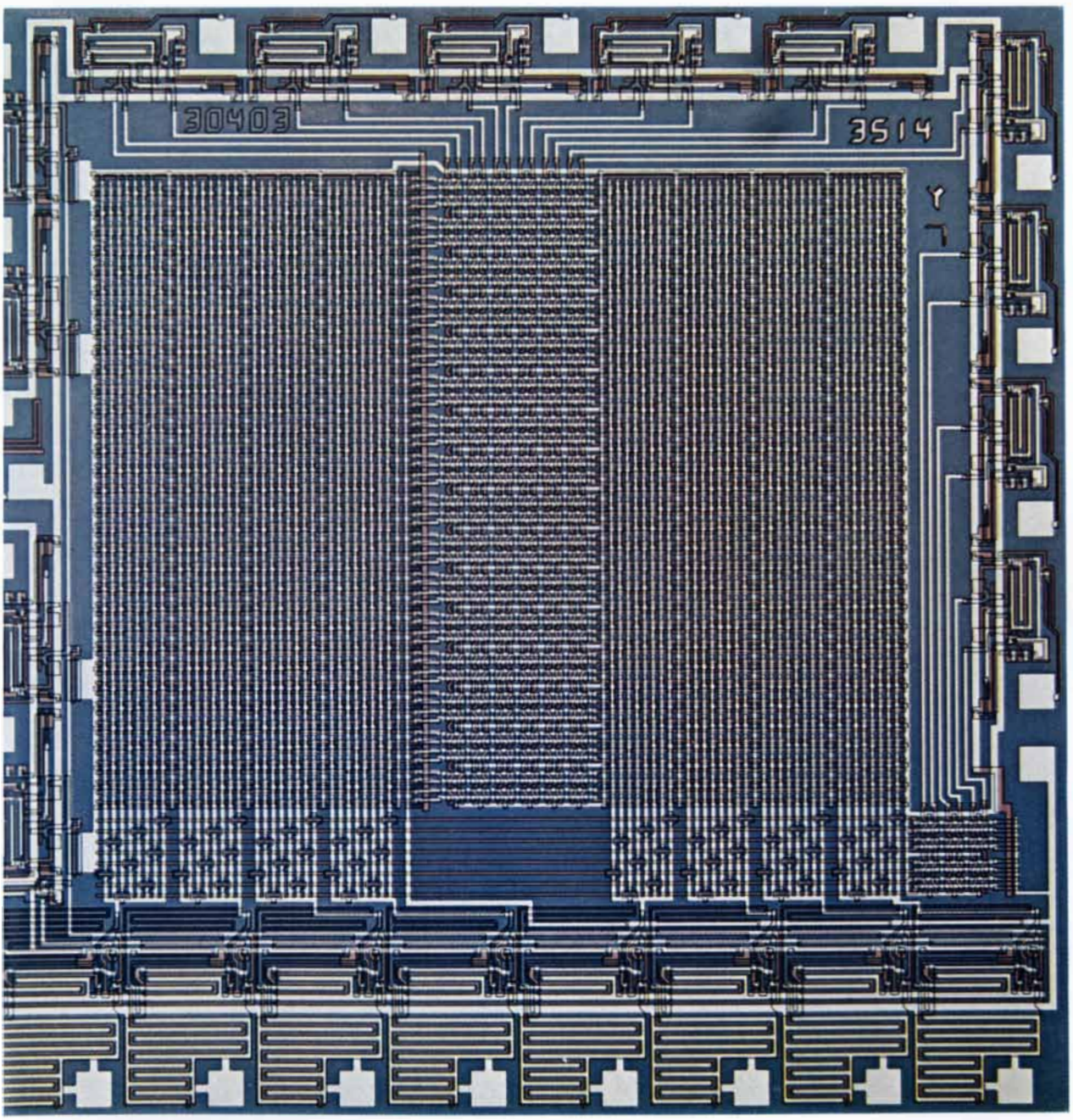
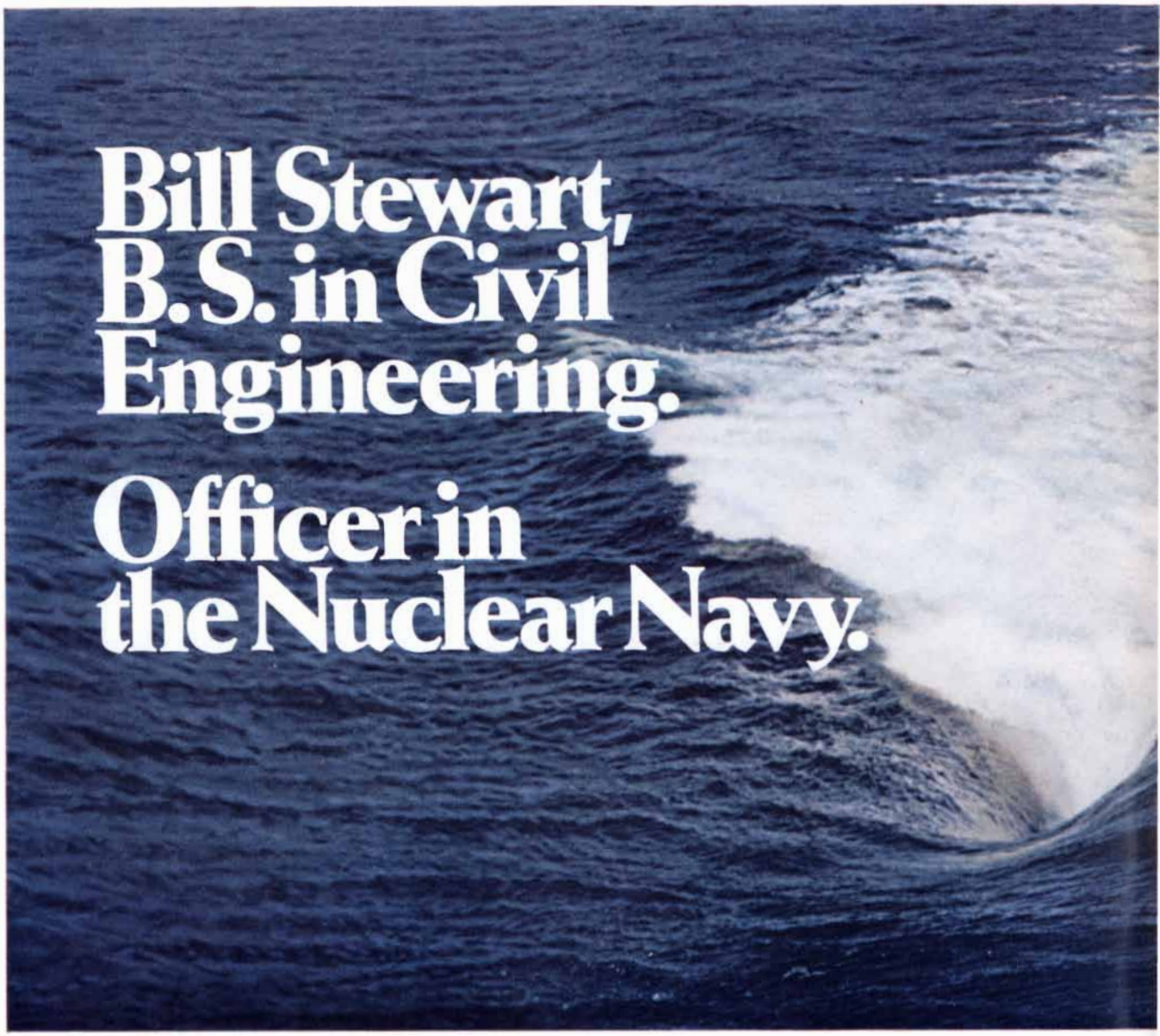
SCIENTIFIC AMERICAN



