G. Webster; SCIENTIFIC AMERICAN, December, 1977]. The effectiveness of a flu vaccine is short-lived. The symptoms of influenza are widely misunderstood, so that laymen "and perhaps half the doctors in the country" apply the term incorrectly to various gastrointestinal disturbances. The flu virus may not even be the major cause of aches and fever originating in a respiratory infection. Even a successful flu vaccine may therefore seem ineffective. Finally, the year-to-year impact of the influenza virus is hard to estimate. For all these reasons influenza is far more "slippery" than the viral diseases hitherto successfully attacked by mass immunization, such as smallpox, poliomyelitis and measles. To be sure, "where risks are high and countermeasures readily available," even a slippery disease might properly be combated by a Federal immunization program, but only if the risk is a risk of death, if the preventive is effective for the people who are most at risk and if it can substantially increase their survival. In 1976, the authors believe, "in the absence of *manifest* danger, all-out action was a mistake."

Onward and Upward

It is hard to determine whether or not a large integer is a prime number, that is, a number with no divisors other than 1 and itself. Any number can be tested



for primality by dividing it by all smaller primes, but as the size of the number to be tested increases implementing this procedure on even the fastest computer soon becomes prohibitively time-consuming. There are certain types of numbers, however, for which more efficient algorithms of proving primality have been developed. For example, in 1971 Bryant Tuckerman of the Thomas J. Watson Research Center of the International Business Machines Corporation proved that the number $2^{19,937} - 1$ is prime. This 6,002-digit number was larger than any other known prime, and empirical evidence suggested that the next prime that could be identified by a similar computing technique might be much larger. Nevertheless, in 1975 two 15-year-old high school students in Hayward, Calif., set out to find a new largest prime. Three years later they succeeded: in October of last year Laura Nickel and Curt Noll proved that the 6,533-digit number $2^{21,701} - 1$ is prime.

Prime numbers that can be expressed in the form $2^p - 1$, where p is a prime integer, are called Mersenne primes, after the 17th-century mathematician who studied them. Over the past century most of the large primes discovered have been Mersenne primes because there is an efficient algorithm for determining whether any number $2^p - 1$ is prime. (Since $2^n - 1$ can only be prime if n is prime, it is not necessary to consider numbers of this form with composite exponents.) The algorithm, called the Lucas-Lehmer test, was devised by the French mathematician Édouard Lucas in 1876 and then improved on by D. H. Lehmer of the University of California at Berkeley in 1930. To test $2^p - 1$ for primality in this way, a sequence of numbers is generated, beginning with 4, in which each term (u_i) is equal to the square of the preceding term minus 2 ($u_{i-1}^2 - 2$), reduced modulo $2^p - 1$. If the pth term of this sequence is zero, then $2^p - 1$ is prime; otherwise it is a composite number.

The number of computational steps required to carry out this procedure in a straightforward way is proportional to the cube of the exponent p. The number of steps required to prove a number prime by the method of dividing by smaller primes is proportional to the square root of the number itself, since division by the primes up to the square root of the number will disclose all its factors. As p increases, the cube of p grows much more slowly than the square root of $2^p - 1$. In fact, for even moderately large values of p the Lucas-Lehmer test is substantially more efficient, making the discovery of large Mersenne primes possible.

When the exponent p is close to say 20,000, applying the Lucas-Lehmer test presents a complicated programming problem, involving the multiplication and subtraction of numbers with thousands of digits. (It is convenient, however, that in binary notation $2^p - 1$ is a string of p 1's.) Nickel and Noll, working at the computer center of the California State University extension at Hayward, wrote five different versions of their computer program before they began testing numbers last September. Tuckerman had tested all the prime exponents up to 21,000, and so it was at that point that Nickel and Noll began. Some 350 hours of computer time later the program identified 21,701 as the exponent of a Mersenne prime.

The discovery of this new largest prime is of particular interest. First of all, since it has never been proved that there are an infinite number of Mersenne primes, the discovery of a new one is always significant. Moreover, it has been observed that on occasion Mersenne primes occur relatively close together in pairs known as twins, for example $2^{2,203} - 1$ and $2^{2,281} - 1$. The new largest prime is quite close to the preceding one. If the two are twins, then the next Mersenne prime may well be at the limit of current computing capability. On the other hand, the two may not be twins, and $2^{21,701} - 1$ may have a larger twin close by. To clarify these matters Nickel and Noll are now revising their computer program for the testing of numbers with even larger exponents and hope to find yet another largest prime.

Why O_2 electrodes limit fuel cell voltage.

What suppresses fuel cell output more than anything else? The oxygen electrode. Mysteriously, it refuses to develop the electrical potential that thermodynamics says is possible.

In the H_2-O_2 fuel cell, for example, theory promises 1.23 volts. Yet, reality produces only 1.06.

Many explanations have been proposed to account for this anomaly. But none have withstood scientific scrutiny. Here at the General Motors Research Laboratories, however, we've devised a hypothesis that we're convinced will hold up. It stems from our fundamental studies of the electrochemistry of oxygen.

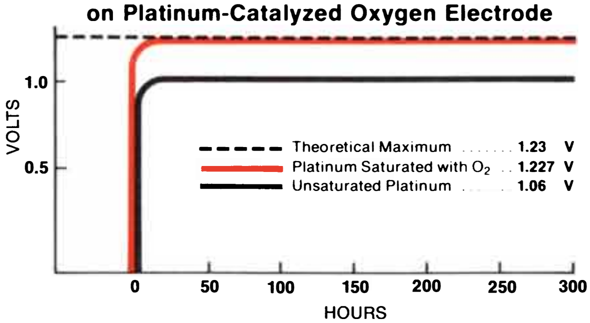
Swelling (area within circular impression) caused by oxygen dissolving into platinum.

Our reasoning: The O_2 electrode is catalyzed with platinum, whose surface adsorbs oxygen. But the adsorbed oxygen continuously dissolves into the Pt lattice, limiting adsorption to about 30% of the Pt surface. Consequently, minute corrosion cells form in the remaining areas, and their combined potential bucks the 1.23 volts.

To test the logic, we charged a pure platinum diaphragm with oxygen. The oxygen did indeed adsorb on the surface and dissolve into the metal, as x-ray analysis and the expanded diaphragm (photograph) later showed.

When it became saturated, the diaphragm could no longer dissolve any oxygen. This allowed adsorption to spread over the entire Pt surface, thus preventing corrosion. Now unopposed, the potential between the platinum and a reference electrode climbed to 1.227 volts, the highest ever reported for oxide-free platinum.

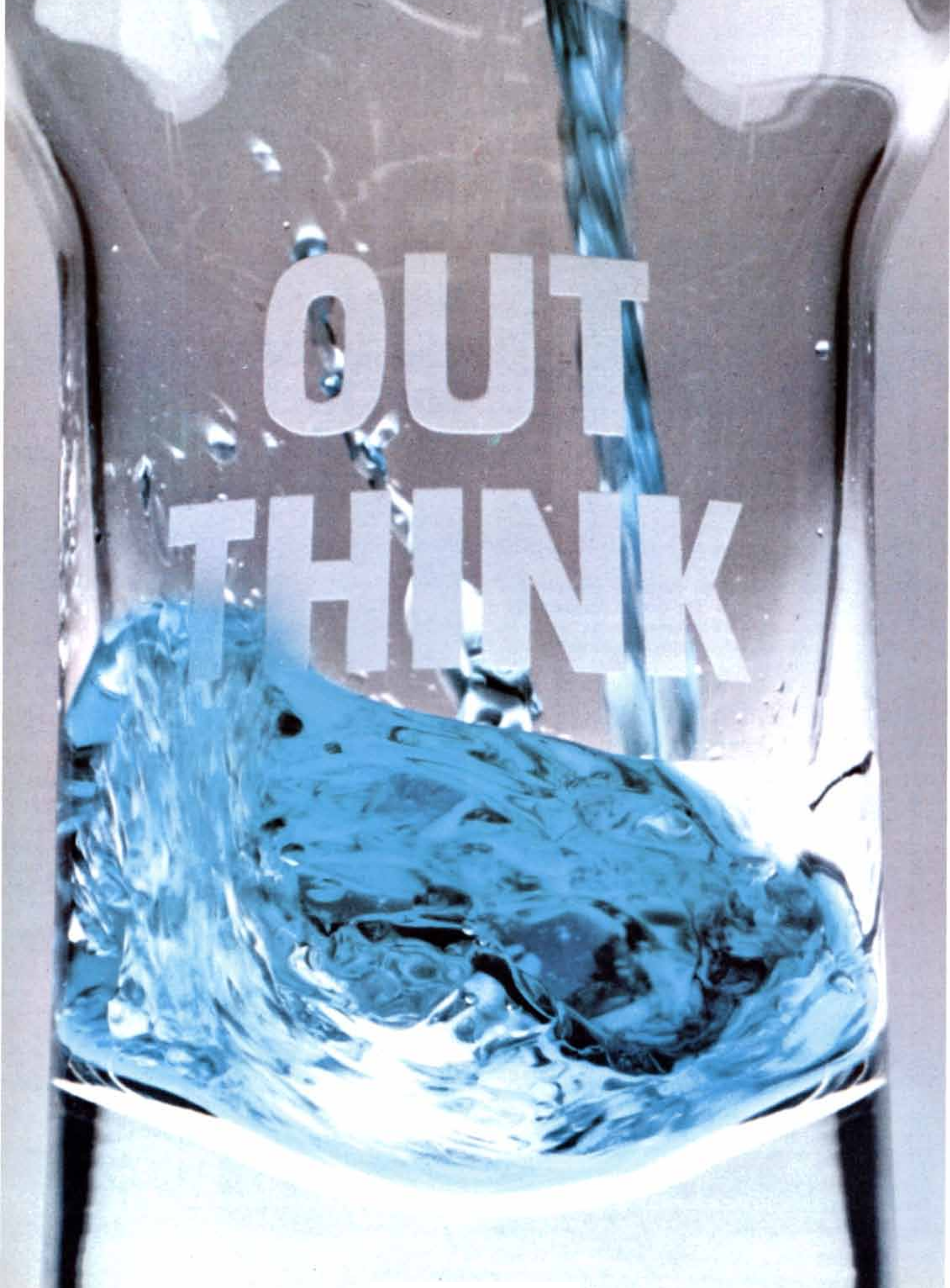
Effect of Oxygen Saturation on Platinum-Catalyzed Oxygen Electrode



Granted, understanding the problem doesn't necessarily solve it. But . . . first things first.

We currently have openings for Ph.D.s in engineering or the physical, mathematical, or biomedical sciences. If interested, please send your resume to: GMR Personnel, Dept. 110. An Equal Opportunity Employer.

General Motors Research Laboratories
Warren, Michigan 48090


OUT
THINK

Liquids, the most adaptable of all matter, shape themselves exactly to fit their containers.

In much the same way, a Datapoint ARC™ system adapts itself to fit the needs of a business.

In this day when competitive edges are frequently slight and often difficult to come by, you have to out-think your competition to gain an advantage. The Attached Resource Computer system can help by putting powerful data processing capability at the hands of those who need it most, wherever they are in the company.

Not just another network

An Attached Resource Computer system is an arbitrary number of small, powerful, inexpensive computers, each performing its assigned tasks independently, yet all are totally interconnected to share common resources and draw from a common database. For high reliability, the system is composed of proven Datapoint processors and peripherals.

One of the ARC system's strengths is its specialization of components. One or more processor in each system is dedicated to file processing and management. The others are applications processors. Internal communications are achieved by an Interprocessor Bus, using dedicated Resource Interface Modules (RIMs).

This capability frees a computer from juggling all three functions. Consequently, processing can be carried out at much higher throughput. Even though the processors are functionally dispersed, the system is a single, unified computing facility, with the processing speed and capability of a large, conventional computer.

Solves growth problems

This new, modular architecture offers easy-to-implement solutions to the problems of system planning and growth. The system manager can add (or delete) processors or file storage as workloads change.

Expansion can occur in modular, incremental steps, at predictable, affordable costs. Only enough processing power is required at a given time to accomplish the immediate workload. Important capital is not tied up in costly overcapacity, yet the work is not delayed by a bogged-down, overworked computer.

Protects software investment

The ARC system can use your present Datapoint system software. When additions are made to the system, current software is not affected. All the Datapoint resources suitable for participation in an ARC system are completely compatible.

While direct hardware cost comparisons are difficult because of differences in architecture, our studies indicate that the cost of an ARC system may be only 50% to 70% of the cost of a more conventional system with equivalent capabilities.

Total adaptability

The applications at work in an ARC system can be as varied as the diverse functions of the business environment. The concept applies to any type of firm engaged in one or multiple endeavors. Its usefulness is equal whether it's in a corporate home office, a division headquarters, or a remote office. Or all three can simultaneously employ their own ARC or other Datapoint systems and interconnect them through telecommunications links.

A nationwide support network

There are more than 500 Datapoint Customer Service Representatives in the field, stationed in over 130 local service area offices. The system is designed to place customer service and support as near as possible to Datapoint's 11,000 installations.

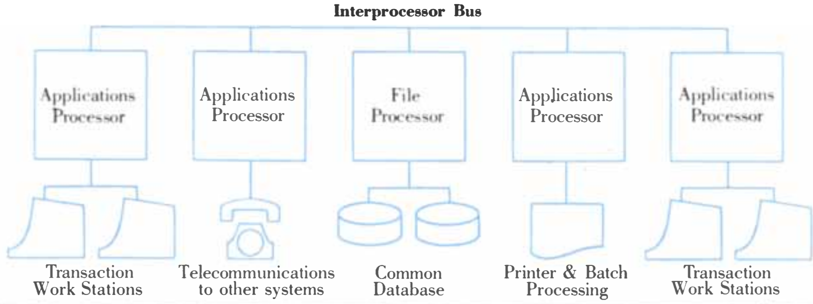
Write for more information

A Datapoint Attached Resource Computer system is the ideal solution to your business data processing needs. Datapoint has to out-think its competition. The ARC system puts us another big jump ahead.

A booklet describing the features and advantages of the ARC system is available by writing to Datapoint Corporation, Marketing Communications Dept. (M62), 9725 Datapoint Drive, San Antonio, TX 78284. Or call the Datapoint sales office nearest you.

ONE ARC™ SYSTEM!

One of an almost limitless number of configurations, this ARC system contains up to 480K of on-line user memory, 200 MB of disk storage, and 48 transaction work stations, as well as facilities for batch processing, printing and telecommunications.



◀ *The adaptability of an Attached Resource Computer™ system is another example of how Datapoint out-thinks its competition to help you out-think yours.*

DATAPOINT CORPORATION
The leader in dispersed data processing™

The Deep Structure of the Continents

The oldest parts of the continents appear to have deep root zones that travel along with the continents as the tectonic plates move. The zones may be chemically distinct from the surrounding rock

by Thomas H. Jordan

Within the past two decades earth scientists have assembled vast amounts of geological, geophysical and geochemical evidence to develop a new picture of the history, structure and dynamic behavior of the outer layers of the earth, informed and rationalized by the theory of plate tectonics. The theory has been particularly successful in describing how the basaltic oceanic crust, some seven kilometers thick, is continually produced at the crests of mid-ocean ridges by the rise of molten magma from the mantle below, how it moves across the surface at the rate of centimeters per year as the top layer of huge, rigid plates 10 times as thick and how it is consumed back into the mantle in subduction zones at the deep ocean trenches. The theory has been less successful in explaining the formation and structure of the continents.

The continental crust, which averages 35 kilometers in thickness, is lighter than the oceanic crust and richer in silicon and potassium. In plate-tectonic theory the continental crust is considered to be a buoyant product of melting and remelting that has accreted in the course of sea-floor cycling over long periods of time; continental drift, the well-documented movement of the continents across the earth's surface, has been explained simply as the passive rafting of this light crust as part of a moving plate. Because the continental crust is buoyant it cannot be mixed back into the mantle in large amounts by tectonic processes such as subduction; it tends instead to float on the denser mantle somewhat like a slag on molten iron. Except for this distinction, plate tectonics postulates no essential differences between the motions of the continents and of the ocean floor, or between the subcrustal structures of the continental plates and of the oceanic plates.

Seismological data, however, reveal substantial contrasts between continental and oceanic structures extending well below the base of the crust. It appears that under the oldest parts of the continents there are deep root zones several hundreds of kilometers thick, which travel along coherently with the continents as the plates move. The discovery of these root zones has challenged some of the basic tenets of plate-tectonic theory. The emerging picture of the deep structure of the continents is providing new insights into the mechanical and chemical processes that control continental evolution and tectonics.

Cratonic Stability

The oldest rocks of the continental crust are found in the basement complex of the continental shields and the continental platforms. The shields are extensive uplifted areas that are essentially bare of recent sedimentary deposits. The platforms are broad, shallow depressions of the basement complex filled by nearly flat-lying sedimentary rocks. Together the shields and platforms constitute the cratons, the stable blocks that are the nuclei of present-day continental masses. Although most of the basement rocks exposed on the shields or buried under platform sediments have been metamorphosed during ancient episodes of orogeny, or mountain building, they have remained undisturbed for very long periods, typically a billion years or more.

The continents comprise more than their stable nuclear cratons, however. In contrast to the cratons are the continents' modern orogenic zones, huge areas that have been pervasively deformed quite recently by tectonic activity resulting from the convergence of two opposing plates. The crust in an active orogenic zone does not behave rigidly; plate tectonics fails to describe its motions. The contrast between cratons and orogenic zones is seen today in southern Asia, which is being deformed by the collision of the northward-driving Indian and Arabian cratons with the Eurasian continent. This violent collision has crumpled Asia to produce a great mountain belt extending from the Anatolian peninsula across the Middle East, the southern U.S.S.R. and most of China to the marginal seas of the Pacific Ocean [see "The Collision between India and Eurasia," by Peter Molnar and Paul Tapponnier; SCIENTIFIC AMERICAN, April, 1977].

A remarkable aspect of this ongoing orogenic event is that the Indian craton continues to drive headlong into Asia at the rate of five centimeters per year, upthrusting high mountains over an area of 10 million square kilometers, without itself deforming. To understand the dynamics of this perplexing tectonic behavior, to account for the violent deformation of Asia and the stolid persistence of an undeformed Indian subcontinent, is a challenging problem for geophysics. Any valid model of continental structure and tectonic behavior must account for this contrast, and it must be based on information from the depths of the mantle, where the forces that have shaped the continents originate.

Continental tectonic behavior is largely controlled by the mechanical structure of the lithosphere, which is functionally defined as the earth's strong outer shell, composed of the crust and the upper part of the mantle. The lithosphere endows the plates with their rigidity. Studies of the earth's gravitational field and of the earth's response to the rapid loading and unloading of its surface by large masses such as glacial ice sheets and mountains suggest that the lithosphere has an average effective thickness of about 100 kilometers. Below it is the asthenosphere, or weak layer, a region of the mantle where even small stresses cause material to flow. According to the plate-tectonic model the lithosphere constitutes the plates and the plates move with respect to the asthenosphere, within which the shear strains associated with plate motion are concentrated.

The strength of the lithosphere is derived largely from the upper layer of the mantle, and the strength of the mantle material varies with temperature. Measurements of thermal gradients near the surface show that temperatures within the lithosphere increase rapidly with depth and that the rates of increase vary widely with geographical location. Laboratory studies and evidence from the field indicate that at a given temperature crustal rocks, ranging in composition from basalts to granites, are significantly less strong than peridotites, the major rocks of the mantle. Because the oceanic crust is thin, mantle peridotites are present at shallow depths, where the temperatures are low. It is apparently this near-surface layer of cold peridotite that gives the oceanic lithosphere most of its strength to resist tectonic deformation. The continental crust, in contrast, is thick, and so the temperatures at its base are higher than those at the base of the oceanic crust. The uppermost mantle under the continents is therefore weaker than its oceanic counterpart, and that weakness seems to explain, at least in part, why the orogenic zones are so easily deformed.

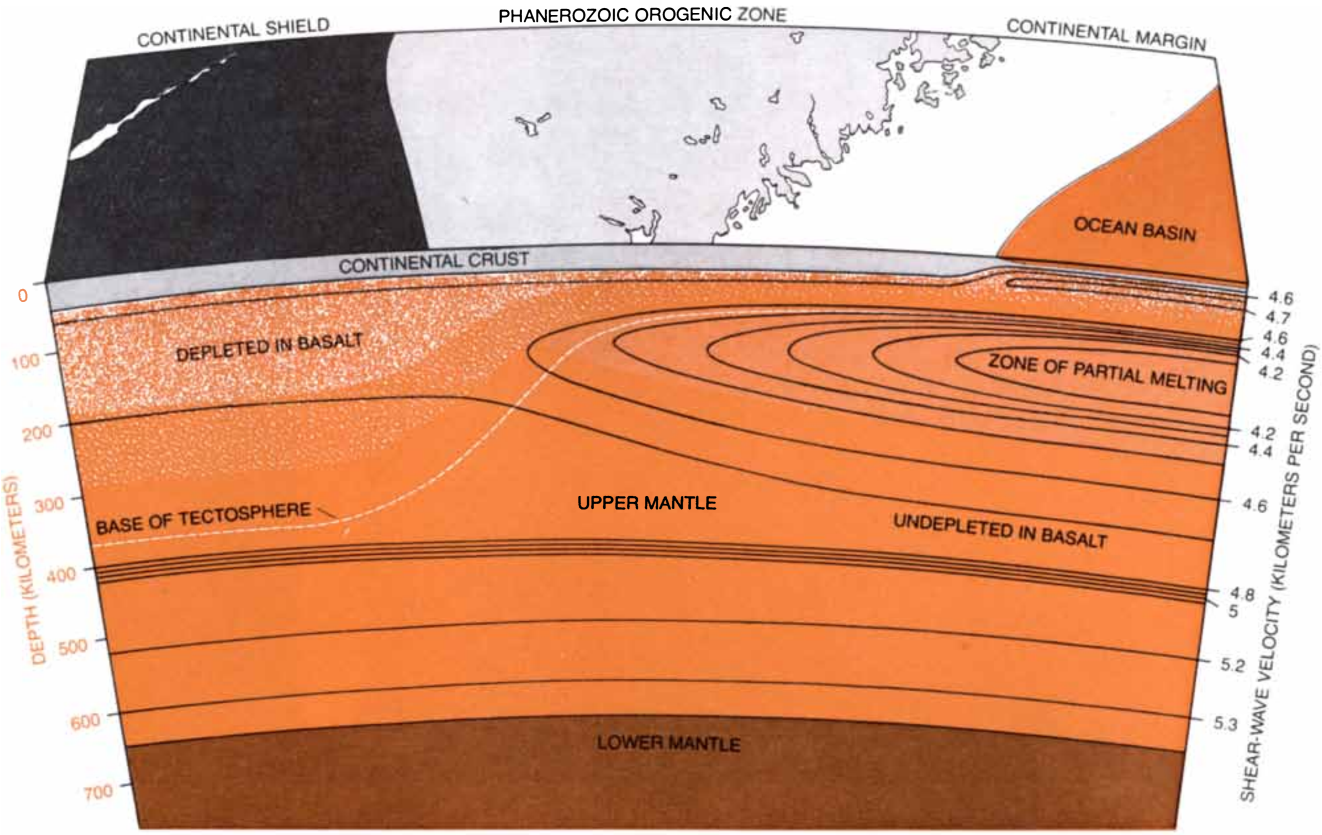
If that is the case, why have the cratons remained stable and undeformed over such great spans of geological history? The answer is closely linked to the problem of deep continental structure. One can argue that although the temperatures at the base of the continental crust generally exceed those at the base of the oceanic crust, the thermal gradient is almost always lower (the temperature increases more slowly with depth) in the cratonic mantle: the temperatures characteristic of flowing asthenosphere are not reached under the cratons until depths greater than those at which they are reached under the oceans. Therefore the integrated temperature of the upper mantle—and hence the effective thickness of the lithosphere—might be greater under the cratons than under the oceans or the orogenic zones. Geophysicists have presumed that the increased lithospheric thickness caused by lower temperatures at greater depths explains cratonic stability.

As we shall see, it is not that simple. Recent work indicates that in addition to the thermal contrasts there are chemical differences within the cratonic roots of the upper mantle that are also important in regulating continental tectonics. Understanding the new and still controversial model of the continental upper mantle that incorporates these compositional differences requires a detailed exploration of the deep structure of the continents.

The Earth's Elastic Structure

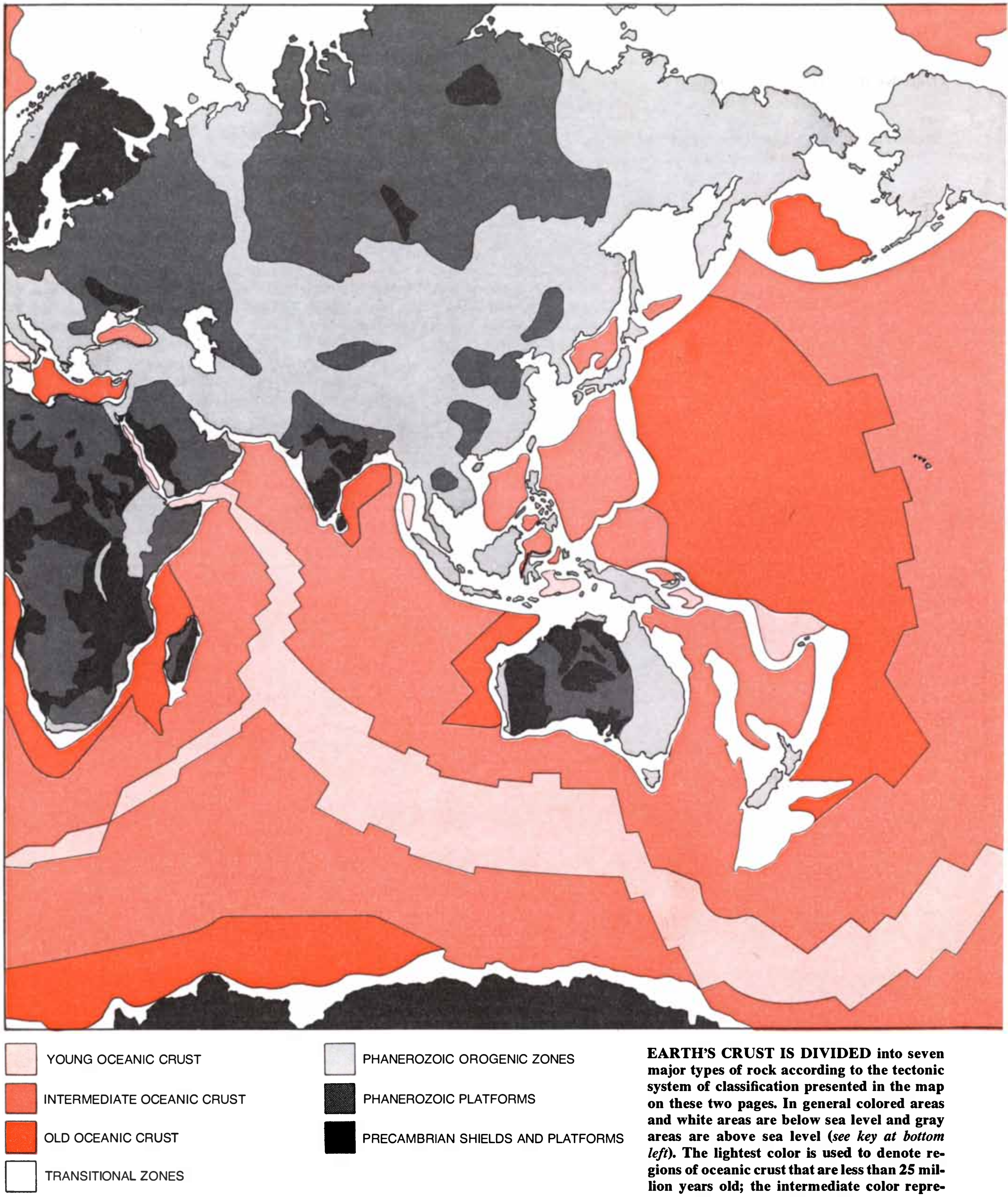
Most of what is known of structure within the earth's interior has been revealed by seismology. When powerful sources of elastic waves—earthquakes and large explosions—"illuminate" the interior, the images recorded by seismographs indicate spatial variations in the earth's elastic properties. The elastic response to a seismic source usually is adequately described by three parameters: the density, the speed of compressional waves (whose motion is polarized along the wave path) and the speed of shear waves (whose motion is polarized transverse to the path). Each elastic parameter is a function of pressure, temperature and chemical composition.

For most seismological purposes the earth can be assumed to be a spherically symmetric structure whose pressure, temperature and composition (and hence elastic parameters) vary only with distance from its center. That is the assumption, for example, in the computer algorithms that locate large earthquakes by fitting the observed arrival times of



HYPOTHETICAL MODEL of the subsurface transition zone under a typical continental margin shows some of the features that might characterize the earth's upper mantle along a section extending, say, from the Canadian shield (*left*) to the Atlantic Ocean basin (*right*), crossing the New England coast in the vicinity of Portland, Me. The large lateral variations in seismic shear-wave velocities observed in recent measurements are indicated by the black contours. According to the author's basalt-depletion hypothesis, the earth's upper mantle is composed of a mineral assemblage called peridotite (*solid color*), which under the continents is depleted in certain basaltic components; the density of the white dots signifies the degree of depletion. The broken white line represents the approximate minimum depth to the base of the tectosphere, term adopted by the author to define the volume of tectonic plates. Zone of partial melting lies under oceanic crust.

seismic waves. Once an earthquake has been located in space and time, the seismic-wave travel time to any station can be computed. If the earth were really spherically symmetric, that is, if its elastic properties varied only with radius, these travel times would depend only on the great-circle distance between source and receiver and not on their specific coordinates. In fact, however, small but significant geographical variations are observed in the travel times to various stations, typically with magnitudes of less than 1 percent of the total travel time. Analysis of these small regional differences indicates that structures associated with the cratons persist down to at least several hundred kilometers, that is, to depths an order of magnitude greater than the thickness of the continental crust.



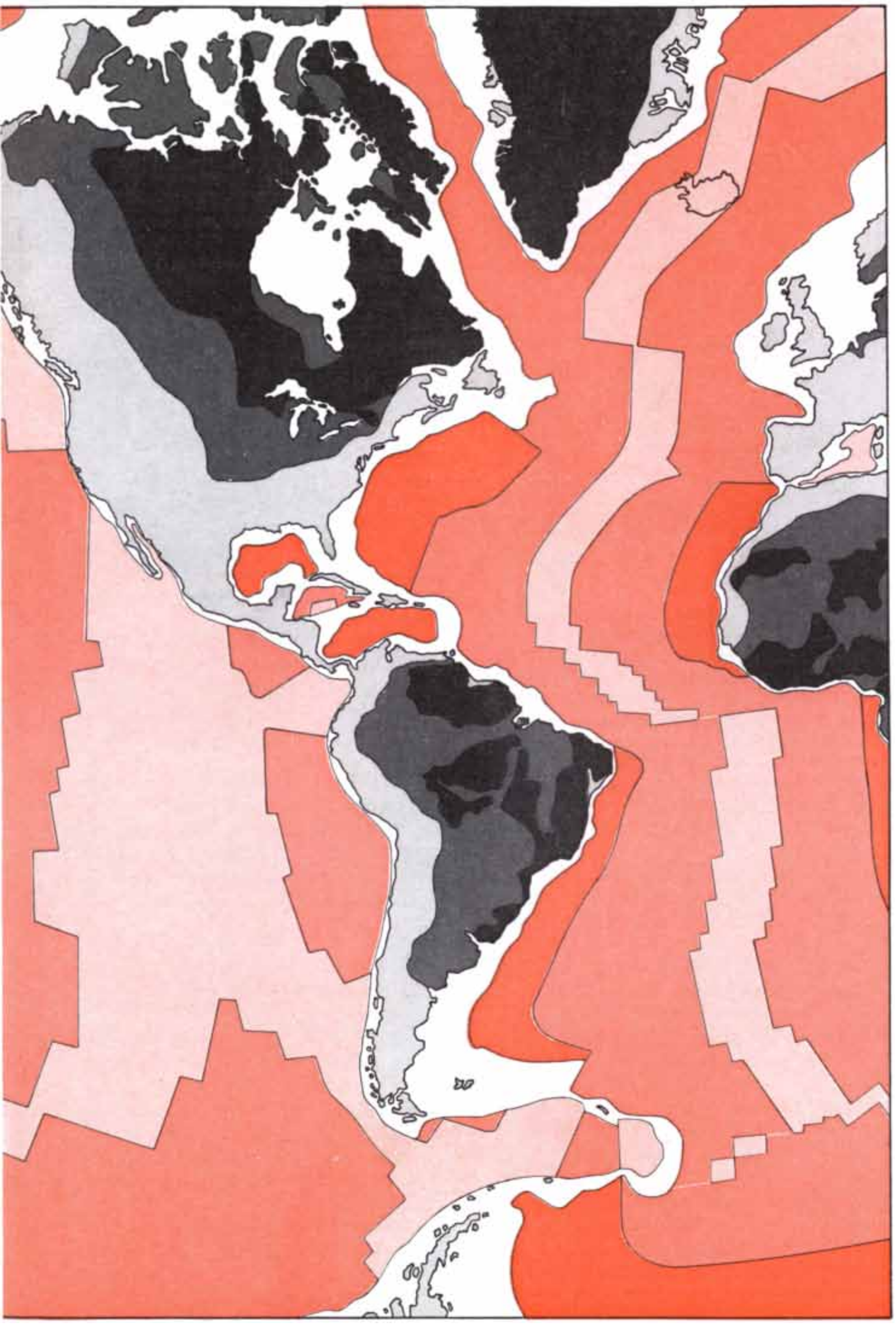
EARTH'S CRUST IS DIVIDED into seven major types of rock according to the tectonic system of classification presented in the map on these two pages. In general colored areas and white areas are below sea level and gray areas are above sea level (*see key at bottom left*). The lightest color is used to denote regions of oceanic crust that are less than 25 million years old; the intermediate color repre-

The seismological study of continent-ocean heterogeneity is hampered by the dearth of seismic recording stations in the ocean basins. To investigate the upper mantle in regions that lack seismic stations seismologists generally depend on surface waves, whose energy is trapped near the surface and whose travel times depend on the properties of the crust and upper mantle along the entire path from source to receiver. The propagation speeds of the surface waves' low-frequency components are more sensitive to deep structure than those of the higher frequencies, and from measurements of this frequency dependence the variations of the upper mantle's elastic properties can be estimated for both continental and oceanic paths. It was surface-wave data collected in the 1920's by the pioneering seismologists Beno Gutenberg and Robert Stoneley that showed the crust under the oceans to be much thinner than the crust under the continents, but systematic study of large-scale lateral variations below the crust did not begin until the late 1950's and early 1960's, when the first global networks of standardized broad-bandwidth seismometers were installed.

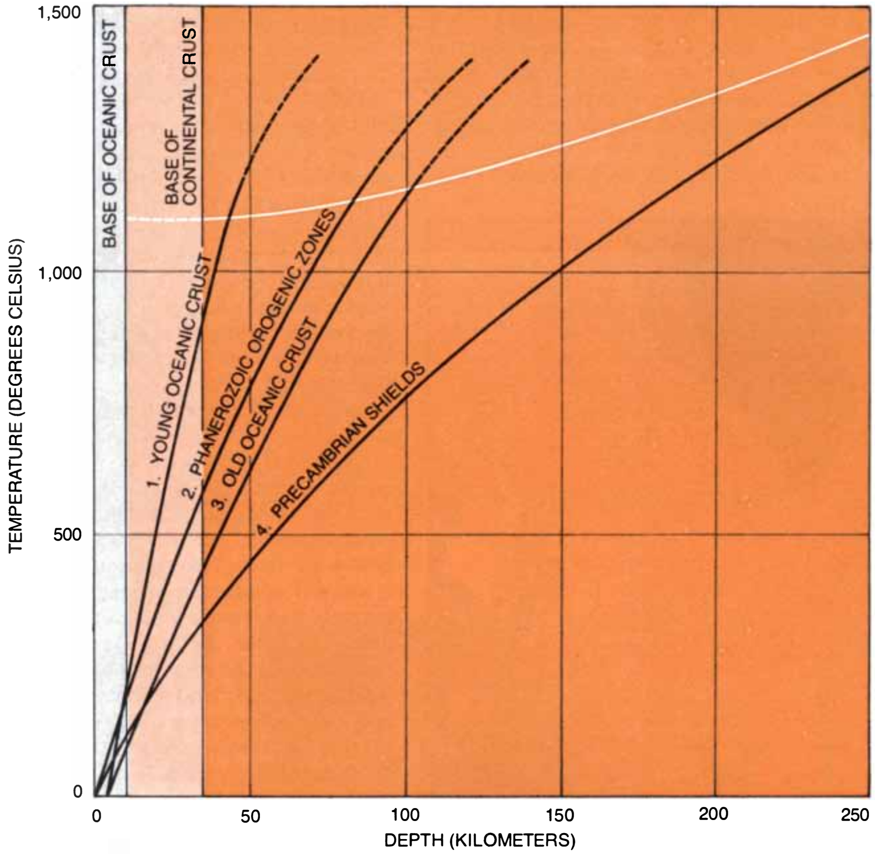
Surface-wave data from these networks quickly convinced seismologists that the structural differences between continents and oceans must extend far below the crust. Throughout most of the crust and mantle the elastic parameters (density and wave speeds) increase with depth because the pressure increases and squeezes the rock into tighter, more rigid structures. Under the ocean basins and most active orogenic zones, however, the shear-wave speed was found to decrease sharply with depth in a transition region that begins about 50 or 100 kilometers below the surface and forms a "low-velocity zone" approximately 100 kilometers thick. Under the cratons, on the other hand, this low-velocity zone is absent or lies deeper and is less prominent.

These seismological findings accorded well with the accepted models of the mantle's thermal structure. Wave speeds generally vary inversely with temperature; a small amount of melting results in a dramatic decrease in velocity. Laboratory studies and theoretical considerations indicated that the upper mantle's low-velocity zone is probably caused by a small amount (1 percent or less) of melting in mantle peridotites owing to the high temperatures about 100 kilometers below the oceans and the orogenic zones. Because the thermal gradients under the cratons are more gradual, the peridotites there melt only at greater depths, if at all.

This structural picture of the upper mantle, which emerged in the early 1960's, also made considerable sense in terms of the rapidly developing theories of sea-floor spreading and plate tectonics. Seismologists were quick to identify the easily deformed asthenosphere with the partially molten material in the low-velocity zone and to identify the rigid lithosphere with the cooler material above the low-velocity zone. It appeared that by mapping the depth to the low-velocity zone seismologists were actually mapping the geographical variations in plate thickness. A corollary followed from the model, however: Any structural variations below the top of the low-velocity zone should not be coherent with the positions of the conti-



sents regions of oceanic crust that are between 25 and 100 million years old, and the darkest color corresponds to regions of oceanic crust that are more than 100 million years old. The white areas are regions of transitional or submerged continental crust, including the continental margins, the island arcs and the oceanic plateaus. The lightest gray denotes regions of continental crust that have been affected by orogenic (mountain-building) activity in the Phanerozoic era, that is, within the past 600 million years or so. The intermediate gray represents stable continental platforms with a history of sedimentary deposition in the Phanerozoic era, and the darkest gray corresponds to stable continental shields and platforms with no accumulated sediments since the Precambrian era, which came to an end approximately 600 million years ago.



TEMPERATURE of the crust and the upper mantle varies systematically according to the tectonic classification of the crustal rock. The temperature rises most rapidly with depth under the young oceanic regions (*curve 1*) and least rapidly under the Precambrian continental shields (*curve 4*). These variations in the temperature-depth relation appear to be responsible for the observed geographical differences in the velocities of seismic waves. Under most tectonic regions the temperature curves rise rapidly enough to intersect the threshold at which the mantle begins to melt (*white curve*), causing the elasticity of the mantle to decrease suddenly at a certain depth and thereby creating a seismic low-velocity zone. Under the continental shields, however, the temperature curve rises more gradually and melting may not occur. This conclusion is consistent with the view that under shields seismic low-velocity zone may be absent.

nents; plate motions should continually be rearranging the relative locations by moving the lithospheric material with respect to the asthenosphere.

The Depth Issue

From that point of view the seismic evidence for consistently higher shear-wave speeds under the cratons at depths greater than 200 kilometers, below the lithosphere and below the depth of the oceanic low-velocity zone, was disturbing. In 1963 Gordon J. F. MacDonald of the University of California at Los Angeles reviewed the problem of deep continental structure. His models based on the surface-wave data and on information about the mantle's thermal state implied that the lower temperatures and higher wave speeds characteristic of the cratons at shallow depths must extend to depths on the order of 500 kilometers. MacDonald was convinced that such deep structure was incompatible with the notion of a mobile, convecting mantle, and so he concluded that the relatively rapid motions of the continents required by sea-floor spreading and continental drift were improbable, if not impossible.

MacDonald's arguments against drift could not, however, withstand the tide of new information that swept across the earth sciences in the ensuing decade, persuading all but the most skeptical that plate motions are indeed real. In time the data on which MacDonald had based his thesis were reinterpreted by advocates of this "new global tectonics." In 1970 J. G. Sclater and Jean Francheteau, working at the Scripps Institution of Oceanography, formulated thermal models of the oceanic and continental lithospheres that were consistent with plate-tectonic hypotheses. They showed that the differences between continental and oceanic thermal profiles could be confined above a depth of 200 kilometers, rather than extending to 500 kilometers or more as MacDonald had argued, and they proposed a model with an oceanic lithosphere approximately 100 kilometers thick and a 200-kilometer continental lithosphere. Their model received support from Adam M. Dziewonski of Harvard University in 1971. He demonstrated that even if the surface-wave data did allow structural differences between continents and oceans to exist at the great depths advocated by MacDonald and others, the data certainly did not require such variations. In fact, the data could be satisfied by models where the significant structural contrasts were concentrated above 200 kilometers, a value consistent with the conclusions of Sclater and Francheteau and compatible with the general idea of the lithospheric plates. The evidence for MacDonald's deep continental roots appeared to have evaporated.

Some seismologists remained dissatisfied with this picture of upper-mantle structure, however. I. Selwyn Sacks of the Carnegie Institution of Washington analyzed data from very deep earthquakes on the west coast of South America and concluded that the thickness of the plate constituting the South American craton must exceed 300 kilometers. A similar conclusion was reached by Shelton Alexander of Pennsylvania State University, whose analysis of surface-wave dispersion across stable continental blocks led him to argue that structural contrasts between the cratons and ocean basins extend to a depth of at least 400 kilometers. Dziewonski had shown, however, that surface-wave data could not conclusively distinguish structures at these depths, and so most geophysicists remained skeptical.

The evidence that finally confirmed the idea that continental structures are significantly thicker than 200 kilometers came not from the surface-wave experiments but from observations of the more unusual *ScS* waves. These shear waves travel nearly vertically downward from the source (an earthquake) and are reflected from the sharp discontinuity between the solid mantle and the liquid core.

The phase that is denoted ScS_1 (or just *ScS*) is reflected once from the core-mantle boundary to the receiver; the phase ScS_2 is reflected twice from the core and once from the free surface, and so on for the higher-order phases. *ScS* phases have many properties that make them useful for the study of upper-mantle heterogeneity. For example, the travel-time difference between ScS_2 and *ScS* is sensitive to variations in shear-wave speed in the upper mantle under the ScS_2 surface-reflection point, but not to variations near the source and receiver, where ScS_2 and *ScS* follow almost the same paths. By varying the recording site (and thus the location of the surface-reflection points of the *ScS* phases) one can therefore detect lateral heterogeneity in regions where seismic stations are either sparse or absent, such as the deep ocean basins.

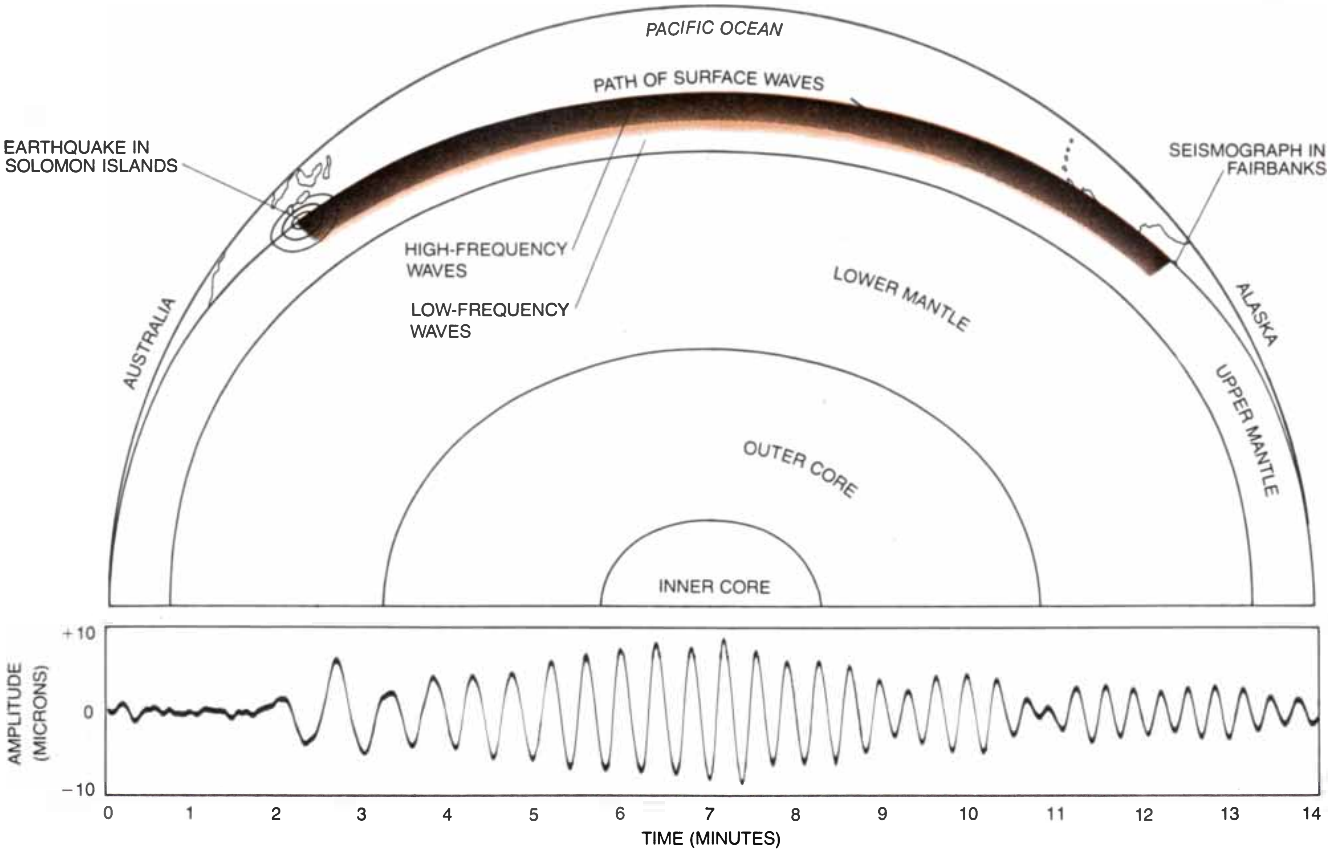
In 1975 and 1976, first at Princeton University and then at the Scripps Institution of Oceanography, Stuart A. Sipkin and I published two studies of global velocity variations in the upper mantle based on *ScS* phases. We found that the average travel times of shear waves moving vertically through the mantle and crust under the ocean basins are about four seconds greater than the corresponding travel times for the cratons. Since the oceanic low-velocity zone was known from the surface-wave data to be more accentuated than its continental counterpart, the mere fact that the shear waves moved more slowly under the ocean came as no surprise. It was the large magnitude of the observed difference that was unexpected, because the existing models of structural contrasts between the continents and the ocean predicted much smaller differences; for example, the vertical shear-wave travel times that were computed from Dziewonski's surface-wave models of cratonic and oceanic structures differed by less than one second rather than by four seconds.