METAL-OXIDE-SEMICONDUCTOR CHIP

ONE DOLLAR

August 1973



LTJG Stewart, University of Kentucky selected in 1969 for the Navy's

He is already qualified to operate and maintain one of over 100 Navy Nuclear Propulsion Plants.

In 1973, the Navy will accept about 500 outstanding college graduates like Bill Stewart for its Nuclear Propulsion Officer Program. Each one of these young men will have an incomparable opportunity to become a leader in the Nuclear Navy.

For its Nuclear Program, the Navy wants only men who can master advanced technology. While they do not necessarily have to be engineers, many of them will be. They must have a solid foundation in mathematics and physics. They must demonstrate superior ability. High standards limit selection opportunity to only the most qualified candidates.

If you qualify, you will spend a solid year learning all aspects of nuclear propulsion plant theory and operation. All of this training will build upon your ability to solve difficult problems through the application of fundamental engineering principles.

While this program is tough, the Navy pays well for the kind of ability and dedication that is

68, is one of only 429 officer candidates
Nuclear Propulsion Program.

required. Starting salary in the Nuclear Program is comparable with salaries given to junior executives by private companies; and career advancement is rapid. There are many other benefits, including, of course, the opportunity to see much of the world.

If you think you have the abilities and desire to become a leader in nuclear engineering, call the Nuclear Desk, toll-free, 800-841-8000. A nuclear-trained officer will contact you and answer your questions about the program.

Do something special with your life.

TO: LTJG William Stewart, USN, Building 157, Fourth Floor
Washington Navy Yard, Washington, D.C. 20374. 344

Yes, I'm interested in becoming a Nuclear Propulsion Officer Candidate. Please send me more information on the Nuclear Program as it applies to:

☐ nuclear-powered submarines.
☐ nuclear-powered surface ships.

If you really don't want to wait, call us toll-free, **800-841-8000.** In Georgia, call 800-342-5855. Call collect: in Alaska—272-9133, in Hawaii—533-1871.

Name_____ Age_____
Address_____ City_____
State_____ Zip_____ Phone_____
College_____ Major_____ Class of_____

Be someone special in the Nuclear Navy.

The real
Man in his ignorance has always assumed that the earth would rejuvenate herself, no matter what he did.

mother earth father sky navajo sandpainting artist alfred dihyja

The earth can lose her resiliency. Her waters are polluted, her soil is becoming dust, some of her trees are poisoned. Someday she may be beyond her ability to heal herself. But what we do to her we do to ourselves.

The Indians call the earth their mother
We must protect the earth. We must learn to live in harmony with her; we must learn her moods and rhythms. We must love the earth.

The ideal
When man and nature are in harmony, the land is happy. There is great strength in this idea.

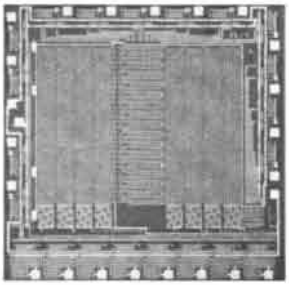
AtlanticRichfieldCompany

Established 1845 **SCIENTIFIC AMERICAN** August 1973 Volume 229 Number 2

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: $10 PER YEAR, U.S., POSSESSIONS AND CANADA: $12.50 PER YEAR, ALL OTHER COUNTRIES. EUROPEAN SUBSCRIPTION COPIES MAILED C/O INTERNATIONAL DISTRIBUTORS, KOEPOORTBRUG 1, B 2000 ANTWERP, BELGIUM.



THE COVER

The photograph on the cover shows an integrated circuit fabricated by the MOS (metal-oxide-semiconductor) technique (see "Metal-Oxide-Semiconductor Technology," by William C. Hittinger, page 48). The entire circuit is laid down on a silicon "chip" less than a quarter of an inch on a side. The MOS technique is characterized by the unipolar transistor, which differs from the bipolar transistor of other microelectronic technologies. It also achieves the highest density of components on a single chip. The circuit shown in the photograph is a 4,096-bit read-only memory manufactured by Fairchild Semiconductor. The low cost per component of MOS devices has had a heavy impact on such products as pocket calculators.

THE ILLUSTRATIONS

Cover photograph by Steve Allen

Page	Source
12–18	Allen Beechel
19	Boeing Aerospace Company
20	Pieter C. Wensink, Carnegie Institution of Washington
21–26	Bunji Tagawa
27	Arleen Forsheit and Norman Davidson, California Institute of Technology (*top*); Bunji Tagawa (*bottom*)
28	Pieter C. Wensink, Carnegie Institution of Washington (*top*); Bunji Tagawa (*bottom*)
29	Bunji Tagawa (*top*); Mary Lou Pardue, Massachusetts Institute of Technology (*bottom*)
31	National Accelerator Laboratory
32–35	Ilil Arbel
37	Argonne National Laboratory
38	National Accelerator Laboratory
49	RCA Corporation
50–53	Ilil Arbel
54	Fairchild Semiconductor
55	Texas Instruments Incorporated
56	RCA Corporation
57	Martin Marietta Corporation (*top*), RCA Corporation (*bottom*)
60	Earth Resources Technology Satellite
62–68	Eric Mose
71	David Bygott
72–78	Gabor Kiss
80–86	Dan Todd
88	Henryk M. Wisniewski, Albert Einstein College of Medicine
89	Robert D. Terry and Martin Weiss, Albert Einstein College of Medicine
90	George N. Wise, Albert Einstein College of Medicine (*top*); Neal Weinreb, Mount Sinai Hospital (*bottom left*); Robert S. Hillman, University of Washington Harborview Medical Center (*bottom right*)
91–97	Jerome Kuhl
98–100	Ilil Arbel
104	Gerard Schol
105–109	Roger Hayward

When you want to talk to the telephone company, here's who's on the other end of the line.

First of all, what you're not going to get is a shuffle from one person to another.

What you'll get every time you call your local telephone company's business office is a service representative.

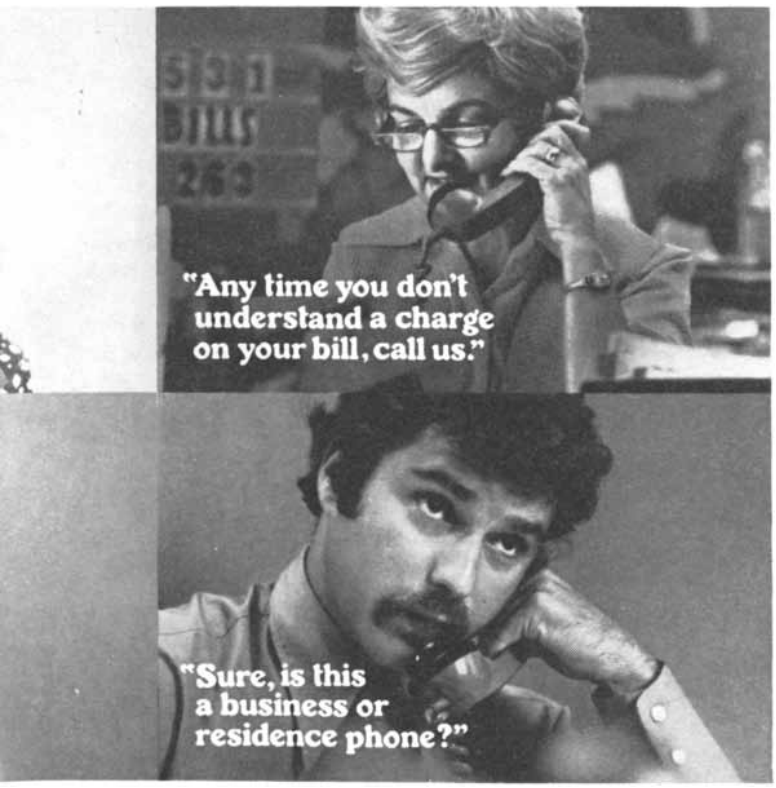
And every service representative has one job, and one job only. To help you get your problem solved. Quickly. Politely. And to your best possible satisfaction.

Naturally, you aren't the only person with something to talk over with us.

Each month, over 12 million of you call us. So AT&T and your local Bell Company have more than 37,000 service representatives to listen and to act.

(The time and money to train them runs into the millions. But when you consider the job they're doing, every dollar is well spent.)

Even with all these 37,000 people, we can't promise to solve every telephone problem immediately. But we can promise a service representative will try.



We hear you.

LETTERS

Sirs:

The article on the role of the heartbeat in infant-mother relations by Lee Salk [SCIENTIFIC AMERICAN, May] offers a partial explanation for the obscure interpersonal process postulated by Harry Stack Sullivan to explain the communication of anxiety between mother and infant. He writes in *The Interpersonal Theory of Psychiatry* (page 41):

"The tension of anxiety, when present in the mothering one, induces anxiety in the infant. The rationale of this induction—that is, how anxiety in the mother induces anxiety in the infant—is thoroughly obscure.... I bridge the gap simply by referring to it as a manifestation of an indefinite—that is, not yet defined—interpersonal process to which I apply the term *empathy*."

It seems reasonable as a result of Salk's work to regard the heartbeat as one of the important signals in this interpersonal process.

This is not the only form of subliminal communication associated with interpersonal processes for which there is good scientific evidence. McClintock (*Nature*, Volume 229, pages 244–245, 1971) verified the old folk tale that menstrual periods of females with close social ties tend to become synchronized and attributed this to some interpersonal physiological process, possibly pheromonal.

ROGER W. PEASE, JR.

Senior Editor, Science
G. & C. Merriam Company
Springfield, Mass.

Sirs:

Did the heartbeat serve to make us a largely right-handed race? Lee Salk has ably presented his fine research in showing the mother's left-arm preference for holding a baby, and the "why" of it as being due to a baby's need to hear the mother's heartbeat.

Let us go further. First, nature evolved a heart most strongly heard on the left. Second, as Salk shows, *both* left-handed and right-handed mothers have the high preference for left-arm baby-holding. Third, the primitive mother, having few if any conveniences, had many tasks to perform with the right hand while baby-carrying. Fourth, right-handed women were then more effective as workers and childbearers. Fifth, selection took over, making most of us right-handed.

Salk shows that in cases of premature birth the preference disappears. This could not influence selection, because in a primitive society premature babies could not survive.

Could we speculate further, in respect to the asymmetry of the human brain as told by Doreen Kimura [SCIENTIFIC AMERICAN, March]? The baby held at the left often has the right ear against the mother, with the left ear exposed to extraneous sounds. Could this have had to do with giving us right-ear superiority for speech and left-ear superiority for other sounds?

A. D. MOORE

Department of Electrical
and Computer Engineering
College of Engineering
University of Michigan
Ann Arbor

Sirs:

We recently had a science fair at Waunakee High School, where I attend school and am a freshman. When looking for a project, I became quite interested in Lee Salk's article titled "The Role of the Heartbeat in the Relations between Mother and Infant," so I based my project on it.

My partner and I selected a baby doll that was as close to a real infant as possible. We asked all women who came by our table to help us. We held the doll right in the center so as to indicate no preference when we asked the visitor to hold it.

A total of 45 women helped us with our survey. Eighty percent held the doll on their left side and 20 percent held it on their right. As Dr. Salk pointed out in his article, most people thought the way they held it was because they were right- or left-handed. I thought you might be interested in our results.

CONNIE GROSS

Waunakee, Wis.

Sirs:

Mr. Pease's comments concerning the communication of anxiety between mother and infant are in support of the findings that are coming out of the studies by Sperber and Weiland that I referred to in my article.