These large *ScS* travel-time variations, combined with the surface-wave data, require that significant structural contrasts between oceans and cratons persist to depths that certainly exceed 200 kilometers and probably exceed 400. The surface waves are most sensitive to the variations in elastic parameters near the surface, whereas the *ScS* phases sample the upper mantle almost uniformly at all depths. If one attempts to construct cratonic and oceanic models in which the large contrasts in shear-wave speed demanded by the *ScS* data are concentrated above 200 kilometers, then the surface-wave dispersion curves calculated from the models invariably conflict with the dispersion curves that are actually observed. Indeed, it can be shown with a fair amount of rigor that no reasonable model subject to this depth constraint adequately satisfies both the *ScS* data and the surface-wave data. It appears that only models with significant variations at or below a depth of 400 kilometers can successfully explain both sets of data. Such models can be constructed [*see illustration on page 93*], but they are not unique. Experiments aimed at refining the models are in progress in my laboratory and elsewhere, and more definitive results should be forthcoming.

Continental Tectosphere

If at this stage one cannot specify with much certainty how the contrasts between continental and oceanic elastic structures vary with depth or even to what maximum depth they extend, it is nonetheless clear that contrasts in shear-wave speed persist considerably below a depth of 200 kilometers. This fact alone forces one to reconsider a concept that has been central to the plate-tectonic model: the notion that the lithosphere everywhere constitutes the plates.

Lithospheric material is by definition strong; it can support the shearing forces induced by surface loads such as mountains for millions of years without much permanent deformation. The determination of lithospheric thickness from observations of loading phenomena and the gravity field was first undertaken by geophysicists in the early part of this century, and modern refinement and reanalysis of the basic data have confirmed one of their original conclu-



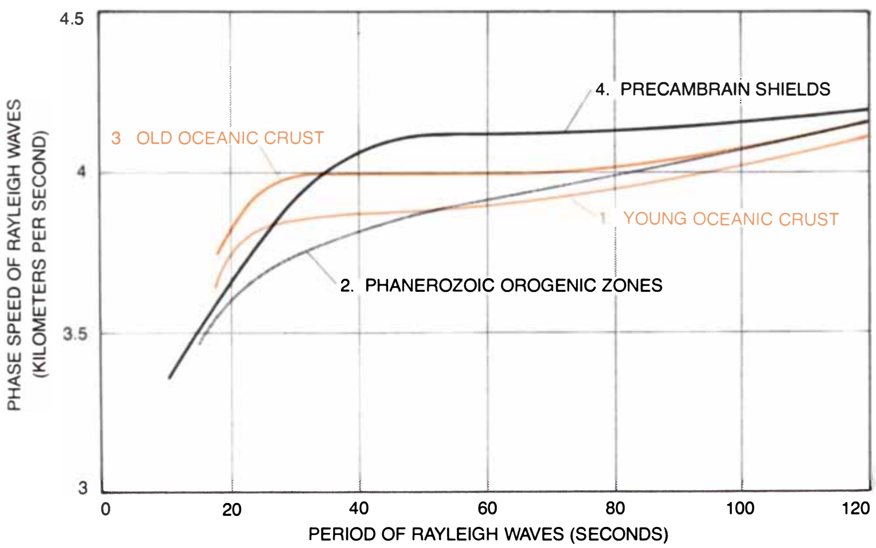
SURFACE WAVES, seismic waves whose energy is trapped near the earth's surface, provide seismologists with a great deal of information about variations in the elastic structure of the upper mantle. The idealized drawing at the top, for example, shows the path of a surface-wave train recorded at a station in Fairbanks, Alaska, following an earthquake in the Solomon Islands in June, 1970. The portion of the actual seismogram reproduced at the bottom shows a characteristic trace made by a Rayleigh wave, a type of seismic wave whose motion is elliptical in the plane of propagation. (Only the vertical component of the wave's motion is represented here.) Such surface waves are said to be dispersive, that is, their speeds of propagation depend on their frequency. Lower-frequency components (*color*) are more sensitive to structural variations at greater depths in the mantle than the higher-frequency components (*gray*), and they generally travel faster.

sions: Under even the most stable cratons such as the Fennoscandian and Canadian shields the effective thickness of the lithosphere is less than 200 kilometers and probably not much more than 100. In fact, the average thickness of the cratonic lithosphere seems to be not much greater than that of old oceanic lithosphere; for example, R. I. Walcott of the Canadian Department of Energy, Mines and Resources has estimated that the lithosphere in Canada, presumably representative of the stable continental areas, is about 110 kilometers thick, whereas the lithospheric thickness under Hawaii, and probably elsewhere in the older oceanic basins, is greater than 75 kilometers. The material below these depths is part of the asthenosphere; when it is acted on by even small shearing forces, it quickly deforms.

The seismic data, in other words, imply that the subcratonic regions of anomalously high shear-wave speeds extend below the lithosphere and into the asthenosphere. As a matter of fact the deep lateral variations in elastic parameters turn out to be remarkably coherent with all the crustal-age classifications shown in the illustration on pages 94 and 95. Two explanations are possible: either the continental deep structures characterized by high shear-wave speeds are more or less statically coupled to the overlying continents—are actually parts of the continental plates—or they represent thermal variations that are somehow dynamically maintained by flows of mass in a convecting mantle. For the latter mechanism to be viable essentially all cratons would have to be the sites of a downward flow of cold material (because low temperatures are necessary to explain the high shear-wave speeds); such flows would have obvious effects on the earth's gravity field, however, and no such effects are observed. Moreover, a mass-flow hypothesis imposes severe geometrical restrictions on the flow field that appear to be inconsistent with other constraints on mantle dynamics. It seems, therefore, that deep continental structures cannot be explained as features that are dynamically maintained by convection.

One is therefore forced to conclude that these deep structures do indeed constitute the lower portions of the continental plates and that they have been translating coherently with the crust for hundreds or even thousands of millions of years, in spite of the fact that the material forming these deep structures is part of the asthenosphere. Faced with this situation some geophysicists have preferred simply to redefine the lithosphere to include these deep structures. The inconsistent usage can lead to confusion, and so I have advocated instead the use of the term tectosphere to denote the region occupied by the coherent entities we call plates, retaining for the lithosphere its classical definition as the layer of significant strength. (This usage is not without precedent: in two of the original papers outlining the plate-tectonic theory, Walter M. Elsasser and W. Jason Morgan of Princeton applied the word tectosphere in this context.)

The tectosphere is defined by its kinematic behavior, or purely by its motions; the lithosphere and asthenosphere, on the other hand, are defined by their dynamic behavior, that is, by the way they respond to imposed forces. Under the oceans the tectosphere and the lithosphere are, for most practical purposes, identical in spatial extent, attaining maximum thicknesses of about 100 kilometers in the oldest ocean basins. Under the continents, however, the tectosphere and the lithosphere are not the same: the cratonic tectosphere extends below the lithosphere, perhaps to depths of 400 kilometers or more.



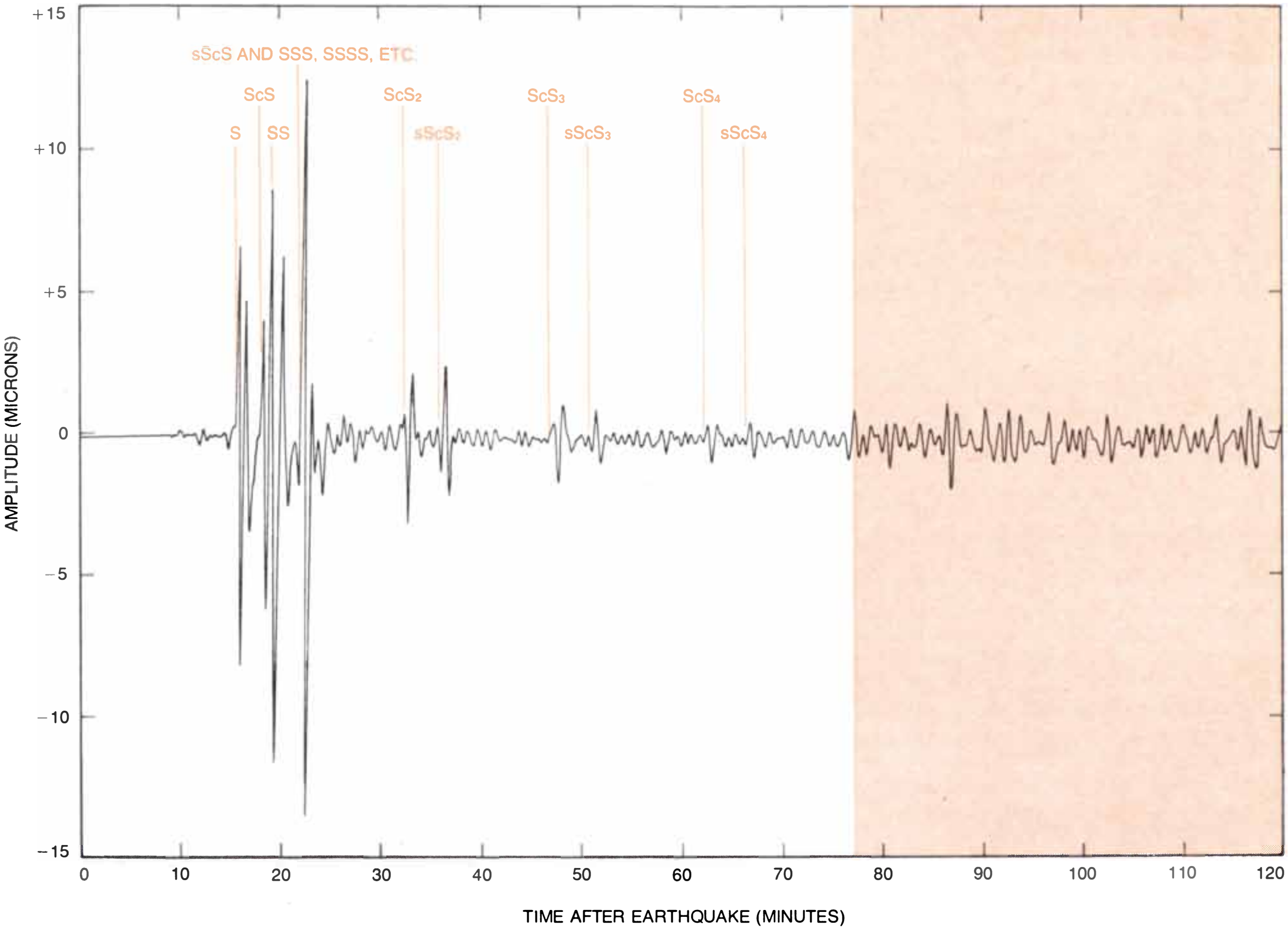
FREQUENCY DEPENDENCE of surface-wave speeds can be used to estimate the variation of the mantle's elasticity with depth. The typical dispersion curves plotted here correspond to regions of crust with different tectonic classifications. In the case of low-frequency surface waves (that is, those with periods longer than about 40 seconds) the highest speeds are observed for waves propagating across the Precambrian continental shields (*curve 4*), whereas the lowest speeds are observed for waves propagating across comparatively young oceanic crust (*curve 1*).

Continental Thermal Evolution

The variations in tectospheric thickness inferred from the surface-wave and *ScS* data correlate rather well with estimates of the heat flux coming from below the crust, which suggests that the thickness of the tectosphere is controlled by the temperature structure of the mantle. Tectospheric thickness also correlates with crustal age: generally speaking, the thinnest tectosphere underlies the youngest oceanic crust, whereas the thickest tectosphere underlies the oldest continental crust. In the oceans the plate-tectonic model has provided a simple explanation for these correlations. As new sea floor spreads laterally away from its site of creation at the crest of an oceanic ridge it loses heat by conduction to the surface, and so the temperatures within a surface "thermal boundary layer" decrease with time. Consequently the thickness of the boundary layer increases with the age of the crust, and the heat flux from its surface decreases. Modeling studies indicate that the thickness of this oceanic thermal boundary layer is essentially the same as that of the oceanic tectosphere (and lithosphere); it averages about 70 to 100 kilometers.

The continents also show a systematic decrease in surface heat flow and increase in tectospheric thickness with "crustal age," provided the age is defined not as the time since the original formation of the crust but as the time since the last major orogenic event. This behavior has encouraged some geophysicists to extend the thermal-boundary-layer hypothesis to the continents. According to this hypothesis, the subcontinental mantle is strongly heated during major orogenies and subsequently cools by conduction of heat to the surface. The cratons are therefore explained as regions that have been cooling for a very long time and where the thermal boundary layer has grown to extreme thickness. The times for this thermal decay are actually about right: the thickness of a thermal boundary layer that has cooled for one or two billion years should be nearly 400 kilometers, which could explain the seismic data.

In spite of this agreement and its attraction as a unified theory of tectospheric thermal evolution, there are se-



VARIETY OF SHEAR WAVES, seismic waves whose motion is polarized at right angles to the wave paths, are represented in this complex seismographic trace, recorded at a station on the Hawaiian island of Oahu following a deep-focus earthquake near the South Pacific island of Tonga in October, 1974. The various types of transverse waves responsible for the pulses in the seismogram, labeled by the colored symbols at the top, are explained in the illustration on the next page. In general waves that arrived at the station more than about 77 minutes after the earthquake (*colored area*) traveled the long great-circle route around the earth between Tonga and Oahu.

rious difficulties in applying the thermal-boundary-layer hypothesis to the continents. The most obvious problem concerns predictions made by the principle of isostasy. Because no significant shear stresses can be statically supported by the asthenosphere, the pressures exerted on the asthenosphere must be equalized at some minimum depth in the upper mantle called the level of compensation. To a good approximation isostasy requires that the mass contained in columns of equal cross section of ocean, crust and upper mantle be equal; in effect, the sum of the products of the layers' thicknesses and densities must be the same.

Under the oceans as the mantle cools and the thermal boundary layer grows the density within the boundary layer increases. To maintain isostasy the crust and upper mantle subside and the depth from sea level to the ocean floor increases. Hence mantle rock is displaced below the level of compensation and replaced by water, which is much less dense, allowing the mass columns to remain in balance. This mechanism explains quite accurately the shapes of the mid-ocean ridges and the location of the centers of sea-floor spreading at the crest of the ridges: the mantle under newly formed crust is hot and light; it cools and subsides as it moves away from the spreading centers.

On the continents, however, the history of vertical movements does not agree nearly so well with the predictions of the thermal-boundary-layer hypothesis. As the continental crust and mantle cool following an orogenic heating event they should also subside, and this subsidence should be accompanied and recorded by the deposit of sediments. (Otherwise all the cratons would be under water!) Although large accumulations of sediments are observed on some continental margins and in certain continental basins, the well-documented history of the cratons is quite different. For example, the growth of a thermal boundary layer 300 kilometers thick—probably the minimum necessary to explain the seismic data for the cratons—predicts crustal subsidence, and thickening by sedimentation, of nearly 20 kilometers. The cratons, however, are actually characterized by their notable lack of sedimentary cover; huge tracts of basement rock are exposed in the shields, and even the platforms are rarely covered by more than a few kilometers of sediments. Furthermore, cratonic crusts are no thicker than younger continental crusts. It is quite apparent, therefore, that the thermal-boundary-layer hypothesis fails to explain the existence and development of deep continental structure.

What explanation can be offered in its stead? In 1975 I proposed that the thermal evolution of the continental tectosphere is intimately tied to its chemical evolution and that the deep continental structures implied by seismology are the visible expressions of compositional as well as thermal variations in the upper mantle. I was led to this hypothesis by the necessity of constructing a model that meets a particular constraint: a model of a subcontinental tectosphere that is stable over long periods of time in a mantle that must be presumed to be convectively unstable.

Let me explain the reasoning. Below the lithosphere one condition for long-term stability is hydrostatic equilibrium: a rest state in which surfaces of constant pressure and surfaces of constant density coincide and are horizontal. The reason is that in the asthenosphere—below 100 kilometers or so—even small departures from this ideal rest state would cause material to flow. Yet, as I have indicated, the high shear-wave speeds characteristic of deep continental structures seem to require anomalously low temperatures under the continents, and hence lateral thermal gradients in the mantle under the oceans and under the continents, at depths well in excess of 100 kilometers. In a chemically homogeneous mantle such temperature gradients would induce lateral density gradients; surfaces of constant density and surfaces of constant pressure would not coincide. The consequence would be convective instability, and the resulting flow of material would quickly disrupt and eventually destroy the sublithospheric portions of the tectosphere. The deep continental tectosphere is not in fact disrupted and destroyed, however; it is notable instead for its long-term stability.

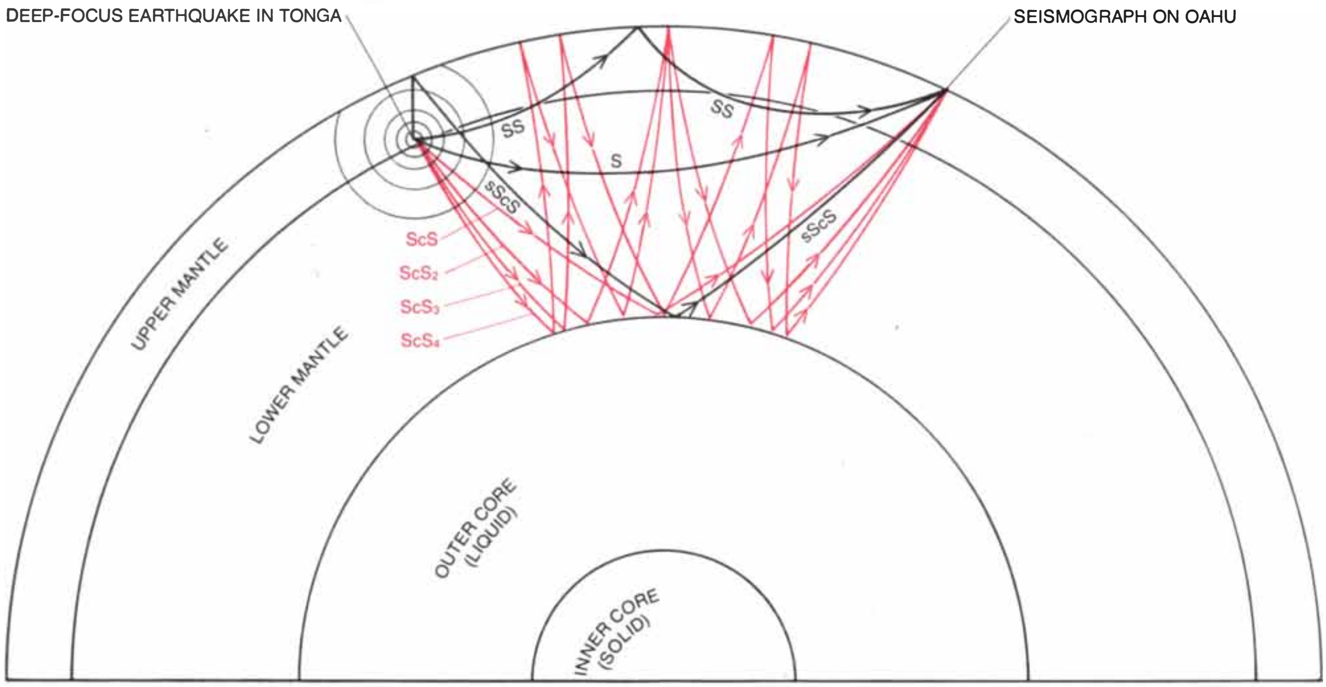
I proposed, therefore, that the mantle is not chemically homogeneous, and that the lateral thermal gradients associated with continental deep structures are stabilized against convective disruption by differences in composition. The compositional gradients were assumed to be dynamically adjusted in such a way that the tectosphere extending below the lithosphere is in approximate hydrostatic equilibrium with the warmer mantle surrounding it. That is to say, if the mantle were chemically homogeneous, there would be excess mass under the continents owing to the higher densities caused by lower temperatures; according to the hypothesis of inhomogeneous composition, these mass excesses are locally compensated by chemically controlled deficiencies in mass. In other words, the mineralogical assemblages within the thick cratonic root zones are actually equal in density to those at the same level under the oceans because the subcratonic assemblages are cooler; they would be less dense if the continental and oceanic temperatures were the same.

Basalt Depletion

The deduction that continental deep structures are compositionally distinct from their surrounding mantle was based purely on geophysical reasoning, without specific geochemical considerations. Further investigation, however, quickly revealed a simple and geochemically attractive mechanism for generating compositional heterogeneities: the density of the continental tectosphere could be lowered by the removal of a basaltic component from the mantle.

Basalt is the most common magma erupted on the earth's surface: about 20 cubic kilometers of basaltic magma is removed from the mantle every year and added to the crust. Most of this volcanism occurs in the oceans along the ridge-crest spreading centers or at "hot spots" such as Hawaii, but rocks of basaltic composition are also important constituents of the continental crust. When basalt is crystallized near the surface, it is composed primarily of the minerals plagioclase (a calcium-sodium-aluminum silicate) and clinopyroxene (a dark-colored, chemically variable silicate containing iron, magnesium, calcium and sodium). At subcrustal depths, however, basalt crystallizes into a mineral assemblage called eclogite, which also contains clinopyroxene but which has the dense mineral garnet as its aluminous phase instead of plagioclase.

Until 20 years ago some geochemists thought the upper mantle might be composed entirely of eclogite, but now the



PATHS OF SHEAR WAVES represented in the seismogram on the preceding page are shown in this diagram, along with their conventional seismological designations. The paths labeled *S* and *SS* denote transverse waves that travel through the mantle from source to receiver without reachingt he earth's core. The waves that travel along the path labeled *ScS* are reflected once from the sharp discontinuity that separates the solid mantle from the liquid core before they return to the earth's surface. The waves that travel along the paths labeled ScS_2, ScS_3 and so on are reflected more than once from the mantle-core boundary and are also reflected from the surface. The multiple *ScS* waves are particularly useful for studying the structure of the upper mantle along their paths between source and receiver. Unlike surface waves, shear waves of this type are sensitive to variations in the elasticity of the mantle at great depths. Accordingly they provide critical information about the depth to which certain structural differences between the continental crust and the oceanic crust extend.

Let it snow, let it snow, let it snow...

Button up your overcoat and join us making fun of winter...at the carnivals, follies, sleigh rides, ice slides and polar bear dips – all across Ontario. Take to the woods, on hundreds of miles of pristine cross-country trails, where the world is white, the silence golden. Or head for the hills and test your stuff on some of the longest and best maintained runs east of the rockies. Après ski, come in from the cold...our first-class resorts offer saunas, pools, firelit lounges, gluhwein and good times, and we treat you royally. This year you'll enjoy a substantial premium on every dollar – and the removal of Ontario's 7% room tax is an added plus. For information call us COLLECT (416) 965-4008, or write: Ontario Travel, Dept. L.S., Queen's Park, Toronto, M7A 2E5.

Ontario Canada

Harris technology on the job

Harris 1680 distributed data processing system at General Foods

In information systems, Harris produces data processing terminals, general-purpose computers, word processing systems and supervisory control systems.

Multi-function distributed data processing systems from Harris provide remote access to central computers *plus* on-site computing power, helping business and government handle growing data processing workloads.

Harris technology works worldwide—in communication equipment, information systems, government systems, semiconductors and printing equipment. For information, write: Harris Corporation, Melbourne, Florida 32919.

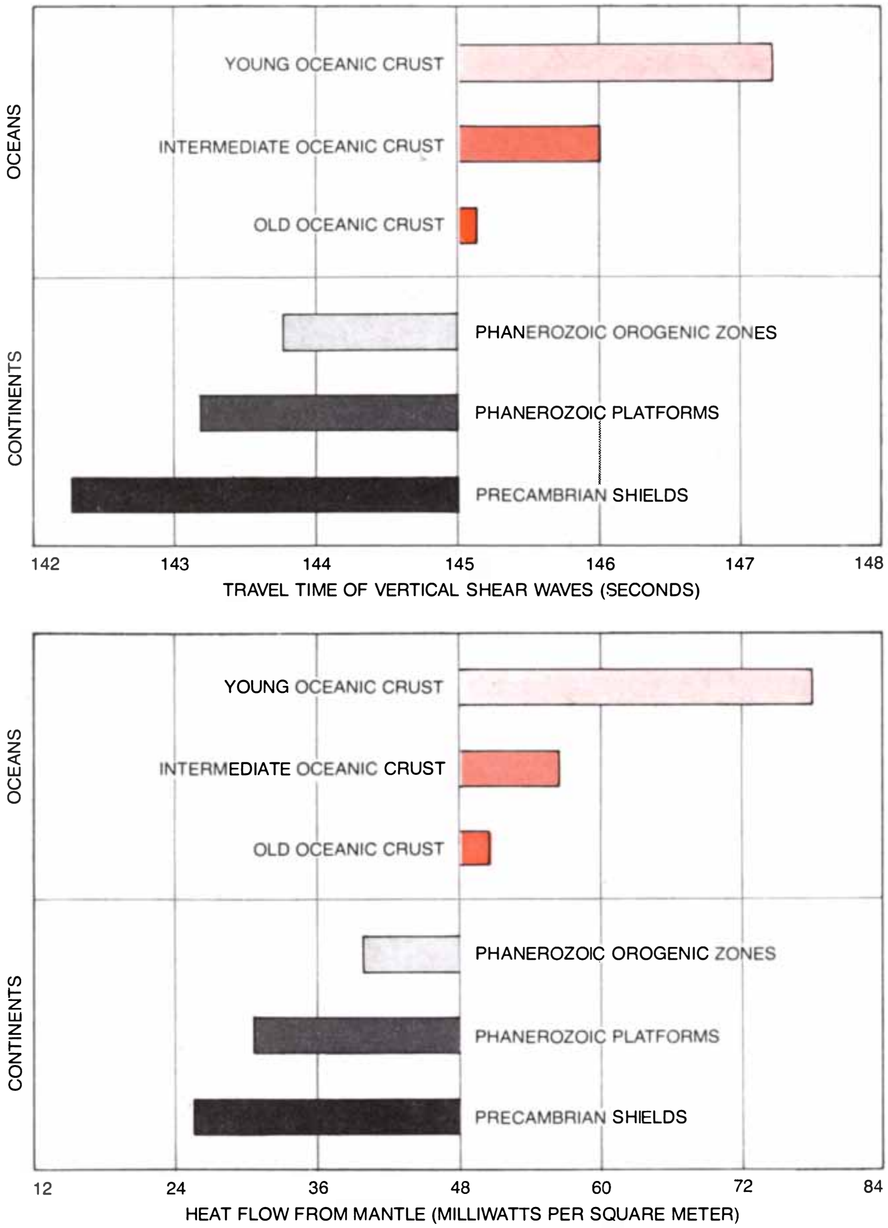
HARRIS

COMMUNICATIONS AND INFORMATION HANDLING

bulk of the upper mantle is thought to be a peridotite, which at depths greater than about 70 kilometers forms an assemblage of four minerals: olivine (60 percent), orthopyroxene (12 percent), clinopyroxene (15 percent) and garnet (13 percent). Such a four-phase peridotite is called a garnet lherzolite. Each of the minerals in a garnet lherzolite has a different melting temperature. The basaltic components (clinopyroxene and garnet) melt at lower temperatures than the more refractory components (olivine and orthopyroxene). If the temperature is high enough so that some fraction of the mantle—say 10 or 20 percent—is melted, the molten fraction has a basaltic composition. If this melt is then removed, it leaves behind a residual rock that is depleted in basalt and consists primarily of olivine and orthopyroxene.

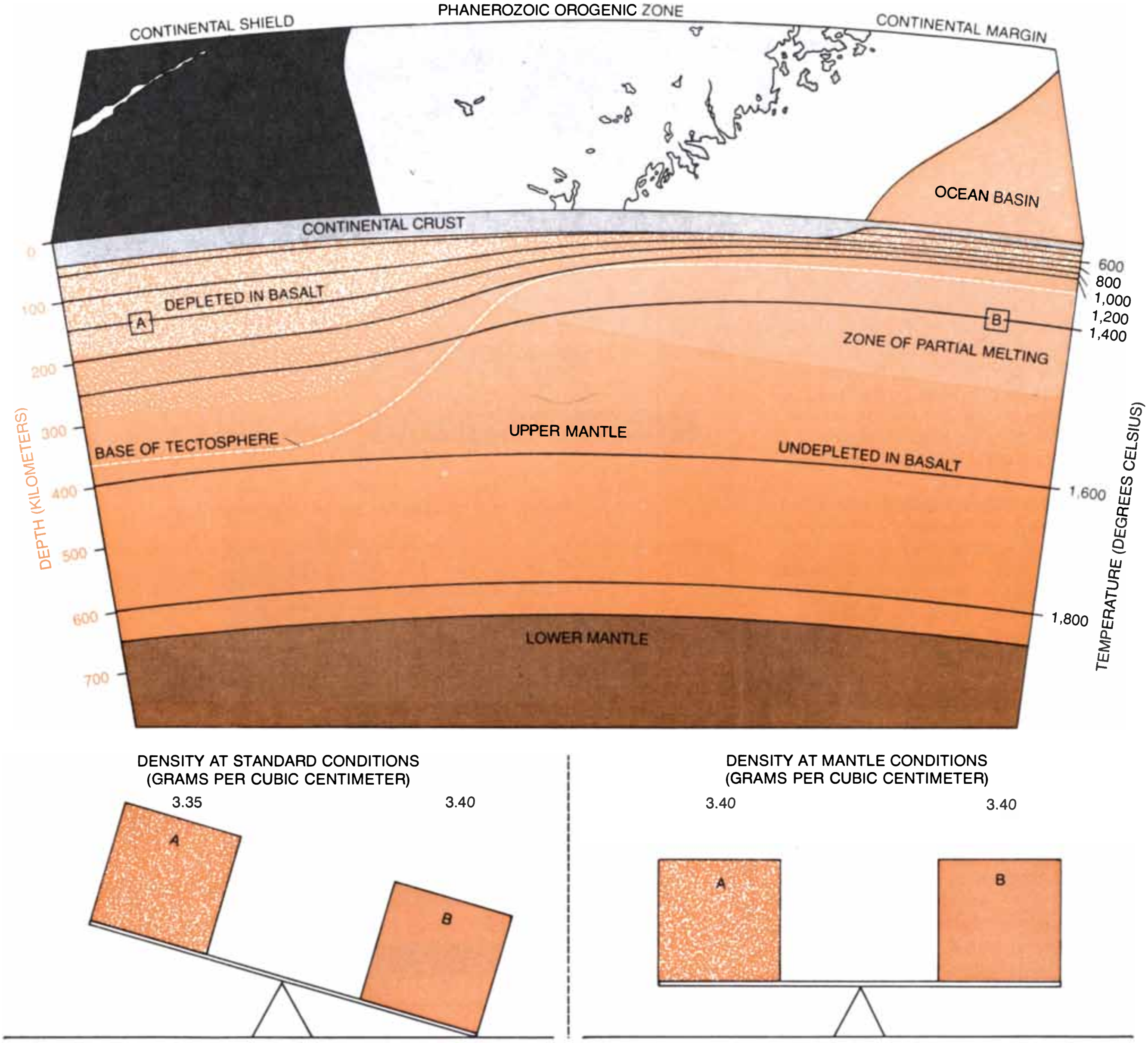
The important fact is that at a given temperature and pressure the density of this depleted residual rock is somewhat lower than that of the parental garnet lherzolite; that is, the removal of a basaltic component from the mantle reduces the mantle material's density by a small amount, typically 1 percent or so. The reason for the decrease in density is easy to understand. The removal of basalt leaves the mantle with a smaller proportion of garnet, which is substantially denser than the other minerals in a garnet lherzolite (3.7 grams per cubic centimeter compared with 3.3 grams per cubic centimeter) and also lowers the mantle's content of iron (the heaviest abundant element in the mantle). Surprisingly, these density relations were not generally appreciated by geophysicists and geochemists until quite recently; the first clear, quantitative statements concerning them were published in a short paper by Michael J. O'Hara of the University of Edinburgh in 1975. O'Hara's conclusion that depletion in basalt lowers the density of the mantle has been subsequently confirmed by the experimental work of Francis R. Boyd and R. H. McCallister of the Carnegie Institution of Washington and by my own extensive numerical calculations.

In 1976 I proposed that the compositional variations required to stabilize the continental tectosphere were induced by basaltic depletion of the tectosphere. My calculations showed that the effect should be quantitatively sufficient. At depths of from 150 to 200 kilometers the subcontinental mantle is estimated to be colder than the suboceanic mantle by some 300 to 500 degrees Celsius. In a chemically homogeneous mantle such temperature differences would produce differences in density of from 1 to 1.5 percent. Density differences of that magnitude could be exactly compensated if the peridotite that is now 150 to 200 kilometers below the cratons had been melted to the extent of 10 to 20 percent at some time in its history, and if the basaltic molten fraction had been removed by migration to the crust.



CLOSE CORRELATION is observed for different types of crustal rock between the time it takes for a shear wave to travel vertically from a depth of 700 kilometers to the surface (*bar charts at top*) and the amount of heat flowing from the mantle (*bar charts at bottom*). The vertical travel times are averages constructed from a large number of multiple *ScS*-wave travel times; the mean travel time for the entire globe is about 145 seconds. The characteristic heat-flow values shown are also averages based on the measured heat flow for the various types of crust; the mean global heat flow is 48 milliwatts per square meter. The correlation supports the view that the seismic differences associated with different types of crust arise from differences in the thermal structure of the underlying mantle. Colors are keyed to map on pages 94 and 95.

This basalt-depletion hypothesis can be checked by direct geochemical observations. Rocks that were once 150 to 200 kilometers below the continents are actually found on the surface, as rounded pebbles and boulders called xenoliths (foreign rocks) in formations known as kimberlite pipes: the eroded necks of certain peculiar volcanoes found only in stable continental regions [see "Kimberlite Pipes," by Keith G. Cox; SCIENTIFIC AMERICAN, April, 1978]. These pipes have been studied extensively because they are the ultimate surface sources of all diamonds. (The type locality of kimberlite—the peridotitic volcanic rock that fills the pipe—is the famous diamond-producing area of Kimberley in South Africa.) The xenoliths, some containing diamonds, apparently were ripped from the walls of the volcanic



DENSITIES OF MANTLE ROCKS at a given level under the continents and the oceans are assumed to be nearly equal in the author's model, even though the temperatures under the continents are lower. To compensate for an increase in density due to thermal contraction he postulates that the mantle rocks are differentially depleted in certain basaltic components, which under the conditions present in the mantle are represented by the minerals clinopyroxene and garnet. At standard conditions the peridotites that constitute the cold continental root zones would accordingly be less dense than their suboceanic counterparts. At a depth of 150 kilometers, for example, the subcontinental mantle (*A*) is estimated to be about 400 degrees Celsius cooler than the suboceanic mantle (*B*), requiring a decrease in density by the basalt-depletion mechanism of approximately .05 gram per cubic centimeter (*shown in the scales at bottom*). Standard conditions here mean a temperature of 25 degrees C. and a pressure of one atmosphere; mantle conditions are temperatures of 1,000 and 1,400 degrees respectively for subcontinental and suboceanic mantle rocks and a pressure of 50,000 atmospheres for both. Diagram at the top is identical with the one on page 93, except for the substitution of the temperature contours (*black*) for the seismic-velocity contours.

conduit during the rapid ascent of gaseous kimberlite magma from a great depth and were transported quickly to the crust, in most cases without significant chemical alteration. Geochemical and petrological studies have shown that many of the xenoliths are samples from depths of between 100 and 250 kilometers.

Most of the mantle xenoliths are garnet lherzolites. Petrologists such as P. H. Nixon of the University of Papua New Guinea and Boyd of the Carnegie Institution have noted that the xenoliths are apparently depleted in a basaltic component: they typically contain less garnet and clinopyroxene than does the mineral assemblage estimated to constitute the bulk of the oceanic upper mantle. To demonstrate the depletion I constructed a model composition of the cratonic upper mantle in the depth range from 150 to 200 kilometers by averaging together the available chemical analyses of garnet lherzolite xenoliths from kimberlite pipes. The depleted nature of this average continental garnet lherzolite is shown in the illustration on page 106. At a standard temperature and pressure the density of this continental material turns out to be 1.3 percent less than the density of the best available model for the oceanic upper mantle—or right in the middle of the 1-to-1.5-percent range specified by the basalt-depletion hypothesis!

Closer examination of the compositions and densities of the garnet lherzolite xenoliths provides a more stringent test of the basalt-depletion hypothesis. The differences in composition between the continental and the oceanic mantle are hypothesized to be proportional to their temperature differences; the temperature differences can be presumed to decrease with increasing depth, vanishing near the base of the tectosphere. The amount by which the continental rocks are depleted in basalt should therefore decrease with depth. As was first pointed out by Nixon and Boyd, that is indeed the case. The density contrast appears to decrease with depth just as predicted by the basalt-depletion hypothesis.

The basalt-depletion hypothesis was motivated indirectly by seismological data: the high wave speeds that indicate low temperatures under the cratons. Basalt depletion may play a more direct role in explaining those seismic-wave speeds, and particularly the shear-wave speed. Two effects are important. First, basalt removal lowers the iron content of the mantle, and experiments have shown that seismic-wave speed rises with decreasing iron content. Second, basalt removal increases the melting temperature of the mantle (because the components that melt at low temperatures are withdrawn) and that also increases the wave speeds, even at temperatures below the melting point. Basalt depletion thus helps to explain why the shear-wave speeds under the cratons are so high.

In summary, the basalt-depletion hypothesis has several attractive features: it provides a mechanism for stabilizing the sublithospheric portions of the continental tectosphere against convective disruption, it explains at least two salient features of kimberlite xenolith petrology and it is consistent with the available seismological constraints. Much more extensive testing is nonetheless needed before the model can be accepted as a foundation on which to construct a theory of continental development. Diagnostic tests of the model do not come easily; the phenomena it represents are buried under hundreds of kilometers of rock. Yet the constraints on the nature of the deep continental root zones are multiplying at such a rapid rate that if the model is fundamentally wrong, we should know it very soon.

Cratonic Stability Reassessed

In the meantime some tentative speculations on the nature of cratonic stability and continental evolution are warranted. In the context of the model I have proposed, continental drift is dynamically complex, more than simply a passive rafting of light crustal material on an otherwise undistinguished piece of lithosphere. The basalt-depletion model suggests that continental tectonic behavior is actually regulated by chemi-

cal variations extending deep into the upper mantle.

Under regions of active tectonism at or near plate boundaries, such as the western U.S. and southern Europe, the layer of depleted peridotite is apparently thin. As the mantle under this thin layer cools beyond some critical point, undepleted mantle material sinks, causing convective overturn and the transport of more heat to the surface. Because they are heated by this convective overturn, the crust and the uppermost mantle in such a region are weak and easily deformed. In contrast, the depleted layer under the cratons is hypothesized to be much thicker. The low density of this residual peridotite prevents it from being easily mixed back into the mantle by convective action, so that the material in the depleted zone is stabilized in nearly hydrostatic equilibrium at lower temperatures than its surrounding mantle. Because its temperature is lower and its low-temperature melting fraction has been removed its viscosity is higher, and so the zone of horizontal shearing associated with plate motions is confined below the depleted zone. The crust and uppermost mantle are isolated from the thermal perturbations and mass motions associated with small-scale convective disturbances, and so cratonic stability is maintained.

How is a thick, cool, depleted subcrustal tectosphere formed in the first place, however? Where does all that basalt come from and where does it go? Geologists have deduced that the basic plate-tectonic processes observed today—sea-floor spreading, continental drift, subduction and so on—were active in the Precambrian era, which ended about 600 million years ago. Some geologists believe there is evidence for plate tectonics in rock formations three billion years old or more. It is reasonable to assume, therefore, that tectospheric development has been governed by the plate-tectonic cycle, at least for the past three billion years.

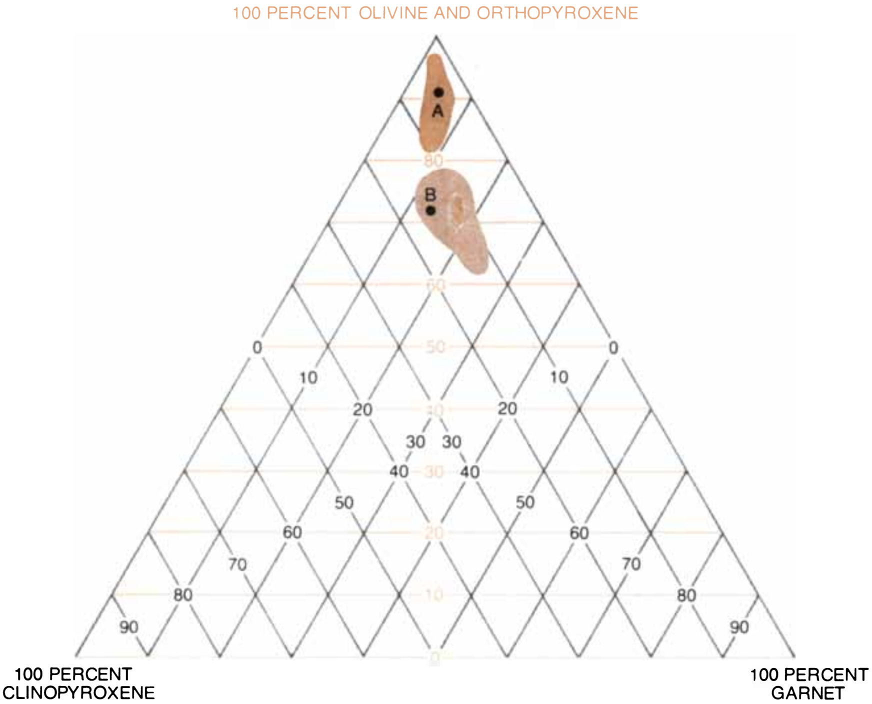
What particular processes may have been important in depletion? At present the continental upper mantle is being depleted in basalt in two primary settings: at sites of the magmatic episodes associated with continental rifting and in the wedge of mantle above the subduction zones. The former are characterized by huge volumes of "flood basalts" extruded during periods of crustal extension preceding the breakup and drifting of continents. (For example, the Paraná basin flood basalts of South America, which were erupted just before the opening of the South Atlantic Ocean, have a total volume of nearly two million cubic kilometers; the volume of mantle depleted by this episode must have been at least several times greater.)

Even more voluminous are the basalts emplaced in the crust above subduction zones, landward of the oceanic trenches along the "active" continental margins and island arcs. These sites are usually associated with the more silica-rich andesitic volcanism, but basalt is actually the more common magma type, particularly in juvenile island arcs. Petrologists and geochemists are trying to identify the sources of these magmas. The andesites may come from the subducted oceanic crust, but most of the basalt is apparently generated from the mantle wedge above the descending oceanic lithosphere. Hence a major portion of the residual depleted mantle eventually incorporated into thick tectospheric root zones may be originally depleted in regions of rapid island-arc formation such as the present-day western Pacific.

Melting and basalt extraction are processes confined mainly to the mantle above a depth of 200 kilometers, however; below that the pressure is so great that the melting required to produce basalt is rarely achieved, if ever. A depleted tectosphere thicker than 200 kilometers therefore cannot be generated by melting processes alone. E. R. Oxburgh and E. M. Parmentier of the University of Oxford have recently proposed another source. They suggest that some of the peridotite that is depleted as it rises and melts at a mid-ocean ridge and is transported by sea-floor spreading to a site of subduction may in the course of subduction attach itself to the tectosphere.

However it is generated, any protocontinental tectosphere in an island-arc environment or along an active continental margin would be thin, chemically heterogeneous and poorly consolidated. The thickening comes with time. First the dispersed regions of depleted mantle and their overlying crust are swept up and accreted by plate motion to the primary continental masses. Although such terrain is somewhat consolidated, it still has a thin tectosphere and hence may be a site of orogenic activity; good examples of continental tectosphere in this stage of development are found today in Indonesia, western Canada and Patagonia.

Further consolidation and thickening of the tectosphere apparently result from the major compressive events at



EVIDENCE supporting the basalt-depletion hypothesis has been obtained from xenoliths (literally "foreign rocks") brought to the surface during the eruption of basaltic magma from kimberlite pipes. These rocks, which presumably are representative samples of the composition of the upper mantle under the continental crust at depths ranging from 100 to 250 kilometers, are found to have smaller concentrations of clinopyroxene and garnet (*upper shaded area*) than those estimated for the upper mantle under the oceanic crust (*lower shaded area*). The density of the kimberlite xenoliths is on the average about .05 gram per cubic centimeter less than the estimated density of the peridotite in the oceanic upper mantle, in agreement with the value predicted by the basalt-depletion hypothesis for the average density of the peridotite under the continental crust. The composition of the basalt-depleted peridotite at point *A* in this illustration and in the one on page 104 is 67 percent olivine, 23 percent orthopyroxene, 4 percent clinopyroxene and 6 percent garnet; the undepleted peridotite at point *B* in both illustrations is 60 percent olivine, 12 percent orthopyroxene, 15 percent clinopyroxene and 13 percent garnet. The basaltic components are shown in black; nonbasaltic components are shown in color.

convergent plate boundaries, particularly episodes of collision between continents. The history of the crust following a continental collision has been analyzed in detail by John F. Dewey and Kevin C. A. Burke of the State University of New York at Albany. During very violent collisions, such as the ongoing one between India and Asia, the former continental-margin crust is compressively thickened by a factor of two or more, causing the uplift of a plateau (such as the Tibetan plateau, where surface-wave data indicate that the crustal thickness is about 70 kilometers). Dewey and Burke think this process of compressive crustal thickening involves deformation over a broad zone. Such deformation should be concomitant with a proportional thickening of the subcrustal tectosphere by the lateral and downward advection of depleted peridotite. In this way the thick root zones of the cratons could be formed.

Depletion, consolidation and thickening, then, are thought to be the important constructive events in the development of a continental tectosphere. To stabilize a thick tectosphere in any particular region, these processes may have to be repeated over and over again because they must often compete against destructive events caused by fluid-mechanical instabilities in the upper mantle. Whereas large-scale basaltic depletion is nearly irreversible, consolidation and thickening of the tectosphere surely are not. Any significant heating of a cool, thick, stabilized tectosphere by heat conduction from its periphery, by intrusion or by internal heat sources results in its thinning and dispersal by the upward and lateral flow of low-density depleted mantle. Such motions may play an important role in episodes of continental rifting and breakup. A heating event of this kind may have destabilized the once cratonic portions of the western U.S., causing the widespread extension and rifting now taking place in the Basin and Range province of Utah, Nevada and surrounding states.

The rough sketch of continental tectospheric structure and development presented in this article is no more than an incomplete and highly speculative working model, shaped by many disputable generalizations and colored by the author's own prejudices. Much closer scrutiny is necessary to evaluate its feasibility and to validate the hypotheses on which it is based. Regardless of its eventual fate, however, the model does emphasize the maturity of modern earth science. Any detailed picture of deep continental structure and its evolution must be painted on a cloth woven from many threads of geological, geochemical and geophysical information. Only such a unified picture will reveal the true nature of the dynamic processes that have shaped the continents.

The Trajectories of Saccadic Eye Movements

In which a close examination of the rapid, sharp rotations of the eyeball known as saccades provides new information concerning the way the human brain controls eye movement

by A. Terry Bahill and Lawrence Stark

As a person reads, looks at a picture, drives a car or even stares at a spot on the wall his eyes make a multitude of movements, the great majority of which are the staccato flicks of the eyeball known as saccades. Saccadic eye movements serve to place the small area at the center of the retina called the fovea on different parts of the visual field. The fovea is the part of the retina where visual receptor cells are most densely packed, and so it is the part most capable of mediating detailed vision. Information from the periphery of the retina is used to direct the various saccadic movements, most of which range in magnitude from four minutes of arc to 15 degrees. (Larger shifts of gaze are usually accomplished by a combination of head and eye movements.)

The trajectories of different saccadic eye movements are unusually closely related to the patterns of motoneural firing, or the control signals, that generate them. In the eye-movement system the load the eyeball presents to the muscles is small and constant, and the muscles are strong and able to contract rapidly. As a result the precise neurological control signals that drive the muscles are faithfully reflected in the movements of the eyeball. In other words, the shape of the signals can be deduced directly from the saccadic trajectories. Once the control signals have been identified it is possible to obtain a better understanding of the nature of saccades and of the mechanisms that generate them.

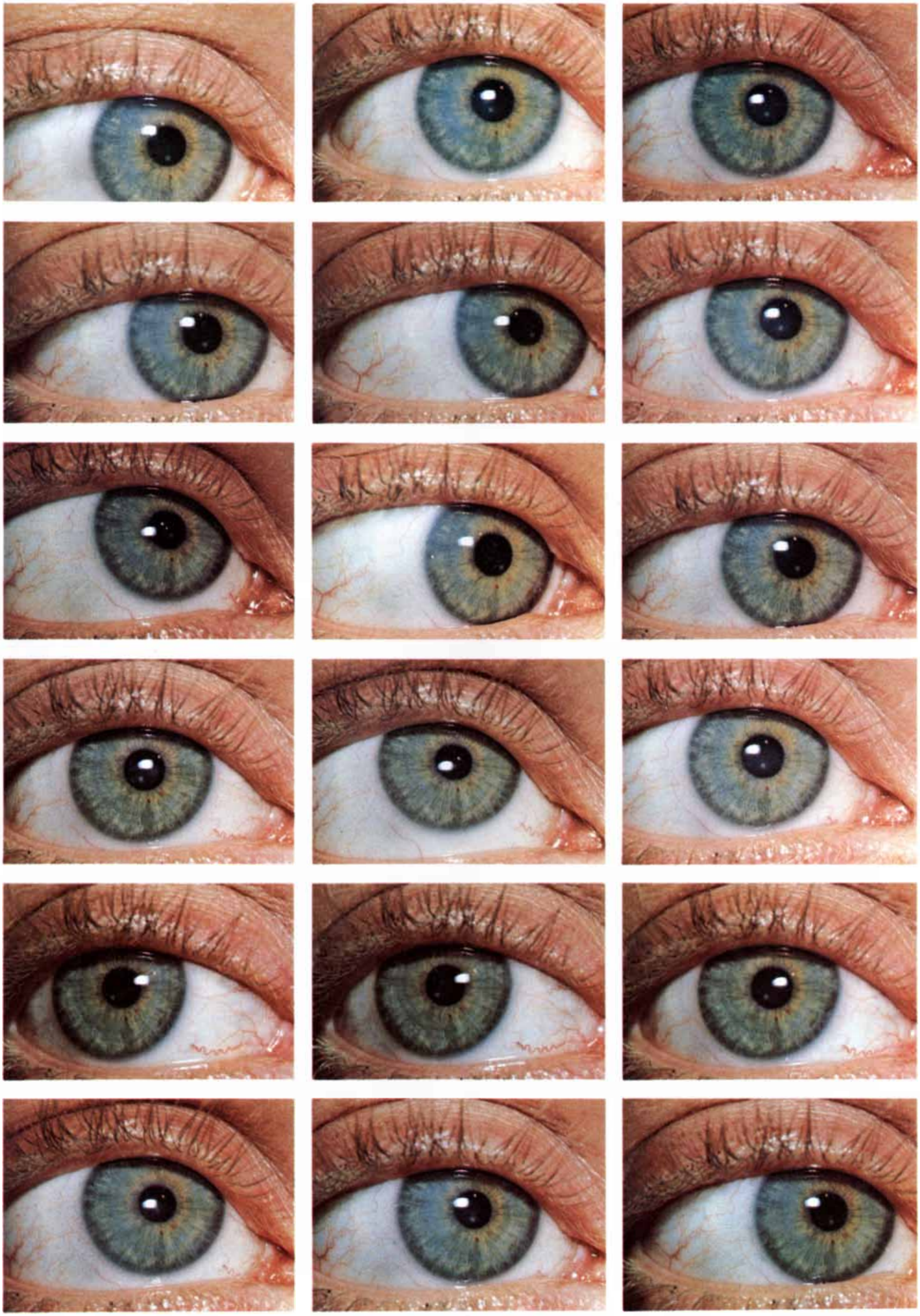
Here we shall describe the application of this method to several different kinds of saccadic eye movements: saccades that overshoot (or undershoot) their targets, double saccades and obliquely directed saccades. (Although much of the information that has been gathered on eye movements pertains to horizontal movements, most eye movements have a vertical component as well as a horizontal one.) First, however, we should describe our method of gathering eye-movement data and our tools for analyzing the data.

We used a photoelectric method to record the eye movements. In this method a pair of photodiodes mounted on spectacle frames are aimed at the border between the colored iris and the white sclera on each side of one of a subject's eyes. A small dot, the target, is made to move abruptly on a screen in front of the subject. When the eye turns, say toward the nose in response to a particular target jump, the photodiode nearer the nose is exposed to more of the dark iris and less of the light sclera, so that its photocurrent decreases; simultaneously the photodiode nearer the temple is exposed to more of the sclera and less of the iris, so that its photocurrent increases. The difference between the two currents provides an accurate measure of eye position. In our experimental system the photocurrents were converted into voltages, amplified and recorded on a computer disk-memory unit. The velocities of the recorded movements were computed by the computer itself. (Vertical movements were measured with a slightly modified apparatus.)

It is a common practice in recording eye movements to filter out high frequencies in order to eliminate instrumental noise and irregularities, but we maintained a large bandwidth: from zero to 500 hertz (cycles per second). This modification was important because high-frequency filtering would have removed the important fine details of the saccadic trajectories and also would have reduced the computed velocities. We were able to reduce instrumental noise to a very low level, so that it was much smaller than the biological signal we were measuring. Moreover, by continually calibrating and adjusting the instrumentation we were able to maintain a linear input-output relation for saccades of up to 20 degrees. Our experimental apparatus is quite inexpensive, and recently, in cooperation with William F. Hoyt of the University of California Medical School in San Francisco and B. Todd Troost of the University of Pittsburgh School of Medicine, we have employed it in gathering quantitative information to aid in the diagnosis of patients with various eye-movement disorders.

As our measuring instrumentation sampled the trajectory of the eye more finely, or frequently, we came closer to identifying the control signals driving each individual eye movement. At 10 samples per second (corresponding to a bandwidth of five hertz) we could study the fixation periods between saccades but not the rapid saccadic movements themselves. This rate is adequate for studying the eye movements in reading and the various paths the eye follows in scanning different scenes or objects. At 100 samples per second we could get information about the visual system's strategy for controlling the production of saccadic eye movements. For example, at that rate normal saccades can be

SACCADES SERVE TO MOVE the fovea of the eye, the small high-resolution area at the center of the retina, to different points in the visual field, as is depicted by the sequence of photographs shown here. The trajectories of these staccato rotations of the eyeball can be seen when a bright, undiffused light is shined on a subject's eye. As the subject looks at different points along a horizontal line (or even stares at a fixed point) the bright spot where the light reflects off the cornea provides a fixed reference point for viewing the saccadic eye movements that are generated. A person with normal vision typically executes about two saccades per second. Most of these eye movements range in magnitude from four minutes of arc to 15 degrees.



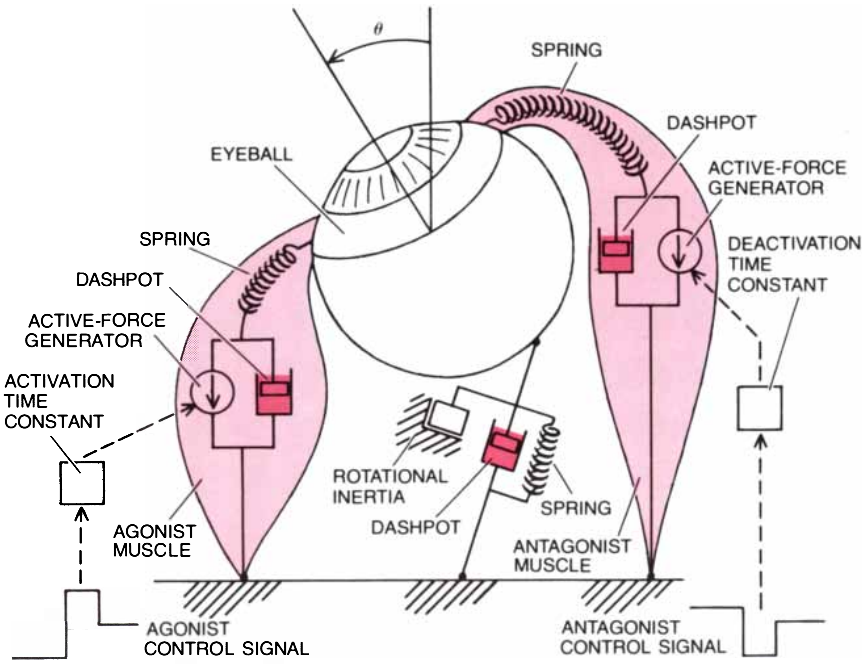
observed to occur at intervals as short as 200 milliseconds. In 1962 Laurence R. Young of the Massachusetts Institute of Technology and one of us (Stark) first suggested that the control system for the generation of saccades can be described as a sampled-data control system, that is, a control system that delivers output discontinuously. At 1,000 samples per second our instruments yielded a wealth of detailed quantitative information on the parameters of saccades, in particular peak velocity, duration and magnitude. And we could also see the minute variations in the shapes of the saccades, some of which are described below.

Two new conceptual tools were developed to help explain the variations in the saccadic trajectories. The first tool makes use of a standard engineering technique: when bioengineers want to analyze a very complicated system, they often resort to a model to describe the system in a less complicated way. Models are simpler and easier to deal with than real systems. For example, a street map is a model of the streets of a city, and it is easier to give directions by referring to the map than by referring to the real world. Many models have been developed to help study the neurological control of human movement. One of the earliest models of the oculomotor control system was developed by René Descartes, who first recognized that movements of the eyeball require the coordinated activity of at least two muscles: an agonist muscle that shortens and pulls against the eyeball to provide the turning torque for the eye movement, and an antagonist muscle that relaxes and lengthens. Hence normal eye movements can be made only through the reciprocal innervation of pairs of muscles. Descartes proposed this principle of reciprocal innervation in 1626, but it was not until 1963 that Gerald Cook of M.I.T. and one of us (Stark) incorporated it into a model of saccadic eye movement. Subsequently, in association with Michael R. Clark and Frederick K. Hsu of the University of California at Berkeley, we were able to develop a reciprocal-innervation model for all human eye movements.

The reciprocal-innervation models are homeomorphic, in the sense of the word employed by the biomathematician Richard E. Bellman: there is a direct, detailed relation between the components of each of the models and the components of the real system. The advantage of working with a homeomorphic model of saccadic eye movements is that we were able not only to simulate the input of neurological signals and the output of eye movements, as might be done with a black-box model, but also to study how varying the individual parameters of the model affects output and how the physiological properties of the real muscles and tendons of the eye-movement system affect those parameters.

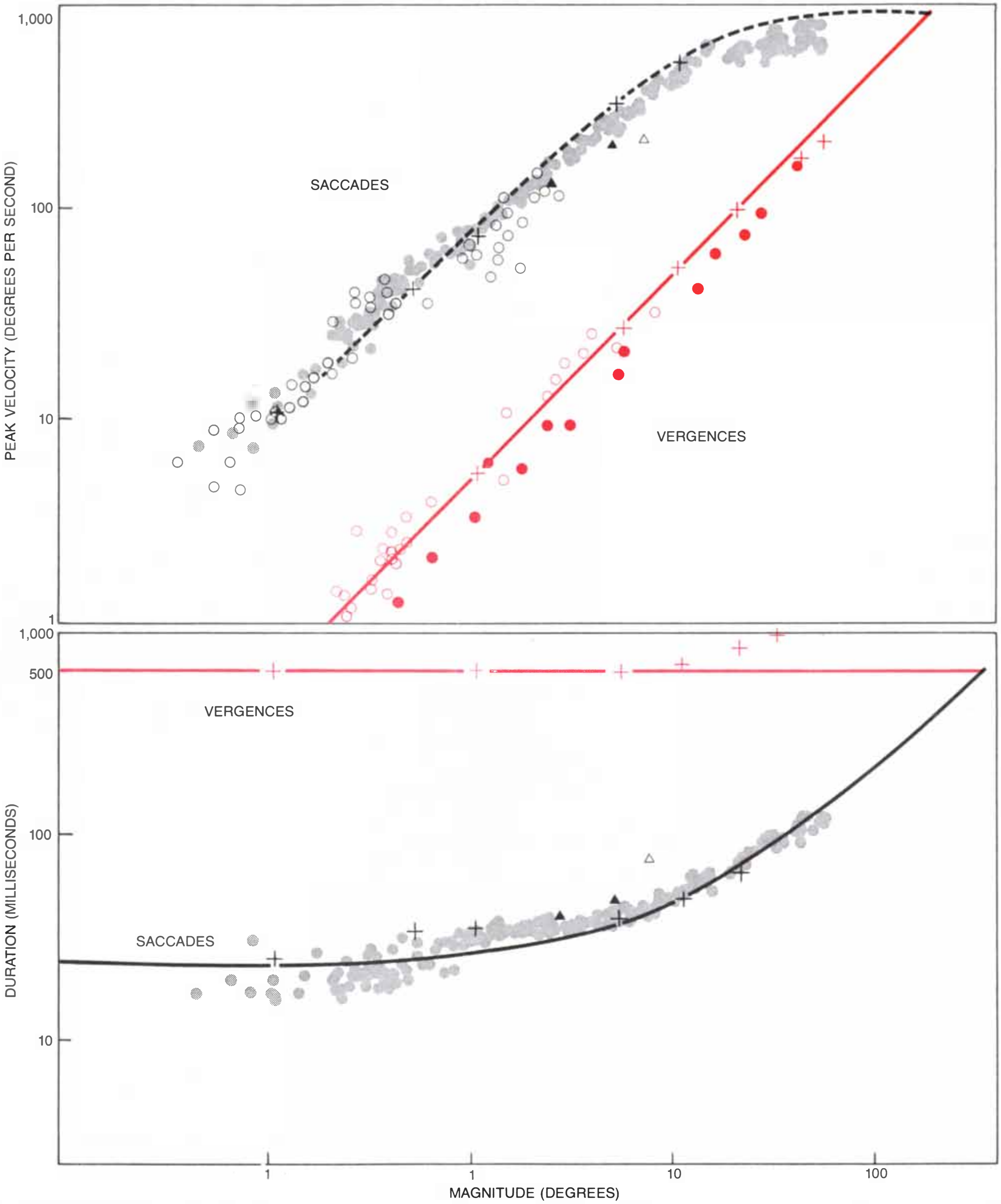
The parameters of our reciprocal-innervation model represent the inertia of the eyeball, its viscoelasticity on being rotated in the orbit of the eye, the elasticities of the ocular muscles and tendons and the apparent viscosity of the ocular muscles. (An inherent property of muscle, known as the Fenn-Hill-Katz force-velocity relation, makes the force-generating mechanism of muscle appear to be damped, as if by a shock absorber, or viscous damping element.) The model was implemented on a digital computer connected to a storage oscilloscope (an oscilloscope with a memory). Hence in our simulations of human eye movements we could vary the input of neurological control signals and immediately observe their effect on the output of saccadic trajectories. In particular we could experiment with differently shaped control signals until we found the one that generated the trajectory with parameters most like those of a particular type of saccade.

An important aspect of modeling is the validation of the model. We validated our model in several ways: qualitatively, by comparing the shapes of real saccades with those of model saccades; quantitatively, by comparing the parameters of the real saccades with those of model saccades; analytically, by performing a sensitivity analysis of the model (systematically varying individual parameters of the model in order to observe their influence on the total eye-movement system), and heuristically, by simulating eye movements the model was not designed to simulate.



RECIPROCAL-INNERVATION MODEL the authors used to simulate various eye movements is based on the concept that eyeball rotation is effected through the innervation of at least two muscles: an agonist, which shortens and pulls against the eyeball, and an antagonist, which lengthens and relaxes. In the model neurological control signals, or patterns of motoneuron firing, such as the ones shown at the lower left and lower right, are given as input. Model elements known as activation and deactivation time constants reproduce the smoothing of the signals that occurs in the real eye system as the signals are delivered to and processed by the agonist and antagonist muscles. The inertia of the eyeball in rotation and the viscosity and elasticity of the eyeball-and-eye-muscle system are represented by the inertia of the rotating mass, the dashpot and the spring shown below the eyeball. (A dashpot, or shock absorber, is a viscous damping device.) The active-force generators in the model simulate the production of muscle force not as it is measured at the tendon of a real muscle but as it originates within the muscle. Each active-force generator is connected in series with a spring, which represents muscle elasticity. The output of the model is the eye position θ, the angle at which the eyeball is rotated away from the straight-ahead position. It is a special property of the real eye system that the trajectories of the eye are faithful reflections of the neurological control signals that generate eye movements. By experimenting with differently shaped control signals in the model, the authors were able to identify signals that generate particular variations in saccadic trajectories.

The concept of our second tool comes from the Hertzsprung-Russell diagram, which astronomers use to classify and compare diverse types of stars. The diagram plots the luminosity of the stars

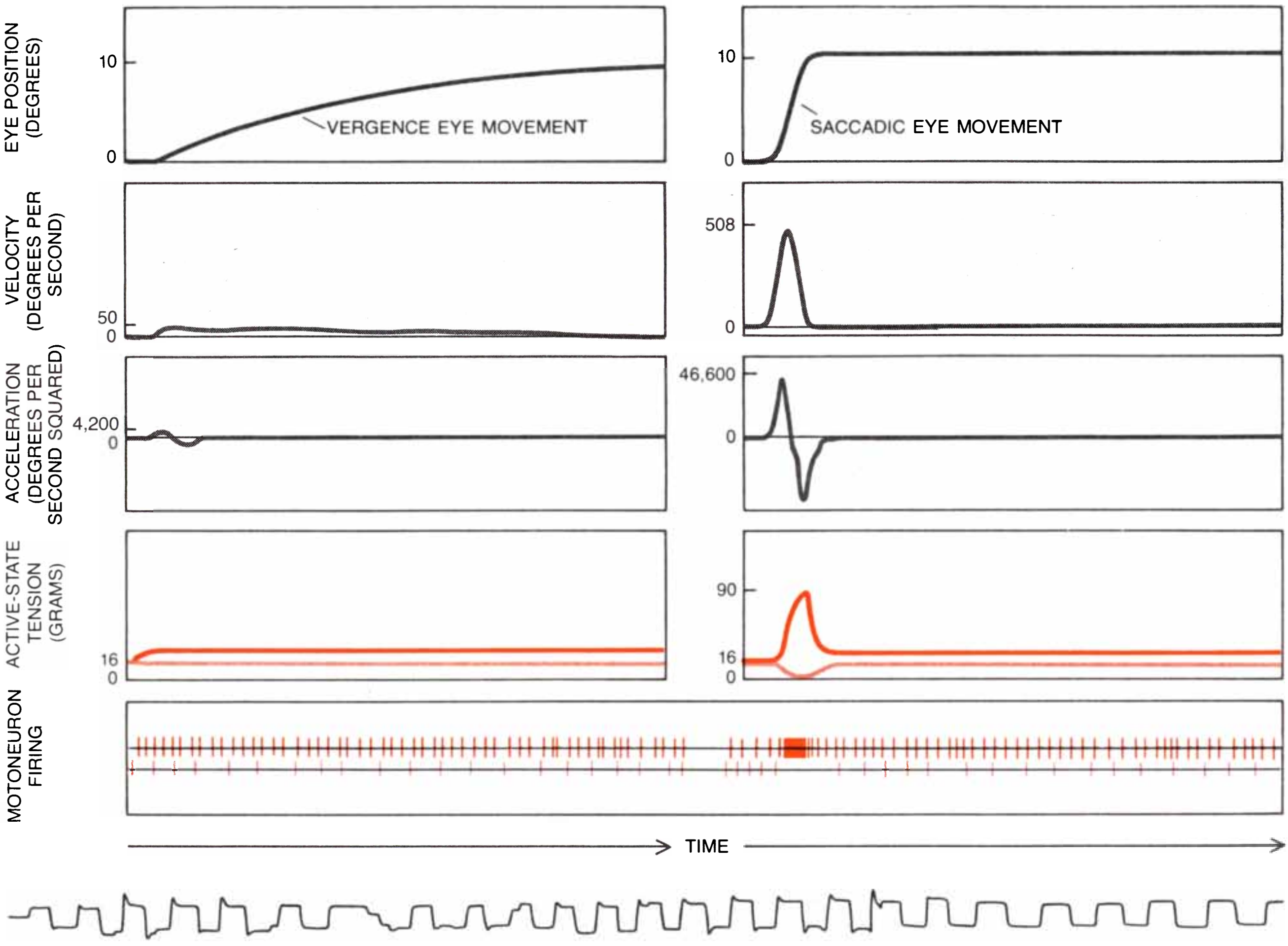


MAIN-SEQUENCE DIAGRAM for human eye movements plots the peak velocity (*top*) and the duration (*bottom*) of various eye movements as functions of their size. Experimental data on eye movements fall into two main groups on the diagram: the fast, staccato saccadic eye movements (*gray dots*) that make detailed vision possible and the slow, gliding "vergence" eye movements (*dots in color*) that track between far objects and near ones. There are variations in saccades; for example, some may overshoot their final position and return to it with a fast saccadic motion known as a dynamic overshoot (*black circles*). A slow, gliding movement known as a glissade (*circles in color*) may also return the eye to its final position. In some cases the eye may execute two small, closely spaced saccades instead of one large saccade to get from one position to another. As the diagram shows, it is not the overall movement (*open triangles*) that has the parameters of a normal saccade; rather it is the individual small movements (*filled triangles*) that are saccadic. The black crosses on the diagram represent saccades that were simulated on the reciprocal-innervation model (*see illustration on opposite page*); the crosses in color represent simulated vergences. The solid and broken lines shown in the illustration are analytical functions that fit the distribution of the experimental data.

against their surface temperature, or color. The majority of normal stars lie on a "main sequence": the swath of stars that runs diagonally across the diagram [see "The New, or Modified, Form of the Theory of Stellar Evolution," by Henry Norris Russell; SCIENTIFIC AMERICAN, October, 1925]. We have developed a similar diagram that enables us to classify and compare diverse saccadic phenomena. In our main-sequence diagram the peak angular velocity and the duration of eye movements are plotted as functions of the magnitude of the eye movements [*see illustration on preceding page*]. Our eye-movement data fall on two separate branches in the main-sequence diagram: one branch is made up of the rapid saccadic eye movements and the other is made up of the slow "vergence" eye movements that are utilized in tracking between distant objects and close ones. (The two other major types of eye movements—the vestibulo-ocular movements, which maintain visual stability during head movements, and the smooth-pursuit eye movements, which follow a moving object such as a flying bird—do not appear on the main-sequence diagram, since neither of these types of movement is elicited by target jumps.)

The main-sequence diagram demonstrates that there is little variation in the parameters of saccadic eye movements. For example, for a normal unfatigued subject the range in the duration of 10-degree saccades is from 38 to 45 milliseconds and the range in peak velocity is from 420 to 520 degrees per second. Similarly, a saccade covering 10 minutes of arc would last between 15 and 22 milliseconds and would attain a peak velocity of between 11 and 15 degrees per second. These ranges are small compared with those of other biological systems. We maintained small ranges of variation for the main-sequence data by (1) rejecting data from fatigued subjects, (2) plotting the parameters as functions not of the size of the target movement but of the size of the actual eye movement and (3) defining the size of the saccadic eye movement as being the size of the initial saccade and ignoring drift and noise at the end of movements. As a result, although the saccades of different human subjects have slightly different parameters, saccades whose parameters fall outside the main-sequence ranges are usually indicative of some pathological condition.

In 1965 observations of similar main-sequence data led Cook, Bert L. Zuber and one of us (Stark) to postulate that there is no basic difference between saccades. In other words, the parameters of all saccades fit on the main-sequence diagram, no matter how large or small the saccades are, what their purpose is or what kind of stimulus elicits them. Sub-



ABRUPT TARGET JUMPS generate two types of eye movements: slow, smooth vergence eye movements (*left*) and sharp, rapid saccadic eye movements (*right*). A vergence is generated by a step signal to the eye muscles: a small, sustained increase in motoneuron firing activity for the agonist muscle (*dark color*) and a similar decrease in firing activity for the antagonist muscle (*light color*). In this type of movement the difference between the actual eye position and the eye position coded by the new steady-state firing level provides the driving force that moves the eye slowly into its new position. A saccade is generated by a pulse-step signal to the eye muscles: a high-frequency burst of motoneuron firing receding to a new increased step level of firing for the agonist muscle (*dark color*), and a sharp drop in firing followed by a rise to a decreased step level of firing for the antagonist (*light color*). In this case it is the pulse that provides the driving force, moving the eye quickly to its new position; the step signal holds the eye in place. Irregularities in the pulse-step signals create differently shaped saccadic eye movements, as in successive movements between two targets 10 degrees apart shown at the bottom.

sequent experimentation has supported this theory of the unified nature of saccades.

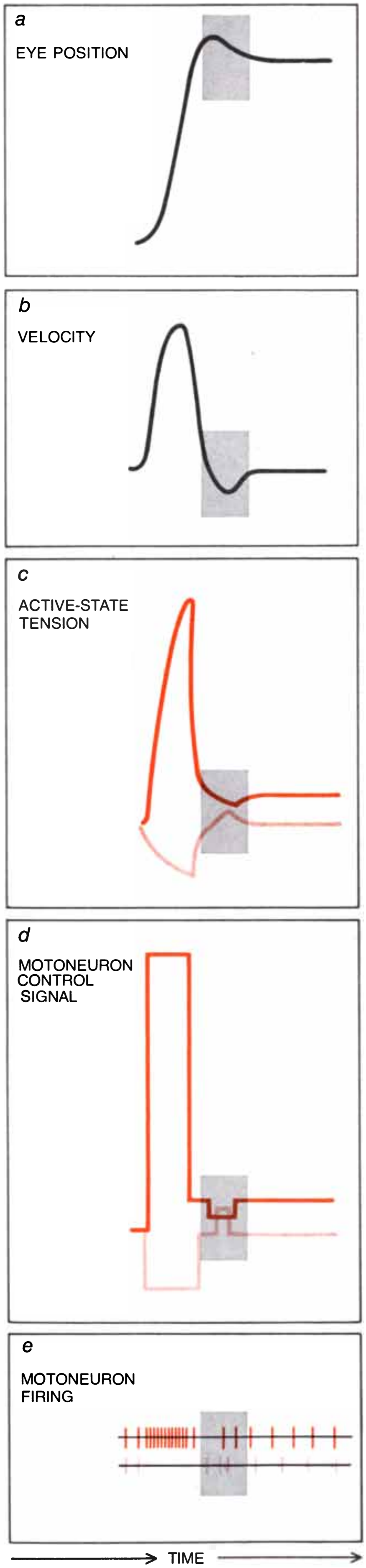
Experimental data also reveal, however, that although the main-sequence parameters and the general shape of saccades vary little, the structure of saccades varies greatly, that is, there are many small deviations from the normal saccadic trajectory. For example, some saccades drift past their target, others do not complete their movements with a single standard saccade and so on. These deviations appear quite irregularly, changing from day to day or from eye movement to eye movement, and they are often monocular, or limited to only one eye. (Indeed, the variations in the saccadic trajectories of a subject's two eyes are often quite different.) The explanation of this seemingly paradoxical phenomenon—the presence of trajectory variations in eye movements with stereotyped main-sequence parameters—is found by taking a closer look at the saccadic eye movements.

So far we have considered saccades as units, but in actuality each saccade is made up of several dynamic components, or smaller eye movements. It is the variation in the number and arrangement of these smaller components that gives rise to the variation in saccadic eye movements. We have been able to determine that all the variations in saccadic trajectories—all the changes in the configurations of the dynamic components of saccades—are due to variations in the neurological control signals of saccades. (In people with weakened eye muscles, of course, that is not the case.)

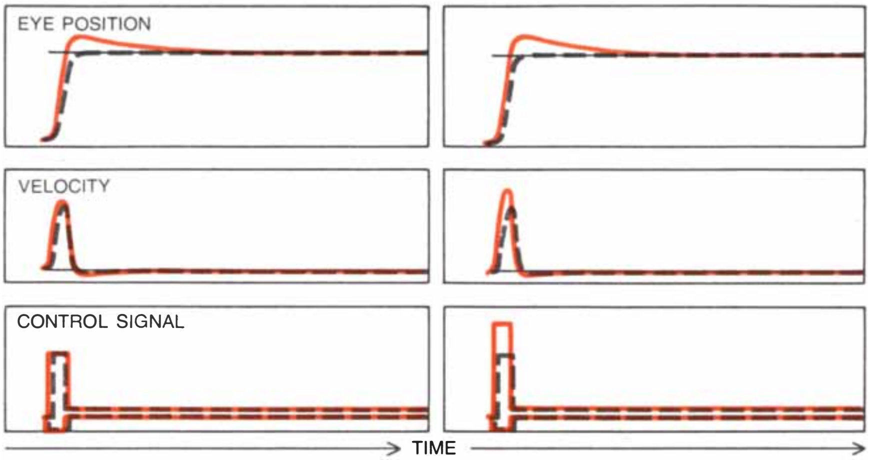
The reasons for the irregularities in normal control signals are not well understood; they do not depend on the amplitude of the saccades, the nature or visibility of the targets, the general lighting conditions or the purpose of the saccades. We do know that fatigue is responsible for some deviations, and it appears that psychological stress can also play a role. In any case the control signals do vary, and they give rise to the variations found in saccadic trajectories. As we have explained, the trajectories reflect the signals so faithfully that it is possible to see exactly how the changes in the signals affect the shape of the saccades.

First, consider the control signals that generate normally shaped eye movements. Saccadic eye movements and vergence eye movements, which form the two branches of the main-sequence diagram, are generated by two different types of neurological control signals. A slow, smooth vergence eye movement (in particular a divergence in which one eyeball rotates away from the other) is the result of a step change in the overall firing rate of the motoneurons that innervate the eye muscles: a sustained, low-frequency increase in the motoneuron activity for the agonist muscle and a similar decrease for the antagonist muscle produce a new steady-state firing level. The driving force for the eye movement is the difference between the actual eye position and the eye position coded by this new level of firing. A saccadic eye movement is generated by a two-part signal known as a pulse-step control signal. In this case the driving force for the eye movement is the pulse: a short burst of high-frequency motoneuron firing for the agonist and a corresponding pause in motoneuron activity for the antagonist. The pulse serves to move the eye rapidly from one point to another, and the step signal that follows serves to hold the eye in its new position. A pulse signal moves the eye much faster than a simple step signal; for example, a 10-degree vergence eye movement lasts for about 500 milliseconds, or about 10 times longer than a saccade of the same magnitude.

To understand how irregularities in pulse-step control signals create the minute variations in saccadic trajectories consider an example: the variation called saccadic overshoot, in which the eye travels beyond its ultimate fixation, or final position, and then returns, coming to rest on the fixation. Most saccades of the human eye display some type of overshoot. There are three distinct types, distinguished according to the way the eye moves back to its final position. In dynamic overshoot there is a fast return phase; for example, in a saccade with one degree of dynamic overshoot the return usually takes about 20 milliseconds, with the eyeball attaining a peak velocity of about 60 degrees per second. In glissadic overshoot there is a slow, gliding return, lasting for more than 200 milliseconds and for one degree of overshoot achieving a peak velocity of about five degrees per second. In static overshoot the eye remains fixed in the off-target position for between 150 and 200 milliseconds, until finally visual feedback elicits a corrective sac-



MODEL SIMULATION of a 10-degree saccade displays the variation in trajectory called dynamic overshoot: a normal saccade is followed by a small movement in the opposite direction that carries the eye back to a position it overshot (*a*). As the time-velocity relation (*b*) suggests, the return phase of a saccade with dynamic overshoot is itself a saccade. The driving force for the second small saccade is a momentary role reversal for the agonist muscle (*dark color*) and the antagonist muscle (*light color*) near the end of the saccadic eye movement (*c*). This change in the active-state tensions of the two muscles is in turn generated by a reverse pulse in the middle of the pulse-step signal generating the saccade (*d*), that is, there is a temporary reversal in the motoneuron activity for the saccadic movement. The firing patterns of a typical agonist motoneuron and a typical antagonist motoneuron are displayed at bottom (*e*).



SLOW, GLIDING GLISSADES at the end of a saccade can result from an oversize pulse component (*color*) in an otherwise normal pulse-step control signal (*broken black line*). When the pulse signal is too high (*right*), either because too many motoneurons were recruited for firing or because their firing rate was too high, the difference between the eye position at the end of the pulse and the eye position coded by the step signal generates a slow, smooth movement, following the primary saccade, that brings the eye back to the position it overshot. The same type of movement is generated when motoneuron firing goes on for too long, so that pulse is too wide (*left*). Thin black lines mark final eye position and velocity. Quantitative analysis of model simulations suggests that in subjects with normal vision pulse-width errors are commoner.

cade. Static overshoot is the only type that can be perceived with the unaided eye, and so it is the one most frequently observed by clinicians. Excessive static overshoot is often symptomatic of a disorder of the cerebellum.

Most saccades show dynamic overshoot. This irregularity seems to appear quite capriciously, so that a subject may exhibit dynamic overshoot in most of his saccades one day and in only a few the next day, and different subjects exhibit very different patterns of the variation. Dynamic overshoot seems to be independent of the size of the saccade; we have recorded saccades of all sizes (from a few minutes of arc up to 50 degrees) with and without dynamic overshoot. Moreover, dynamic overshoot does not seem to be limited to any initial eye conditions or direction of travel; it has been detected in saccades that begin with the eye looking straight ahead and in others that begin with the eye rotated as much as 35 degrees away from that position, and it has been detected in saccades directed toward the nose and in others directed away from it. The size of the overshoot increases with the size of the saccade, but in a normal saccade of, say, 10 degrees with dynamic overshoot the magnitude of the return is typically .25 degree. A comparison of the return phases of saccades having dynamic overshoot with the main-sequence data reveals that these return phases have the same main-sequence parameters as small saccades.

The eye system—the eye muscles, the eyeball and the surrounding tissue—is overdamped, as though there were large shock absorbers throughout the system to slow movement and eliminate oscillations of the eyeball. Therefore the fast return phase of a saccade with dynamic overshoot, which looks like an oscillation, cannot originate in the internal mechanics of the system; it must be of neural origin. Since the return phases are shaped like saccades, it might be expected they would be generated by saccadic motoneuron signals. To investigate this possibility we simulated saccades with dynamic overshoot on the reciprocal-innervation model.

The simulations show that a saccade with dynamic overshoot is generated when there is a reversal of motoneuron activity—an extra, reverse pulse—in the course of a standard saccadic pulse-step control signal. For the agonist muscle, then, the control signal consists of a sharp rise in motoneuron firing, followed by a sharp drop and then a rise to the new steady-state level determined by the step signal. Conversely, the signal to the antagonist muscle consists of a sharp drop in firing (the original pulse for the relaxing muscle) followed by a sharp rise and then a drop to the new steady-state level. As a consequence of these firing patterns the direction of the force generated by the active-state tensions of the two muscles is reversed in the middle of the saccadic movement, pulling the eyeball back toward a position it overshot and finally holding it there. (The active-state tension of a muscle is the force developed by the muscle not as it is measured at the tendon of the muscle but as it originates in protein interactions within the muscle.)

What is special about the control signal with the reversal of motoneuron control activity that creates saccades with dynamic overshoot? Clark and one of us (Stark) have shown that, to use the terminology of engineering optimality control, this saccadic pulse-step control signal is time-optimal, that is, there is no shape for a control signal that would move the eye faster between two points in the visual field than the one used by the brain for saccades. Recent electromyographic data obtained from human subjects by Alan Scott, Robert Kenyon and one of us (Stark) have verified that the control signal reversals predicted by the reciprocal-innervation model are indeed found in human beings. The decision to generate a saccade with dynamic overshoot would have to be made relatively early, before the signal is actually constructed. We found, however, that not all variations originate that early; for example, glissadic overshoot is generated when there is a mistake in the actual construction of the pulse-step control signal.

Remember that a saccadic eye movement with glissadic overshoot is composed of a normal saccade that overshoots its final position followed by a slow, gliding movement back to that position. The slow return movement was named a glissade by Robert Daroff and his co-workers at the University of Miami. Glissadic overshoot appears much less often than dynamic overshoot in the saccades of normal, unfatigued subjects. Its frequency is greatly increased, however, by fatigue and certain pathological eye conditions. When fatigue can be ruled out as a controlling factor, analysis of glissades can aid in the diagnosis of internuclear ophthalmoplegia, a syndrome often observed in multiple sclerosis.

We began our study of glissadic eye movements by considering how the glissades fit into the main-sequence diagram. We observed that the velocity and duration of glissades do depend on their magnitude and that those parameters fall on the vergence branch of the diagram. It seemed likely, then, that like vergence eye movements they would be driven by the force generated by a difference between an actual eye position and an eye position coded by a step change in motoneuron firing. To test that possibility we simulated saccades with glissadic overshoot on the reciprocal-innervation model.

The simulations suggested that a glissade results in instances where the pulse and step components of a saccadic control signal are mismatched. For example, glissadic overshoot is caused when the step component of a control signal is correct with respect to a particular target distance but the pulse component is too high (either because too many motoneurons have been recruited for firing or because their firing frequencies are too high) or too wide (because the duration of the pulse is too long). The pulse then propels the eye into a new position beyond the target position, but since the step signal is coded for the earlier position, the eye slowly drifts back until an equilibrium is reached. Similarly, if the pulse of a signal is too small, the eye stops short of its target position and is pulled slowly forward by the step signal. This phenomenon is called glissadic undershoot.

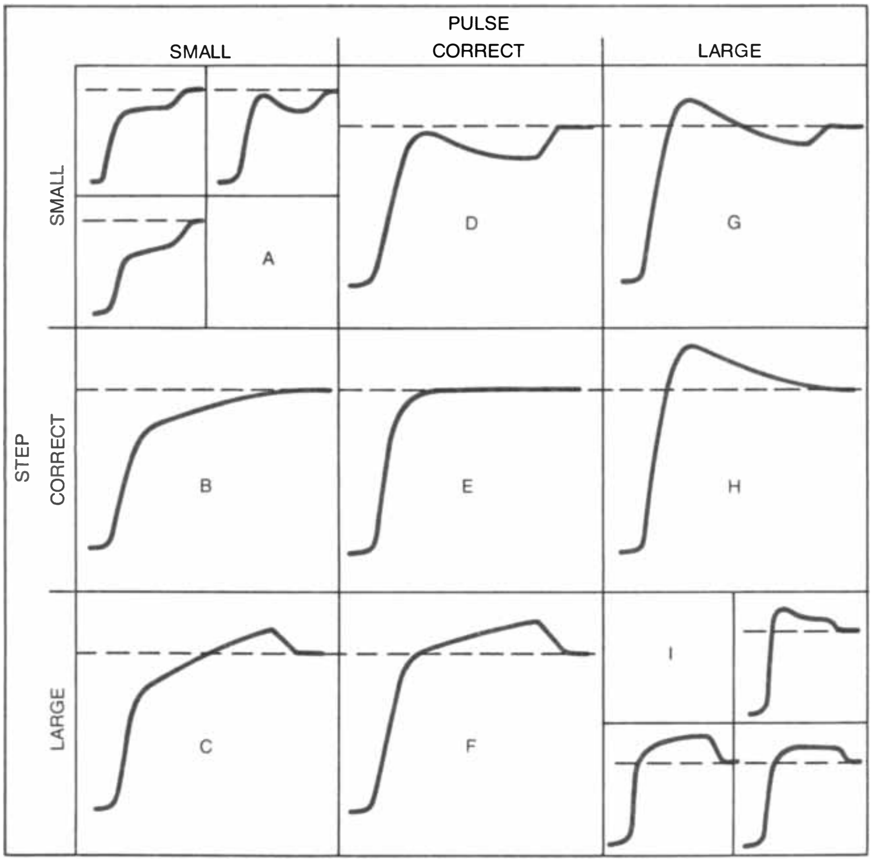
It is important to note that glissadic eye movements are not vergence eye movements. In spite of their similarities, the two types of movement differ in several ways. For example, vergences are a binocular phenomenon and glissades are usually monocular. Furthermore, there is a pause between a target jump and the resulting vergence eye movement, but there is no such pause between the saccadic error and the glissade that corrects it.

There are many different ways in which the pulse and step components of a control signal can be mismatched. In any particular saccade either one or both of the components can be undersized, oversized or correct. When it is the step component that is incorrect, being either too large or too small for a particular target position, a steady-state position error is established; the eye remains off-target for hundreds of milliseconds, until visual feedback elicits a corrective saccade. When both the pulse and step components are either too large or too small, the resulting saccadic eye movement will have a glissade to either the left or the right or will have no glissade at all, depending on which component has the larger error.

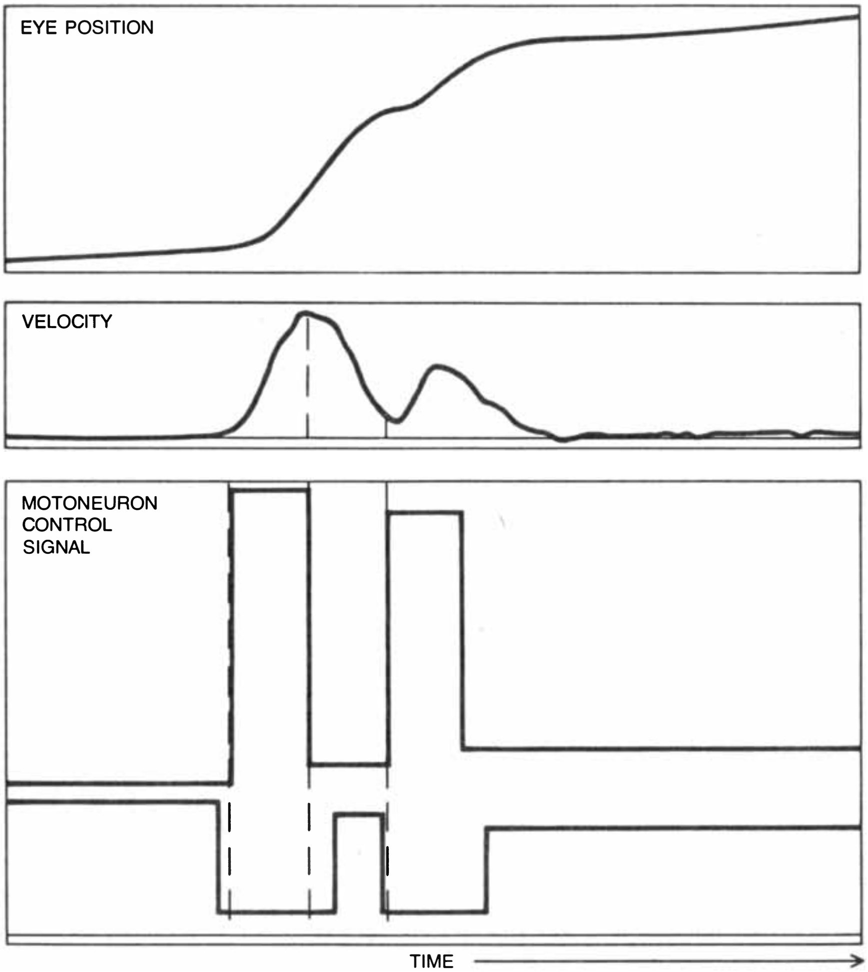
Glissades are only one of the many types of saccadic variation whose frequency is increased by fatigue. Fatigue creates irregularities in the saccadic control signals, so that saccades are slowed down and even broken up into multiple movements. After a subject has executed, let us say, 500 10-degree saccades, his fixations become more inaccurate, requiring more corrective saccades. Aberrant saccades, whose parameters do not fit on the main-sequence diagram, begin to appear. In some cases the eye changes its fixation not with one large saccade but with two smaller saccades, called double saccades. At that point if the subject is asked to make "more accurate" saccades, he can do so. After about 1,200 10-degree saccades the subject will still be able to make some normally shaped saccades, but their maximum velocities will be some 10 percent lower than those of the subject's first saccades of the day. It is at this point that overlapping saccades begin to appear: double saccades so close together that their velocity profiles overlap. All saccades the subject produces after this time will exhibit signs of fatigue.