I am delighted with Professor Moore's

thoughtful comments, since they give me another view on the possible explanation of my findings. His explanation is perhaps the most parsimonious one I have encountered since it includes rationale for the findings I have obtained in the various studies conducted. I was not aware that we have right superiority for speech and left superiority for other sounds. This is indeed enlightening.

It was good of Miss Gross to write me about the results of her study. It is always encouraging for someone doing research to see the results of a colleague. Miss Gross seems to have validated some of my findings, and I am delighted that my study motivated her to do her own independent exploration. I hope that she continues in her scientific career, and that she will pass along to me any other evidence she may feel is helpful.

LEE SALK

Division of Pediatric Psychology
Department of Pediatrics
The New York Hospital–
Cornell Medical Center
New York

Scientific American, August, 1973; Vol. 229, No. 2. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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

Advertising correspondence should be addressed to Harry T. Morris, 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 25 cents.

Change of address (or other subscription correspondence) should be addressed to Circulation Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Please notify us four weeks in advance of change. If possible, kindly furnish an address imprint from a recent issue. Send both old and new addresses or complete and mail Post Office Form 3578.

☐ Please enter change of address.

☐ Please enter subscription.
U.S. and Canada:
☐ 1 year, $10 ☐ 2 years, $18 ☐ 3 years, $25
All other countries:
☐ 1 year, $12.50 ☐ 2 years, $23 ☐ 3 years, $32

NAME

NEW ADDRESS

OLD ADDRESS

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

AUGUST, 1923: "During the year 1922 there were 134 airplane accidents involving 69 fatalities, and 107 persons were injured through airplane accidents. An analysis of this record by the Aeronautical Chamber of Commerce of America, as given in its annual report to the Secretary of Commerce, strongly stresses the fact that, if casualties are to be reduced, the Government must provide air laws and exercise jurisdiction over all civil flying. Year after year our congressmen have been urged to pass laws for the regulation of flying and the licensing of pilots, and year by year they have been neglecting this duty they owe to the nation. It is not only that there is a deplorable and unnecessary loss of life but also that these ever recurring tragedies disturb the public mind and prevent people from investing in a system of transportation that they wrongly believe to be not yet safe and practicable."

"Transmission at 220,000 volts is now an accomplished fact. According to *Electrical World,* the switches on the 220,000-volt Big Creek lines of the Southern California Edison Company have been closed. The event marks the culmination of years of research, experiment and study of the problems of higher-voltage transmission by some of the most prominent electrical engineers in the profession. The Big Creek power lines had been operating for some time at 150,000 volts. Before being changed over to 220,000-volt operation additional insulation units were added, making a total of 11 units on suspension and 13 on dead ends. Each line has a capacity of 125,000 kilowatts at 220,000 volts."

"According to experiments that have been made in Mannheim, Germany, the Germans have succeeded in making gasoline from coal. A report from the American Consular Service indicates that a plant has been established for making about 60 tons of gasoline a day from coal. The conversion of the coal into gasoline is accomplished by hydrogenating the coal, whereby its chemical nature is changed. About 90 per cent of the coal is thus converted into a liquid that is similar to fuel oil in its chemistry and fuel value."

"As usual, the compilation of the automobile license figures for the old year shows that there has been a very large increase in the number of cars in use. This year it is for the first time reasonably possible to accommodate the entire population of the U.S. in our cars, if every one of the cars were available for the purpose; each car would have to carry only eight passengers to achieve that result. The richest state in automobiles is California, where there is one car for every four people; the leanest state is Alabama, where there is one car for every 12 people. The contrast between these two states is an indication that there is still much more room for cars."

SCIENTIFIC AMERICAN

AUGUST, 1873: "Metal products, iron ores and iron and steel particularly are exhibited by many nations at the great Vienna Exposition. The steel is usually the product of the Bessemer process, and the facility with which every grade can be obtained is illustrated not only by beautiful specimens of every degree of carbonization but also by a wonderful variety of manufactured articles. The fact that we are rapidly passing from the iron age to the age of steel is here most fully and convincingly exhibited. Bessemer steel is already becoming nearly as inexpensive as the better grades of iron, and it cannot be long before the rapid extension of Bessemer manufactures, and the still rapid succession of improvements in the details of the method and of the apparatus, shall so far reduce its cost as to permit its substitution for iron for nearly all uses."

"Quite recently M. Fizeau has published the particulars of a long series of experiments on the velocity of light, made between stations about six miles apart, using the rays from an oxyhydrogen light. He gives as the mean of 650 good observations a velocity of 186,326 miles per second. The result obtained by Roemer is usually given in text books, and in fact is commonly quoted as the correct velocity of light. But the close agreement of the more recent researches of MM. Foucault and Fizeau, and the elegant methods used by these philosophers in their researches, render it nearly certain that the velocity of light in the air is between 185,177 and 186,363 miles per second."

"The highest railway speeds in the world are attained in England. The highest railway speed in England is attained on the Great Western Railway, and this speed may be taken roundly as 50 miles per hour. Mr. Stirling has run with one of his great outside-cylinder express engines a train of 16 carriages at 70 miles per hour on the Great Northern, on a level or with a slightly falling gradient. On a first-class line there can be no question, therefore, but that a speed of 65 to 70 miles per hour may be available. We believe that it would be possible to lay permanent way so well, and to maintain it in such excellent order, that trains might travel on it with perfect safety at 100 miles per hour."

"A recent improvement, announced by Mr. Burgess, a photographic artist of Peckham, England, consists in sensitizing gelatin by means of bromide of silver. The mixture is applied warm to the glass plate, and the picture may be taken with the plate either wet or dry. The time of exposure is the same as for the ordinary wet collodion plates. The alkaline-pyro developer is used, the picture making its appearance rapidly with any required degree of intensity. The new process promises to compete sharply with the ordinary collodion process."

"Within the past few weeks a new section of the Pennsylvania oil region has been tapped by enterprising well drillers, and their labors have been rewarded by the opening of flowing fountains of the unctuous commodity. The result of these new petroleum supplies is the overstocking of the market and the decline in price to the insignificant sum of 75 cents per barrel. The principal use of petroleum at the present time is in the form of illuminating oil. Various attempts have been made to employ it as a substitute for bituminous coal in the manufacture of illuminating gas, and if this could be accomplished with economic advantage, the demand for crude petroleum would soon be equal to the supply, and steady, remunerative prices might always be expected. The discovery of new uses to which this abundant article can be put likewise presents itself as an excellent subject for research."

THE AUTHORS

RICHARD D. ENGLISH and DAN I. BOLEF ("Defense against Bomber Attack") are at Washington University; English is a graduate student in mathematics and Bolef is professor of physics. They met when English, as an undergraduate, enrolled in an interdisciplinary course on war and peace that Bolef was teaching. English writes that his special interest "is in mathematical analysis and how one applies mathematics to the real world," and that his leisure time "is divided among walking, model-building and science fiction, with some history on the side." Bolef was graduated from Pennsylvania State University in 1946 and obtained his master's degree and his doctorate at Columbia University. He has been active in organizations concerned with the interaction of science and society.

DONALD D. BROWN ("The Isolation of Genes") is a member of the staff of the Department of Embryology of the Carnegie Institution of Washington. "I went to Dartmouth College for three years, receiving no degree," he writes, "and then to the University of Chicago Department of Medicine. This wonderful school gave me the chance to grow up, and in doing so I discovered biological research." Eventually he decided "to amalgamate molecular biology with my minor interest in medicine; what came out was the molecular biology of higher organisms." Brown obtained his M.D. and spent a year as an intern, but since then he has been involved in research. He was with the U.S. Public Health Service for two years and the Pasteur Institute in Paris for one year before joining the Carnegie Institution in 1960.

BARRY C. BARISH ("Experiments with Neutrino Beams") is professor of physics at the California Institute of Technology. He writes that he obtained his undergraduate and graduate degrees at the University of California at Berkeley, "where I met my wife," who is now "a practicing psychiatric social worker." He adds: "Having become interested in high-energy physics while at Berkeley, I have continued the pursuit since coming to Cal Tech 10 years ago. This requires somewhat of a gypsy existence. In spite of all this traveling our first choice when we have free time is to take a trip. I am also involved in reading, spending time outdoors and enjoying the company of friends."

WILLIAM C. HITTINGER ("Metal-Oxide-Semiconductor Technology") is executive vice-president, RCA Consumer and Solid State Electronics. He was graduated from Lehigh University in 1944 with a degree in metallurgical engineering; this year the university gave him an honorary doctorate in engineering. After military service he spent 22 years with companies of the Bell telephone system, including several years with Bell Laboratories, which he joined in 1954 as a member of the technical staff. In 1958 he was named director of the semiconductor-device laboratory of Bell Laboratories, and in 1962 he became executive director of the semiconductor-device and electron-tube division. From 1966 to 1968 he was president of Bellcomm, Inc., a company jointly owned by the American Telephone & Telegraph Company and the Western Electric Company and engaged in systems engineering for the manned spaceflight program of the National Aeronautics and Space Administration. From 1968 to 1970 he was president of the General Instrument Corporation.

DAVID E. JAMES ("The Evolution of the Andes") is a member of the staff of the Department of Terrestrial Magnetism of the Carnegie Institution of Washington. He obtained all his degrees from Stanford University, beginning with his bachelor's degree in 1962 and ending with his Ph.D. (in geophysics) in 1966. He joined the Carnegie Institution as a postdoctoral fellow in 1966. "I typically spend about two months a year in South America," he writes. "Once one has adjusted to the wonderfully low-pressure Latin-American life-style, South America not only is fascinating scientifically and culturally but also provides relief from the depressing flow of 'world news' of the Northern Hemisphere. My wife, Odette, a geologist vastly superior to me but occupied exclusively with lunar research, enjoys our occasional vacations in South America but expresses little interest in Andean rocks—too altered and messy compared with the pristine lunar samples. Aside from my work my almost all-consuming activity is bicycle racing, a sport that is finally receiving some well-deserved attention in the U.S. Skiing and tennis get squeezed into off-season gaps in the bicycle training schedule."

W. H. THORPE ("Duet-singing Birds") is emeritus professor of animal ethology at the University of Cambridge. He was also director of the subdepartment of animal behavior of the university's zoology department from its inception until 1969. He has been a Fellow of the Royal Society since 1951 and is a fellow (formerly senior tutor) of Jesus College, Cambridge. Thorpe writes that he was "originally an entomologist, studying the behavior and respiratory physiology of parasitic and aquatic insects," but "switched gradually from insects to the behavior of birds." He is the author of five books, including *Bird Song: The Biology of Vocal Communication and Expression in Birds; Science, Man and Morals,* and *Animal Nature and Human Nature.*

GRAHAM WALKER ("The Stirling Engine") is professor of mechanical engineering at the University of Calgary in Canada. In addition to his teaching he has had military and industrial engineering experience. "Outside of engineering," he writes, "my principal interests are farming by summer and oil painting by winter. My farm is an increasingly decrepit establishment well removed from the city and fast returning to nature. The principal incumbent is an elderly tractor of limited agricultural value but an appreciating historical significance. I started painting in the early 1960's to fill the interregnum between manipulations during overnight runs on the slow computers of those days. I find it interesting to reflect on the reaction that C. P. Snow, in his concern with the 'two cultures,' would have to the prospect of a mechanical engineer doing paintings in a mathematical computing laboratory." Walker's Ph.D. is from the University of Durham.

ROSCOE O. BRADY ("Hereditary Fat-Metabolism Diseases") is chief of the developmental and metabolic neurology branch of the National Institute of Neurological Diseases and Stroke. After taking his bachelor's degree in chemistry at Pennsylvania State University he obtained his M.D. at the Harvard Medical School. He spent a year as an intern at the hospital of the University of Pennsylvania, four years in postdoctoral training at the University of Pennsylvania School of Medicine and two years at the National Naval Medical Center before joining the National Institutes of Health. "I was recently married," he writes, "and my major interests in addition to my lovely wife include teaching at the local medical schools, tennis, squash, piano, bridge and trying to keep things on an even keel in the stock market."