It should be noted that larger saccades are more fatiguing than smaller ones. After making only 30 50-degree saccades or 80 30-degree saccades a typical subject cannot make normally shaped saccades. Moreover, a subject who becomes fatigued on small target jumps still shows signs of fatigue if he switches to larger target jumps, whereas a subject who becomes fatigued on large target jumps is sometimes able to execute smaller normal saccades before the effects of fatigue set in again.

We are using the word fatigue in a very general sense, because in our experiments we have no way of differentiating among muscle fatigue, sensory adaptation to stimuli and central-nervous-system habituation to impulses. In any case fatigue does bring about deviations in the neurological control signals and so creates variations in saccadic trajectories. Overlapping saccades may be the most irregular of these saccadic variations. Simulations on the reciprocal-innervation model suggest that overlapping saccades are two saccades so close together that the eyeball does not have time to decelerate to zero velocity before the second saccade begins. This does not mean the pulse portions of the two control signals overlap, however; the pulse of a normal saccade is usually completed about halfway through the movement, so that there is a brief drop in the high-frequency motoneuron activity between the two saccades in an overlapping pair. The main-sequence parameters of these pairs of saccades indicate that their generation depends on more than just a breakdown in the motoneuron signal for one saccade that would have covered the same distance as the two small saccades. If the saccades were simply a temporal decomposition of the larger saccade, then the sum of their parameters, but not the individual parameters themselves, would fit into the main-sequence data. Such is not the case. Rather, it is the individual parameters that fit into the main sequence, implying it is the magnitude and not the duration of the larger saccade that has



PULSE AND STEP COMPONENTS of a saccadic control signal can be mismatched in a variety of ways, resulting in the 13 trajectories shown here. For example, when the step is correct but the pulse is too large (*H*), a glissadic overshoot is generated (*see illustration on opposite page*). When the pulse is too small (*B*), the resulting saccade stops short of its final position and a glissade finishes the movement. When the step component is incorrect (either too strong or too weak), the saccade moves the eye to an off-target position, where it remains until visual feedback instigates a corrective saccade, either forward to the target (*D*) or backward to it (*F*). When both the pulse and the step are too large (*I*) or too small (*A*), the primary saccade may be followed by a rightward glissade, a leftward glissade or no glissade at all, depending on the relative sizes of components. Broken line through each trajectory marks final, or target, eye position.



OVERLAPPING SACCADES (*top*) are two normal saccades spaced so close together that their velocity profiles overlap (*middle*). Model simulations indicate there is no such overlap, however, in the motoneuron signals that generate the saccades (*bottom*). In most saccades pulse ends about halfway through eye movement, so that in overlapping saccades there is a drop in motoneuron activity between two signals. Fatigued subjects often show overlapping saccades.

been divided into two parts. Therefore each saccade in an overlapping pair is generated by a normal pulse-step signal for a saccade of its magnitude. Since the overlapping saccades are not the result simply of a pause in motoneuron activity, the motor control center that constructs the neurological control signals must make an early decision to abandon the single-saccade, time-optimal control strategy and construct a pair of overlapping saccades, each of which is time-optimal for its magnitude.

Let us now take up the oblique saccades, which have both horizontal and vertical components of movement. Main-sequence analysis of these components reveals that the movement in both directions is basically saccadic. Vertical saccades are somewhat slower than horizontal ones, so that they are shifted slightly from the horizontal-saccade data on the main-sequence diagram, but otherwise vertical and horizontal saccades are similar in their overall dynamical structure. The same types of irregularities appear in horizontal and vertical saccades with about the same frequency. Since there seems to be no intrinsic difference between the trajectories of the two components, there is no need to discuss the neurological origins of the vertical-saccade variations: dynamic overshoots, overlapping saccades and so on. There is, however, a most interesting question still to be answered: What is the relation between the trajectories of the two components of oblique saccadic eye movements?

The horizontal and vertical components of oblique saccades are physiologically independent, that is, they are implemented by separate neural channels with different motor control centers, different motoneurons and different eye muscles. We believe that in addition the components are temporally and dynamically independent, with unrelated saccadic trajectories generated by distinct patterns of neurological control signals. This thesis is supported by two main observations: there seems to be no relation between the timings of the two components, and the particular variations in the trajectory of one component seem to be unrelated to the particular variations in the trajectory of the other.

The temporal independence of the components is demonstrated by the fact that oblique saccades are infrequently straight. A saccade can be straight, of course, only if its horizontal and vertical components begin and end simultaneously. Main-sequence analysis of the components of typical oblique saccades shows, however, that in most cases the components of a saccade have different durations. For example, a typical five-degree saccade inclined at a 53-degree angle from the horizontal has a three-degree, 30-millisecond horizontal component and a four-degree, 40-millisecond vertical component, so that it is clearly impossible for the two components to begin and end simultaneously. The shorter component of an oblique saccade may start at any time in the course of the longer component, and on occasion the end of one component may occur before the beginning of the other, giving rise to an *L*-shaped trajectory.

The dynamic independence of the horizontal and vertical channels is demonstrated by the curvedness of oblique trajectories. In other words, the irregularities in the shape of an oblique saccade result from the differences in the dynamic structures of the two components. For example, there seems to be no connection between the appearance of, say, dynamic overshoot in the trajectory of one component and its appearance in the trajectory of the other.

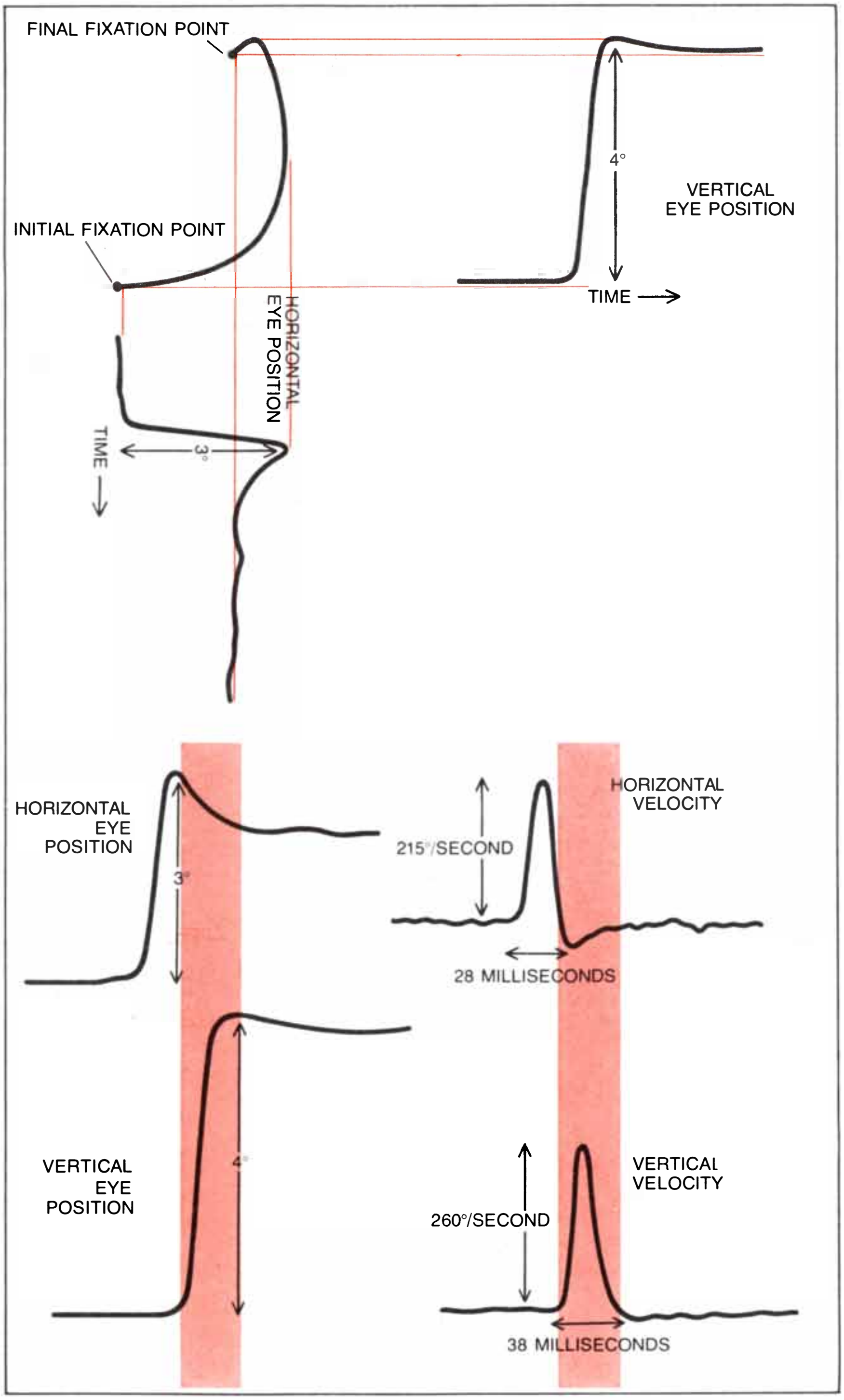
As we have shown, the variations in saccadic trajectories are generated by diverse types of irregularities in the pulse-step control signals: motoneuron-firing reversals, pulse-step mismatches and so on. Hence the dynamic independence of the components' trajectories further confirms our belief that they are generated by control signals that are in turn generated by independent motor control centers in separate parts of the brain. We have observed that the trajectories of oblique saccades are much more varied than those of horizontal saccades, which might be expected in movements where irregularities can be introduced through deviations in two different sets of neurological signals.

It should be noted that in fact some information—perhaps only noise—is exchanged between the neural channels that implement the horizontal and vertical saccadic eye movements. For example, purely horizontal target jumps rarely elicit totally linear saccades. In most cases the horizontal saccade elicited by such a target jump is accompanied by a transient vertical movement, as was noted by Raymond Dodge of Wesleyan University some 80 years ago. It appears that noise from the horizontal-movement channel affects the vertical-movement channel as well, provoking the extraneous vertical component.

Hence when the vertical components of eye movements as well as the horizontal ones are considered, a greater variability in the trajectories of saccades is observed. Successive saccadic eye movements between two fixed points usually have different trajectories, and often they have dramatically different trajectories. Even when succeeding oblique saccades have identical horizontal and vertical components, the dynamic and temporal relations of the components in each movement (and by inference the relations of the neurological control signals behind each pair of components) are almost always different. If the eye-movement system is further enlarged to include both eyes, an even greater variability is found. In particular, dynamic overshoot, glissades, double saccades and overlapping saccades are monocular phenomenona.

Surprisingly, however, the variability in saccadic trajectories does not seem to be a matter of importance to the brain. Vision is not disturbed by it, perhaps because vision is suppressed before and during a saccade. (This suppression can be demonstrated by switching one's gaze back and forth between two objects in a room. Although the image on the retina clearly moves, the room is not seen in swirling motion.) Furthermore, before each saccade the brain computes the expected change in the eye's frame of reference, so that the viewer is not subsequently aware of the shift in his retinal image. When the movement has been completed, a comparison is made between the expected image and the actual image. It seems likely that if the trajectories of saccadic eye movements became too extreme, the brain would interfere and modulate them, interposing some form of control or recalibration, possibly accomplished in the cerebellum, to make the trajectories less bizarre.

In our investigation of the trajectories of saccadic eye movements we have made extensive use of the techniques of bioengineering, as is demonstrated by our experimental apparatus, which was explicitly designed to match the high velocity of the eye; our utilization of a digital computer for data collection, data analysis and simulation; our application of the main-sequence concept to encompass the diverse data on eye movements, and our development of the reciprocal-innervation model as a means of conceptualizing the control of eye movements. With these tools we have gained a much deeper understanding of the nature and control of various eye movements. Indeed, it has been possible to identify the level of neurological control activity at which many specific types of variation in eye movement originate. As a result we have in many instances been able to show precisely how fatigue, stress and disease affect the neurological control signals for eye movement and therefore the normal functioning of the eye. Clinical applications of these advances are now being pursued, and we believe they will lead to better diagnosis and treatment of patients with a variety of physical disorders.



OBLIQUE SACCADIC EYE MOVEMENT (*top*) has a vertical component of movement as well as a horizontal component, each of which is itself a saccade. As the position and velocity plots of the two components (*bottom*) indicate, the components of oblique saccades seem to be dynamically independent, with different durations and desynchronized trajectories. In this case the horizontal saccade has dynamic overshoot but the vertical saccade does not; moreover, the horizontal saccade is nearly finished by the time the vertical saccade (*band in color*) begins.

The Coupled Motions of Piano Strings

Most of the notes on a piano are sounded by two or three strings. The strings are not tuned to precisely the same frequency, a fact that contributes in unexpected ways to the tone of the instrument

by Gabriel Weinreich

In 1709 the Italian harpsichord maker Bartolommeo Cristofori built the prototype of the modern piano in which hammers strike the strings. Since that time pianos have always been made with more than one string corresponding to most notes. The keyboard of the modern Steinway Model B has 88 keys, with the 68 highest ones setting in motion triplets of strings and the 20 lowest ones setting in motion pairs of strings or single strings. Although the tripling and pairing of strings was introduced in keyboard instruments in the middle of the 17th century to increase the volume of the sound, recent acoustical work demonstrates that in the piano, at least, the tripling and pairing also affects the quality of the sound. The work reveals that when the strings that constitute a single triplet or pair are made to sound separately, they differ slightly in frequency in a seemingly random way, even after the best piano technicians have tuned them. Remarkably, listeners have turned out to prefer the sound made by a key when there was a small discrepancy among the individual frequencies to the sound made by a key when there was no discrepancy.

My own work at the University of Michigan has focused on how these small "mistunings," or frequency discrepancies, contribute to the sound of the piano. I have also investigated other ways the tripling and pairing of strings has changed the sound from what it would have been if pianos had been built with only one string per note. Out of my work and that of other investigators has emerged a detailed picture of how the sound at first decays rapidly (the "prompt sound") and then decays slowly (the "aftersound").

When a piano key is depressed, a felt-covered hammer strikes the corresponding unison group of three or two strings. At the same time the block of soft felt called a damper is lifted from the strings so that they can vibrate freely. Releasing the key actuates the damper that stops the strings from vibrating. The sound will also cease if the vibrations are allowed to die away to inaudibility of their own accord.

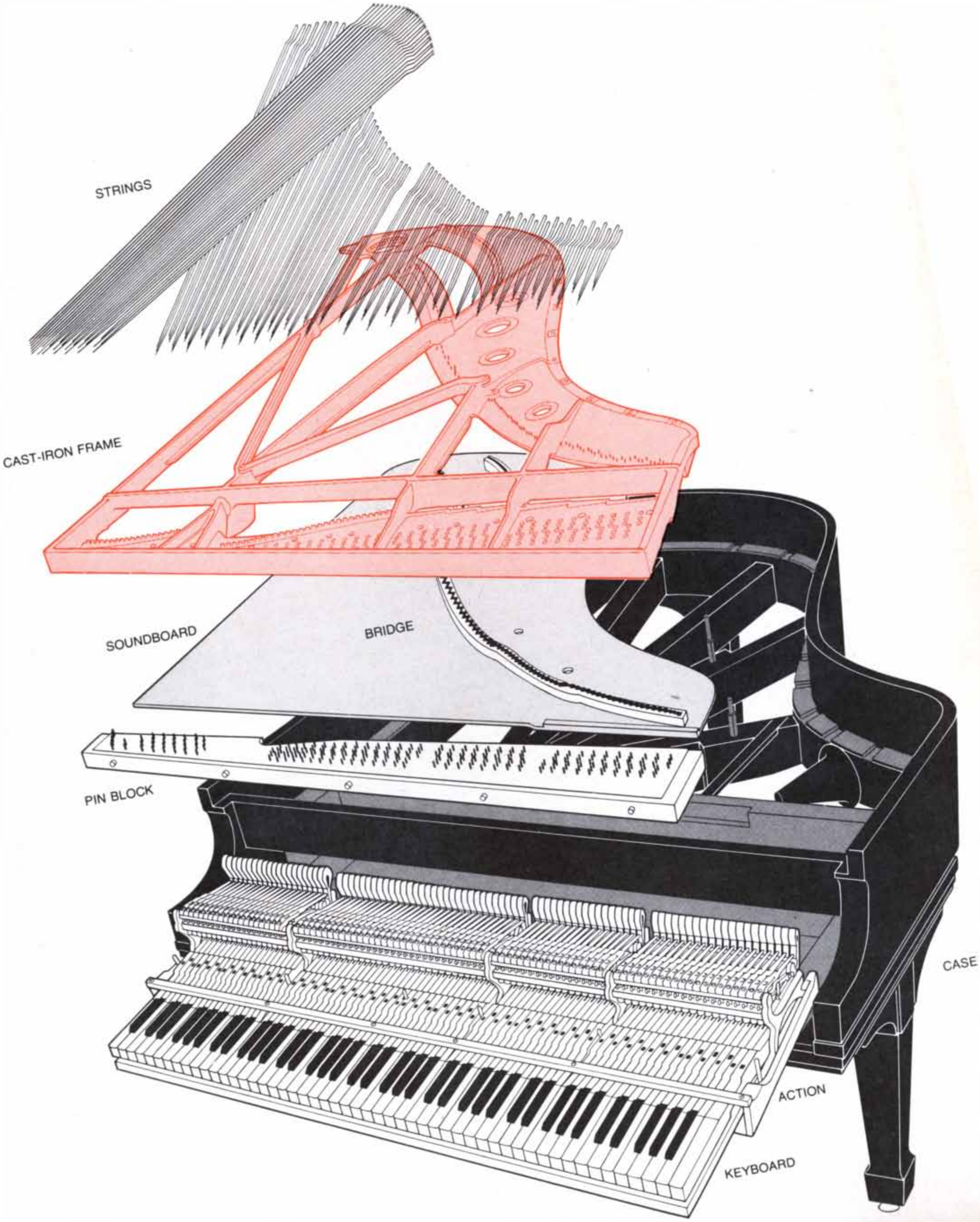
Every musical sound ultimately originates with internal vibrations of the instrument. In wind instruments such as the flute and instruments with bowed strings such as the violin the vibrations are "sustained": energy is fed into them in the course of each oscillation. In the piano, on the other hand, the vibrations are "free": no energy is fed in after the initial impact of the hammer. As a result the musical characteristics of the piano depend mainly on how the energy in the strings is dissipated.

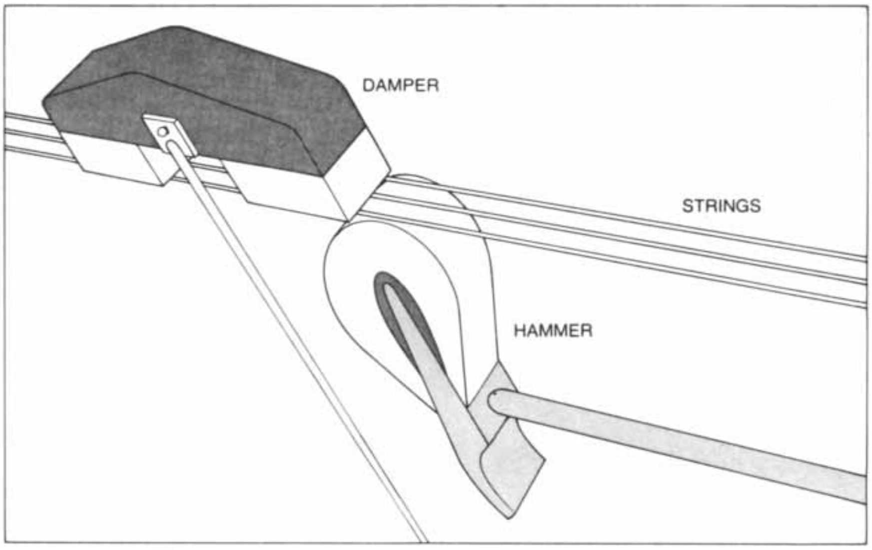
The energy is dissipated as a result of frictional forces that manifest themselves in various parts of the piano. When a string vibrates with small amplitudes, the energy dissipation is "linear": the rate at which the string loses energy is proportional to the amount of energy it contains. In such a linear system the decay of the vibration is exponential: equal fractions of energy are dissipated in equal intervals of time. The radioactive decay of carbon 14 is a familiar example of a linear system that is exactly analogous to the vibrational decay of a piano string. The rate at which a material loses carbon 14 is proportional to the amount of carbon 14 it contains. The fact that half of the carbon 14 decays every 5,730 years means the rate of decay is exponential, just as it is for the energy of a piano string.

The ratio of the pressure amplitudes of two sounds is measured in decibels, with every 20 decibels corresponding to a change in the sound pressure by a factor of 10. Hence a decay of 40 decibels corresponds to a drop in the sound pressure by two factors of 10, or a factor of 100, and a decay of 10 decibels corresponds to a drop in the sound pressure by a factor of $\sqrt{10}$, or a factor of 3.16. When the sound pressure is specified by giving its ratio (in decibels) to some standard fixed value, it is called the sound-pressure level. One advantage of such an approach is that for exponential decay a plot of the sound-pressure level as a function of time turns out to be a straight line. Sound pressure is a physical quantity that is similar to loudness but not exactly the same. Sound pressure is strictly a physical phenomenon, whereas loudness is a psychophysical phenomenon that depends not only on the physical properties of the sound but also on how the ear and the brain respond to them. For example, a doubling of sound pressure would not necessarily be perceived as a doubling of loudness. For qualitative purposes, however, it is safe to think of sound pressure and loudness as being almost the same thing.

The decay of the sound of a piano string (when the other strings in its unison group are not allowed to vibrate) is actually more complicated than the picture of a single straight line suggests. The plot of sound-pressure level as a function of time turns out to be not one straight line but a kinked line consisting of two straight segments, the first starting at a high level and decaying quickly,

EXPLODED VIEW OF STEINWAY MODEL B PIANO in the illustration on the opposite page shows the relations of the main components. The keyboard has 88 keys that are divided into seven and a third octaves. The 68 highest keys (*on the right*) set in motion triplets of strings, and the 20 lowest ones (*on the left*) set in motion pairs of strings or single strings. Connected to the keyboard is the action, which includes hammers and dampers that determine the string motions. When a key is depressed, a hammer sets the strings vibrating. The strings cross a wood bridge that transmits the vibrations to the soundboard, from which the vibrations are radiated into the air. Sound is also radiated to a lesser degree from other parts of a piano.


STRINGS
CAST-IRON FRAME
SOUNDBOARD
BRIDGE
PIN BLOCK
CASE
ACTION
KEYBOARD



HAMMER HITS THE STRINGS that correspond to one note with the same strength and at the same time. Because the hammer strikes the strings in the vertical direction, they move mostly in that direction. They also move, however, a little in the horizontal direction. This motion could be caused by small irregularities in the face of the hammer or the position of the strings.

and the second taking over at a lower level and decaying slowly [*see upper illustration on opposite page*]. This means that although the sound is always decaying exponentially, the rate of exponential decay changes fairly abruptly at one point. The initial fast decay characterizes the prompt sound and the final slow decay characterizes the aftersound. The presence of these two kinds of sound is a characteristic feature of the piano tone. The prompt sound has something of a "ping," similar to that of the xylophone, but whereas after a few seconds a xylophone becomes silent, a piano is still singing away. It is this singing that constitutes the aftersound and enables sustained melodies to be played on the piano that cannot be played on the xylophone.

The presence of prompt sound and aftersound is not the result of nonlinearity: proportionately larger frictional forces at higher sound pressures causing a higher rate of decay at the beginning. Such nonlinearities are commonplace in many other physical systems. For example, a pendulum that is subject to only a small amount of friction at its pivot point will be slowed by air friction that is nonlinear. At large amplitudes the pendulum agitates the air violently, whereas at small amplitudes it moves the air smoothly. The air friction is nonlinear because the turbulent air at large amplitudes exerts a proportionately larger force on the pendulum than the smoothly moving air at small amplitudes.

That a similar phenomenon is not the cause of the aftersound of a piano can be demonstrated experimentally. If the frictional forces were nonlinear, the break from fast decay to slow decay would take place at the same amplitude regardless of the initial displacement of the string. The initial displacement would affect only the length of the prompt sound; the larger the displacement, the longer the prompt sound. This means that plots of sound-pressure level as a function of time made for various initial displacements would be horizontal translations of one another [*see top illustration on page 122*].

In fact the break from fast decay to slow decay actually takes place at different sound pressures but at the same number of seconds after the initial displacement. In other words, the prompt sound lasts for the same amount of time regardless of how much the string is initially displaced. The sound-pressure plots are simply vertical translations of one another, not the horizontal translations indicative of nonlinear friction. In a linear system such as a piano string the amplitude of the motion can be uniformly enlarged or diminished without changing the quality of the motion.

What is it, though, that causes the change in the rate of decay? It turns out that for a single string the change results from the existence of two polarizations of vibration in the string.. I shall refer to them as vertical and horizontal, corresponding to the actual directions in a grand piano. Other motions, including circular and elliptical ones, can be thought of as superpositions of the two basic polarizations. Since the hammer hits the string in the vertical direction, it may seem strange that the string acquires any horizontal motion at all. Indeed, initially the string is moving mostly in the vertical direction. What little horizontal motion there is initially could be the result of small irregularities in the face of the hammer or the position of the string. In other words, the slight horizontal motion comes from the fact that the hammer does not strike the string precisely in the vertical direction.

I used a sensitive electronic probe to separately measure the vertical and horizontal motions of a single string. My measurements showed that each separate polarization decays exponentially but that the vertical one decays much faster [*see upper illustration on opposite page*]. This means that although the vertical motion is initially much stronger than the horizontal one (probably by at least a factor of 10), the horizontal motion eventually comes to be dominant. Hence the vertical motion gives rise to the short-lived prompt sound and the horizontal motion gives rise to the long-lived aftersound.

Why does the rate of decay differ for the two kinds of motion? The answer depends on the ways the string can lose energy. When the string loses vibrational energy, the energy can be lost to heat inside the string (internal friction), to the motion of the adjacent air (viscosity and sound radiation) or to the string's supports. In the piano this last mechanism is the dominant one. The keyboard end of the string is fixed against an iron frame, whereas the far end goes over a wood "bridge" that is glued to the soundboard. The bridge is not totally rigid because its function is to make the soundboard vibrate synchronously with the string. Most of the sound is radiated into the air from the soundboard, although it is also radiated to a lesser degree from other parts of the piano. The vertical motion of the soundboard turns out to have much more "give" than the horizontal motion, and so energy is easily transferred from the vertical motion of the string to the vertical motion of the soundboard. This accounts for the faster decay of the string's vertical motion, which is responsible for the prompt sound. In due course I shall modify this simple picture, because the mere presence of "give" does not necessarily lead to energy loss.

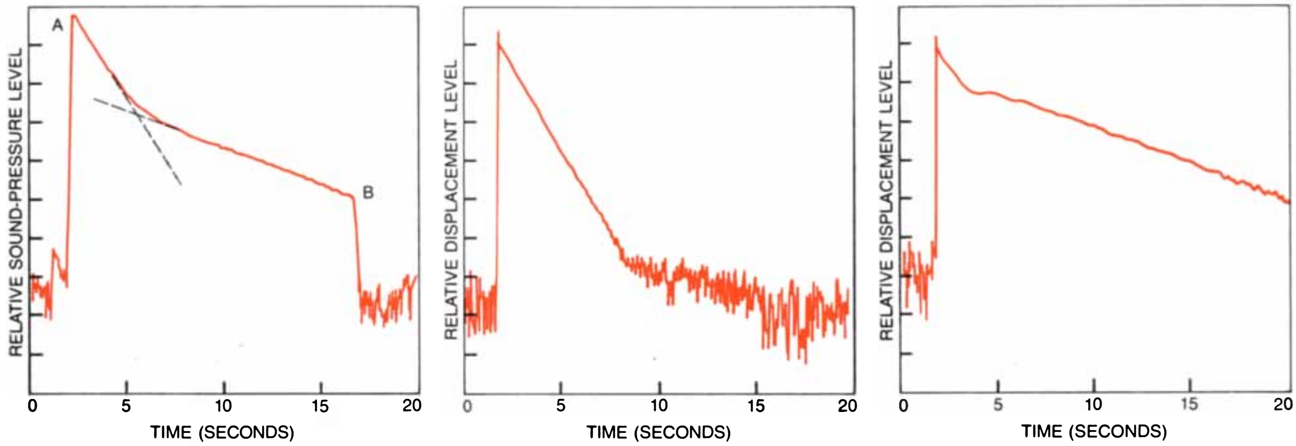
Acoustical physicists have not yet investigated in detail how the horizontal vibrations radiate. I performed a simple experiment to see whether or not the horizontal sound waves and the vertical ones emanate from the same source in the piano. (Strictly speaking, the sound waves themselves are not horizontal and vertical. Here these terms refer to the string polarizations to which the sound waves correspond.) If they emanate from different sources, then when they are of comparable strength, they interfere, or combine, to form a sound wave whose amplitude at any given point in the room equals the sum of the amplitudes of the component waves.

Where the waves are exactly in phase (crests coinciding and troughs coinciding) the amplitude is increased, and where they are exactly out of phase (crests coinciding with troughs) the amplitude is decreased. To see if I could detect this interference phenomenon I put a microphone at different places around a piano and examined the behavior of the sound-pressure level during the transition from prompt sound to aftersound. Sure enough, at the time of transition I found the points where the component waves reinforced each other and the points where they canceled each other. This means that the relative phase of the component waves is different at different points, indicating that the horizontal and vertical waves emanate from different "antennas."

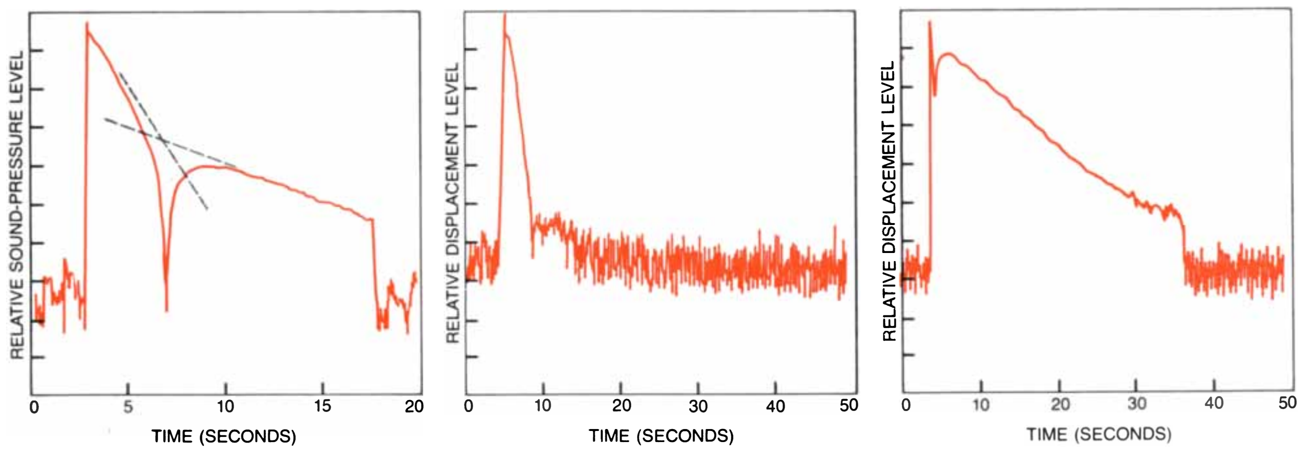
The existence of horizontal motion constitutes only one contribution to the aftersound, a contribution that is independent of the fact that the strings come in unison groups of three and two. Thus an aftersound due to horizontal motion would still be heard if pianos were built with only one string per note. The tripling and pairing of the strings, however, also contributes to the aftersound. The strings that make up a unison group cross the bridge close to one another. As a result their motions are coupled: when one string vibrates, the bridge vibrates with it and transmits the motion to the other strings in the unison group. Experiments show that the vertical displacement of one of these coupled vibrating strings sometimes decays at the slow rate characteristic of aftersound. Hence uncoupled horizontal motion and coupled vertical motion independently contribute to the aftersound.

Why does the coupled motion decrease the decay rate rather than increasing it or leaving it the same? To explain this phenomenon I must introduce the fact that the rate at which energy is dissipated through the motion of the bridge is determined not by whether the bridge can move but whether it does move. Specifically, if two strings cross the bridge at the same place and vibrate with the same frequency and amplitude but with opposite phase (one string going up while the other is going down), the net force on the bridge will be zero. Therefore the bridge will not move at all. To each string separately the bridge will seem to be completely rigid. If, on the other hand, the strings vibrate with the same phase (as well as the same frequency and amplitude), the bridge motion will be double what it would be if only one string were vibrating. Hence the decay rate also doubles. The same principles that govern the dissipation of energy in two strings govern it in three strings, although in the latter case the difficulty of following three phase factors can obscure what is going on.

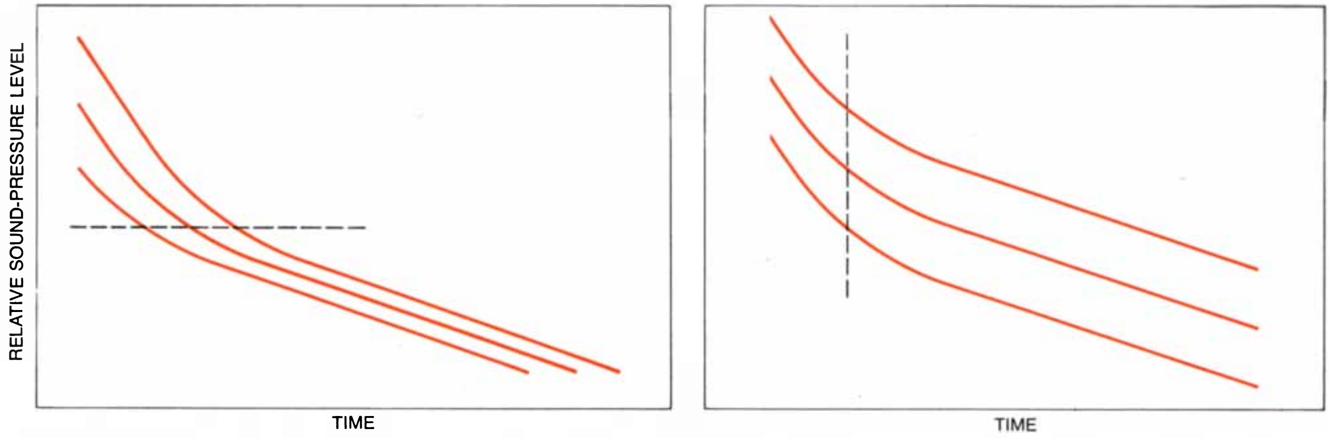
In most situations in acoustical physics



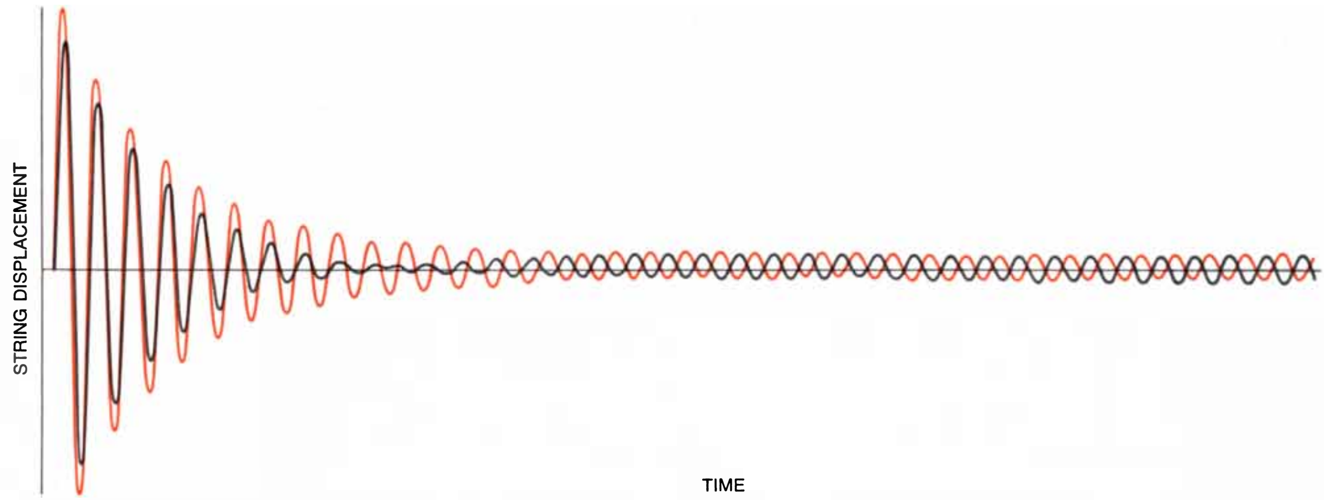
RELATIVE SOUND-PRESSURE LEVEL as measured in decibels is plotted as a function of time for a single vibrating piano string (*left*). Each division of the vertical scale corresponds to 10 decibels. At time *A* the key is depressed and the hammer strikes the strings, and at time *B* the key is released and the damper stops the motions of the strings. The plot can be resolved into two straight segments, of which the broken lines are extensions. The initial short-lived rapid decay constitutes the prompt sound, and the final long-lived slow decay, which gives the tone of the piano its characteristic singing quality, constitutes the aftersound. Noise in the electronic recording apparatus is responsible for jagged curves before *A* and after *B*. It turns out that when only one string of a unison group is vibrating, the vertical motion of the string (*middle*) gives rise to the prompt sound and the horizontal motion of the string (*right*) gives rise to the aftersound.



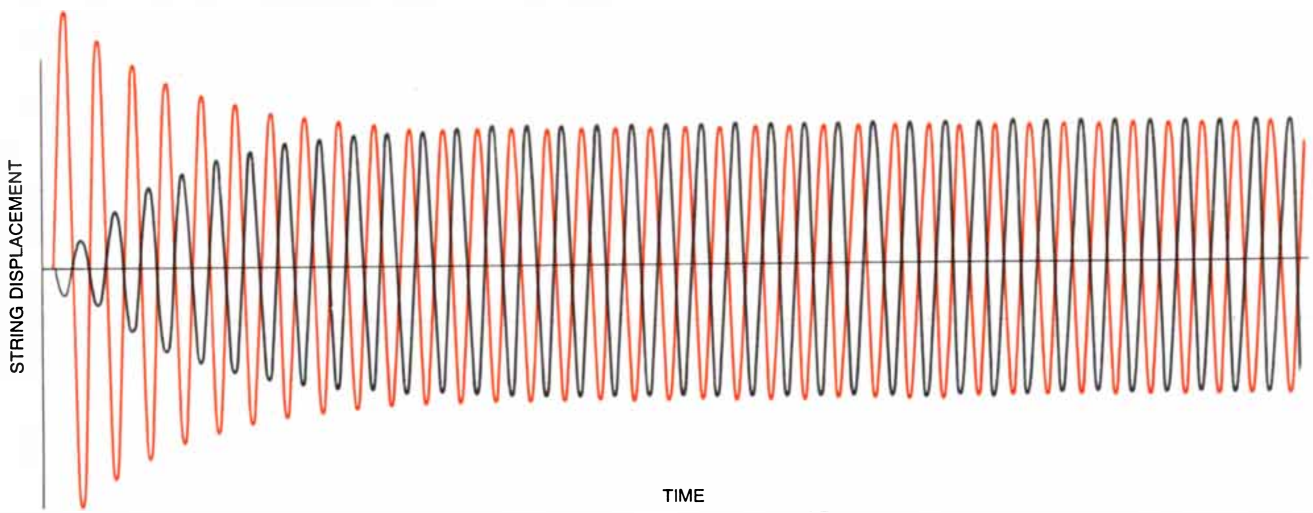
INTERFERENCE PATTERN between sound waves corresponding to vertical and horizontal vibrations of one string indicates that such waves emanate from different sources in the piano (*left*). This plot of relative sound pressure was made under the same conditions as the plot at the left in the upper illustration on this page except that the position of the recording microphone was changed. At this location sounds from the vertical and horizontal vibrations arrive with opposite phase, so that they cancel each other when their amplitudes are the same. When two strings in a unison group vibrate (*right*), the motion lasts much longer than when only one of them vibrates (*middle*).



FRICTIONAL FORCES that are proportionately larger at higher sound pressures are not responsible for the transition from prompt sound to aftersound. Such nonlinear frictional forces would cause the change to take place at the same sound pressure regardless of the string's initial displacement. In that case plots of sound-pressure level as a function of time made for various initial displacements would be horizontal translations of one another (*left*). It turns out, however, that the plots are vertical translations of one another (*right*), which is indicative of linear friction. Thus the prompt sound lasts the same amount of time regardless of the string's initial displacement.



HAMMER IMPERFECTIONS can result in string amplitudes that are not absolutely equal. Here two strings are set in motion at the same time but with the colored string having a larger amplitude than the black one. The motions of the strings start to decay, and when the amplitude of the black string approaches zero, the bridge continues to move because it is still being forced to do so by the colored string. As a result the black string not only reaches zero amplitude but also goes "beyond" it, building up a vibration of the opposite phase by absorbing energy from the bridge. Ultimately the motions are exactly antisymmetric. Such antisymmetric motion gives rise to aftersound.



UNA CORDA PEDAL, or soft pedal, increases the ratio of aftersound to prompt sound by shifting the entire keyboard so that a hammer strikes only one string of a pair. The unstruck string (*black*) starts to absorb energy from the bridge, which is vibrating synchronously with the other string (*colored*). The unstruck string immediately begins to move in phase opposite to the phase of the other string. As a result there is antisymmetric motion from the start, allowing the tone of the piano to retain its singing quality for quiet passages.

the motions of two strings are neither exactly the same (symmetric) nor exactly opposed (antisymmetric). In a piano the motions of the strings in one unison group will initially be almost perfectly symmetric, since the hammer apparently strikes the strings with the same strength at the same time. Minor imperfections in the hammer, however, will result in string amplitudes that are not absolutely equal. Consider, then, the case of two strings moving in phase but with the first string having a larger amplitude than the second one. At first both strings lose energy, and each string loses it faster than it would if it were vibrating alone, since the other one is "helping" the bridge to move.

When the amplitude of the second string (the one with the smaller amplitude) approaches zero, the bridge continues to move because it is being forced by the first string. As a result the second string not only reaches zero amplitude but also goes "beyond" it, building up a vibration of the opposite phase by absorbing energy from the bridge. Since the two strings are now moving in opposite phase, the bridge motion is less than it would be if one string were moving in the absence of the other. The two amplitudes asymptotically approach each other with opposite phase. Ultimately the motions of the two strings are exactly antisymmetric. It is the initial symmetric motion of the strings that constitutes prompt sound and the later antisymmetric motion of the strings that constitutes aftersound.

A useful way of looking at the situation I have just described is to think of the original motion as a superposition of two kinds of motion: a symmetric motion and an antisymmetric one. Let me illustrate what I mean by superposition. If I have 10 apples and you have six apples, we can describe the situation as a superposition of a symmetric state where we each have eight apples and an antisymmetric state where I have two apples and you have minus two apples. This is silly in the case of apples but useful in the case of vibrations, because it enables us to think of the symmetric component decaying at its characteristic rate, and the antisymmetric one decaying at a much lower rate (or, in the ideal case, not decaying at all). The algebraic sum of the two string amplitudes drops toward zero, but the difference remains constant for a long time.

Because the most general motion of two piano strings can be expressed as the superposition of symmetric motion and antisymmetric motion, the two kinds of motion are the normal modes of the piano-string system. It is interesting to note that the break in the decay of piano-string vibrations is not a unique phenomenon in physics. In fact, the decay characteristics are precisely analogous to those of such elementary subatomic particles as neutral kaons. Two kinds of kaon (K^0 and $\bar{K}^0$), a particle-antiparticle pair, can be formed through the strong interaction in nuclear collisions. In studying kaon decay physicists have identified two other varieties of kaon: K^0_S, which decays rapidly, and K^0_L, which decays slowly. It turns out that K^0_S and K^0_L are respectively symmetric and antisymmetric superpositions of K^0 and $\bar{K}^0$, just as the prompt sound and aftersound are respectively symmetric and antisymmetric superpositions of the motions of two piano strings. As a result a beam that initially consists only of K^0 particles will also have a kink in its decay curve.

The phenomenon of antisymmetric motion in a piano also accounts for the function of the una corda pedal, or soft pedal. The normal aftersound is about 20 decibels below the initial level of the prompt sound, a ratio that is apparently pleasing to the ear. This ratio, however, is not adequate for very quiet passages. When the piano is played softly, so that the amplitude of the prompt sound approaches the amplitude of the background noise in the concert hall, the aftersound becomes inaudible. If the notes are long, the piano will lose its sustaining quality and sound like a xylophone. To prevent this the piano is equipped with the una corda pedal, whose mechanical function is to shift the entire keyboard so that a hammer strikes only two strings of a unison triplet. Instead of exciting almost exclusively the symmetric motion with only a trace admixture of the antisymmetric motion, the una corda pedal excites both kinds of motion almost equally.

Why is this? The third string that was not hit by the hammer starts to absorb energy from the bridge, which is vibrating synchronously with the other two strings. The third string begins immediately to move in a phase opposite to the phase of the other two strings. As a result there is antisymmetric motion from the start. Therefore the level of aftersound with respect to prompt sound is markedly increased and the singing quality of the piano is restored.

The aftersound that comes either from antisymmetric motion or from horizontal polarization is quite soft compared with the prompt sound. And since such aftersound arises out of structural irregularities, it probably varies erratically from note to note. The mistuning of strings that constitute a unison group is a third mechanism that contributes to the aftersound. This mechanism is adjustable, however, and a skilled piano technician probably varies the mistuning to compensate for the erratic effects of the structural irregularities in order to equalize the strength of the aftersound from note to note. To analyze how mistuning affects the aftersound, a distinction must be made between this phenomenon and the phenomenon of "beats." If two independent oscillations whose frequencies differ slightly are added together, they will alternate slowly between a state of reinforcement (when they have the same phase) and a state of cancellation (when they have opposite phase). To the listener this sounds like a steady pitch with a pulsating loudness, which is what the word beats refers to. In a piano, however, the two strings do not vibrate independently. The motion of the bridge causes the vibration of one string to affect the vibration of the other. As a result not only the frequencies but also the decay rates are markedly affected.

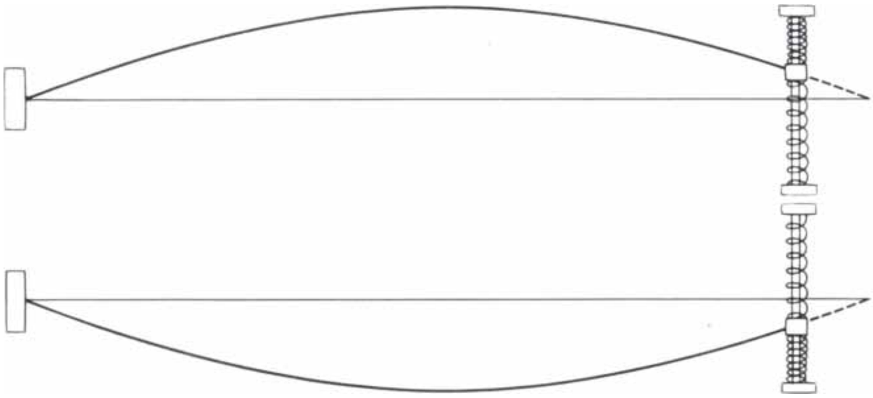
The mere motion of a support does not automatically lead to the dissipation of energy. In certain physical systems energy is not dissipated but is simply transferred back and forth between various subsystems. Consider a string attached to a ring that can slide up and down without friction on a fixed rod, and assume that the ring is sandwiched between two coil springs that act to keep the ring in its central position. When the string pulls up on the ring, the ring moves up, and when the string pulls down on the ring, the ring moves down. The motion of the support simulates the motion that would be executed by an extra piece of string attached in turn to a perfectly rigid support. Hence the effect of a "springy" support is to make the string move as if it were longer than it really is, and so to lower the frequency of the string.

A springy support does not, however, damp the motion of the string, because in the course of each complete cycle energy that flows into the support flows back into the string. As the string pulls the ring away from its central position against the force of the springs the string is doing work on the ring. On the other hand, as the ring returns to its central position assisted by the restoring force of the springs the ring is doing work on the string. Therefore there is no net energy transfer.

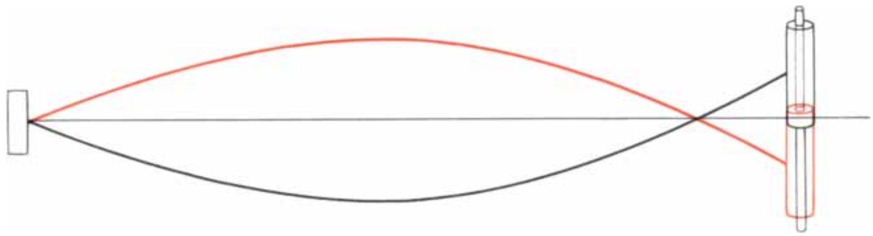
There would also be no net energy transfer in the case of a string attached to a massive block that can slide up and down without friction on a fixed rod. Here the motion of the block is governed not by a restoring force (since there are no coil springs) but by inertia. It is assumed for the sake of simplicity that gravity plays no role. Inertia acts to keep the block moving in whatever direction it is going. When the block reaches its maximum displacement in one direction, the string pulls back on it against inertia in order to slow it down and start it moving the other way. Inertia then propels the block through its central position to its maximum displacement in this direction. Once again the string pulls back on the block against inertia and sends it moving in the original direction, and the cycle continues to repeat itself.

The fact that the string is often pulling back on the block makes the string "think" it is shorter than it really is. Thus a massy support raises the frequency of the string. Like a springy support, a massy support does not damp the motion of the string. The work the string does on the block while pulling it back against inertia to reverse its direction is equal to the work the block does on the string while pulling the string along as inertia propels the block toward its central position.

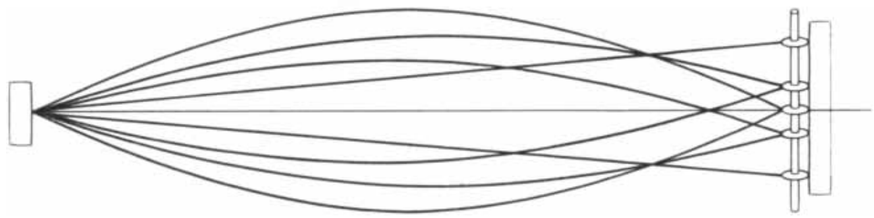
The idealized cases of a perfectly springy ring and a perfectly massy block indicate that supports can move without dissipating energy. Therefore the mere motion of the bridge in a piano does not indicate that the strings are losing energy. The bridge actually resembles a third idealized case: a "resistive" support where the phase difference between the displacement of the support and the force on it is a quarter of a cycle. In this case the frequency of the string remains the same but its motion is damped. An example of a perfectly resistive support is a ring whose motion is governed not by coil springs or by inertia but by friction. To overcome friction the string is constantly doing work on the ring, and so the string's energy is dissipated. That the phase shift is a quarter of a cycle is a shorthand way of saying that when the ring reaches its maximum displacement in either direction, the string reaches zero displacement (its central position), and when the string reaches a maximum displacement in either direction, the ring reaches zero displacement.



SPRINGY SUPPORT lowers the frequency of a string without damping the motion. An example of an ideal springy support is a ring sandwiched between two coil springs that slides up and down without friction on a fixed rod. The support lowers the string's frequency because it simulates motion that would be executed by an extra piece of string: the ring reaches maximum displacement when the string does and reaches zero displacement when the string does.



MASSY SUPPORT raises the frequency of a string without damping its motion. An example of an ideal massy support is a massive block that can slide up and down without friction on a fixed rod. Here the motion of the block is governed not by the restoring force of coil springs but by the effects of inertia. At the positions of the block shown here the string must pull back on the block against inertia in order to reverse its direction. The fact that the string is often pulling back on the massive block means the string "thinks" of itself as being shorter than it really is. As a result the ideal massy support has the effect of increasing the frequency of the string.



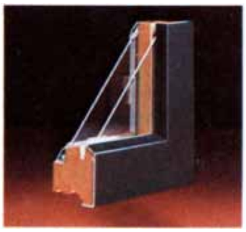
RESISTIVE SUPPORT leaves the frequency of the string undisturbed but damps its motion. An example of a perfectly resistive support is a ring that slides up and down on a rod but whose motion is retarded by friction against a wall. The string's motion is damped because the string is constantly doing work on the ring to overcome the friction between the ring and the wall. The phase difference between the displacement of the support and the force on it is a quarter of a cycle: when the ring reaches a maximum displacement, the string reaches zero displacement; when the string reaches a maximum displacement, the ring reaches zero displacement.

Let me explain how the ideal resistive situation exhibits the characteristics of the aftersound of mistuned piano strings. What happens if two strings are started in exactly antisymmetric motion but with frequencies that are not quite identical? At first the support is not moving, since the initial string motion is antisymmetric. The string with the higher natural frequency begins to advance in phase over the other string, and so their motion is no longer purely antisymmetric. As a result the strings exert a small force on the bridge. The phase difference between the force and the motion of each string is indeed a quarter of a cycle [*see top illustration on page 126*]. The strings reach points of maximum displacement when the force is smallest and points of minimum displacement when the force is largest. The former relation holds because at the points of maximum displacement the strings' amplitudes are opposite, and so they cancel each other to produce the smallest force on the bridge. The latter relation holds because at the points of minimum displacement the strings' amplitudes are of the same sign, and so they add together to exert the greatest force on the bridge.

On the assumption that the bridge is a purely resistive support the bridge develops in turn a small motion that is a quarter of a cycle out of phase with the force. Hence the motion of the bridge is in phase with the motion of one of the strings and in opposite phase with the motion of the other. The in-phase string "sees" the bridge as a springy support, whereas the opposite-phase string sees the bridge as a massy support. This means that the frequency of the in-phase string is raised and the frequency of the opposite-phase string is lowered. It turns out that the string with the lower original frequency will have its frequency raised and the other string will have its frequency lowered, so that both end up vibrating at precisely the same frequency. The decay rate, however, is no longer zero, as it was for pure antisymmetric motion where the strings vibrate at exactly the same frequency. In other words, the mistuning generates a sound of a single frequency that decays slowly.

If two strings are started with perfectly symmetric motion but with frequencies that are not quite identical, one string will begin to fall behind the other in phase. As a result the bridge motion will not be exactly a quarter of a cycle out of phase with the motion of either string, as it would be if the strings had continued to move in a perfectly symmetric fashion. The frequency of each

So easy to wash. Thanks to Pella's unique hinging system, the sash pivots straight out and moves toward the center of the frame providing plenty of room to clean outside glass from inside the room.



Aluminum exterior needs no painting. Pella offers an optional jacket of protective aluminum cladding in a choice of colors outside, with the beauty and natural insulation of wood inside.

Instant privacy or shade. Another energy-saving option, the exclusive Slimshade® fits in the dust-free space between the panes of the Pella Double Glass Insulation System . . . helps reduce heat loss and solar heat gain.

Only Pella offers it all.

FREE. 16-page "Window & Door Ideas" book for traditional and contemporary homes. Tells you everything you need to know about window types and how each works, window treatment in various architectural styles, and important considerations in window placement and selection. Also contains information on how to eliminate screen and storm chores, reduce fuel costs and outside maintenance. Because Pella Windows are unique, they're sold only through Pella Planning Centers. The one closest to you is in the Yellow Pages under "windows".

Please send me a copy of "Window & Door Ideas".

☐ I plan to build. ☐ I plan to remodel.

☐ I would like an appointment at the Pella Planning Center nearest me; have someone contact me.

Name

Address

City

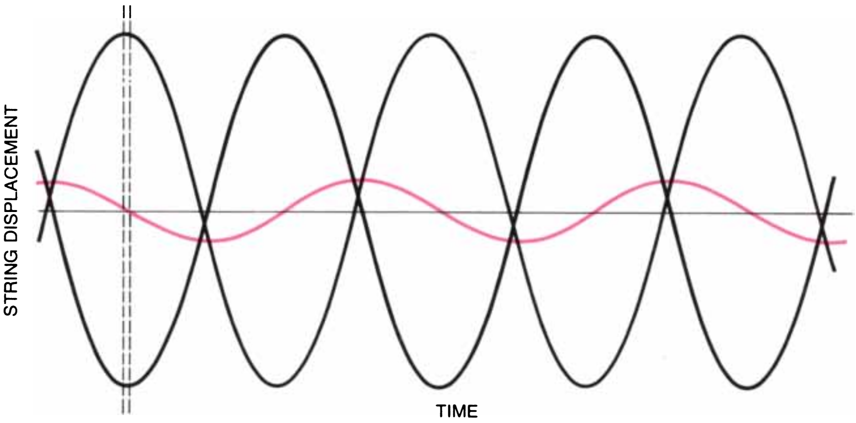
State Zip

Telephone

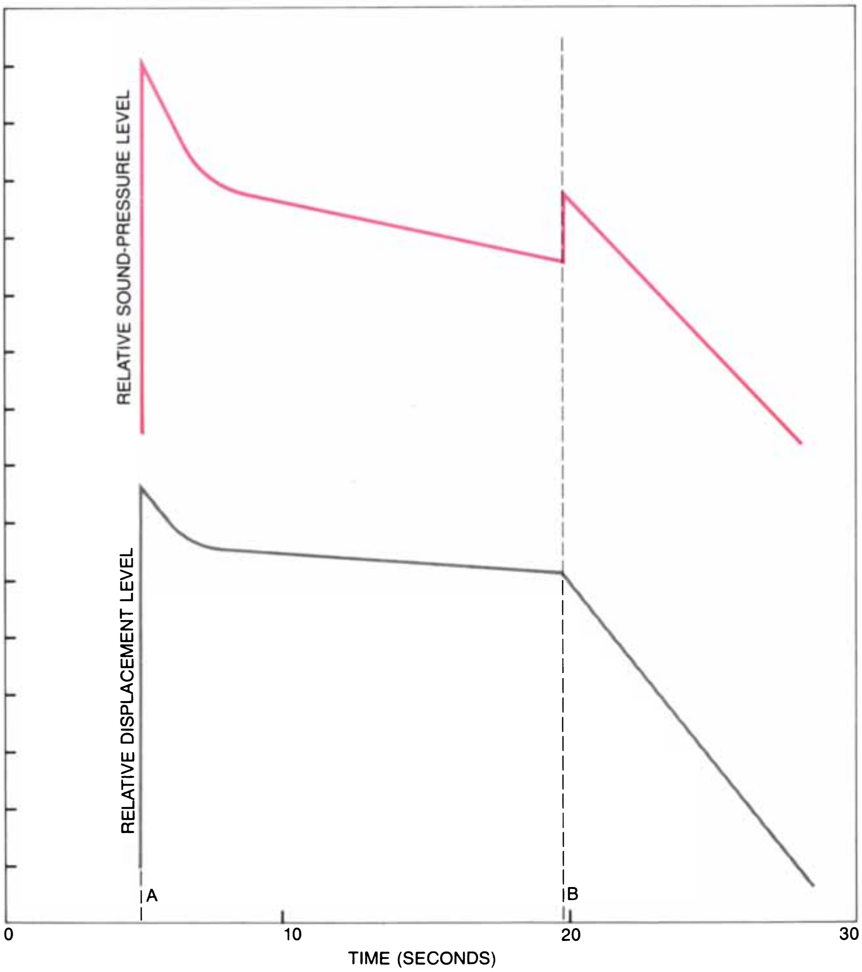
Mail to: Pella Windows & Doors, Dept. C17A9. 100 Main St., Pella, Iowa 50219.

Also available throughout Canada. Coupon answered within 24 hours.

© 1979 Rolscreen Co.



RESULTANT FORCE (*colored curve*) **on the bridge of a piano is proportional to the algebraic sum of the strings' displacements. When the motions of the two strings (*black curves*) are almost perfectly antisymmetric, the resultant force on the bridge is about a quarter of a cycle out of phase with the motion of either string. That the phase shift is a quarter of a cycle is a shorthand way of saying that the resultant force is smallest when the strings reach points of maximum displacement and greatest when the strings reach points of minimum displacement.**



KNOWLEDGE OF ANTISYMMETRIC MOTION makes it possible to construct a piano that could introduce an accent into the middle of an otherwise sustained note. The piano could be made with split dampers that would separately stop the motion of each string in a unison group. At time *A* a hammer sets in motion the two strings and at time *B* one of the strings is damped. The top plot shows the relative sound-pressure level as a function of time, and the bottom plot shows the relative displacement level of the undamped string. From *A* to *B* the curves are typical ones that decay rapidly at first and slowly at the end. When one of the strings stops moving at *B*, the strong antisymmetric motion of the two strings abruptly stops and the undamped string immediately starts to decay at the original rapid rate. The sound pressure experiences a sudden increase because there is scarcely any antisymmetric motion to retard the motion of the bridge. The sudden increase gives rise to an accent in the middle of the note.

string is raised or lowered, depending on whether the phase difference between the motion of the string and the motion of the bridge is closer to the phase difference that characterizes a springy support or closer to the phase difference that characterizes a massy support. Once again no beat is heard as the frequencies are brought together. Since the slight mistuning introduces a trace of antisymmetric motion, the damping is a little smaller than it is in the perfectly symmetric case where the precisely tuned strings cooperate fully in moving the bridge.

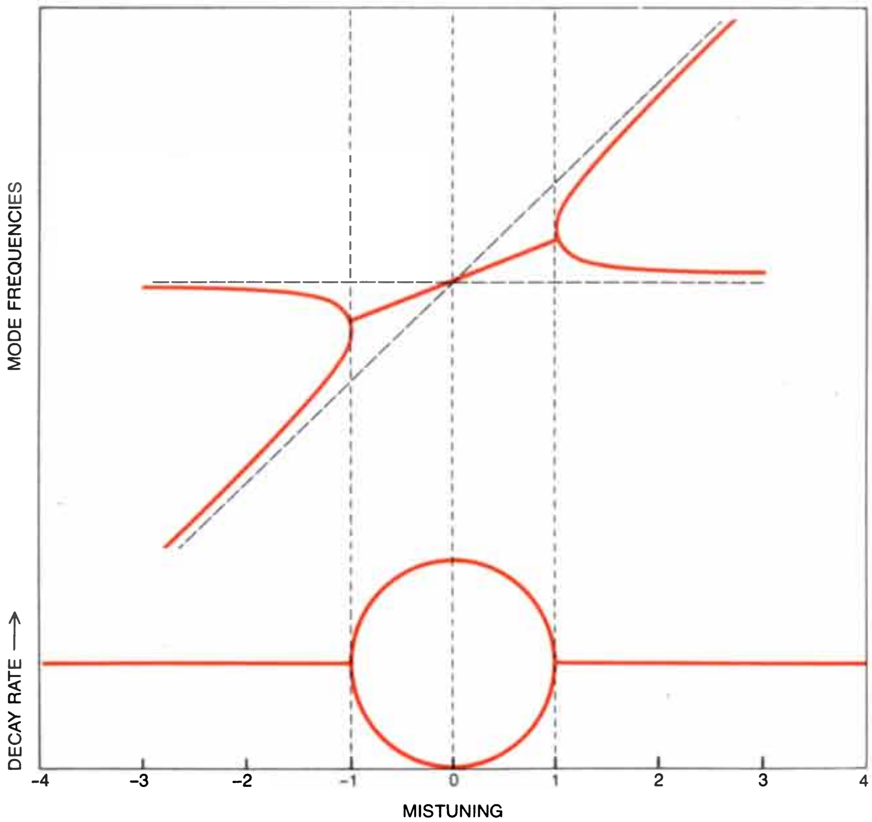
In the cases of initial symmetric motion and initial antisymmetric motion the presence of resistive coupling tends to lock together the frequencies of the two strings but to alter the decay rates. Of course, there is a limit to how far apart the original frequencies can be. As the mistuning is increased the phase difference between the strings increases until it reaches a quarter of a cycle, where the frequencies break apart. At even greater phase differences beats are heard, and the decay rates of both the symmetric motion and the antisymmetric motion become equal to the decay rates for uncoupled strings.

When two strings are in tune, their motion can always be expressed as the superposition of symmetric and antisymmetric modes. When two strings are slightly mistuned, their motion can still be expressed as the superposition of two modes: an almost antisymmetric mode whose damping is small, although not quite zero, and an almost symmetric mode whose damping is large, although not twice as large as the single-string rate. In both modes the amplitudes of the two strings are equal. If a hammer strikes the strings at the same time and with the same strength, the exactly symmetric motion the hammer excites is not a normal mode; rather, it must be viewed as a superposition consisting mostly of the rapidly decaying (almost symmetric) mode but containing an admixture of the long-lived (almost antisymmetric) mode as well. The amount of this admixture depends on how different these normal modes are from perfect symmetry and perfect antisymmetry, which depends in turn on the extent of the mistuning.

The major difference between the contribution to the aftersound due to the mistuning and the contributions due to the horizontal polarization and the antisymmetric motion is that a skilled piano tuner can adjust the former but not the latter. I think this explains seemingly random variations in unison tuning that were observed by Roger E. Kirk of the D. H. Baldwin Company. A skilled piano tuner varies the mistuning in such a way as to make the aftersound uniform and smooth from note to note by compensating for the irregular effects of the horizontal polarization and the antisymmetric motion. In this way the piano attains its characteristic beauty of tone that less skilled piano tuners are unable to induce. To test my hypothesis I would have one tuner tune the same piano a number of times, with someone else detuning it in between. If the same "random" mistunings manifested themselves each time, it would prove that the mistunings were not random at all.

Piano physics has now reached the stage where each step forward raises more questions than it answers. For the investigator this is an extremely exciting stage. The trial-and-error method that has historically characterized the development of musical instruments is particularly inefficient for such a huge acoustical structure as a piano, where the investment required for a new design is so large that it discourages experimentation. For this reason the emergence of a detailed physical picture of the workings of the piano promises to have a tremendous impact on piano technology. Even the present incomplete picture suggests innovations. For example, the understanding of antisymmetric motion points to the construction of a piano that could introduce an accent into the middle of an otherwise sustained note.

Picture a piano with split dampers that could separately stop each string in a unison group. Perhaps a special pedal would control the split dampers. Now consider a unison group of two strings. When the corresponding key is depressed, a note is heard that has a typical mixture of prompt sound and aftersound. After a few seconds the symmetric component of the motion has completely died away and only aftersound can be heard. At this point the special pedal is depressed that damps the motion of one of the strings. As a result the strong antisymmetric motion of the two strings abruptly stops and the undamped string immediately starts to decay at the original rapid rate. The sound pressure suddenly increases, as the amplitude of the motion of the bridge soars in the absence of the retarding effect of the antisymmetric motion of the strings. In this way a sharp accent is introduced into the middle of the otherwise sustained note. Other ways of controlling tone quality will become apparent once the physics of the piano is completely worked out.



FREQUENCIES OF A PAIR OF STRINGS lock together when the motions of the strings are coupled through a purely resistive support. The mistuning, or difference between the uncoupled frequencies, is given in "natural units," which are related to the single-string damping rate. For a typical pair of strings in the middle of the keyboard, one natural unit is about a third of a vibration per second. The broken lines in the top graph indicate the frequencies in the absence of coupling. The point where the broken lines cross each other is where the two strings have exactly the same frequency. In a piano the presence of a purely resistive support causes frequencies with a mistuning of either +1 or −1 natural unit to come together and lock at a common frequency. For smaller mistunings the frequencies stay locked but the decay rate, which equals the single-string rate for larger mistunings, splits for the two strings (*bottom*).

The Head of the Sperm Whale

It can represent a quarter of the animal's length and a third of its total weight. The oil-filled spermaceti organ housed within it may keep the whale neutrally buoyant during dives

by Malcolm R. Clarke

Among the great whales the sperm whale is most clearly recognizable in having a head that seems disproportionately large. The sperm whale's head can make up more than a third of the animal's total weight (50 tons for the average adult male) and more than a quarter of its total length (an average of 60 feet). There is good reason for this apparent disproportion. The sperm whale's skull accounts for perhaps 12 percent of the weight of the head. The other 88 percent consists mainly of a peculiar anatomical feature located in the whale's snout above the upper jaw: the spermaceti organ. The organ is a complex mass of muscle and oil-filled connective tissue. The oil is what gave the sperm whale its name; in a large male the organ may hold four tons of spermaceti oil.

Such a mighty organ clearly must have a very important function in the life of the sperm whale. The purpose of the great "case," as whalers often call it, has long been a subject of speculation. Even the structure of the spermaceti organ was scarcely known until the past decade or so. Before then guesses about its function could rest only on generalities and on what was known about similar but far smaller organs in a few other toothed whales. Today enough is known about the organ to make it possible to suggest its main purpose: to enable the whale to remain neutrally buoyant when it is submerged.

Why were the anatomical details of the organ not known much earlier? Part of the answer is that pioneers in the field gathered their data by dissecting sperm-whale fetuses. It happens, however, that both the skull and the snout of the fetus differ greatly in proportion from the same components in the head of the adult whale. Another part of the answer is that it is no small task to dissect an adult sperm whale. Without such commercial whaling facilities and tools as flensing platforms, steam winches, five-meter steam-driven saws and razor-sharp flensing knives the dissection would be quite impossible. Even with such aids and the cooperation of the commercial whalers this is not a trivial task.

On the flensing platform half-ton masses of fat, flesh and fiber are cut, rolled and pulled off this way and that until the observer's sense of orientation is easily lost. It is no wonder that even in this decade a book has been published that presents the principal structures of the sperm whale's head upside down. Only after watching and photographing many sperm whales being cut up on the flensing platform could one hope to clarify the anatomy of the adult whale's snout, and that is the task I undertook. My work was greatly advanced by the capture of one small adult whose head the commercial whalers cut for me in a series of transverse sections 20 centimeters thick. I was then able to photograph the sections and measure them in detail.

Many functions for the spermaceti organ have been suggested. Among them are that it is a means of generating and focusing sounds (and for receiving them), a means of moving air between the whale's lungs and its nostrils when the whale is deep underwater, a means of opening and closing the whale's long nasal passages, a means of absorbing nitrogen from the bloodstream in the course of deep dives and even a means of attack and defense.

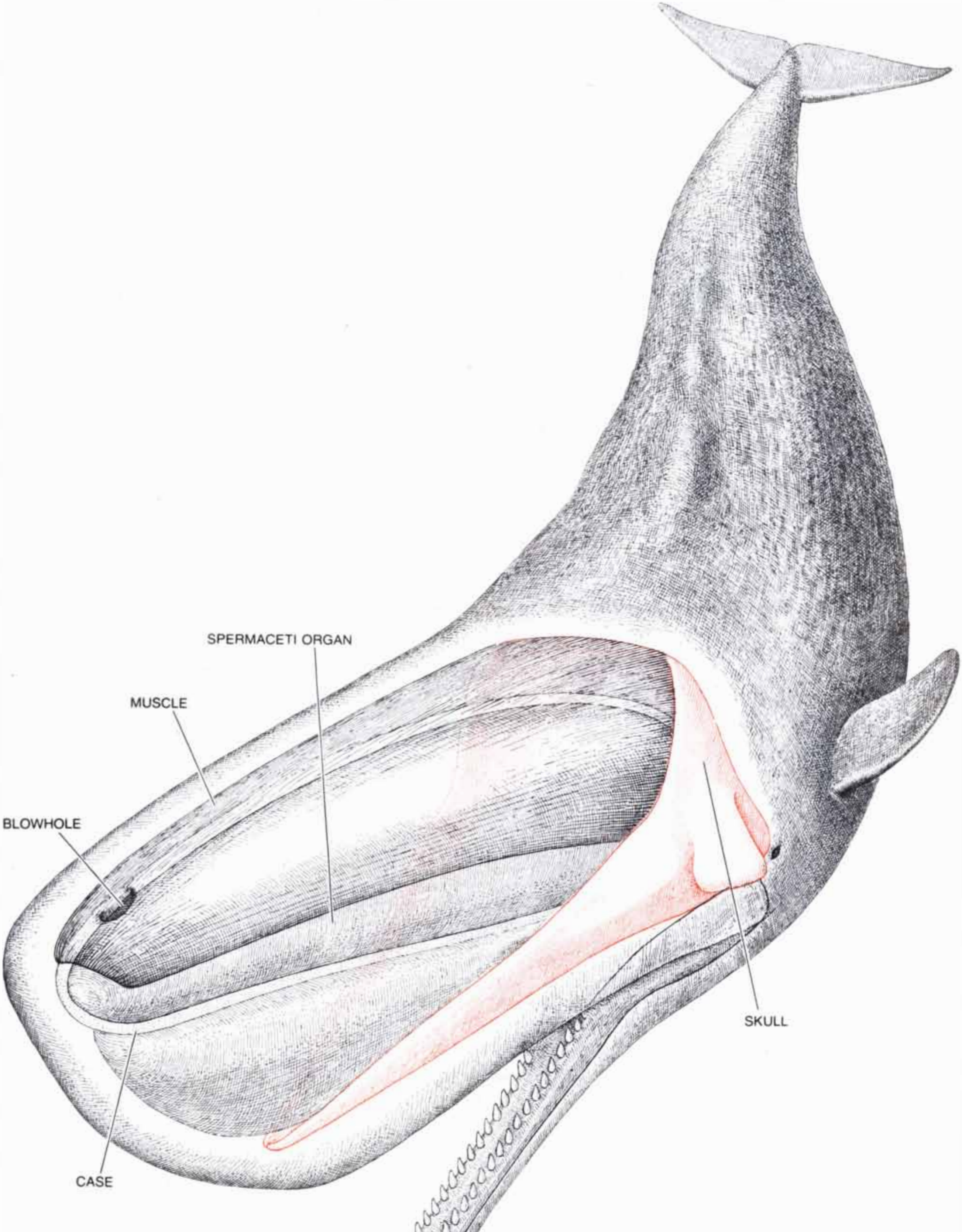
Certainly this complex organ may have more than one function. Variations in the chemical composition of the spermaceti oil in different parts of the organ suggest that one function may well be the channeling (or focusing) of sound generated by the whale. At the same time it is hard to accept the suggestion of one worker that such focusing can concentrate sound intensely enough to stun the squids that are the sperm whale's main prey. More than one of the other suggestions, however, seem reasonable, if difficult to demonstrate.

Although an understanding of the structure and proportions of the head of the adult sperm whale is a necessary preliminary to studying the function of the spermaceti organ, one must also know something about the biology and behavior of sperm whales, in particular how these toothed whales differ from other whales that do not have a large snout. Moreover, the large quantity of oil contained in the spermaceti organ suggests that the oil itself must serve some special function; one needs to know something of the physical properties of the oil to understand what role it plays in the life of the sperm whale.

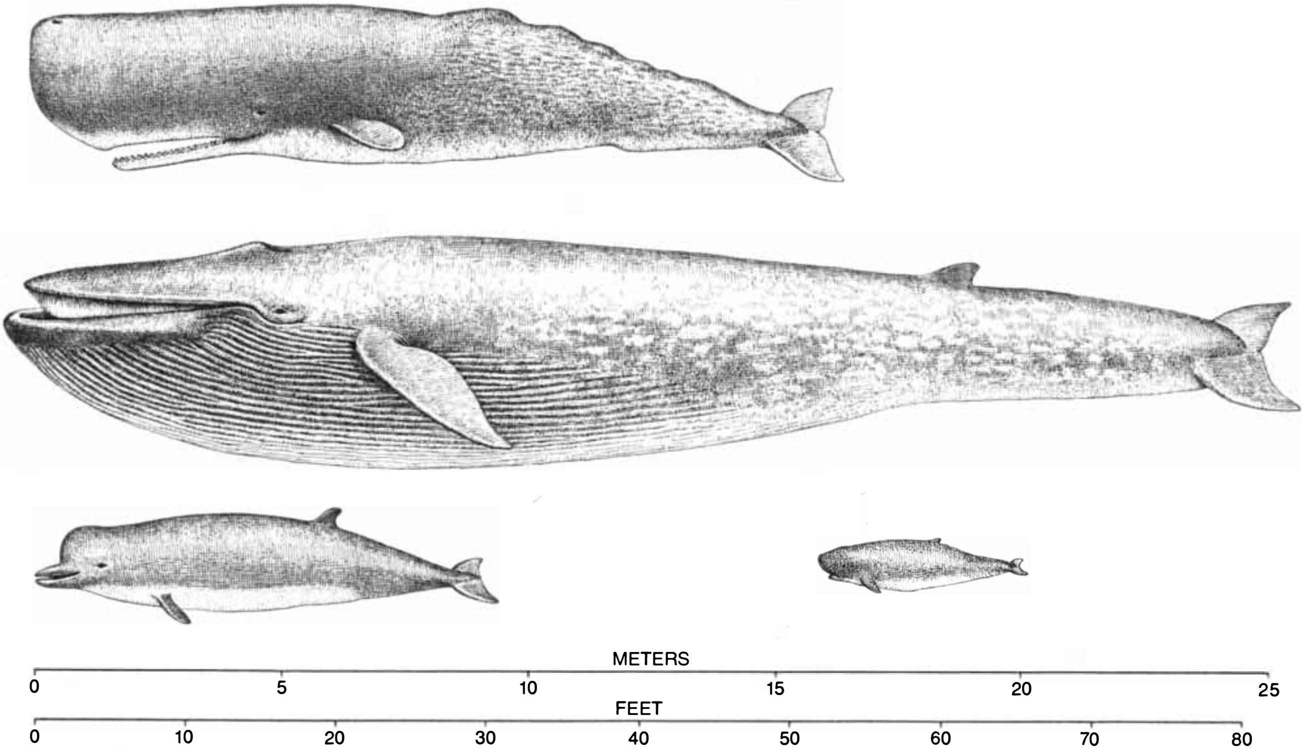
Let us begin by reviewing the biology and behavior of the sperm whale in search of clues to the function of the animal's snout. One immediately apparent fact is that sperm whales are unusual, although not unique, on the roster of toothed whales in being distributed worldwide. They have been hunted for centuries in every ocean from as far as 60 degrees north latitude to 40 degrees south. Since the development of modern whaling techniques the sperm whale has also been pursued to the high latitudes of the Antarctic. Together with its baleen-whale cousin the right whale, the sperm whale was of particular importance when the whaling industry depended on hand harpooning from open boats: unlike many other whales, these two stayed afloat after being killed.

The food of the sperm whale consists almost entirely of that Concorde of the snail family, the squid. To catch these speedy denizens of the depths the sperm whale dives deep and stays down for long periods. A large sperm whale is typically submerged for 50 minutes of an hour-long diving cycle. During a 10-minute surface interval between dives the whale will take 50 to 60 breaths of air. Dives longer than 50 minutes have been observed; the record dive is somewhat more than 80 minutes.

Sperm whales not only stay submerged for long periods but also frequently go deeper than 1,000 meters. One sperm whale, watched by sonar, was observed to go below 2,250 meters; even deeper descents can be inferred from the presence of bottom-dwelling



SPERM WHALE'S SNOUT, which can weigh more than 12 tons, is largely occupied by the spermaceti organ, a complex oblong mass of oil-filled connective tissue that is surrounded by layers of muscle and blubber and is enclosed by a membrane called the case. The organ is cradled in a long conical depression largely formed from the upper jawbones of the sperm whale's elongated skull (*light color*).



THREE TOOTHED WHALES that possess spermaceti organs are compared in this drawing with the largest of all whales. They are the sperm whale *Physeter catodon* (*top*), the pygmy sperm whale *Kogia* (*bottom right*) and the bottle-nosed whale *Hyperoodon* (*bottom left*). The largest whale is the blue, or sulfur-bottom, whale *Balaenoptera musculus,* one of the suborder Mysticeti: the toothless baleen whales.

sharks in the stomach of a sperm whale captured in an area where the bottom depth was greater than 3,000 meters.

Why should a sperm whale need to dive so deep? Squids are found at all depths in every ocean. They are a principal food of many other air-breathing vertebrates: seabirds as well as such marine mammals as seals and lesser toothed whales, including porpoises. Perhaps the reason for the sperm whale's diving behavior is that the deeper a squid-eater can descend, the farther it can outdistance its competitors and the greater the stocks of squids within its reach are. Certainly the sperm whale can catch deep-living squids that are beyond the reach of seabirds, seals or porpoises. The whale can also reach bottom in the zone where the sea floor drops steeply to the abyssal plain from the edge of the continental shelf. Here on the continental slope, at depths of between 200 and 3,000 meters, many squids lay their eggs and can be found in large numbers and dense concentrations. Few air-breathing animals other than the sperm whale can hope to reach such easy pickings.

A peculiarity of the sperm whale's diving cycle is that the animal frequently surfaces within a few hundred yards of the point where it began its dive. The reason is not that its underwater time has been occupied by a slow descent and ascent. The whales are known to descend at a speed of about four knots (120 meters per minute) and to ascend at a speed of about five knots. Hence a round trip to a depth of 1,000 meters would not take more than 15 minutes. As we have seen, the duration of a deep dive is some three times longer. Both the duration of the dive and the fact that the sperm whale's place of emergence is close to its place of submergence suggest that when the whale comes to the bottom of its dive, it must sometimes lie almost still in the water.

Many, although by no means all, squids are fast swimmers over short distances. One may therefore wonder why a sperm whale would lie still at depth instead of actively pursuing its swift prey. The whale's lower jaw is long and narrow. It is the jaw of a snapper; even when it is open, it offers little water resistance. Perhaps the sperm whale's hunting strategy relies less on active pursuit and more on silent hovering followed by a quick pounce into a passing shoal of squids. Little or no daylight penetrates these hunting depths, but most of the squids on which the whale preys are luminescent. In its efforts to catch these speedy invertebrates a still, silent whale may well have the advantage over a swimming one.

A whale can lie still underwater only by being very nearly neutrally buoyant, that is, by having the same density as the surrounding water. In the older classes of man-made submersibles that displaced from 1,600 to 2,000 tons buoyancy had to be controlled within 40 liters of water (between two and three hundred-thousandths of the displacement weight) to enable the vessel to lie still in the water for listening purposes. Many water-dwelling animals can also achieve buoyancy within very fine limits and so can lie still at the depth where they live. For example, some fishes counter the sinking effect of those body tissues that are denser than water by storing low-density fats; other fishes manage neutral buoyancy by means of an air-filled swim bladder. Many squids do the same by replacing the dense sodium ions in their body with less dense ammonia ions.

In its biological and behavioral aspects, then, the sperm whale exhibits some unusual features. Among them are migratory behavior that takes the males from equatorial waters to polar ones, deep dives of long duration, the ability to lie still when submerged and the property of floating when dead. No one feature is unique to the sperm whale, but only the sperm whale is known to combine them all.

Of these four features two concern buoyancy and the other two—great range both horizontally and vertically—involve changes in ambient water conditions that are accompanied by changes in buoyancy. Although the sperm whale could carry the right amount of fat or air to be neutrally buoyant at a particular geographical location and depth, if the whale depended on such a static system for the control of buoyancy, the different water densities at other depths and geographical locations would push it either down or up.

Is this the clue we are seeking? Is the spermaceti organ a device for controlling buoyancy over a wide range of conditions? If it is, the organ must be able to vary its density. How could it do so? Only one substance is present in the organ in large quantities that is also able to undergo a substantial change in density: the spermaceti oil itself. In point of fact this oil has long been known to have properties different from those of other whale oils. When the liquid oil is dipped out of a dead whale's head and exposed to ambient air temperatures, it soon loses its clarity and becomes a soft crystalline solid.

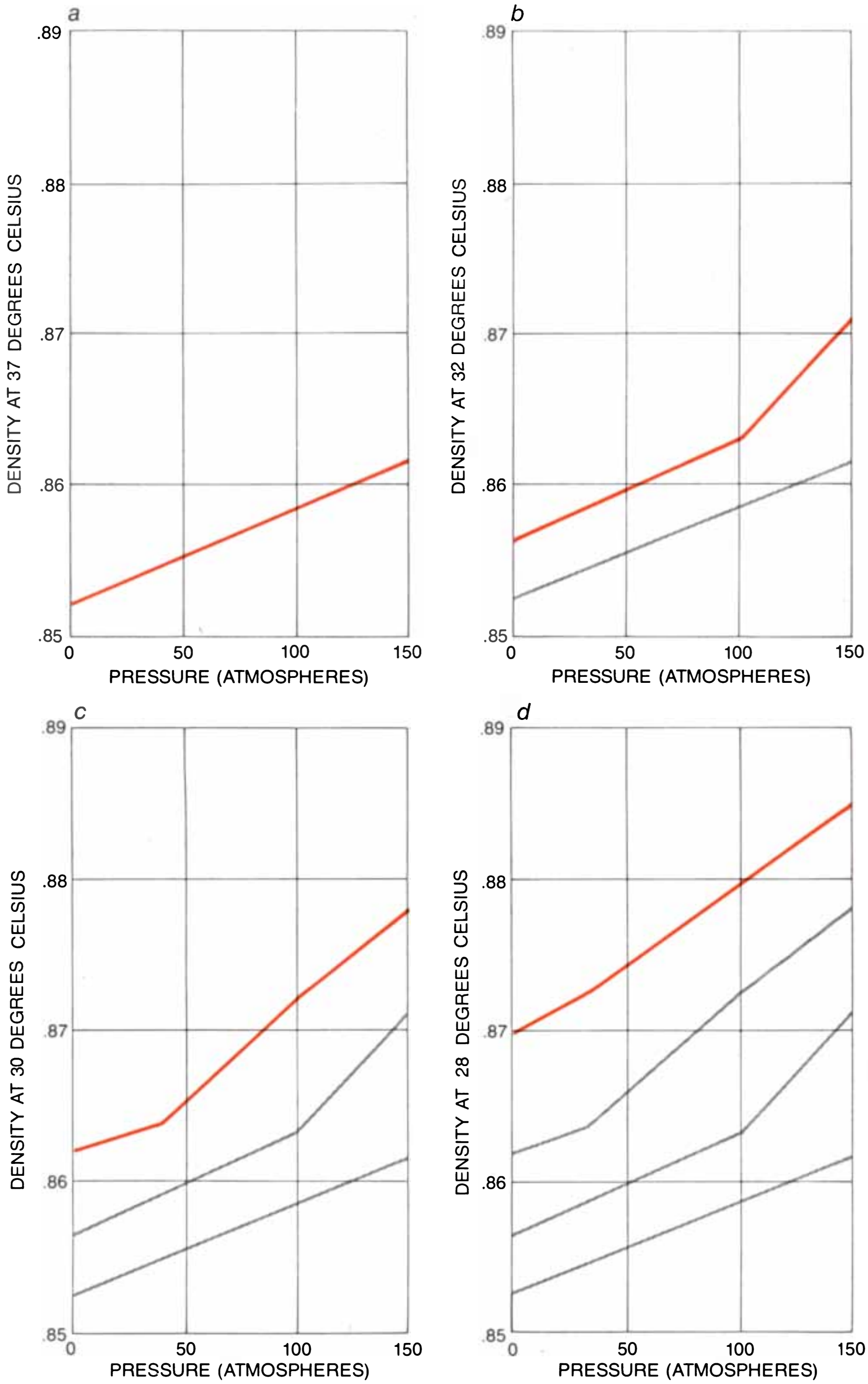
Temperature probes of freshly killed sperm whales show that when the whales are resting on the surface, the temperature of the spermaceti oil is 33 degrees Celsius (90 degrees Fahrenheit). The oil begins to crystallize, or congeal into a solid, when its temperature drops below 31 degrees C. Unlike the crystallization of water, which is almost instantaneous at the freezing point, the crystallization of spermaceti oil is a gradual process that is not completed until the temperature drops several degrees. When spermaceti oil freezes, it becomes denser and therefore occupies less volume. And occupying less volume, it displaces less of the ambient seawater and is less buoyant.

If the temperature of the spermaceti oil could be varied, then, the changes in density that accompany changes in temperature might be enough to let the sperm whale control its own buoyancy. Are such changes in temperature physiologically possible? The problem is of course one of the loss and gain of heat. With that in mind let us consider the anatomy of the spermaceti organ.

The tissues that house the spermaceti oil in the whale's snout have a dense network of capillaries supplied with blood by large arteries that enter the snout at the rear. The circulation of the arterial blood is therefore the principal means of conveying heat to the oil. The same circulation at the capillary level is also the main distributor of heat within each block of spermaceti tissue; when the tissue is cooled locally, the movement of blood through the capillaries helps to spread the cooling effect.

In addition the larger arteries and veins in the snout of the sperm whale lie side by side; this countercurrent system assists the cooling of the spermaceti tissue by the exchange of heat between the warmer incoming arterial blood and the cooler outgoing venous blood. The arteries that supply blood to the snout are surrounded, particularly at the point where they pass through the skull, by a dense network of veins carrying cooler blood away. The heat exchange can maintain a sharp difference in temperature between the snout (where the blood is normally below 34 degrees C.) and the rest of the whale's body (where the blood is normally above 37 degrees C.).

The sperm whale can lose heat through the surface tissues of the snout, either by "passive" conduction through the blubber and skin or by "active" heat transport: circulation of the blood to the papillae, minute fingerlike structures within the skin. As we shall see, passive conduction by itself is much too slow a process for the achievement of neutral



SPERMACETI OIL varies in density according to its temperature. These graphs show the density of the oil at progressively lower temperatures: at 37 degrees Celsius (*a*), at 32 degrees (*b*, *with the 37-degree reading included for comparison*), at 30 degrees (*c*) and at 28 degrees (*d*). Pressure also affects the density of the oil. Each 10 meters of additional depth adds one atmosphere (14.7 pounds) of pressure; thus at a depth of 500 meters the pressure is 50 atmospheres. The slopes of the four graphs trace the increasing density of the oil up to the equivalent of 1,500 meters. The findings are from a study made at the British National Physical Laboratory.

buoyancy in the time available. Active heat transport is a more promising possibility.

The anatomy of the sperm whale's head suggests still another means of heat loss: the sperm whale's nasal passages. Asymmetrical nasal passages are characteristic of toothed whales, but no other toothed whale has passages like those of the sperm whale. The simpler of its nasal passages, the left one, runs back from the cavity under the animal's single blowhole, curves to pass the left side of the spermaceti organ and enters the skull just in front of the brain case. The sperm whale breathes through this muscular tube, which can be expanded until it becomes circular in cross section.

The right nasal passage is totally different. Although it too begins in the cavity under the blowhole, it at first runs forward and is a tubular passage of small diameter. Then it widens and flattens out as it passes downward to form a broad, flat chamber, the vestibular sac, located at the front of the snout just under the blubber. A broad horizontal opening in the rear wall of the sac, often called the "monkey's mouth," gives entry to the continuation of the nasal passage, a wide, flattened tube that runs back the length of the snout to a point just in front of the brain case. There the wide passage narrows as it enters the skull and meets the left nasal passage in a common cavity. Just forward of this narrowing the right nasal passage opens upward to connect with a second sac, the nasofrontal sac, located above the central part of the skull crest.