Your body is 8 hours slow

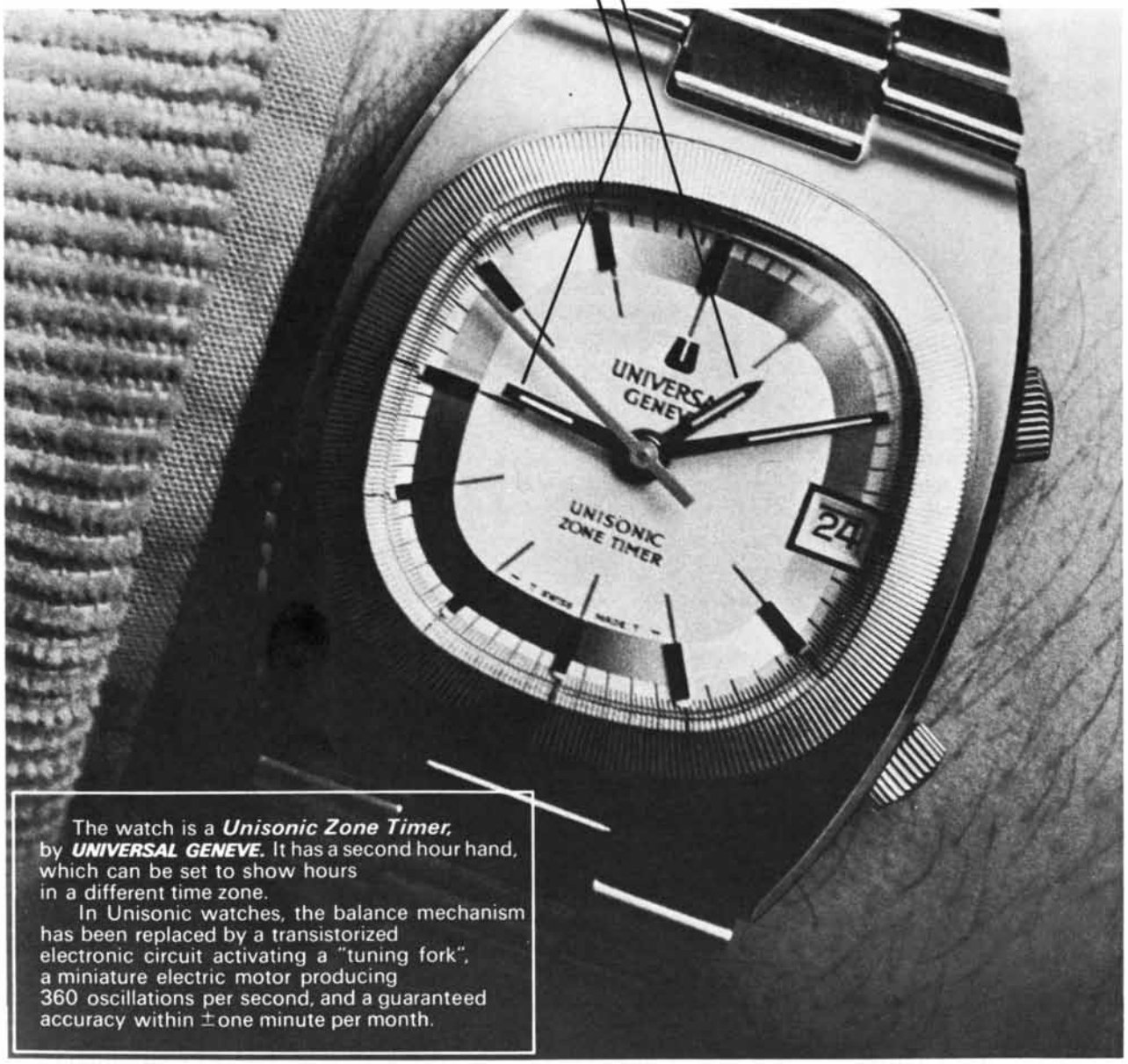
(09.10) This is the actual time
(01.10) This is the time your body brought along from home.

You have just arrived in Tokyo, flying from some place in Europe. Tokyo time is 8 hours ahead of European time, and your body cannot immediately adjust to such a change.
You feel it – you are not in your best shape, no matter how pleasant and comfortable your flight was.
There is one single remedy: rest. Take 24 hours before you start working – negotiating, talking business, whatever you came to Tokyo for.

Our Trans-Siberian Express gives you those 24 hours. For it leaves Europe on Saturday, at noon, and arrives in Tokyo Sunday morning, at 9.10.
You have 24 hours to get your "body clock" on time again before Monday morning.
And another thing: Our Trans-Siberian gateway is Copenhagen. And it is shorter to Tokyo via Copenhagen than over any other gateway in western or central Europe.

Going to the Far East, you have a choice:
FLY VIA COPENHAGEN –
OR FACE A DETOUR

SAS SCANDINAVIAN AIRLINES



The watch is a ***Unisonic Zone Timer,*** by ***UNIVERSAL GENEVE.*** It has a second hour hand, which can be set to show hours in a different time zone.
In Unisonic watches, the balance mechanism has been replaced by a transistorized electronic circuit activating a "tuning fork", a miniature electric motor producing 360 oscillations per second, and a guaranteed accuracy within ±one minute per month.

Established 1845 SCIENTIFIC AMERICAN August 1973 Volume 229 Number 2

Defense against Bomber Attack

Among the weapons not covered by the strategic-arms-limitation agreements are antiaircraft systems. The Department of Defense has now proposed that the nation's air defenses be modernized

by Richard D. English and Dan I. Bolef

A nuclear attack on the U.S. would probably consist of an initial volley of ballistic missiles, followed some hours later by bombers, which would assess the damage and attack any areas that had escaped destruction. At the strategic-arms-limitation talks in 1972 the U.S. and the U.S.S.R. virtually renounced defenses against the missile attack. Weapons that would detect and destroy enemy bombers, however, are not proscribed by the SALT agreement, and the Department of Defense has recently proposed construction of a new bomber-defense system.

The new system would "modernize" the Aerospace Defense Command, a branch of the Air Force charged with defending the nation against aerial attack. It would have four components: a new early-warning radar array, an improved interceptor aircraft, a new surface-to-air missile and the equipment to launch it, and an airplane capable of coordinating an attack on hostile bombers from aloft. The Department of Defense has requested $409.2 million for the program during fiscal year 1974, following appropriations totaling slightly more than $1 billion in previous years. These are relatively small sums within the defense budget, but they are primarily for development, not procurement or construction. The entire project could cost $10 billion or more over the next several years; operating expenses and additional armament proposed for the 1980's would add further to the cost.

Unlike the Safeguard anti-ballistic-missile system, a proposal discussed extensively in the press before it was repudiated by the SALT accords, the Air Force bomber-defense plan has received little public debate. Before the aircraft and other equipment are procured, and before forces are committed to their maintenance and operation, three questions should be answered affirmatively: Is there a reasonable threat of a bomber attack on the U.S.? Would the proposed system repel such an attack, and would it do so more reliably or more efficiently than weaponry now available? Finally, is it worthwhile to defend the nation against a bomber attack when such an attack is likely to occur only in the aftermath of a nuclear missile strike? We shall address these questions here.