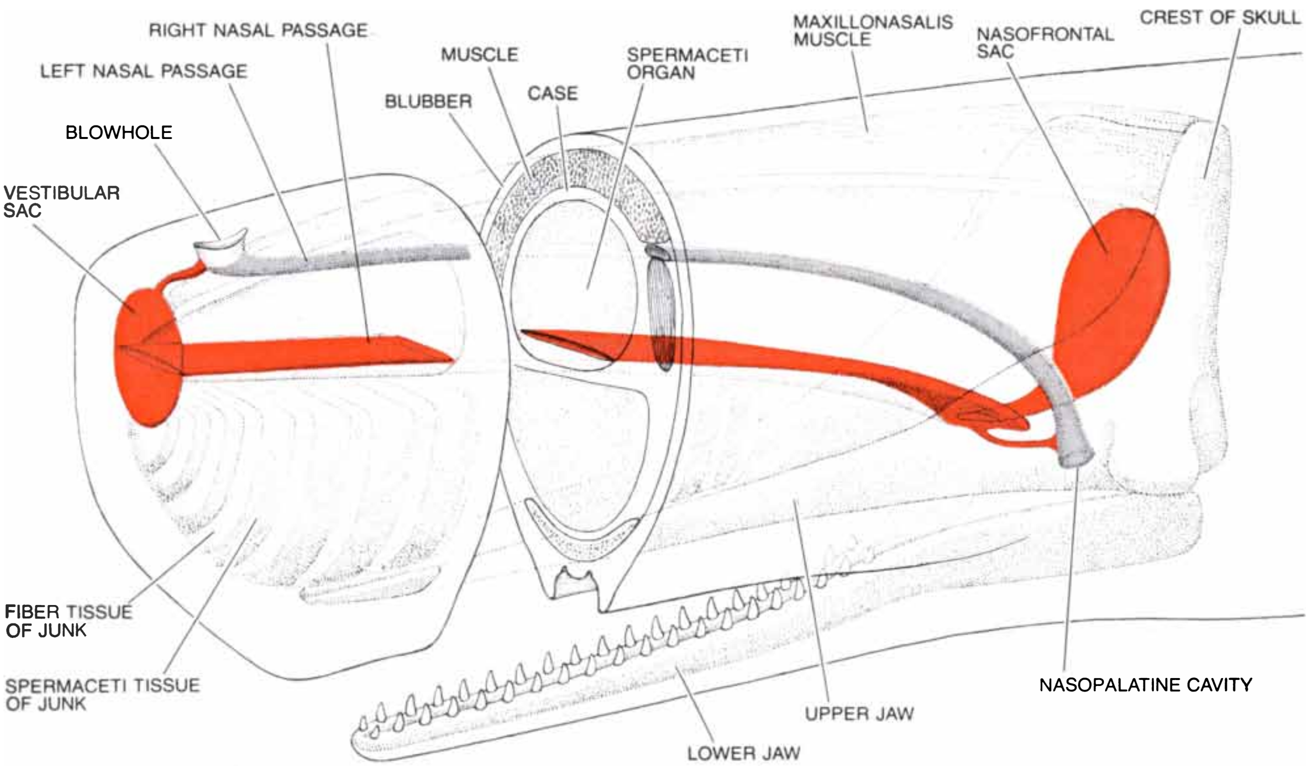
The right nasal passage is thus often more than a meter wide; its roundabout course through the snout is five meters or more long. Not only does it pass through the core of the spermaceti organ but also its two sacs cover the front and rear ends of the organ.

The interior of the right nasal passage is lined with a delicate layer of black tissue. Under the black layer are two layers of white tissue, first an elastic layer and then a fibrous one; the three layers together form a wall that is between .6 millimeter and one millimeter thick. Spermaceti tissue lies directly in contact with the wall of the nasal passage on all sides, and capillaries from the spermaceti tissue enter the white elastic layer.

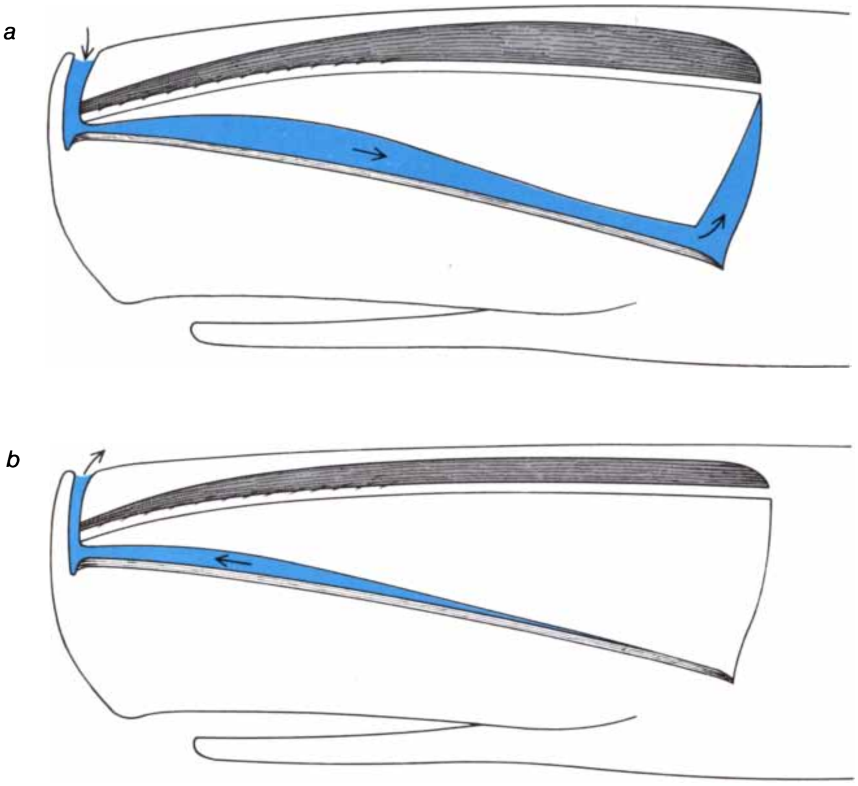
The intimate relation between the sperm whale's right nasal passage and the spermaceti organ is such that if seawater enters the nasal passage, the spermaceti oil will be markedly cooled. One can calculate the rate of oil cooling for any given temperature of seawater (on the basis of heat exchange between the blood in the capillaries and the cooler seawater) from the total area and the thickness of the nasal-passage wall. One can go on to calculate the rate of heat loss via the skin of the whale's snout, via the right nasal passage or via both heat-exchange areas combined and so determine the time required for the whale to reach neutral buoyancy over a range of selected depths in either an Antarctic environment or an equatorial one.

In either environment, with minimum values in the calculation, the right nasal passage proves to be a slower heat exchanger than the skin of the whale's snout. The calculated difference between the two surfaces, however, is probably not significant; the nasal passage is elastic, and its wall could be expanded to present a greater heat-exchange area than the minimum I have calculated.

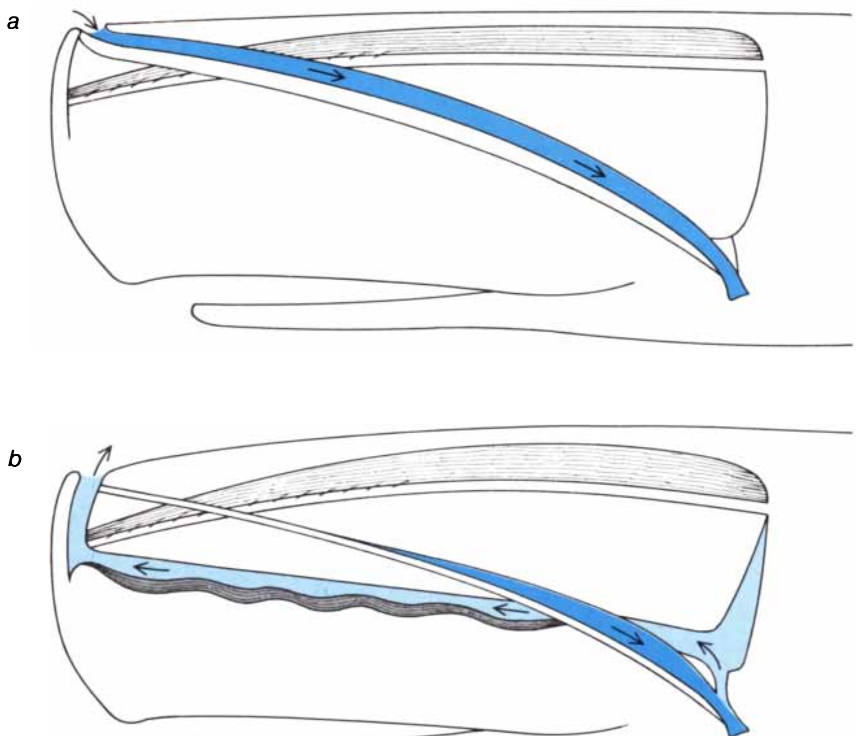
My calculations show that a 30-ton



NASAL PASSAGES of the sperm whale are not only asymmetrical but also intimately associated with the spermaceti organ. The left nasal passage (*dark gray*) is the simpler of the two. Beginning in a cavity under the blowhole, the passage curves to pass along the left side of the spermaceti case and terminates in the nasopalatine cavity of the skull. The right nasal passage (*color*) also begins under the blowhole; it then runs forward and widens to form a vertically oriented sac, the vestibular sac, at the forward end of the spermaceti case. A narrow horizontal opening at the back of the sac provides for the continuation of the passage rearward through the interior of the spermaceti organ until the passage approaches the scooplike crest of the whale's skull. There the nasal passage gives rise to a second vertically oriented sac, the nasofrontal sac, located at the back of the spermaceti case, before narrowing to enter the skull cavity that is shared in common with the left nasal passage. The complex route followed by the right nasal passage is some five meters long, and the passage itself is in places more than a meter wide. Over most of its length vessels from the network of capillaries in the spermaceti-oil tissue extend into the wall of the nasal passage. A major muscle of the snout, the maxillonasalis, runs from the crest of the skull to the forward half of the spermaceti case.



WATER INTAKE through the right nasal passage (*a*) may be accomplished by contraction of the major snout muscle (*dark gray*). Muscle action would widen the passage (*color*) so that water could enter; small muscles within the spermaceti tissue would aid the process. Relaxation of these muscles would allow the passage to narrow again (*b*), thereby expelling the water.



ALTERNATE INTAKE ROUTE is the left nasal passage (*color*). Contraction of a minor muscle, the nasal-plug muscle, would widen the left passage and draw water into the nasopalatine cavity (*a*). Thereafter (*b*) the action of muscle fibers in the floor of the right nasal passage (*gray*) could pump water forward through the spermaceti organ and out through the blowhole.

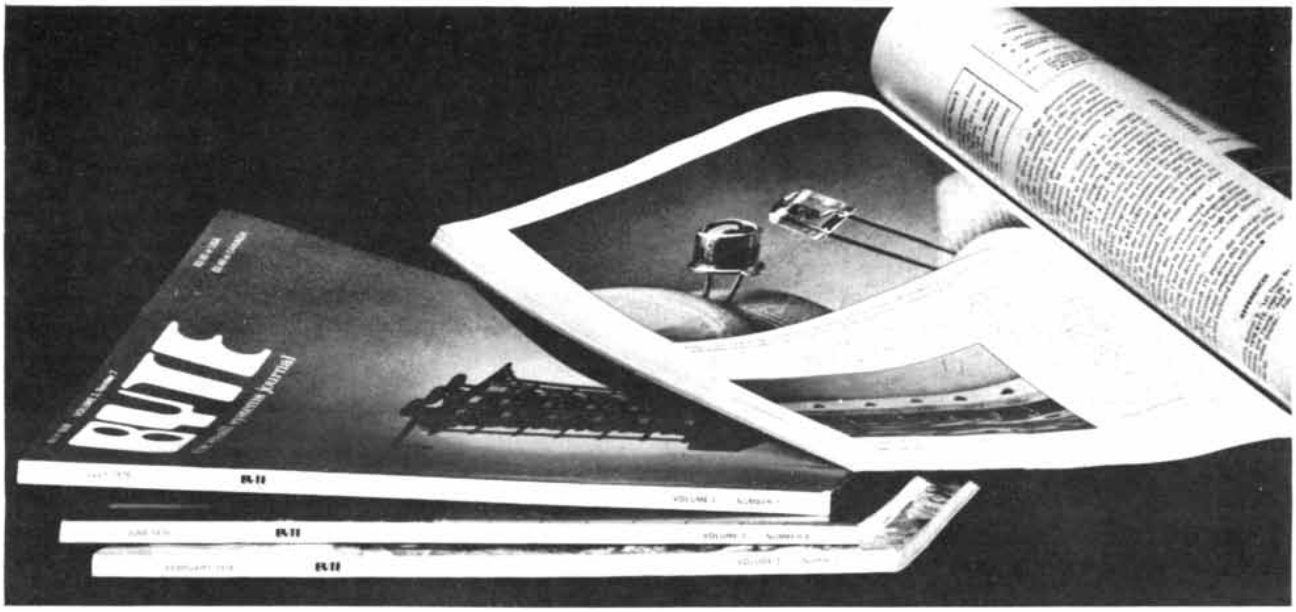
sperm whale, exploiting both heat-exchange areas simultaneously, could adjust to neutral buoyancy in less time than it normally takes to swim to a depth of 500 meters. In dives from 200 to 1,000 meters deep, exchanging heat via the snout skin alone, the whale would reach neutral buoyancy within five minutes of attaining the desired depth. If both heat-exchange areas come into play, the interval would be shortened to three minutes.

In the Antarctic, because lower water temperatures mean greater buoyancy at and near the surface, the calculated time needed to reach neutral buoyancy at a depth of 100 meters, with either heat exchanger in play alone, comes to about 20 percent of the total submersion time of even a prolonged dive. At first this may appear to be too great an investment of time to make the attainment of neutral buoyancy worthwhile. There is, however, a counterbalancing factor: swimming in the cold surface water of Antarctic latitudes, the whale may maintain the temperature of its spermaceti oil at a level lower than the 33 degrees C. characteristic of equatorial waters. If it does, cooling to neutral buoyancy would be quicker. In addition, when the whale was submerged during a dive, its resorting repeatedly to the nasal heat exchanger could hasten the achievement of neutral buoyancy. My calculations show that below 200 meters a filling of the nasal passage with seawater twice rather than once would be more than enough to exchange the required quantity of heat.

Just how can seawater be drawn into the whale's right nasal passage? Surrounding the outer wall enclosing the spermaceti tissues are large muscles. They run from the front half of the "case" to attachments on the crest of the skull. Their contraction would suffice to raise the front end of the case, thereby lifting the upper half of the nasal passage. The same contraction would also open the front end of the nasal passage and draw water in from the cavity under the blowhole. The relaxation of the muscles would expel the water.

How far might the seawater travel along the right nasal passage? There are additional muscle fibers in the floor of the passage and still other fibers that run forward within the spermaceti tissue from the front wall of the nasofrontal sac. The contraction of these fibers would hold down the bottom half of the right nasal passage, ensuring that the water would travel at least as far as the nasofrontal sac.

One cannot exclude the possibility that seawater also reaches the spermaceti organ by a different route: the left nasal passage. Once drawn into this shorter tube from the cavity under the blowhole, the water could be pumped to the cavity where both nasal passages meet;



Get the inside story...

on software, hardware, simulations, computer games, robotics, computers and calculators, languages and compilers, custom systems design and scores of applications.

Read BYTE.®

BYTE, the leading magazine in the personal computer field, will keep you aware of fast-paced changes in the ever growing world of microprocessors. You'll find **BYTE** tutorials invaluable, reports on home computer applications instructive, and the reviews of computer products stimulating.

Tutorial information in **BYTE** is of interest to both the beginner and experienced computer user. **BYTE**'s editorial content explores the fun of using and applying computers for personal satisfaction in activities from electronic music to electronic mail, from games to pragmatic personal business. In **BYTE** is found authoritative yet easily read coverage of principles of computer hardware and software design, approaches to novel applications, and essentials of proven interest to personal computer enthusiasts.

Each month serious computer users read **BYTE.** They're rewarded with timely articles by professionals, computer scientists and competent amateurs. Isn't it time you got the inside story? Subscribe now to **BYTE,** the Small Systems Journal.

Fill in and mail the coupon today. Read your first copy of BYTE. If it is everything you expected, honor our invoice. If it isn't, just write "Cancel" on the invoice and mail it back. You won't be billed and the first issue is yours at no charge.

Allow 6 to 8 weeks for processing. If you have any questions, dial toll free 800-258-5485. This is the number of the BYTE Subscription Department.

© BYTE Publications, Inc. 1978

BYTE Subscription Dept. P.O. Box 590 Martinsville, NJ 08836
Please enter my subscription for

☐ One year $15 (12 issues) ☐ Two years $27 ☐ Three years $39

☐ Check enclosed entitles me to 13 issues for price of 12 (North America only)
☐ Bill Visa ☐ Bill Master Charge ☐ Bill me (North America only)

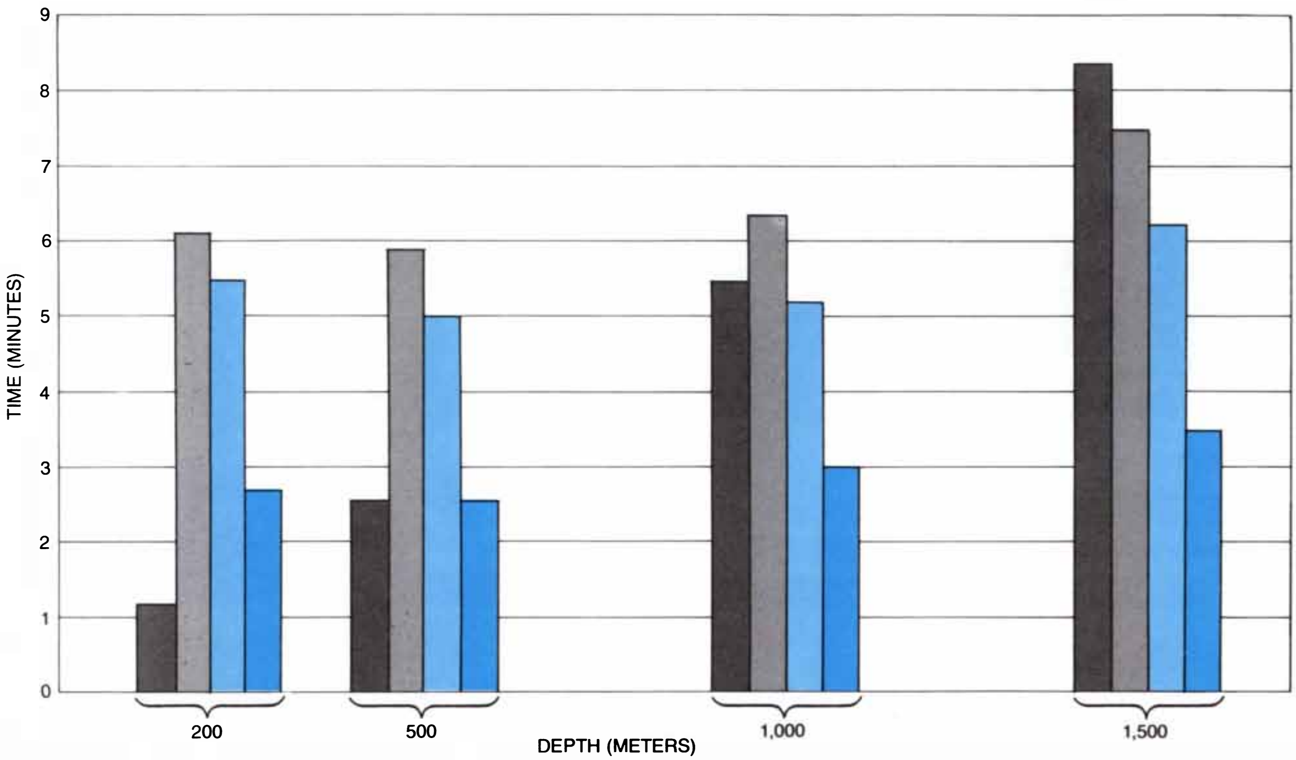
Card Number ______ Expiration Date ______

Signature ______ Name (please print) ______

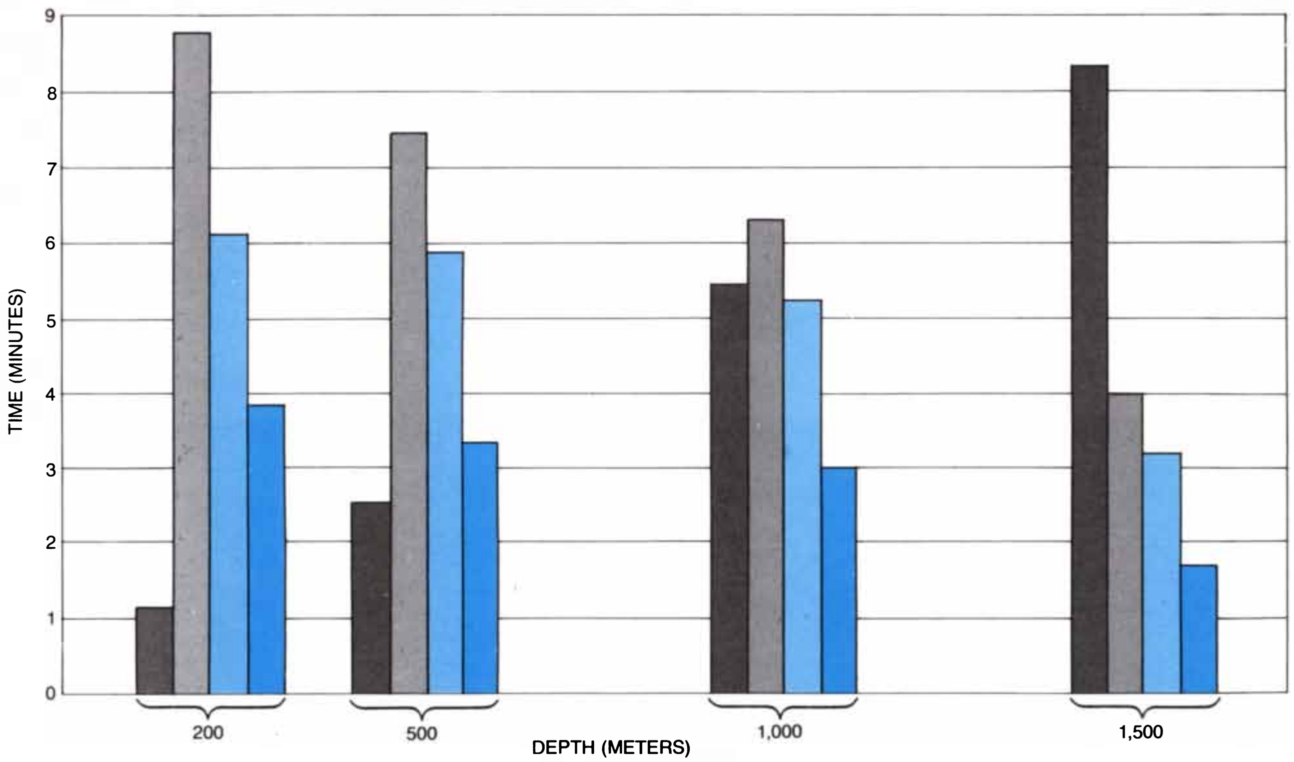
Address ______

City ______ State/Province/Country ______ Code ______

Foreign Rates (To expedite service, please remit in U.S. Funds)
☐ Canada or Mexico ☐ One year $17.50 ☐ Two years $32 ☐ Three years $46.50
☐ Europe, one year (air delivered) $25
☐ All other countries, one year (surface delivered) $25. Air delivery available on request. 7119



THREE ALTERNATIVES for cooling spermaceti oil when a whale is diving at the Equator take different lengths of time, as is apparent in this graph. The calculations are for a 30-ton whale diving with full lungs. The first bar (*gray*) in each of the four sets shows the time needed by the whale to swim to the indicated depth at a speed of five knots. When the dive is to 200 meters, even maximum heat exchange via the right nasal passage and the skin of the snout combined (*color*) must continue for more than an additional minute before neutral buoyancy is achieved. Heat exchange via the right nasal passage (*light gray*) or via the snout skin (*light color*) would have to continue even longer. With deeper dives the trend favors achievement of neutral buoyancy during the time the whale is swimming to depth. In a dive to 500 meters the time required to achieve neutral buoyancy through maximum heat exchange is equal to the time of descent. In a dive to 1,500 meters even the least efficient form of heat exchange, via the nasal passage alone, achieves neutral buoyancy during descent.



SAME ALTERNATIVES are calculated for a dive in polar waters; again the whale weighs 30 tons and is swimming down with full lungs at a speed of five knots. Maximum exchange of heat (*color*) almost achieves neutral buoyancy during a descent to 500 meters. Heat exchange via the snout skin alone (*light color*) achieves neutral buoyancy during a descent to 1,000 meters. Even the least efficient form of heat exchange, via the nasal passage only (*light gray*), achieves neutral buoyancy four minutes before the whale reaches 1,500 meters.

the water could then enter the rear of the right nasal passage and be moved forward by the contraction of the muscle fibers in the passage floor. If that is the case, one wonders why a relatively major role has been assigned to these few fibers and a relatively minor one to the very large muscles attached to the case. Nevertheless, on the two occasions when dissection revealed the presence of seawater in a sperm whale's right nasal passage, I also found water in the left passage, and so this alternate route cannot be ruled out.

To find out what changes in the density of spermaceti oil accompany changes in temperature and pressure I asked members of the staff of the British National Physical Laboratory to conduct density measurements over a range of temperatures and pressures. They found that whereas changes in density are mainly a function of temperature, pressure also has a marked effect, particularly near the onset of freezing.

At this point in my study I faced two related questions. First, what changes in buoyancy would a sperm whale encounter over its range of diving depths in both a polar and an equatorial environment, supposing it had no means of controlling buoyancy? Second, is the quantity of spermaceti oil in the head of an adult sperm whale large enough to achieve neutral buoyancy by means of density changes over the same range of depths and environments?

Turning to the first of these questions, the factors that influence sperm-whale buoyancy include the density of the seawater, the effect of pressure on both the liquid and the solid components of the whale's tissues and the effect of pressure on the volume of the whale's lungs. Variations in these factors all follow established physical laws, and therefore their effect on a whale of any particular size can be calculated. The calculations, of course, depend on reasonably accurate estimates both of the proportions of liquid, solid and gas in the diving whale and of the buoyancy of the whale when it is floating on the surface. With such a large animal it is not easy to make precise estimates of this kind. Those used in our study have therefore included minimum and maximum values as well as mean values. The mean values are used in this discussion, but the conclusions remain valid even when the extreme values are used instead.

In moving on the surface between equatorial and polar latitudes a sperm whale experiences a change in surface-water temperature of as much as 26 degrees C. Moreover, if the whale dives to a depth of 1,000 meters at the Equator, the change in temperature may be as much as 23 degrees C. These changes in temperature are accompanied by changes in the density of the seawater. Density is mainly dependent on temperature, but salinity and pressure also enter into the calculations. For example, if a whale that is neutrally buoyant on the surface at the Equator were moved to an Antarctic latitude such as 55 degrees south, the increased density of the surface water would give the whale an increased buoyancy amounting to a little more than .3 percent of its body weight. The increase in water density the whale would encounter during a deep dive, whether at the Equator or at 55 degrees south latitude, would also increase its buoyancy. Because of the greater change in water density with increasing depth at the Equator, however, the increase in the whale's buoyancy there would be about five times more than it would be in the Antarctic: at a depth of 2,000 meters it would amount to almost .5 percent of the animal's body weight.

So much for water density. What about water pressure? Its effect on the water content of the whale's tissues proves to be very slight, as is its effect on the oil content. As for the solid components of the whale's tissues, they are almost incompressible, and so their mass gives the whale added lift as the pressure increases. At a depth of 1,000 meters the lift would equal about .1 percent of body weight. Lumping the effects of water pressure on all three tissue components, the sperm whale at a depth of 1,000 meters would have extra lift equivalent to .07 percent of its body weight.

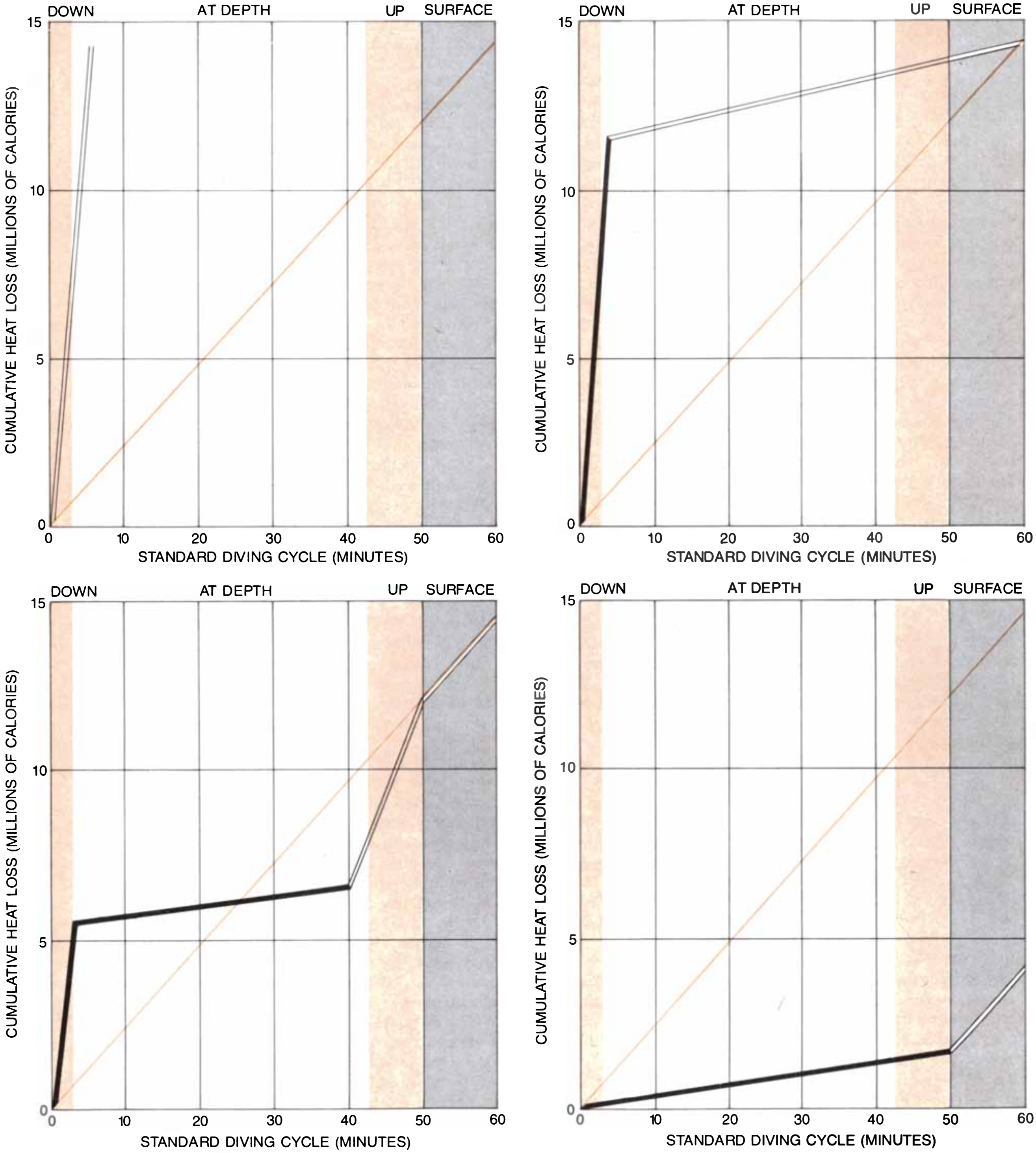
Water pressure also acts on the gas in the whale's lungs during a dive and so affects the whale's buoyancy, particularly at depths of up to 200 meters. When a freshly killed sperm whale is floating on the surface in the Tropics, the volume of its body above the water roughly equals the volume of air in its lungs. By the same token, when a sperm whale exhales on the surface, it comes close to attaining neutral buoyancy. Sperm whales may sometimes exhale before diving; under these circumstances the reduction in buoyancy should facilitate the dive. If instead the whale dives with its lungs full, by the time it reaches 200 meters the effect of the water pressure on the gas in the lungs will have brought about the same reduction in buoyancy as an exhalation at the surface would have. Below 200 meters there is no great difference between diving with full lungs and diving with empty ones.

If one considers the combined effects of seawater density and pressure on the whale, one can calculate the lift or the downthrust the sperm whale will experience at various depths and in various environments. For example, the whale is very positively buoyant on the surface at the Equator. If the animal dives with full lungs, the rapid decrease in lung volume causes a downthrust equal to about .1 percent of its body weight at 100 meters. By the time the whale reaches 200 meters the change in the water density will more than counterbalance this first effect, and the animal will have positive lift. The lift will increase as the whale descends until at 2,000 meters it will equal about .2 percent of the body weight. In Antarctic latitudes, since the cold surface water is denser than the warm surface water at the Equator, the sperm whale begins its dive with greater buoyancy and so does not experience much downthrust in the early part of its descent.

Calculations indicate that throughout its geographical range the sperm whale must adjust its overall density by an amount equivalent to .2 percent of its body weight in order to achieve neutral buoyancy below 200 meters. I have suggested that such an adjustment could be made by a "freezing" of the whale's spermaceti oil and have also described the heat exchangers available to cool the stored oil to a solid. But is the freezing effect adequate to the task? Among the known facts that are useful in answering this question are the findings of the National Physical Laboratory with respect to the density of spermaceti oil at various temperatures and pressures. It is also known that a 30-ton whale will have nearly 2.5 tons of spermaceti oil in its head and that at the surface in equatorial latitudes the oil will be at a temperature of 33 degrees C.

Starting with these facts, we can go on to calculate the spermaceti-oil temperature required to counterbalance the whale's natural lift at successive depths. The calculation shows that below 200 meters the temperature of the oil need be lowered by only a few degrees (and never below 29 degrees C.) to attain the required densities. As an example, if a sperm whale in Antarctic latitudes exhales before diving, the temperature of the spermaceti oil need not fall below 30 degrees in order to counterbalance the whale's natural lift. When the temperature and the density are plotted together, the slopes of the temperature lines and the slopes of the lines indicating the required densities are strikingly similar.

With both the anatomical and the physical data in hand one can consider how the dissipation and the regaining of heat might be timed during the sperm whale's diving cycle. To begin with an example in the equatorial environment, one finds that the heat generated during a diving cycle probably cannot be dissipated during the 10 minutes of each cycle that the whale spends at or near the surface. The reason is that at any depth shallower than 100 meters the difference between the water temperature and the subcutaneous temperature of the whale is less than 2.6 degrees C. Even if the active loss of heat at the surface by vasodilation (that is, an expansion of the blood vessels in the skin) is combined with passive heat loss at depth by conduction, only part of the



HEAT BUDGET OF A DIVING CYCLE calls for a loss of nearly 15 million calories over a 60-minute period that is divided into some three minutes for descent to a depth of 500 meters (*colored band at left*), some 40 minutes of stalking prey at that depth, some seven minutes for ascent to the surface (*second colored band*) and 10 minutes of rest at the surface (*gray band*). The loss of heat is required to balance the body-heat gain resulting from exertion during the diving cycle. Each of the four graphs plots a steady rate of heat gain (*colored diagonal line*) and the rate of heat loss that is achieved by one of four different strategies of heat exchange. Calculations are for a 30-ton whale diving to 500 meters with full lungs at the Equator. The open line in graph *A* represents the result of a dive with continuous heat exchange via body skin and snout skin until the spermaceti oil is cool enough to bring about neutral buoyancy. The heat budget would remain badly out of balance during most of the diving cycle. The line in graph *B* represents the result of a dive with heat exchange via the right nasal passage, body skin and snout skin (*solid portion of line*) until neutral buoyancy is achieved. Heat loss is then reduced by vasoconstriction. Such a strategy would produce a budget that starts badly out of balance but improves steadily thereafter. The line in graph *C* represents the result of a dive with heat exchange via the right nasal passage only during the descent; after that, heat loss is reduced by vasoconstriction until some 40 minutes have passed (*open portion of line*). Heat is then transferred from the body to the spermaceti oil until the oil temperature rises to 33 degrees C. Thereafter vasodilation allows further heat exchange via skin surfaces. This strategy would produce alternating net heat loss and net heat gain until the whale returned to the surface. Strategies involving loss of heat via the snout skin, alone or in combination with the right nasal passage, would give rise to similar curves. One of these three strategies is the likeliest to be adopted by the whale. The line in graph *D* represents a dive with heat exchange minimized by vasoconstriction until the whale returns to the surface. The slight heat loss following vasodilation at the surface in equatorial waters (*open portion of line*), together with limited heat exchange during the dive, would produce an out-of-balance budget; the whale could not dive again within 10 minutes.

heat the whale generates during the active 50 minutes of its diving cycle will be removed. Heat cannot be retained by means of vasoconstriction during the submerged period of the dive because some of the sperm whale's time in deep cold water must be spent in losing heat and achieving neutral buoyancy.

Looking further, we find that the greater part of the heat the whale generates during the active part of the diving cycle may either be lost slowly, by a controlled partial vasodilation that keeps pace with heat production, or be lost rapidly by more extensive vasodilation during one or more short periods in the course of the diving cycle. Brief but extensive vasodilation would not call for prolonged control and would at the same time allow the whale to make use of variations in spermaceti-oil density for buoyancy control. For example, as the whale descends, the oil would be cooled by heat exchange not only via the skin but also probably via the right nasal passage. When the whale has attained neutral buoyancy, soon after reaching the desired depth, any further active cooling of its oil would be halted by means of vasoconstriction. Minor adjustments in buoyancy thereafter could be made by admitting small amounts of seawater into the right nasal passage; that would eliminate any dependence on skin vasodilation for heat loss.

During most of the 50 minutes of the diving cycle spent at depth the body of the sperm whale, and its muscular mass in particular, rises in temperature. The whale's muscles need not increase in temperature more than two degrees C. in order to store all the heat needed to bring the solidified spermaceti oil back to its normal fluid temperature. Just before the whale begins its ascent to the surface the supply of blood to its snout would be increased, carrying heat from its warmer body to the spermaceti organ. Hence at the same time that the whale's body temperature is falling to its normal level the thawing spermaceti oil is rapidly rising to its normal 33 degrees C. Heating the oil of course decreases its density and increases its volume, and so the whale's buoyancy shifts from neutral to positive; this would help to lift the whale toward the surface even if the animal were exhausted. Any muscle-generated heat in excess of that needed to reheat the oil could now be lost by vasodilation.

Using as an example a 30-ton sperm whale diving to a depth of 500 meters in equatorial waters, one can now consider five alternatives with respect to the most efficient means of heat loss. As the first alternative let it be assumed that vasoconstriction does not take place over the entire skin surface until the whale has descended to the selected depth and achieved neutral buoyancy. During the descent heat will be lost over the entire skin surface; the total heat loss will approach the total heat gain as a result of muscular activity during the diving cycle. When the passive loss of heat through the whale's blubber is added to this active heat loss, the total loss possibly exceeds the total gain. This first alternative must therefore be eliminated.

Let it be assumed as the second alternative that vasoconstriction takes place over the whale's entire skin surface as soon as the descent begins, so that almost all the cooling of the spermaceti oil is a result of heat exchange in the right nasal passage alone. Under those circumstances the time required to reach neutral buoyancy will be greater than if heat exchange were simultaneously in progress through the skin. Nevertheless, neutral buoyancy is achieved in about six minutes (12 percent of the 50-minute dive time). The dilation of the blood vessels of the spermaceti tissue before the whale begins to ascend would then shunt heat from its body muscles to its snout until the oil temperature returns to its normal 33 degrees C. Any excess body heat generated during the dive could then be lost by vasodilation.

In the third alternative the assumption is zonal vasoconstriction. The blood vessels of the whale's body skin constrict immediately after submergence but the blood vessels of the snout skin remain dilated until the spermaceti oil is cooled to the point of neutral buoyancy at depth. Such a sequence would lead to neutral buoyancy even more rapidly than the sequence outlined in the second alternative. It requires the further assumption, however, that the sperm whale can dilate and constrict the blood vessels of its snout skin and body skin independently.

In the fourth alternative heat exchange through the right nasal passage and heat exchange through the whale's entire skin proceed simultaneously. This sequence leads to neutral buoyancy faster than either the second or the third alternative. Moreover, the amount of the body heat generated during the dive that remains to be lost to the seawater is smaller.

The fifth and most efficient alternative involves heat exchange via the right nasal passage and the snout skin only.

All except the first of these five alternatives are plausible. The fourth alternative has an advantage over the third and fifth in not postulating a capacity for zonal vasoconstriction and vasodilation. Before the third and fifth alternatives are dismissed on those grounds, however, it should be noted that such zonal dilation is similar to the human facial blush. It seems at least conceivable that sperm whales too can blush.

As evidence for heat exchange via the right nasal passage I have found seawater in this passage on two occasions. Therefore the second, fourth and fifth alternatives are all likely ones, and the fifth alternative, involving both the right nasal passage and the snout skin, is the most likely of all. In this connection, in several recently killed sperm whales the spermaceti tissue closest to the surface of the snout was found to be slightly warmer than more central tissue. This suggests that an efficient cooling system, namely heat exchange via the right nasal passage, had been operating in the center of the snout.

Changes in the density of the spermaceti oil would of course affect the sperm whale's center of gravity. As the oil increases in density during the whale's descent the center of gravity will shift forward; as the density decreases just before or during the ascent the center of gravity will shift backward. These changes would actually be advantageous during both descent and ascent because the whale often moves almost vertically down and up. The shift means, however, that the whale's fore-and-aft trim, to borrow a nautical expression, will not be the same when the whale is submerged as it is when the whale is on the surface. The main factor in the whale's trim while it is on the surface is not the spermaceti ballast but rather the considerable buoyancy provided by its air-filled lungs, so that under these circumstances the role of the spermaceti is trivial. Perhaps, however, the forward shift of the center of gravity at depth actually gives the whale its best trim during the time in the diving cycle when the submerged animal is neutrally buoyant and either lying still or swimming.

Sperm whales less than about 11 meters long have less spermaceti oil in relation to their size than larger ones. The smaller whales are also comparatively shallow divers and do not migrate to such high polar latitudes as the larger whales. As a result they encounter smaller variations in water density, and even though their supply of the oil is smaller, it would suffice to compensate for the smaller range of water densities.

For an air-breathing, warm-blooded, long-diving animal control of buoyancy has numerous advantages. At the surface extra buoyancy allows the whale to relax. At depth the animal's hunting success, which is dependent on silence and good hearing, is not jeopardized by the need to keep swimming in order to remain at a particular level. Moreover, if it were necessary for the sperm whale to keep swimming in order to remain at the same level during its 50-minute submergence, it would have to invest a larger amount of energy in its diving cycle.

These are not the only advantages derived from the control of buoyancy. Imagine for a moment a hypothetical sperm whale that was less buoyant on the surface than is actually the case. Over a part of its range of diving depths and in certain geographical locations this hypothetical whale would actually experience downthrust during the diving cycle. Downthrust, of course, is a hazard to any air-breathing animal that makes deep dives of long duration. By not only having extra buoyancy at all depths but also having control over its buoyancy the sperm whale has a fail-safe diving system. Even if the animal became exhausted during a deep dive, it could still pop up to the surface for air. For example, a sperm whale that has exhausted itself, perhaps by short bursts of high-speed swimming after squids, can shunt the heat that is a by-product of its muscular activity into the spermaceti organ and again become positively buoyant. Such an exhausted whale is just as likely to reach the surface again as a fresh whale that can easily swim up.

What about alternative means of buoyancy control? For example, either collapsing the lungs or emitting air while submerged would affect buoyancy. On close consideration neither stratagem proves to be as efficient as control by means of oil density. The degree of lung collapse depends entirely on water pressure, whereas buoyancy also depends on seawater density, which in turn is dependent on temperature. As diving depths increase, these factors work in opposite directions; only at two points in a deep dive do the two factors exactly match to yield neutral buoyancy.

As for the emission of air, that would be an ineffective means of controlling buoyancy over much of the whale's diving range. For example, below 600 meters even a small emission of air would greatly reduce the volume of air remaining in the whale's lungs for the return to the surface; in some geographical areas a whale that had emitted air at depth would experience considerable downthrust as it swam through the thermocline, the abrupt transition in water temperature that is encountered between 200 and 100 meters. Furthermore, once air is emitted it is lost. The many minor adjustments of buoyancy at different

SPERM WHALE'S HEAD is peeled of its blubber on the flensing platform of a factory ship in the Antarctic; teeth of the lower jaw are clearly visible. Upper-jaw teeth are usually absent.

depths made possible by oil-density control could hardly be matched by a system that depends on an exhaustible reservoir.

Loss of heat from the spermaceti oil as a means of buoyancy control imparts still another distinct advantage. In the Tropics the water temperature at the surface may be less than three degrees C. below the subcutaneous temperature of the whale. At depths shallower than 100 meters the body heat generated by a deep dive could not be lost during the usual 10-minute resting time at the surface between dives. Both the heat exchange that is involved in freezing the spermaceti oil during the descent phase of the diving cycle and the body-heat shunt involved in thawing the oil before and during the ascent phase help to overcome this equatorial handicap.

Few whales are known to dive as deep or for as long a time as the sperm whale. Two other toothed whales—the bottle-nosed whale *Hyperoodon* and the pygmy sperm whale *Kogia*—have spermaceti organs. Both are long-duration deep divers. In view of the complicated structure of their snout it seems probable that both also use the spermaceti organ for buoyancy control. This conjecture cannot be proved, however, until much more is known about the dimensions, anatomy and diving habits of these whales.

Some other whales, for example the fin whale *Balaenoptera physalus*, can make rapid excursions to considerable depths. The dives of the fin whale, perhaps to depths of some 300 meters, seldom last longer than 10 minutes; the whale swims continuously and surfaces some distance from its point of submergence. These baleen whales usually swim faster than the three deep-diving toothed whales. It is clear that they do not submerge for long periods of time or in order to feed as the three deep divers do; their prey is the shrimplike krill found in swarms at or near the surface.

As I have noted, many suggestions have been made regarding possible functions of the spermaceti organ other than buoyancy control. None, however, accounts as adequately for the size and structure of the organ. Such a complex anatomical structure, however, could certainly serve more than one function. Several of the other proposed functions are quite compatible with the buoyancy-control hypothesis I have proposed. Nevertheless, so many peculiarities of the sperm whale's snout and its cargo of oil can be explained by the buoyancy-control hypothesis that it is hard not to accept this as the organ's main function. Final proof of the hypothesis, however, must await measurement of the temperature or the density of the oil within the spermaceti organ in the course of a deep dive. This is a difficult and costly task but not an impossible one.

Robert A. Millikan

A tireless investigator and a Nobel prizewinner at a time when not many Americans were, he had a penchant for controversy in subjects ranging from cosmic rays (which he named) to the support of science

by Daniel J. Kevles

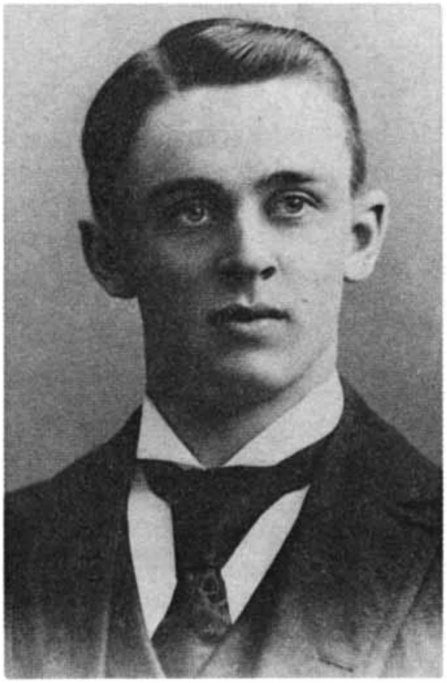
Robert A. Millikan was the most famous American scientist of his day. In 1923 he became the second American (A. A. Michelson had been the first, in 1907) to win the Nobel prize in physics. Millikan is best known to physicists for measuring the charge of the electron with his oil-drop experiment; in the span of a remarkably productive career he also made significant contributions to the study of the photoelectric effect, hot-spark spectra and, above all, cosmic rays. He was more than a research scientist; between the wars he headed the new California Institute of Technology, advised industrial corporations and philanthropic foundations and played a key part in the development of Federal policy for academic science.

The recognition of Millikan's wide-ranging importance has dimmed with the passage of time, in spite of the publication in 1950 of his 300-page autobiography. As such books go, Millikan's is reasonably informative, yet like many autobiographies written at an advanced age (Millikan was 82 when his book was published) it attends more to the earlier events of his life than to the later ones. More important, it omits controversial episodes that were of considerable significance in his career. The controversies centered on cosmic rays and on his attitude toward the relation between academic science and the Government. In both controversies Millikan was on the losing side.

No autobiographer (for that matter no human being) likes to dwell on failure. Millikan went further: he disliked mere suggestions, let alone demonstrations, that he was wrong. Even after he had won the Nobel prize he bridled when the Berkeley physicist Raymond T. Birge proposed that the accuracy of his value for the electron charge might be improved. Edwin C. Kemble of Harvard remarked to Birge: "It's too bad that Millikan isn't a little more of a good sport. A man in his position could well afford to be." Whether Millikan was sporting or not, his false starts and wrong turns, his commitments to the losing side and his defense of ultimately indefensible redoubts reveal no less than his triumphs the significance of his life both as a research scientist and as a public figure.

The son of a Congregational minister and the former dean of women of a small college, Millikan was born in 1868 in Illinois and was raised from the 1870's in Maquoketa, Iowa (population 3,000). He showed no particular scientific inclinations, and neither his family or school nor the agrarian environment stimulated him to move in a technical direction. At Oberlin College (where his mother had gone before him) he pursued a standard classical curriculum; his move toward physics came when his professor of Greek, impressed with Millikan's abilities, invited him to teach an introductory physics course in the preparatory school run by the college. (When Millikan protested that he knew nothing about the subject, the professor replied, "Anyone who can do well in my Greek can teach physics.") After graduation Millikan stayed at Oberlin to teach in the preparatory department for two more years and then went on to Columbia University, where he was the only graduate student in physics. One summer he worked under Michelson at the University of Chicago. Having earned his doctorate in physics at Columbia, Millikan spent a postdoctoral year in Europe, where his teachers included Max Planck, Walther Nernst and Henri Poincaré; he acquired what was on the whole a better than average training for an aspiring American physicist at the turn of the century.

Joining the University of Chicago faculty in 1896, Millikan poured his considerable energies into developing the physics curriculum. At that time Ameri-



PORTRAITS OF MILLIKAN show him (*left to right*) as an undergraduate at Oberlin College in the 1880's; in 1891, when he was graduated from Oberlin and stayed on to teach physics in the

can physics students in both high school and college relied on foreign textbooks. Millikan wrote or coauthored a variety of books and laboratory manuals that became classroom standards. (*A First Course in Physics,* written with Henry G. Gale, sold more than 1.6 million copies between 1906 and 1952.) Recognizing, however, that at the University of Chicago the main rewards were given for research rather than for teaching, in about 1908 he put aside the writing of textbooks to concentrate on his work in the laboratory.

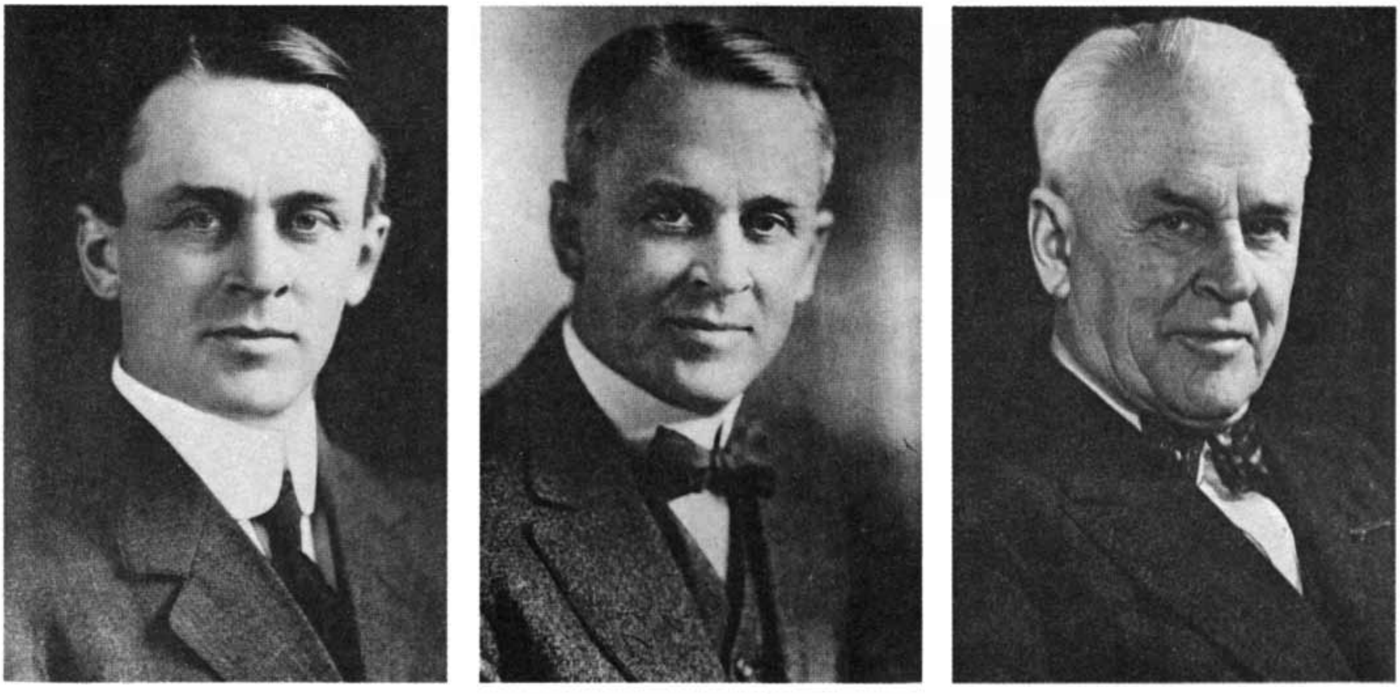
World War I interrupted Millikan's research career. By now one of the nation's leading physicists, he was increasingly active in the professional affairs of his discipline and in the National Academy of Sciences. He was also a pioneer in developing links between industry and academic physics: he became a consultant to the research department of Western Electric, primarily to advise the company on vacuum-tube problems, and he pointed a number of his students toward careers in industry. Early in 1917, to help mobilize science for defense, Millikan went to Washington as a vice-chairman and the director of research of the newly established National Research Council in the National Academy of Sciences. Like many scientists engaged in defense research during World War I, Millikan soon entered the armed services. As a lieutenant colonel in the Army Signal Corps he directed work in meteorology, aeronautical instruments and communications, and in his National Research Council capacity he played an important role in initiating and advancing a major project to develop devices for the detection of submarines.

Millikan's success in the wartime mobilization was no mean achievement. The National Research Council, a private organization like its parent the National Academy, had no Federal appropriation; it had limited private resources and no authority in governmental affairs. Thus disadvantaged in dealing with the Federal bureaucracy, Millikan found that his Army rank did little to ease the difficulty. He was a reserve officer and his leverage with Regular Army officers and career civil servants left a good deal to be desired; to achieve his aims he had to rely heavily on tact, influence and persistence. During the war, as a friend recalled, Millikan learned how to "sell science" to a wide variety of people, military and civilian alike.

Millikan's work in the wartime National Research Council particularly impressed the astrophysicist George Ellery Hale and the physical chemist Arthur A. Noyes, who were inaugurating a venture in education and research in southern California. In 1921, at their urging, Millikan moved to Pasadena to become head (as chief of the executive council) of the new and munificently financed Cal Tech and director of its physics laboratory. ("Just imagine," Wilhelm Röntgen exclaimed, "Millikan is said to have a hundred thousand dollars *a year* for his researches!") Millikan brought scientific acumen and entrepreneurial enthusiasm to the job. He recruited the best faculty possible (including, in spite of a streak of anti-Semitism that was not untypical among academics of the day, such Jewish scientists as Paul S. Epstein and J. Robert Oppenheimer). With a knack for making wealthy southern Californians think it a privilege to be invited to contribute to the institute, he enlarged the Cal Tech endowment and physical plant. Millikan's traits, fused with Hale's vision, Noyes's wisdom and all that money, made Cal Tech virtually an overnight success.

Between the wars Millikan was an influential member of the National Academy and of the National Research Council and its fellowship board, which he helped to administer so as to improve the quality of American physics, particularly in theoretical studies. During World War II he turned over an increasing fraction of his administrative responsibilities at Cal Tech to younger staff members who were running various defense projects. Relinquishing his chief executive's position to Lee A. DuBridge in 1946, Millikan remained at the institute until his death in 1953. Throughout the Cal Tech years, in spite of his administrative commitments, he taught a course in atomic physics and took a keen interest in graduate students. He also maintained an active research program almost to the end.

The investigation for which Millikan is best known, the oil-drop experiment, was undertaken in Chicago in about 1909, soon after he put aside textbook writing to concentrate on research. His research on the electron charge actually began not with oil but with water. The original idea, based on the work of the British-born physicist H. A. Wilson, was first to measure the rate at which a charged small cloud of water vapor fell

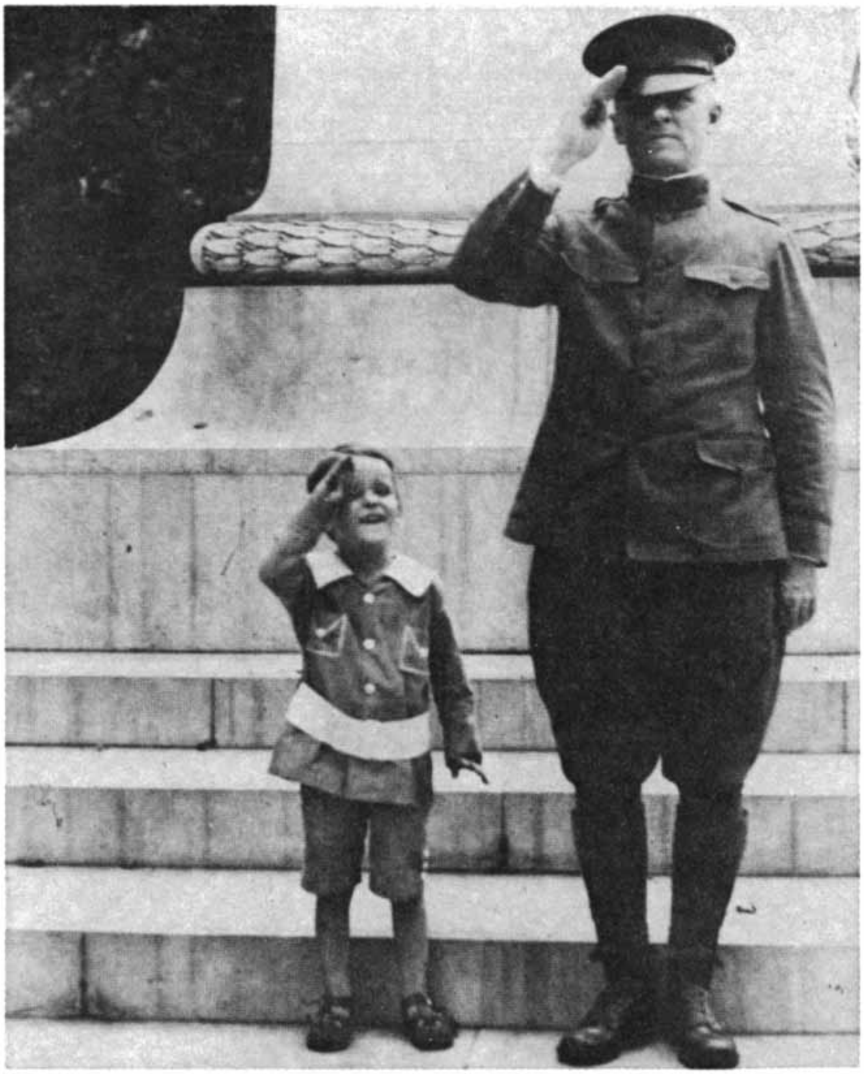


college's preparatory school; as a member of the physics faculty at the University of Chicago soon after the turn of the century; in 1918, when he had made a name as an investigator, and toward the end of his long career as an investigator and an administrator, in about 1940.

under the influence of gravity and then to measure its modified rate of fall under the counterforce of an electric field. Since the mass of the cloud could in principle be determined from Stokes's law of fall, one could calculate the total charge—again in principle. As Millikan recognized, the technique was fraught with uncertainties, including the fact that evaporation at the surface of the cloud confused the measurement of its rate of fall. To correct for this effect he decided to study the evaporation history of the cloud while it was held stationary by a strong electric field.

When Millikan switched on the field, the cloud disappeared, leaving in its place a few charged water droplets moving in a slow, stately manner in response to the imposed electrical force. Millikan quickly realized that he could determine the electron charge accurately by observing the movement of individual charged droplets within an electric field. With this technique Millikan arrived at a rather good value for the electron charge, *e*, which he soon improved by substituting for water a less volatile oil. Since the accuracy of his value was no better than the constants involved in the calculation, he painstakingly reevaluated the coefficient of viscosity of air and the mean-free-path term in the Stokes-Cunningham version of the law of fall. In 1913 he published his value for the electron charge: $4.774 \pm .009 \times 10^{-10}$ electrostatic unit. It was to serve the world of physics for more than a generation, until it was redetermined more accurately in 1928 by Erik Baecklin, who arrived, by an indirect method involving X rays and crystals, at a value of $4.8033 \pm .002 \times 10^{-10}$ electrostatic unit.

No less important than the number itself was the clear-cut implication that could be drawn from the experiment concerning the nature of the electron. Before Millikan the ratio of the electron's charge to its mass was well established as a constant, but the constancy of the ratio did not demonstrate that all electrons were identical particles; some



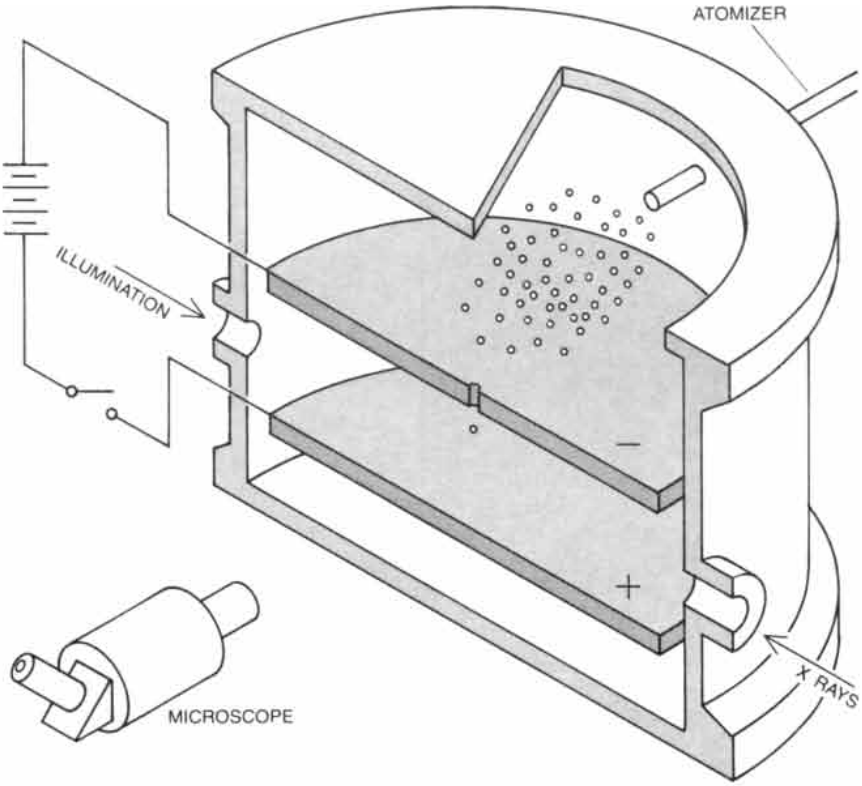
IN UNIFORM AS A LIEUTENANT COLONEL in the Army Signal Corps, Millikan saluted in front of a monument in Washington with his youngest son, Max, in 1918. Millikan was also vice-chairman and director of research of the newly established National Research Council.

antiatomistic European physicists had been maintaining that the electron was not a unique particle and that its observed charge was actually the statistical average of diverse electrical energies. Even in the water-drop stage of his experiment, however, Millikan observed that his measured charge was always an integral multiple—2, 3, 4 and so on—of the same irreducible value. Since the value of the charge, like the ratio of charge to mass, was thereby shown to be a constant, Millikan's experiment provided the conclusive evidence that all electrons were identical fundamental particles.

After completing the electron-charge work, Millikan returned to one of his earliest research subjects: the photoelectric effect. Einstein's quantum-based formula for the phenomenon had not yet been conclusively verified. Taking pains to avoid the mistakes of earlier experimenters, Millikan by 1915 had confirmed the validity of Einstein's equation in every detail, including the linearity of the relation between the maximum energy of the emitted electrons and the frequency of the incident light. In addition Millikan demonstrated that the slope of the line equaled the ratio of Planck's constant h to the electron charge, a result that supplied the best measure of h then available. In spite of his demonstration of the validity of Einstein's equation, Millikan did not believe he had confirmed the quantum theory of light on which it was based. Because of the overwhelming evidence for the wave nature of light he was sure, as many other contemporary physicists were, that the equation had to be based on a false (although obviously quite fruitful) hypothesis.

With the completion of the work on the electron charge and the photoelectric effect, the two investigations for which he was awarded his Nobel prize in 1923, Millikan came to the end of his first research program. He easily found new significant problems to explore. Not long after Millikan came to Chicago, Michelson made him responsible for supervising doctoral research. To find suitable dissertation topics Millikan read *Science Abstracts* regularly, and he became knowledgeable about the rich variety of questions arising in 20th-century physics. Before World War I intervened he managed some preliminary investigation of hot-spark spectra and of the "penetrating radiation" he later christened cosmic rays.

At Cal Tech in the 1920's, working with a hot spark between two electrodes in a vacuum, Millikan embarked with Ira S. Bowen, a graduate student, on a thorough study of the ultraviolet spectra of the lighter elements. By 1924 he and Bowen had extended the observable spectrum down to a wavelength of 1.36 angstrom units, helping to close the last



OIL-DROP EXPERIMENT that established the charge of the electron was conducted by Millikan with the simple equipment shown here very schematically. A droplet from an oil spray introduced into the upper chamber drifts through a pinhole into the space between two charged plates. The charge on the drop can be varied by ionizing the air around it with X rays; an electric field of known intensity can be established between the plates or can be turned off. The rate of rise or fall of the droplet is measured under various conditions by tracking it in the microscope. The various speeds turn out to be proportional to multiples of a single irreducible value: the electron's charge, which Millikan calculated to be $4.774 \pm .009 \times 10^{-10}$ electrostatic unit.

gap between optical and X-ray frequencies. In the course of this work they discovered that, like hydrogen, atoms stripped of their valence (outer shell) electrons had spectra with doublets, or two closely spaced spectral lines. They also learned that certain of their ultraviolet doublets corresponded precisely to various energy levels associated with the X-ray spectra of the heavier elements. The German physicist Arnold Sommerfeld had accounted for these X-ray doublets by a relativistic analysis of atomic orbits. Millikan and Bowen suggested that Sommerfeld's relativistic approach might account for the doublets in the entire field of optical spectra.

They recognized that their results raised a serious difficulty for spectral theory. In Sommerfeld's relativistic scheme different X-ray doublet terms were assigned different quantum numbers, whereas in Niels Bohr's approach to optical spectra the doublets were accounted for by assuming different spatial orientations for orbits with the same quantum number. Since, according to Millikan and Bowen, X-ray and optical doublet terms were indistinguishable, one seemingly had to give up either Sommerfeld's relativistic explanation or the Bohr scheme of spectra. Millikan and Bowen could find no way out of the dilemma, but their forceful statement of it in 1924 contributed, as did that of the German physicist Alfred Landé, to the ultimate resolution of the difficulty through the postulation of electron spin by George E. Uhlenbeck and Samuel A. Goudsmit in 1925.

While pursuing the hot-spark work Millikan also resumed his investigation of penetrating radiation. In 1912 the Austrian-born physicist Victor F. Hess had demonstrated in a manned balloon flight up to 12,000 feet that atmospheric ionization increases with altitude, and on that evidence Hess argued that some type of radiation must be arriving from the heavens. Many scientists nonetheless continued to attribute atmospheric ionization to some terrestrial cause such as radioactivity. Millikan explored the issue by measuring ionization first on top of Pike's Peak and then in the high atmosphere with electroscopes mounted on unmanned sounding balloons.

In the summer of 1925 he proposed to settle the question by measuring the variation of ionization with depth in Muir Lake and Lake Arrowhead in the mountains of California. The two lakes were snow-fed and were separated by many miles and by 6,675 feet of altitude (Muir Lake is the higher); each was likely to be free of both local terrestrial radioactive disturbances and any atmospheric peculiarities that might affect the level of ionization in the other. Millikan's electroscopic measurements showed that the intensity of ionization at any given depth below the surface of Lake Arrowhead was the same as the intensity six feet deeper than that in Muir Lake. Since the layer of atmosphere between the surfaces of the two lakes had the absorptive power of precisely six feet of water, the results decisively confirmed that the radiation was coming from the cosmos.

More penetrating than even the hardest gamma rays known, this cosmic radiation, in Millikan's belief, could not consist of charged particles. Such particles would have to possess energies that were then thought to be impossibly high in order to penetrate, as the cosmic rays did, the combined air and water equivalent of six feet of lead. In Millikan's tentative opinion cosmic rays were likely to be photons, or quanta of electromagnetic radiation. The trajectories of incoming photons would not be affected by the earth's magnetic field; those of incoming charged particles would be affected by it, so that more of them would strike the earth at higher latitudes than would strike it at lower ones. In various experiments conducted in South America and at sea Millikan found no variation in intensity with latitude. In 1929, asked by reporters to comment on the notion that cosmic rays might be likened to charged particles, he remarked: "You might as well sensibly compare an elephant and a radish."

Other workers, notably the Dutch physicist Jacob Clay in 1927, had begun to detect signs of a latitude effect. Furthermore, in 1929 the German physicists Walther Bothe and Werner Kohlhörster found evidence, in coincidence experiments with two particle detectors, that at least a large fraction of cosmic rays consisted of enormously energetic charged particles. Moreover, theorists had predicted that if cosmic rays consisted of charged particles, there should be an intensity difference between those arriving from the east and those arriving from the west. In 1932 Thomas H. Johnson, a young physicist at the Bartol Research Foundation in Philadelphia, detected a slight excess of particles coming from the west on Mount Washington in New Hampshire. Finally, also in 1932, following a worldwide survey at various altitudes, Arthur Holly Compton announced incontrovertible evidence of a latitude effect.

Millikan hotly contested the assertions that cosmic rays were charged particles, in particular Compton's report of a latitude effect. Millikan had repeated his own search for a latitude effect late in 1932 without success. The trouble, he later explained, was that in the longitudinal region of California the dip in cosmic-ray intensity began quite suddenly in the vicinity of Los Angeles and rapidly reached its maximum fall of some 7 percent less than two days' sail south of the city; in Millikan's initial search for the latitude effect his estimated error had been 6 percent. In most of his later searches he went to the north of Pasadena, where the rise in intensity was too small to detect easily. In 1932 he sent H. Victor Neher, a young collaborator at Cal Tech, on a voyage to the south, but Neher could not get his electroscope working until after he had passed the region of the dip.

By 1933, after Neher had found a latitude effect on the return voyage, Millikan was prepared to admit that at least some portion of cosmic radiation must consist of charged particles, but he insisted that any such rays must be secondary ones, produced in encounters between the incoming primaries and the nuclei of atoms in the atmosphere. He continued to maintain that the primaries, or at least some fraction of them, were photons. He clung tenaciously to one or another variation of that belief for 20 years in the face of overwhelming contrary evidence and the bulk of scientific opinion.

One can only speculate about why Millikan insisted on flying in the face of evidence and authority. Possibly it was because with the onset of age (in 1933 he was 65) he had grown scientifically dogmatic and self-centered; he infuriated other cosmic-ray researchers, notably Compton and Clay, who found him cloyingly self-congratulatory about his own work and unblushingly negligent about theirs. Possibly it was because he had committed himself openly and heavily, before the scientific and lay public alike, to the photon theory of cosmic rays. Possibly, too, it was because of his religious views. A Universalist in sectarian affiliation and something akin to a deist in belief, Millikan saw the hand of the Creator beneficently at work in the origin (or at least in his particular theory of the origin) of cosmic rays that were photons.

Millikan's theory proceeded from the fact that no single coefficient of radiation absorption could account for the degree of ionization produced by cosmic rays as a function of depth in the atmosphere. Holding that the absorption curves could be broken down into three independent parts, he inferred that cosmic rays were clustered in three distinct energy bands, at 26, 110 and 220 million electron volts. He recognized that 26 MeV was just about the amount of energy equivalent to the "mass defect" of helium, that is, to the difference between the mass of four hydrogen atoms and the very slightly smaller mass of a helium atom formed by their fusion. Similarly, 110 MeV was close to the mass defects of oxygen and nitrogen, and 220 MeV was the mass defect of silicon.

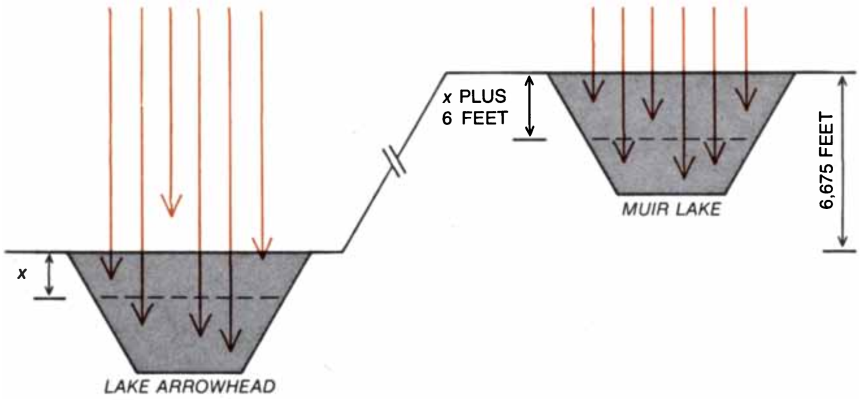
Millikan concluded that the photons striking the earth must be generated when four atoms of hydrogen somehow fused to form helium, when 14 hydrogens fused to form nitrogen and 16 to form oxygen, and when 28 hydrogens fused to form silicon. Cosmic rays, then, were the "birth cries" of atoms, a Millikan phrase that achieved a good deal of currency. Drawing a remarkable religious inference from this theory, he proposed that such creation of the elements, which he called atom building, was proceeding continually, and that the attendant emission of energy arising from the mass defect was saving the universe from the heat death to which the second law of thermodynamics was alleged to have condemned it. The Creator, as Millikan explained the significance of his atom-building hypothesis, was "continually on His job."

Many physicists, even those who for a while preferred the photon interpretation, scoffed at Millikan's hypothesis, not least because of the kinetic difficulties that would have to be overcome for so many hydrogen atoms to collide simultaneously in space. Millikan wrote off the kinetic objections. After all, he declared, modern physics had moved away from a strictly mechanical view of nature. Nevertheless, by the mid-1930's he had to give up the atom-building hypothesis, mainly because experiments in his own laboratory as well as in others had demonstrated that the bulk of cosmic rays have energies far higher than can be accounted for by the mass defects of helium, nitrogen, oxygen and silicon.

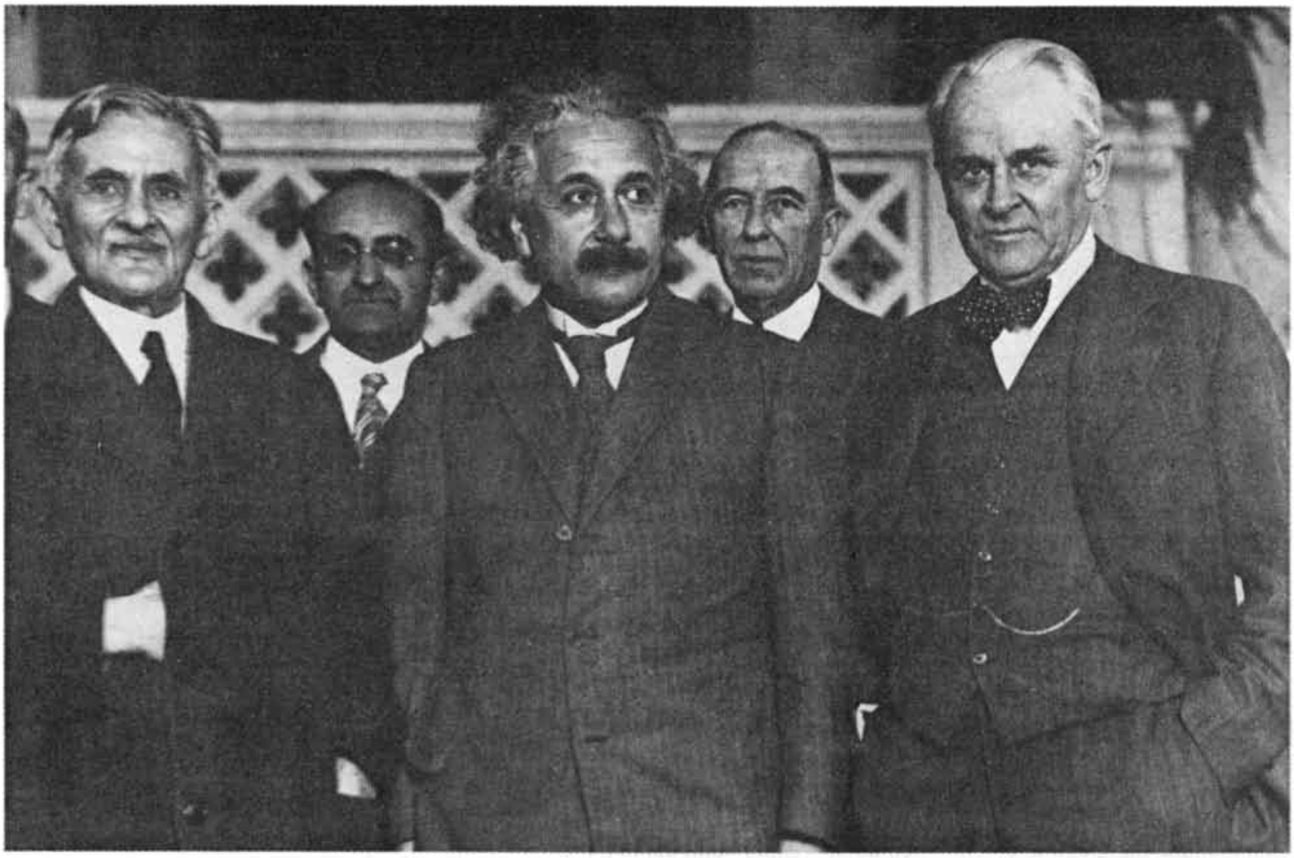
Although he stopped worrying about whether or not the Creator was still on His job, Millikan remained interested in the origin of cosmic rays as a problem in science. In the late 1930's he proposed that cosmic rays originated in the spontaneous annihilation of atoms with an emission of high-energy photons. No more convincing than its predecessor, this hypothesis became untenable (as Millikan himself admitted a few years before his death) after the detection of the pi meson by C. F. Powell and others in 1947 made it clear that the primary cosmic radiation consists almost entirely of protons.

Still, after more than a decade of research, Millikan and his collaborators Neher and William H. Pickering had gathered considerable quantities of important data by measuring cosmic-ray intensities around the world, at sea level, in airplanes at high altitudes and with unmanned sounding balloons at even higher altitudes. In 1934, independently of Clay, Millikan detected the variation of the latitude effect with longitude because of the dissymmetry of the earth's magnetic field. Even the atom-building hypothesis yielded a significant dividend. Eager to obtain a direct measure of cosmic-ray energies, in 1931 Millikan set young Carl D. Anderson to studying the behavior of cosmic rays in a cloud chamber, and Anderson's experiments led to his detection of the positron the following year. As in the case of the positron, Millikan's data had all been obtained in pursuit of mistaken hypotheses, yet the data were no less valuable and the effort was no less striking for the wrongheadedness of their rationale.

Millikan's dispute with Compton, like his "birth cries of atoms" hypothesis, received considerable attention in the press. Millikan decried sensationalism, but the fact of the matter was that he was an avid publicist, one of the best in the scientific community. Articles by or about him were published frequently in both the local and the nation-

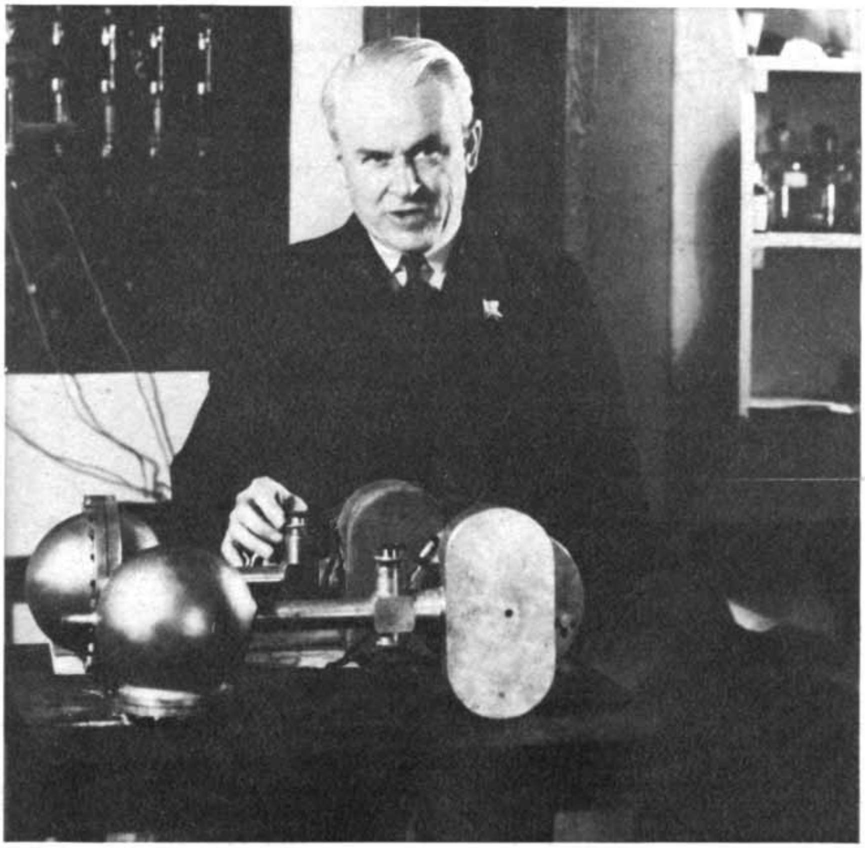


COSMIC-RAY EXPERIMENT devised by Millikan established that the radiation originates in the cosmos, not in the earth or in the lower atmosphere. He measured the ionization produced by cosmic rays in two lakes in California. The ionization intensity at any depth (x) in Lake Arrowhead was the same as the intensity six feet deeper than that ($x + 6$) in Muir Lake; the six feet compensated exactly for the 6,675 feet of air between the surfaces of the two lakes, showing, as Millikan wrote, "that the atmosphere between these two levels acted simply as an absorbing blanket and contributed not a bit to the intensity of the radiation found at the lower level."



THREE NOBEL PRIZEWINNERS in physics were photographed together in January of 1931, when A. A. Michelson (*left*), Albert Einstein and Millikan took part in a conference that was held at California Institute of Technology, which Millikan had headed since 1921.

AT CAL TECH IN 1930'S Millikan was photographed with physicists P. A. M. Dirac (*left*), who was visiting the institute, and J. Robert Oppenheimer, who had come to Cal Tech in 1929 as assistant professor of physics. Dirac won the Nobel prize in physics in 1933.



RECORDING ELECTROSCOPE displayed by Millikan in 1935 was a late model of the kind of equipment with which he had for years been studying ionization caused by cosmic rays.

al press. Radio listeners heard him discourse in his Yankee-sounding twang on the New Deal or the new physics. Irreverent Cal Tech faculty members spoke of the "milli-kan": one thousandth of a unit of publicity.

As a public figure Millikan was most popular and respected in the 1920's, when he won his Nobel prize. (In 1927 he made the cover of *Time,* which noted that he had the face of a "witty and successful banker.") The decade of "normalcy" was also a period of unprecedented interest in science: in science as such, as a source of marvelous technology and also, as the Scopes trial made clear, as a challenge to religious authority. The Reverend Harry Emerson Fosdick was moved to say: "When a prominent scientist comes out strongly for religion, all the churches thank Heaven and take courage as though it were the highest possible compliment to God to have Eddington believe in Him." Millikan spoke out strongly for religion, for moral and spiritual values and for a faith tempered by the open-minded tolerance of the scientific spirit. He also celebrated science as an ally of the laissez faire economic ideas prevalent in the 1920's. The resolution of want, he believed, lay not in governmental economic intervention but in more abundant production by a more abundant industry. "No efforts toward social readjustments or toward the redistribution of wealth," he asserted, "have one thousandth as large a chance of contributing to human well-being as have the efforts of the physicist, the chemist and the biologist toward the better understanding and better control of nature."

Millikan's declarations drew criticism, notably from humanists who chastized him for seeming to argue that morality progressed with the progress of research. While modern man knew more than Socrates, they pointed out, he was neither wiser nor more decent. Industry wrote checks on the intellectual bank account of the sciences, but the sciences did not in turn check the rapacious industrialist. And if science had beneficently enlarged peaceful man's mastery over nature, the humanists stressed, it had also multiplied warlike man's power to kill and destroy. During the Depression the attacks on science became more strident. In response Millikan emerged not only as an exponent of science but also as a staunch defender of it. A conservative Republican, he maintained in the 1930's that investing in basic research was superior to Federal meddling in the economy. The further advancement of science, he argued, would lead to new technology, new industries and ultimately new jobs.

For the sake of economic growth following World War I and of economic recovery in the 1930's, Millikan backed Federal aid to academic science, but he also harbored doubts about it. On the one hand, it was clear that public funds would enlarge the opportunities for investigation and training; on the other, he worried, funds from the Government would inevitably "politicize" science. He was usually vague about whether this politicizing would subject the topics and results of research to intellectual censorship, programmatic control or something else. Whatever he meant specifically, in general he feared Federal interference with the autonomy of academic investigators and their institutions. To avoid that he insisted it was essential to insulate the control of Federal research funds from political manipulation, which in his lexicon meant from politically responsible officials, elected or appointed. To that end Millikan consistently proposed to vest control over public money for academic science directly or indirectly in the National Academy of Sciences. Just as consistently, public officials repeatedly balked at turning over any such public function to a private agency.

Whereas before 1940 most of Millikan's scientific contemporaries had agreed with his view of Federal funding of research, he later found himself increasingly isolated in this respect. After World War II most leading American scientists endorsed the establishment of the National Science Foundation, a governmental body designed with safeguards against crude political control. In spite of the safeguards, Millikan was willing to go along only with a foundation whose membership was controlled by the Academy. He adamantly opposed the establishment of what he called "a new large Federal agency politically responsible and politically controlled." He was appalled that "the great majority of scientists" had evidently lined up behind "pork-barrel science bills." His distress stemmed not only from his old fear that a system of Federal aid to academic institutions would lead to "the ultimate control of education by the party in power" but also from his concern about scientists getting into the business of pleading for public support from Washington. "Despite all our protestations of holiness in motive," he warned, the public would come to see scientists as just another interest group, like the merchant marine, commercial airlines or agriculture.

Most American scientists, however, endorsed the creation of the vast postwar Federal apparatus for aid to academic research, including not only the National Science Foundation but also the Atomic Energy Commission, the expanded National Institutes of Health and the Office of Naval Research. The dissent of Millikan and his few allies was overwhelmed by the flow of Federal funds into university science, the rising affluence of the scientific community and the subsequent American command of world science. Millikan had maintained that private resources could continue to suffice for academic research as they had before 1940. Few scientists agreed with him, and doubtless he was wrong on that score. On the other hand, few thoughtful scientists or academic administrators nowadays would find fault with Millikan for his fears. Even if the Federal Government is far from having seized totalitarian control of education and research, it has begun to exercise a degree of control—some would say excessive interference—in American universities that many academic scientists find uncomfortable, to say the least. And if the public credibility and authority of the scientific community have diminished in recent years, one reason may have been the rising suspicion that scientists are concerned at least as much with self-advancement as with the advancement of knowledge.

In his autobiography Millikan said little, apart from propounding some general strictures against Federal aid to education, about his role in the historical attempts to achieve Government funding of academic science. He devoted only a few pages to the years of labor with cosmic rays. By 1950 he obviously found both subjects awkward to recall, and in the case of cosmic rays perhaps even painful. The autobiography may have dwelt on the earlier chapters of his life not simply because (as is often the case) they loomed larger and sharper in memory but also because they were chapters of unmitigated success.

Yet Millikan could have taken a certain pride in the later chapters of comparative failure. If it is true that he had fallen increasingly out of step on the issue of Federal patronage, it is also true that in his opposition he had raised questions that would return to trouble the scientific community. And if he was on the losing side in the debate over the nature of cosmic rays, in searching for their origin he had tackled a problem of no small difficulty. In 1950 Millikan attached a note at the top of his last paper on the atom-annihilation hypothesis: "New evidence has appeared since this was written which is unfavorable to this hypothesis, but the experimental data herein contained is valid. The actual origin of the cosmic rays is still today an unsolved mystery." The mystery remains unsolved.

BALLOON EXPERIMENT designed to measure cosmic-ray intensities in the stratosphere brought Millikan (*left*) to Fort Sam Houston in Texas in August of 1935. He did valuable research on cosmic rays in spite of the idiosyncratic nature of his hypotheses of their origin.

THE AMATEUR SCIENTIST

How to make dazzling photomicrographs with simple and inexpensive equipment

by Jearl Walker

The hidden world revealed by a simple microscope has fascinated many an amateur scientist. Over the past 20 years a number of amateurs have become notably proficient at photographing what they see in the instrument. This month I shall review some of the techniques that enable the amateur microscopist to make such photographs.

Information on photomicrography is available from several sources. Mine comes from James Bell of Allston, Mass., whose photographs have appeared in *Scientific American* and who now supplies examples (reproduced on the next two pages) of what his photomicrographic techniques can achieve. The variety of specimens you can examine and photograph is almost endless, and so Bell's examples are meant only as an introduction. Although professional photographers often work with elaborate and expensive equipment, Bell's photographs show that even with fairly simple and inexpensive equipment you can make strikingly beautiful and scientifically revealing photographs.

Bell worked with a monocular Bausch and Lomb microscope made in 1915. It was fitted with old Leitz objectives that had magnifying powers of 10 and 100 and with eyepieces that had magnifying powers of 10, 7 and 2.5. He also had a mirror to direct light into a condenser that caused the light to converge on the specimen. He removed the lens from his 35-millimeter Asahiflex camera and attached the camera to the microscope with a commercial lighttight microscope adapter. He did his focusing on the ground-glass back of the camera. To avoid jarring the camera during his exposures Bell relied on a cable release. He made the photographs at shutter speeds between 1/500 and 1/25 second and with Ektachrome-X or Ektachrome-64 film.

Intense light is needed because only a small fraction of it reaches the film. Bell got his light from an ordinary slide projector with a 400-watt bulb. To collimate the light he replaced the lens of the projector with another light condenser. If you work with less light than Bell did, you would have to increase the exposure time by a second or more. Clear photographs would be more difficult to obtain; with moving specimens you would be likely to get only a blur. If the light was too bright for a particular photograph, Bell put one or more neutral-density filters in the light path from the projector. Such filters can be bought from the Edmund Scientific Company (7875 Edscorp Building, Barrington, N.J. 08007) or made by photographing a sheet of white paper and developing the negative.

The infrared component of the light beam may be harmful to the specimens, particularly if they are living organisms. You can remove it from the beam by placing a container of water (the sides should be parallel to avoid unwanted optical effects) or several pieces of optical glass in the light path. A projector may have a built-in heat sink that would serve the same purpose.

Bright-field illumination is the easiest of several kinds of illumination to use but is quite limited in revealing details of the specimen. Bell achieves this kind of illumination by angling his mirror so that the substage condenser (the one below the specimen holder) causes the light to converge on the specimen. When you do this, you see the light transmitted through the specimen. The technique therefore does not work with thick or opaque specimens and does not always show the internal structure of thin specimens. Moreover, the natural colors of a specimen are often lost in the bright light. Still, the technique may be useful in black-and-white photography of thin specimens.

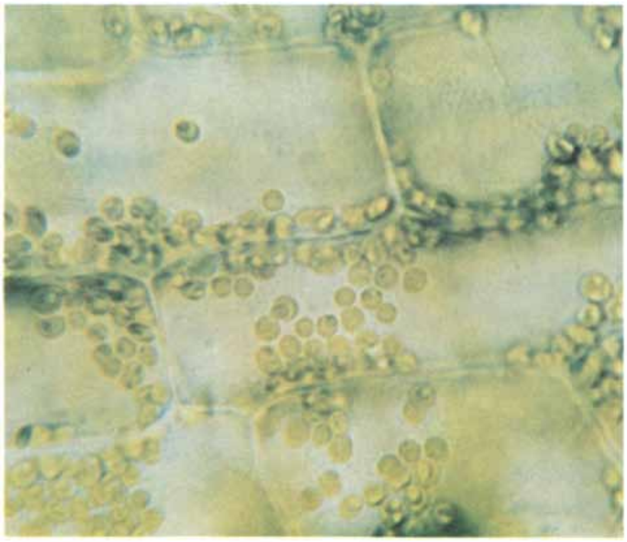
To obtain the highest possible resolution, which is crucial in high-power work, you should follow the directions for Köhler illumination in *Photography through the Microscope,* the bible of photomicroscopy published by the Eastman Kodak Company. In essence the procedure calls for you to adjust the mirror, the collector (the lens between the mirror and the light source) and the condenser so that the filament of the lamp is focused on the condenser and the condenser focuses an image of the collector onto the specimen holder. With this adjustment the specimen is evenly illuminated and you do not see a focused image of the filament.