The bomber threat to the U.S. consists of the 140 heavy bombers in the Russian long-range air force. One hundred of these are Tupolev Tu-95's, turboprop aircraft with a range of 7,800 miles, designated "Bear" in the North Atlantic Treaty Organization code-word system; the rest are Myasishchev Mya-4's, four-engine jets with a 6,050-mile range, designated "Bison" by NATO. Both of these aircraft could deliver nuclear weapons to points within the U.S. and return to bases in the U.S.S.R.

Some calculations of the bomber threat to the U.S. also include the Russian force of medium-range bombers, 500 Tu-16's ("Badger") and 200 supersonic Tu-22's ("Blinder"). Modified versions of these planes ("Badger-C," "Blinder-B") carry air-to-surface missiles instead of bombs, but the number of aircraft so converted is not known. In armament and electronics the medium-range bombers are in some cases more modern than the heavy bombers.

Throughout the 1950's and 1960's the U.S. used aircraft of similar range and performance (B-47's) as a strategic threat against the U.S.S.R. It has been argued that the Russian medium-bomber force should be considered in the same role. The American bombers, however, could fly from bases near the borders of the U.S.S.R. (in Europe, the Mediterranean area, the Far East and elsewhere), whereas the U.S.S.R. has no airfields on the perimeter of the U.S. In addition the U.S. Air Force had a fleet of nearly 1,000 aerial tankers available for in-flight refueling. The Russian long-range bomber force has 50 Mya-4 tankers. For reasons such as these we believe it is misleading to consider the Russian medium-range bombers capable of attacking the continental U.S.

It has been suggested, however, that the medium-range bombers could be used on one-way "suicide" missions against the U.S. Even assuming Russian willingness to sacrifice the planes and their crews, such missions seem unlikely. The Tu-22, under the best of conditions, has insufficient range for even a one-way trip. The Tu-16, if it were based at the tip of the Kola Peninsula in the north-

western U.S.S.R. or the Chukchi Peninsula in the northeastern U.S.S.R., and if it were flown at optimum speed and altitude, could perhaps reach some parts of the U.S. The bomber forces are not based on the Kola or Chukchi peninsulas, however, but farther inland, which increases the distance to their potential targets. Moreover, on combat missions evasive maneuvers necessary to avoid detection and interception would further reduce their range.

For reasons such as these the International Institute for Strategic Studies has concluded that "Soviet [medium-range] bombers can strike at United States allies anywhere on the Eurasian landmass and also at Canada, but not at the United States itself." Similar views have been expressed by recent secretaries of defense, from Robert S. McNamara through Elliot L. Richardson.

Still another threat to the security of the continental U.S. has been perceived, however, in a new Russian bomber, a "variable-geometry wing," or swing-wing, aircraft derived from the Tu-22. This plane, designated "Backfire" by NATO, has wings hinged near the midpoint; in the extended position they make possible lower takeoff and landing speeds, and in the swept-back position they allow higher cruise speeds. The Backfire, designed to approach its target at low altitude in order to avoid radar detection, will have a top speed of more than twice the speed of sound and a range (according to U.S. estimates) of 4,300 miles. It is expected to join the Russian long-range air force next year.

The development of the Backfire has prompted statements of concern from those who advocate strengthened U.S. air defenses. In 1972 Secretary of Defense Melvin R. Laird reported: "There is a probability that it [Backfire] has a capability for in-flight refueling. With refueling it could reach virtually all U.S. targets." Admiral Thomas H. Moorer, chairman of the Joint Chiefs of Staff, also speculating on the possibility of airborne refueling, said in his fiscal-year 1974 Military Posture statement: "The major uncertainty regarding the U.S.S.R. bomber force is still the primary mission of the new Backfire variable-geometry-wing, supersonic bomber.... An intercontinental role for the Backfire cannot be excluded." The threat allegedly presented by the new bomber was in fact cited in Congressional testimony by Major General George J. Keegan, Jr., assistant chief of staff for Air Force intelligence, as a factor justifying the proposed U.S. bomber-defense system.

Independent assessment of the Backfire's potential is difficult because most of the pertinent documents are classified. In the record of Congressional hearings dealing with the airplane all information on its performance was deleted. It is apparent, however, from the published estimate of its range and from the limitations of the small Russian tanker force that the Backfire, like the Tu-16, will not be capable of a conventional, round-trip bombing run, and that a suicide mission, if one were contemplated, would be at least marginal. The International Institute for Strategic Studies, in its publication *The Military Balance, 1972–73,* noted: "Development of new Soviet bombers appears to be limited to prototype testing of the Backfire variable-geometry aircraft, whose limited range seems to restrict its utility as a potential successor to the 140 aging, long-range Soviet bombers in current service."

Whether one assumes an attacking force of 140 heavy bombers or includes some medium-range bombers flying one-way missions, the key question is: Could the proposed defensive system repel the attack? To a large extent that depends on whether it could survive the missile strike that would almost certainly precede the bombers.

The present system could not survive. Continental air defense is now the responsibility of the North American Air Defense Command (NORAD), organized in 1957 and embracing both U.S. and Canadian forces. It has four components: the radars, which detect and track enemy bombers; the command and control centers, known as SAGE (Semi-Automatic Ground Environment) and BUIC (Back-Up Interceptor Control), which direct battle forces against invading aircraft, and two types of weapon, the interceptor and the surface-to-air missile (SAM). The interceptor is a piloted aircraft equipped with air-to-air missiles, sometimes armed with a nuclear warhead. SAM's are usually guided from the ground, although newer types have an on-board guidance system that takes over after the missile has reached the vicinity of the target.

The weakness of NORAD lies in the radars and command centers, which are immobile ground-based installations and thus are ideal targets for ballistic missiles. Because they are "soft" (not encased in concrete or otherwise protected from a near nuclear explosion) even a relatively inaccurate missile could destroy them. Also contributing to the system's vulnerability is the fact that the number of sites is small: there are 14 SAGE units and 56 long-range radar stations. (The BUIC centers are being deactivated.) A fairly small force of missiles (say 210, at three per target) would be more than enough to destroy this network. Since the U.S.S.R. is allowed a total of 2,358 intercontinental ballistic missiles and sea-launched ballistic missiles under the SALT agreement, it could certainly afford to use such a force.

The loss of the radars and the control centers would mean the collapse of NORAD. Without long-range radar it is impossible to detect and to plot the course of attacking bombers. Lacking this information and the command centers to utilize it, it is impossible to guide interceptors to their targets or to alert SAM batteries of the bombers' approach. (The interceptors and the SAM's are themselves also vulnerable to missile attack, a problem that will be taken up below.)