The first of Bell's photographs, which is at the top left on the opposite page, is an example of the bright-field technique. Bell used a magnification of 1,100 to photograph living cells of the freshwater alga *Elodea* to reveal the individual chloroplasts. Moments after the photograph was made the chloroplasts began to circle in the cells as they engaged in photosynthesis in the light from the projector. Although details appear in the photograph, the image is not as sharp as it is in the other photographs, partly because the high magnification reduced the depth of field.

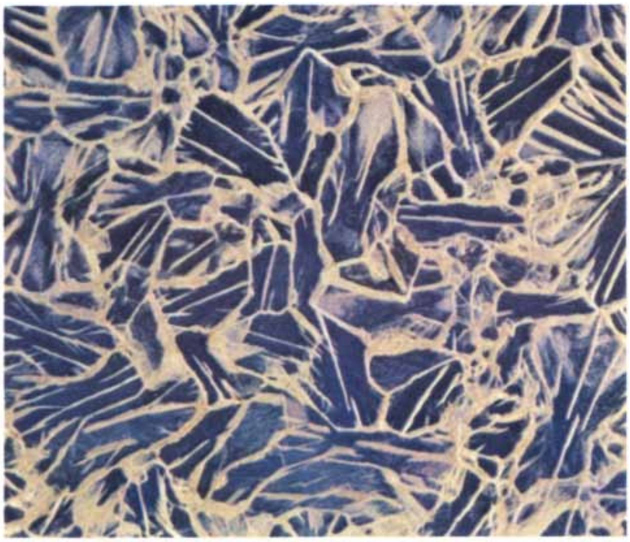
You can get good contrast in your photographs by putting the specimen on a dark background. To create such a background while still illuminating the specimen Bell placed a black dot on a clear glass filter that he slipped into the filter holder just below the substage condenser. The diaphragm of the condenser was opened either completely or almost so to provide a wide cone of light. The purpose of the black dot is to cast a shadow against which the specimen is to be photographed. The light leaking around the edge of the dot illuminates the specimen in white light so that its natural colors are still seen. In the absence of a specimen the crossing light rays would reach the objective of the microscope so diverged that no light would enter the tube, and you and your camera would see only a dark field. Once the specimen is in place, however, light scatters from it into the tube, enabling you to see the specimen. This technique obviously requires a strong light source because only a small fraction of the initial light will reach the camera.

The dark-field filter is made by gluing a small opaque circle of black construction paper on a transparent glass filter. The circle should be about five millimeters in diameter or somewhat larger. Determine the size by trying it in your microscope. The field should be dark when no specimen is in the holder but properly illuminated when a specimen is there. Instructions for determining the size were given in this department some years ago [April, 1968], along with much other information on photomicroscopy.

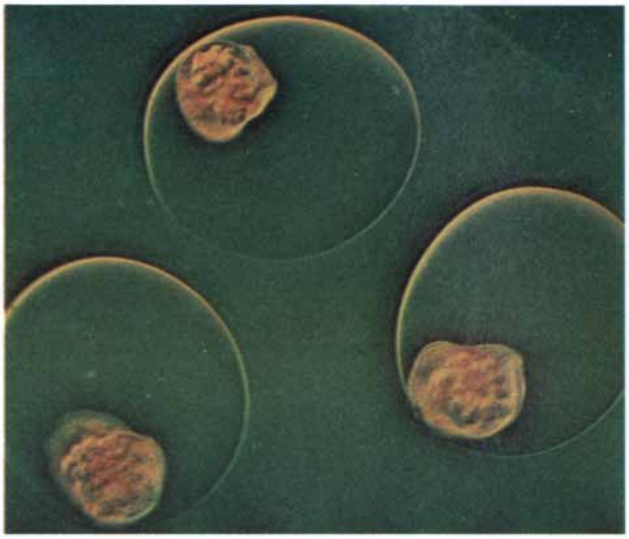
Bell has been able to use dark-field illumination with only his lower-power objective (10×). Greater magnification produces photographs with less contrast because of the difficulty of keeping the objective lens in the dark cone cast by the opaque dot. Objectives of higher power must be brought closer to the specimen for proper focusing, and then the lens begins to intercept some of the light rays coming around the opaque



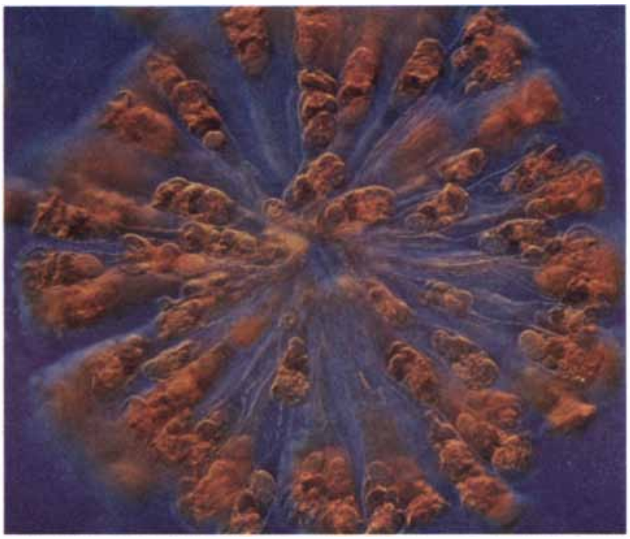
Alga in James Bell's bright-field technique



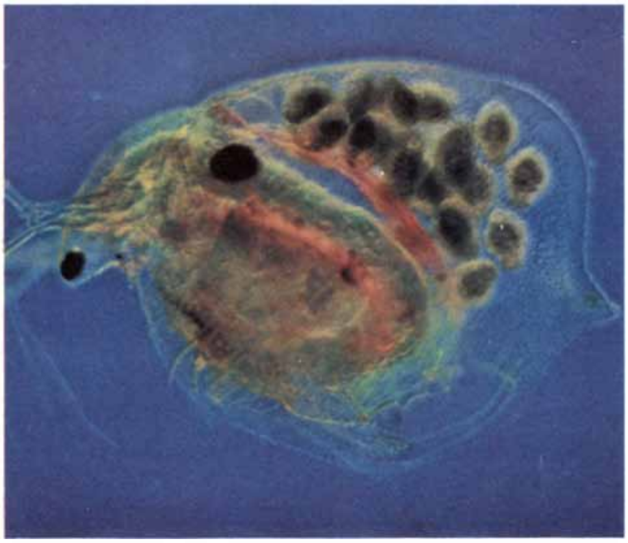
Cholesteric ester crystals



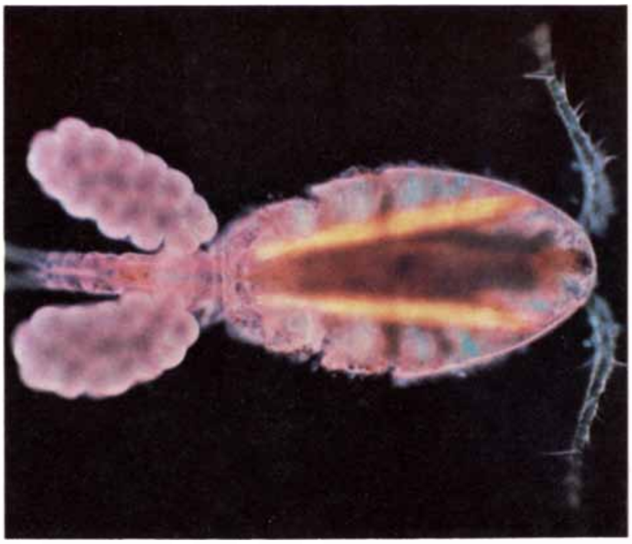
Snail embryos



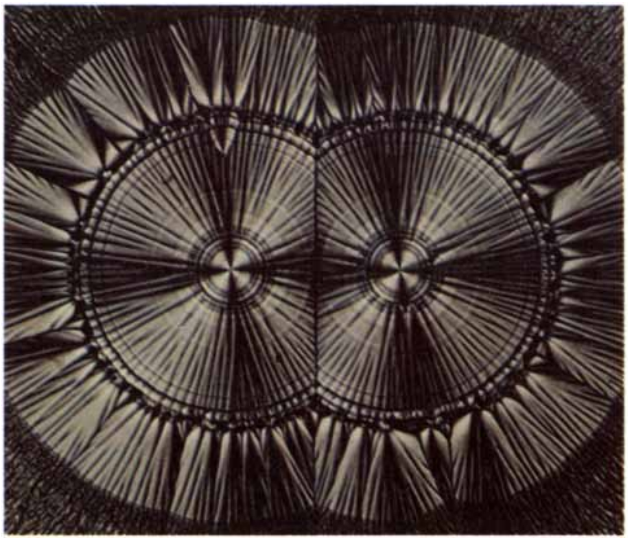
Colony of **Conochilus** *rotifers*



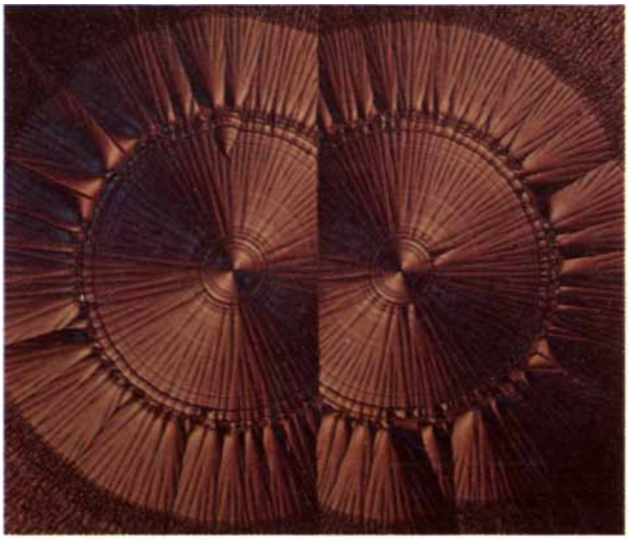
The water flea **Simocephalus**



The copepod **Cyclops**



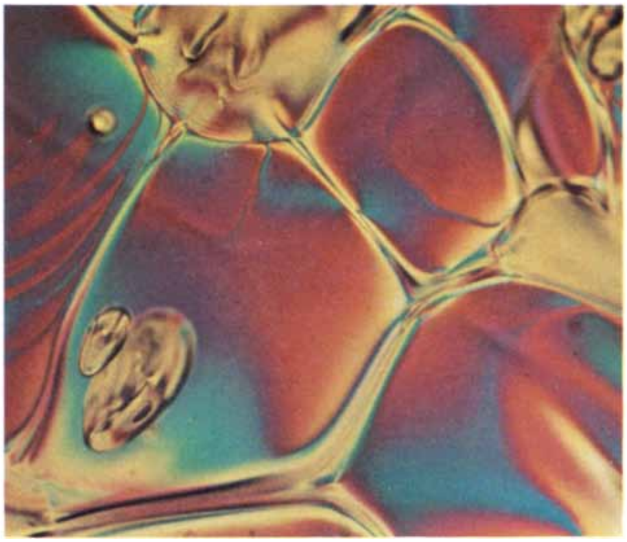
Vitamin C crystal



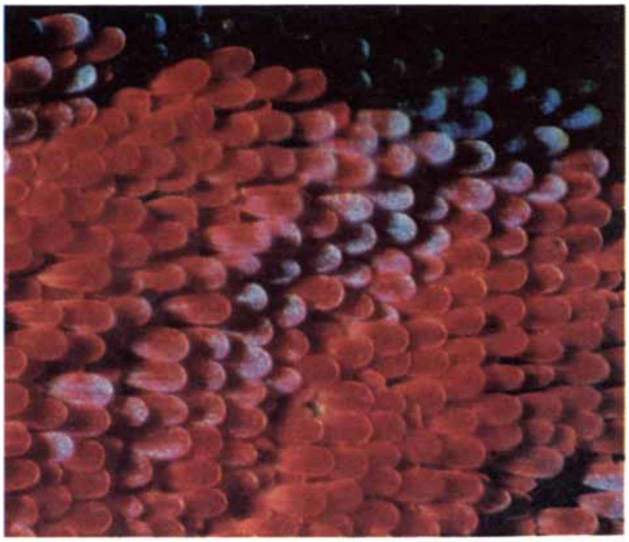
Same crystal through retardation filter

Hippuric acid crystals

Resorcinol crystals



Nematic liquid crystal



Butterfly wing

dot. To go to greater magnifications with the technique Bell would have to buy a dark-field condenser that provides a finer cone of darkness.

With the dark-field technique Bell photographed a specimen of hydra and some volvox colonies. He found these specimens in local ponds, as you can. Descriptions of the rich array of specimens available are given in a number of books of the kind you can buy in a natural-history museum. My favorites are the two children's classics by Richard Headstrom listed in the bibliography for this issue [*page 164*].

Since Bell's kind of photography can be done with live specimens, your photographs (if you are patient enough) may catch the organisms in various types of characteristic behavior. For example, you can photograph a hydra eating. By feeding it small aquatic worms or tiny chunks of raw beef you can induce it to open its mouth and can see how it uses its tentacles. Headstrom notes that with a small amount of acetic acid or methyl green you can provoke the hydra to discharge one of its nematocysts. These barbed tubes are fired with such force that they can penetrate the animal's prey.

Volvox is a gelatinous spherical colony made up of hundreds of flagellate organisms. If you look at a colony closely, you will find that the surface is embroidered with flagella, the whiplike appendages by which individual flagellates propel themselves. Headstrom recommends collecting specimens of this kind with a glass tube. Once you find a specimen in a pond (perhaps with the aid of a magnifying glass) lower a tube into the water while keeping a finger over the top of the tube. Then remove the finger. The specimen will be drawn up into the tube by an influx of water. Replace your finger to retain the specimen and the water in the tube. Carefully release the water over a microscope slide until you have the specimens on it. The best slides are the ones with a concave well to hold the specimen and a little water.

Some specimens contrast better with a colored background than with a dark one. To accomplish this coloring (called partial Rheinberg illumination) Bell replaced the opaque black dot with a transparent colored dot. For example, to provide contrast for a red specimen he would use a blue or green dot cut from colored plastic filters of the type available from Edmund Scientific. The specimen continues to appear in its natural color, since it is still illuminated primarily by the white light leaking around the dot.

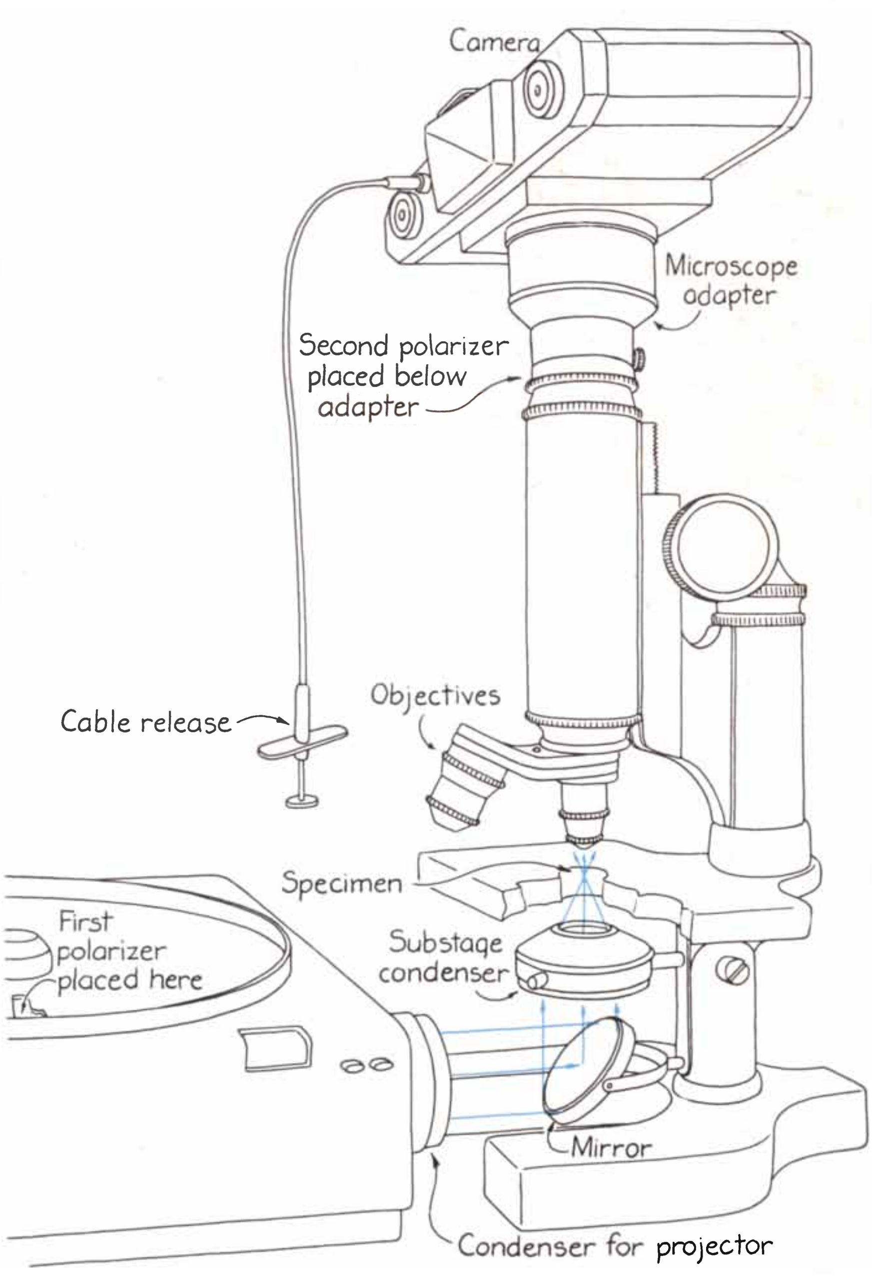
A blue background served in Bell's photographs of the crystals of an ester of cholesterol that he had recrystallized under a glass cover slip. The background enables you to see the interfaces between the crystals and the cracks caused by stress inside individual crystals. To reveal some freshwater-snail embryos still within the egg, Bell made the photograph with a green dot at 80×. This delicate detail would have been totally lost in bright-field illumination.

With full Rheinberg illumination Bell can select any color to illuminate the specimen and can choose a different color for the background. The only change in the procedure is to glue a ring of lightly colored plastic around the darker dot. Then the light transmitted through the dot gives the background color as the light coming through the ring colors the specimen.

Bell photographed the colonial rotifer *Conochilus* with a blue dot and a yellow ring at 120×. He pointed out in a letter to me that the rotifers would have been almost invisible in bright-field illumination; the color contrasts are essential in order to distinguish their edges and internal structure.

Bell finds that the most effective technique for adding color contrast to his photographs is the use of polarized light. The technique works, however, only with birefringent materials such as crystals, muscle and certain small multicellular organisms. A polarizing filter is inserted in the slide projector (just as an ordinary slide would be) to provide linearly polarized light. A second filter, put on the eyepiece of the microscope, can then be turned until the field of view through the microscope is either dark or bright as the filter either blocks or transmits the sense of polarization coming through the instrument.

When a birefringent material is insert-



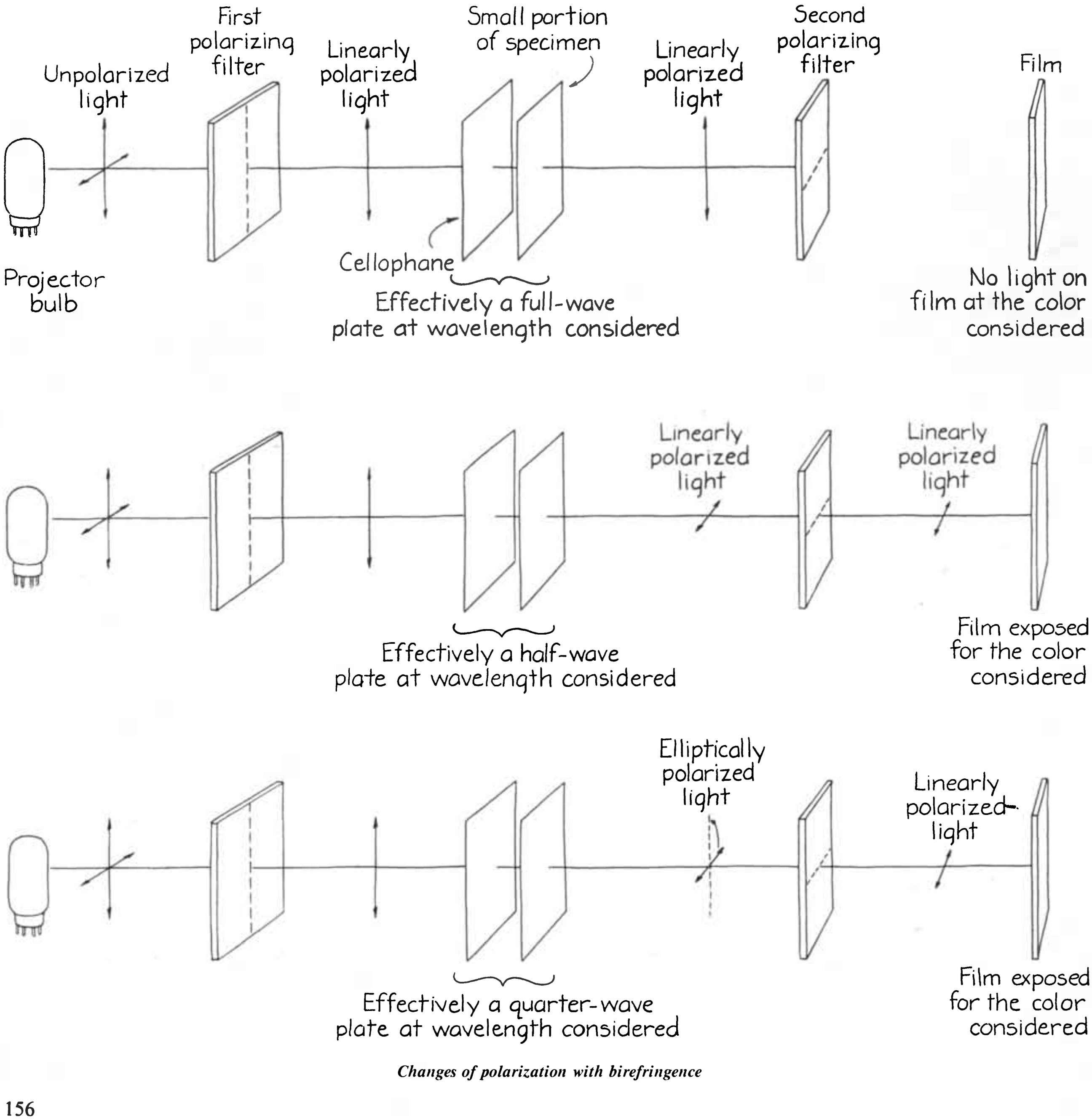
Bell's experimental setup

ed in the specimen holder, the polarized light produced by the first filter is altered in polarization when it passes through the material. Appropriately altered, some of the light can then pass through the second polarizing filter above the eyepiece even though the filter would otherwise block the light. (The details of how the sense of polarization is changed by the birefringent material were covered in this department in December, 1977.) In essence the light is transmitted through the material when part of the light is polarized along one axis and another part is polarized along a perpendicular axis. Both axes are perpendicular to the light ray. The velocity of light is higher for one of the senses of polarization than it is for the other. A result of this difference is that the two senses of polarization can emerge from the material out of step. Since the emerging polarization is determined by the combination of the two senses of polarization, the emerging light could be polarized in several ways.

If the two senses of polarization emerge out of step by half a wavelength, the emerging light is again linearly polarized, but now along an axis perpendicular to the polarization sense of the incident light. This emerging light might be transmitted or blocked by the polarizing filter at the eyepiece, depending on the orientation of the filter.

If the birefringence has forced the emerging senses to be a quarter of a wavelength out of step, the emerging light is elliptically polarized, which means that the tip of the polarization vector rotates about the light ray, mapping out an ellipse. This light will be passed by the filter at the eyepiece regardless of the filter's orientation. If the emerging senses of polarization are in step, the emerging light is polarized in the same way that the incident light is. The orientation of the second filter will then determine whether or not you see any light.

Which of these results occurs depends on the thickness and birefringence of the specimen and on the wavelength of the light. For a particular area of the specimen the red end of the white light incident on the specimen may end up being blocked by the second polarizing filter whereas the blue may be transmitted. That area would appear blue to you. Another area may be colored yellow be-



Changes of polarization with birefringence

cause the other colors of the incident white light are blocked. The advantage in using polarized light in this way is that the color variation produced in an otherwise colorless specimen enables you to distinguish details and internal structure in the specimen.

Bell sent along two photographs made with polarized light. The first one (at 120×) shows hippuric acid crystals that had been dissolved in isopropanol and then allowed to recrystallize on the slide. The sample was sufficiently birefringent and nonuniform to produce many colors. In his other example (at 150×) resorcinol crystals were dissolved in acetone, which was then burned off to facilitate quick recrystallization.

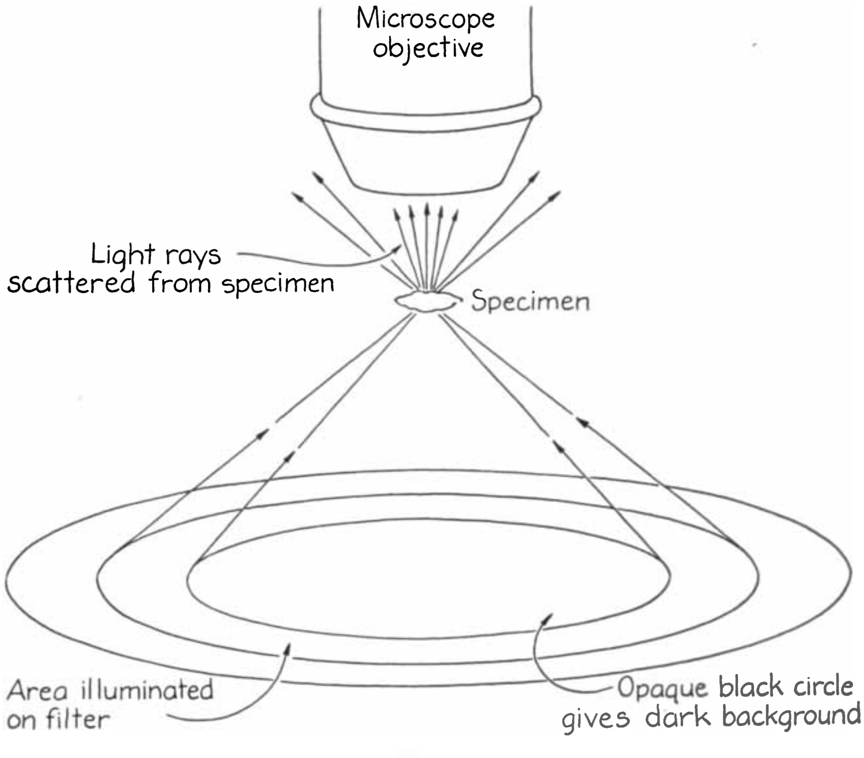
Further color contrast can be produced by inserting one piece or more of clear cellophane or transparent tape in the filter holder below the specimen. Both are birefringent materials that alter the polarization of the light. Instead of cellophane you can use a thin sheet of mica, which can be easily split off from thicker layers of mica with a razor blade. Either material is described as a retardation filter because it forces one sense of polarization to lag behind the perpendicular sense. A retardation filter is useful in enhancing the color variation in a weakly birefringent specimen. Not all kinds of cellophane work, and you will have to experiment to find a kind that does. You could also build up layers of stretched plastic food wrap, following the procedure I gave in my article on birefringence.

Bell has submitted several other examples of his use of polarized light. A living water flea with its young in a brood pouch was photographed at 100× with a blue dot to provide background and with cellophane and polarizing filters to enhance the internal structure of the animal. The eyespot is left black whereas the digestive tract is made orange.

The copepod *Cyclops* was photographed (at 50×) in a similar way except that a black dot was employed to give a dark background. The animal appears with its eyespot red, two muscle bands yellow and its digestive tract brown. The two pear-shaped sacs on the sides contain eggs. A great deal of this detail would be lost in simple bright-field illumination.

The effect of cellophane can be seen in two of Bell's photographs, both of which are of ascorbic acid (vitamin C) that he recrystallized. The photographs show the same two adjacent crystals and were made identically except that cellophane was used in the second photograph. Notice the additional color due to the birefringence of the cellophane.

Bell prepared the crystals by dissolving the ascorbic acid in isopropanol, putting some of the solution on a microscope slide and then burning off the alcohol. The crystals formed within a few seconds. After one set of photographs has been made he can redissolve the ascorbic acid and start over. The technique yields crystals thin enough to be transparent and therefore readily photographed. An alternative method of



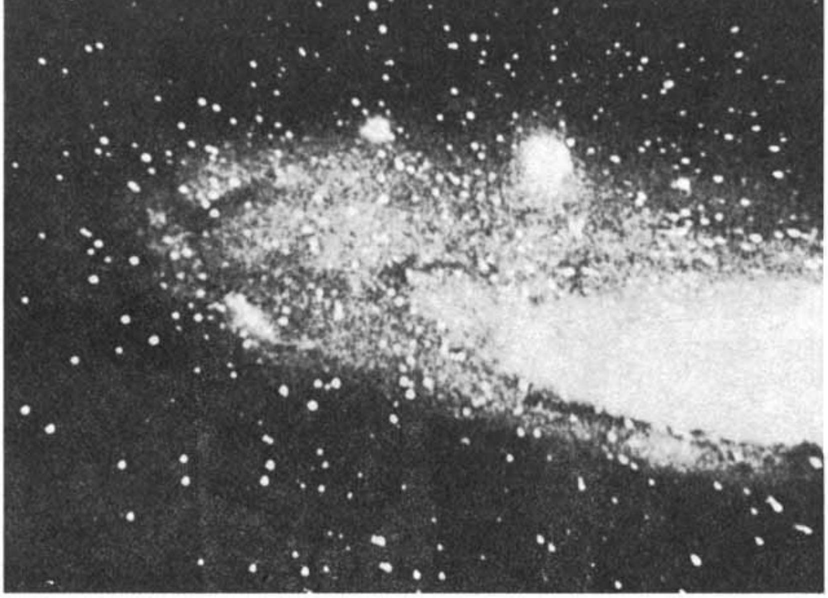
Arrangement for a dark background

forming crystals is to allow the solvent to slowly evaporate.

Another way to make crystals is to melt the substance on a glass slide over a flame. It will crystallize as it cools. Bell does not favor this technique because the melting point of some substances is too high and he invariably ends up breaking too many slides. He does point out that if you want to study certain kinds of liquid crystal, you will have to use the technique, cooling the substance slowly from its melting point in order to get the liquid crystalline state. Bell's photograph of a nematic liquid crystal was made at 100× at room temperature after he had dissolved the material in acetone. The motion in this recrystallization was so rapid that he had to shoot the photograph at 1/500 second to stop the activity.

Some specimens are too thick to be photographed with transmitted light and instead call for reflected light. To accomplish this Bell props his projector up on books and adjusts the angle of the incident light until he gets the right illumination on the specimen. An example of the result is his photograph of a butterfly wing. Many of the colors you see in butterfly wings, on the back of beetles and in bird feathers are due not to pigmentation but to the interference of light waves. Some butterfly wings consist of layers of transparent cuticle. When light is reflected from one of the layers, some of it comes from the top surface and some of it comes from the back surface after passing through the layer. The two emerging rays can interfere with each other to produce colors, just as they do in thin soap films. (I described such interference in this department last September.) You might want to try several different kinds of butterfly wing. They can be ordered from Gilmour-Vendco, Inc. (12685 Highway Nine, Box 196, Boulder Creek, Calif. 95006), or bought at gift shops in many cities. The interference colors would be weak if the wing were illuminated only with transmitted light. Then the colors would have to arise from the interference between light transmitted through the cuticlelike layer (with no internal reflection) and light reflected twice inside the layer before it emerges. This light would be so much weaker than the unreflected light that little interference (and hence little color) would result.

To find the correct exposure you can either bracket your shots with a range of exposure times or attach a light meter to the microscope. Bell employs a homemade light meter featuring a digital millivolt meter and a silicon photocell. He has mounted the cell in a cardboard housing that fits snugly over the viewfinder of the camera. As the light intensity varies with the specimen and the mode of illumination, the resistance of the photocell varies inversely. Through intermediate circuitry these variations



change the readout on the millivolt meter. With some experimentation Bell was able to construct a table of readouts so that he could immediately interpret them in terms of the proper shutter speed. (A design for a simple but highly sensitive light meter employing a cadmium sulfide photoresistor appeared in *Popular Electronics* for June, 1977.)

If you want good reproduction of the natural colors of a specimen, you must be careful about the type of color film you use: it must be matched in color temperature to your lamp. All the lamps look white to the human eye, but the actual distribution of intensities across the visible spectrum will depend on the temperature of the emitting surface in the lamp. A low-temperature lamp of white light has fewer of the low frequencies of visible light (that is, the blue end of the spectrum) than a lamp of higher temperature. The various color films have been adjusted somewhat to compensate for this difference in the color distribution of white light. Nevertheless, there may be some discrepancy between the true colors of a specimen and what you see in a color photograph. You can remedy the situation by putting a color filter in front of the projector (or whatever light source you use). The Kodak book explains which filters should be employed.

Bell has a 400-watt DAT lamp in his projector. The temperature of the emitting filament surface is so high that the light is similar to sunlight in its color distribution. Bell therefore uses daylight color film instead of film that is color-balanced for tungsten lamps operating at a lower surface temperature.

Teresa Owens, as an undergraduate at Reed College in Portland, Ore., has done some careful (and exhausting) work on the rate of cooling of water from various initial temperatures. As I said in this department for September, 1977, hot water will sometimes reach the freezing point before initially cooler water does. The experiment is so rich in variables that verifying the effect and tracking down its cause are challenging.

With an experimental setup similar to mine Owens monitored the temperature of the freezer environment and the temperature of several places in a beaker of water placed in the freezer. Her results indicate that the effect is probably present but can easily be lost in the variation of the environmental temperature between experimental runs. From her data she concludes that the two most important factors determining the effect are the initial environmental temperature and the circulation of air over the top of the container. She believes the temperature gradient inside the water and the mass lost to evaporation are not as important. Much more work could be done on this experiment. If you pursue it, please let me know what you find.

HP measurement and computer advances

The HP 3000 Series 33 computer system. Another innovation that, through advanced LSI technology, delivers major business system capability in a desk-styled package.

With the Series 33, Hewlett-Packard adds a low-cost, compatible member to its 3000 computer family: powerful systems that adapt to your operational requirements for the management of manufacturing and other business information.

A new, dramatic example of HP's advanced silicon-on-sapphire (SOS) technology is the HP 3000 Series 33. Exhaustively tested and refined, SOS semiconductor chips give the Series 33 the related benefits of smaller size, greater performance, and lower cost than computers using traditional technologies.

This general-purpose major business system is especially well suited to on-line transaction processing applications such as order processing, inventory control, and material management. It is a relatively low-cost answer for many companies that can afford only one computer, or for larger organizations implementing distributed processing.

Its power can be simultaneously tapped by users at up to 32 terminals, and the Series 33 runs the same multiprogramming executive operating system—MPE-III, as the larger HP 3000 systems. This means the new Series 33 can immediately share in the thousands of man-years of software development for HP 3000 systems. The Series 33 also provides IMAGE/3000, HP's award-winning data base management software, and its easy-to-use QUERY capability. A new data entry software package, VIEW/3000, makes it easy for experts and beginners alike to create business forms on system video terminals and to enter data into files without writing a program.

The Series 33's system console includes a built-in diagnostic microcomputer for on-the-spot diagnostic testing. With a modem, HP service specialists can even communicate with the Series 33 by telephone for long-distance troubleshooting.

A base system includes the computer—now only three SOS chips on two printed circuit boards instead of nine printed circuit boards—256k bytes of error-correcting main memory, a 1 megabyte flexible disc, a 20 megabyte cartridge disc, and the system/maintenance console. The Series 33 can grow with your needs to 1 megabyte of main memory, and can use up to 8 disc drives with 960 megabytes of disc storage. A wide range of peripherals—magnetic tape drives, printers, and terminals—further increase a user's ability to tailor an HP 3000 Series 33 to a specific application and changing it as operations demand. Prices start at $70,000*.

A working partnership with HP.

This informative management booklet outlines HP's approach to doing business. It summarizes the expertise, resources, support, and computer products we bring to customers. For a free copy write to A.P. Oliverio, Vice President, Marketing, Hewlett-Packard Co., 1504 Page Mill Rd., Palo Alto, CA 94304.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

MATHEMATICAL GAMES

CIRCLES. D. Pedoe. Pergamon Press, 1957.

LOXODROMIC SEQUENCES OF TANGENT SPHERES. H. S. M. Coxeter in *Aequationes Mathematicae,* Vol. 1, No. 1/2, pages 104–121; 1968.

EXCURSIONS IN GEOMETRY. C. Stanley Ogilvy. Oxford University Press, 1969.

WORLD COAL PRODUCTION

ENERGY: GLOBAL PROSPECTS 1985–2000. Workshop on Alternative Energy Strategies. McGraw-Hill Book Company, 1977.

ENERGY SUPPLY TO THE YEAR 2000: GLOBAL AND NATIONAL STUDIES. Workshop on Alternative Energy Strategies. The MIT Press, 1977.

SURVEY OF ENERGY RESOURCES 1976. World Energy Conference. United States National Committee of the World Energy Conference, 1977.

THE ASSEMBLY OF CELL MEMBRANES

THE MOLECULAR BIOLOGY OF CELL MEMBRANES. Peter J. Quinn. University Park Press, 1976.

MEMBRANE ASSEMBLY: SYNTHESIS AND INTRACELLULAR PROCESSING OF THE VESICULAR STOMATITIS VIRAL GLYCOPROTEIN. F. N. Katz, J. E. Rothman, D. M. Knipe and H. F. Lodish in *Journal of Supramolecular Structure,* Vol. 7, pages 353–370; 1977.

RAPID TRANSMEMBRANE MOVEMENT OF NEWLY SYNTHESIZED PHOSPHOLIPIDS DURING MEMBRANE ASSEMBLY. James E. Rothman and Eugene P. Kennedy in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 74, No. 5, pages 1821–1825; 1977.

MEMBRANE ASYMMETRY. James E. Rothman and John Lenard in *Science,* Vol. 195, pages 743–753; February 25, 1977.

SYNCHRONIZED TRANSMEMBRANE INSERTION AND GLYCOSYLATION OF A NASCENT MEMBRANE PROTEIN. James E. Rothman and Harvey F. Lodish in *Nature,* Vol. 269, pages 775–780; October 27, 1977.

GUIDED-WAVE OPTICS

SWITCHING AND MODULATION OF LIGHT IN MAGNETO-OPTIC WAVEGUIDES OF GARNET FILMS. P. K. Tien, R. J. Martin, R. Wolfe, R. C. Le Craw and S. L. Blank in *Applied Physics Letters,* Vol. 21, No. 8, pages 394–396; October 15, 1972.

INTEGRATED OPTICS. P. K. Tien in *Scientific American,* Vol. 230, No. 4, pages 28–35; April, 1974.

SPECIAL ISSUE: LIGHTWAVE COMMUNICATIONS in *Physics Today,* Vol. 29, No. 5, pages 23–60; May, 1976.

LIGHT-WAVE COMMUNICATIONS. W. S. Boyle in *Scientific American,* Vol. 237, No. 2, pages 40–48; August, 1977.

THE DEEP STRUCTURE OF THE CONTINENTS

THE DEEP STRUCTURE OF CONTINENTS. Gordon J. F. MacDonald in *Reviews of Geophysics,* Vol. 1, No. 4, pages 587–665; November, 1963.

THE CONTINENTAL TECTOSPHERE. Thomas H. Jordan in *Reviews of Geophysics and Space Physics,* Vol. 13, No. 3, pages 1–12; August, 1975.

LATERAL HETEROGENEITY OF THE UPPER MANTLE DETERMINED FROM THE TRAVEL TIMES OF MULTIPLE *ScS*. Stuart A. Sipkin and Thomas H. Jordan in *Journal of Geophysical Research,* Vol. 81, No. 35, pages 6307–6320; December 10, 1976.

COMPOSITION AND DEVELOPMENT OF THE CONTINENTAL TECTOSPHERE. Thomas H. Jordan in *Nature,* Vol. 274, pages 544–548; August 10, 1978.

THE TRAJECTORIES OF SACCADIC EYE MOVEMENTS

THE ANGLE VELOCITY OF EYE MOVEMENTS. Raymond Dodge and Thomas Cline in *Psychological Review,* Vol. 8, pages 145–157; 1901.

DERIVATION OF A MODEL FOR THE HUMAN EYE-POSITIONING MECHANISM. Gerald Cook and Lawrence Stark in *Bulletin of Mathematical Biophysics,* Vol. 29, pages 153–174; 1967.

NEUROLOGICAL CONTROL SYSTEMS: STUDIES IN BIOENGINEERING. Lawrence Stark. Plenum Press, 1968.

DYNAMIC OVERSHOOT IN SACCADIC EYE MOVEMENTS IS CAUSED BY NEUROLOGICAL CONTROL SIGNAL REVERSALS. A. Terry Bahill, Michael R. Clark and Lawrence Stark in *Experimental Neurology,* Vol. 48, pages 107–122; 1975.

TIME OPTIMAL BEHAVIOR OF HUMAN SACCADIC EYE MOVEMENT. Michael Clark and Lawrence Stark in *IEEE Transactions on Automatic Control,* AC-20, pages 345–348; 1975.

GLISSADIC OVERSHOOTS ARE DUE TO PULSE WIDTH ERRORS. A. Terry Bahill, Frederick K. Hsu and Lawrence Stark in *Archives of Neurology,* Vol. 35, pages 138–142; March, 1978.

THE COUPLED MOTIONS OF PIANO STRINGS

DECAY RATES OF PIANO TONES. Daniel W. Martin in *Journal of the Acoustical Society of America,* Vol. 19, No. 4, pages 535–541; July, 1947.

TUNING PREFERENCES FOR PIANO UNISON GROUPS. Roger E. Kirk in *Journal of the Acoustical Society of America,* Vol. 31, No. 12, pages 1644–1648; December, 1959.

FUNDAMENTALS OF MUSICAL ACOUSTICS. Arthur H. Benade. Oxford University Press, 1976.

COUPLED PIANO STRINGS. Gabriel Weinreich in *Journal of the Acoustical Society of America,* Vol. 62, No. 6, pages 1474–1484; December, 1977.

THE HEAD OF THE SPERM WHALE

BUOYANCY CONTROL AS A FUNCTION OF THE SPERMACETI ORGAN IN THE SPERM WHALE. M. R. Clarke in *Journal of the Marine Biological Association of the United Kingdom,* Vol. 58, No. 1, pages 27–71; February, 1976.

PHYSICAL PROPERTIES OF SPERMACETI OIL IN THE SPERM WHALE. M. R. Clarke in *Journal of the Marine Biological Association of the United Kingdom,* Vol. 58, No. 1, pages 19–26; February, 1976.

STRUCTURE AND PROPORTIONS OF THE SPERMACETI ORGAN IN THE SPERM WHALE. M. R. Clarke in *Journal of the Marine Biological Association of the United Kingdom,* Vol. 58, No. 1, pages 1–18; February, 1976.

ROBERT A. MILLIKAN

THE AUTOBIOGRAPHY OF ROBERT A. MILLIKAN. Robert A. Millikan. Prentice-Hall, Inc., 1950.

COSMIC RAYS. Bruno Rossi. McGraw-Hill Book Company, 1964.

THE ELECTRON: ITS ISOLATION AND MEASUREMENT AND THE DETERMINATION OF SOME OF ITS PROPERTIES. Robert A. Millikan. The University of Chicago Press, 1968.

THE CONSERVATIVE MODE: ROBERT A. MILLIKAN AND THE TWENTIETH-CENTURY REVOLUTION IN PHYSICS. Robert H. Kargon in *Isis,* Vol. 68, No. 244, pages 509–526; December, 1977.

THE PHYSICISTS. Daniel J. Kevles. Alfred A. Knopf, Inc., 1978.

THE AMATEUR SCIENTIST

PHOTOGRAPHY THROUGH THE MICROSCOPE. Eastman Kodak Company, 1974.

ADVENTURES WITH A MICROSCOPE. Richard Headstrom. Dover Publications, Inc., 1977.

Introducing the Jeep Wagoneer *Limited*
The ultimate in 4-wheel drive

Built for those who demand nothing but the best in 4-wheel drive performance without compromising on luxury. Like all Wagoneers, the Limited handles bad weather and off-road obstacles as easily as it travels pavement. But now in unparalleled comfort and luxury. The seats are of real leather. The spacious and luxurious interior surrounds you with extra thick carpeting. Handsome woodgrain trim inside and out. AM/FM radio with your choice of CB or stereo tape deck. And of course, Quadra Trac,® Jeep's exclusive automatic 4-wheel drive. Automatic transmission, power steering, power front disc brakes and formidable 360 V-8 engine—all standard to make this the most unique and comfortable Jeep Wagoneer ever.

Fleet and leasing plans available.
See your Jeep dealer for details.

Jeep®

We wrote the book on 4-wheel drive

Jeep Corporation, a subsidiary of American Motors Corp.


That long movie, the overtime football game, and 77 other reasons to get a
4 Hour VHS™ Home Video Recorder.
VHS means home video recorders that let you tape four hours of TV on one cassette. And four hours is long enough to tape specials, movies, sports... just about anything including all those shows that run two or three hours. (And last year there were 77 network programs that ran over three hours.)
Panasonic's Omnivision IV™ VHS™ works with almost any TV set. Lets you record on one channel and watch another. Thanks to its built-in timer, you can tape shows even when you're not at home. And unlike some timers that just turn your unit on and let it record until the tape runs out, the Panasonic timer (in Model PV-1100) can be set for 15 minutes, a half hour, an hour...as long or short as you want. Up to four hours.
Panasonic VHS has features like Auto-Stop, tape counters, pause controls, high-performance recording heads, direct-drive motors and more.
Any VHS tape can be played on any four hour VHS machine. And VHS units are available from some of the most famous names in television.
You can even make your own VHS tapes with an optional black & white or color sound camera. And you can also have your slides or home movies transferred to VHS cassettes. Or buy prerecorded tapes.
So when you want four hours of recording, see VHS 4 hour home video recorders.
TV picture simulated
Cabinetry is simulated wood.
CAUTION: Unauthorized recording of copyrighted television programs, films, video tapes and other materials may infringe the rights of copyright owners and be contrary to copyright laws.
Panasonic® VHS™
4 Hour Home Video Recorder