In addition to their vulnerability the radars of NORAD have another weak-

FIBER-GLASS RADOME of the airborne warning and control system aircraft (AWACS) is mounted on struts above the

ness: they cannot track aircraft flying close to the ground in the maneuver called "terrain avoidance." Present Russian bombers can use this technique, and the Backfire is designed expressly for it.

The proposed replacement for the ground-based radar and control stations is the airborne warning and control system aircraft (AWACS). AWACS radar is installed in a modified version of the Boeing 707 commercial jetliner; a rotating radome, or radar pod, is mounted on two struts above the rear section of the fuselage.

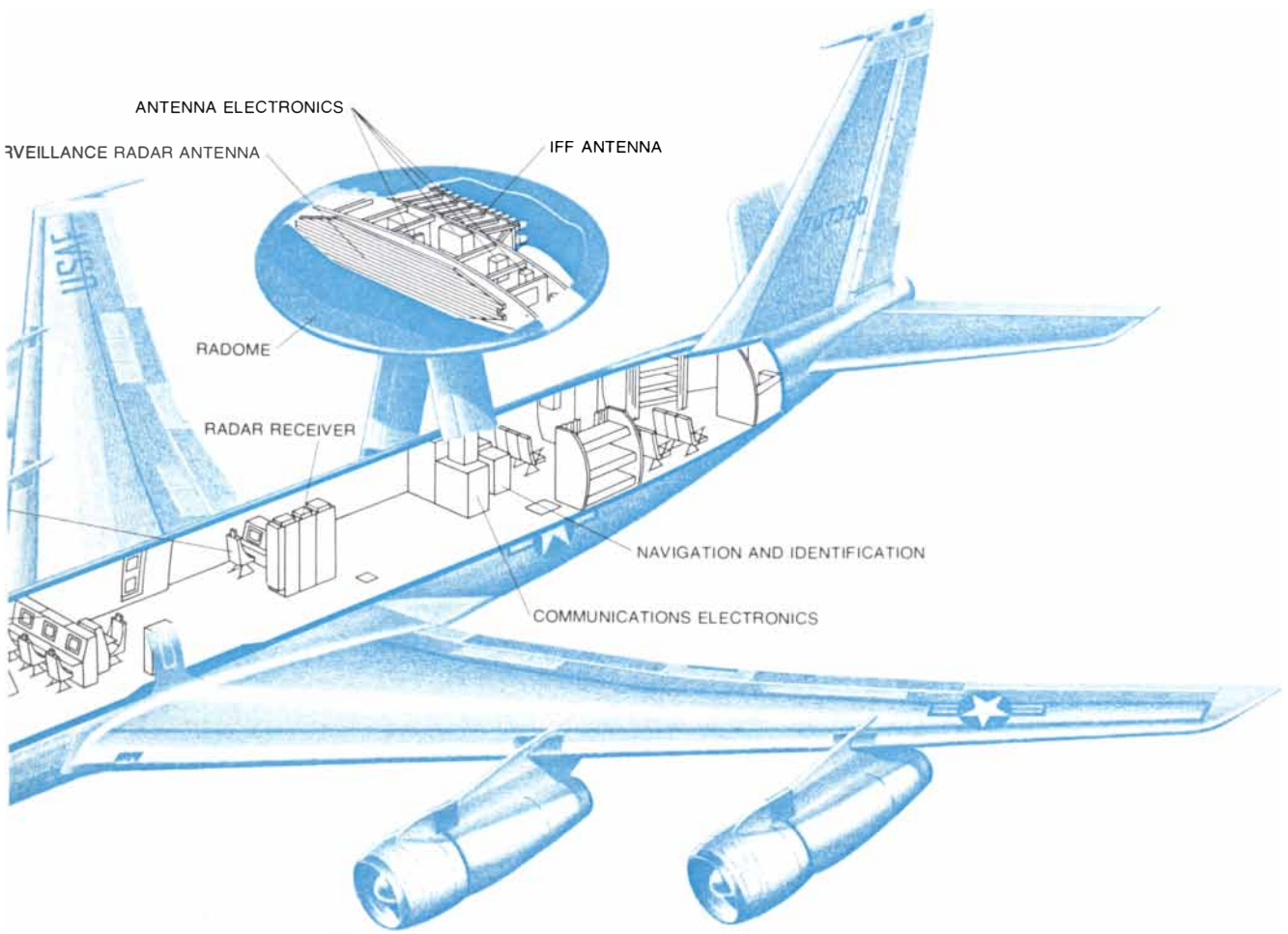
The range of ground-based radar is limited by the curvature of the earth to about 200 miles. According to *Air Force Magazine*, the AWACS aircraft, flying at 30,000 feet, can survey the sky from ground level to about 60,000 feet and has a range double that of ground-based radars. In addition to extending the range of aerial surveillance it provides more reliable tracking of bombers flying in terrain avoidance. It discerns low-flying aircraft against "ground clutter," the radar images of terrestrial obstructions, by using a rapid stream of pulses and by detecting the Doppler shift of signals reflected from a moving target. The Doppler shift caused by the motion of the AWACS aircraft itself is compensated for electronically.

Also in the radome is the antenna of an IFF (for "Identification: Friend or Foe") transmitter and receiver. This system "interrogates" other aircraft within the range of the AWACS radar; transponders in other military craft transmit a coded reply that identifies each airplane. In a combat situation aircraft that did not respond would be considered hostile.

The output of these two systems is processed by computer for visual display. Aircraft being tracked can be plotted on a fixed map grid rather than on the circular sweep-pattern display used in many other radars, which always have the transmitter-receiver at the center. The identity, altitude and velocity, as well as the position, of all aircraft within the range of the radar can be indicated. AWACS will carry a flight crew of four and an additional crew of 13 to operate the surveillance radar and command equipment.

The AWACS aircraft could survive a missile attack simply by staying aloft. With an endurance of at least seven hours, it could probably remain airborne long enough to meet the incoming Russian bombers. If necessary it could be refueled in flight.

Two prototype AWACS aircraft were built by the Aerospace Division of the Boeing Company in 1970. On the basis of flight tests the Westinghouse Electric Corporation was awarded the contract for the surveillance radar, and the Air Force has requested $209.5 million this year for additional development. Pro-



fuselage of a modified Boeing 707. Thirty feet in diameter and six feet thick at the center, the radome contains antennas for surveillance radar and for an IFF (Identification: Friend or Foe) transceiver, which discriminates between friendly and hostile aircraft. Outputs of these systems are displayed on video consoles. While airborne, AWACS would coordinate bomber defenses.

duction, if the project receives continued Congressional approval, will begin in 1975 with an initial order of 42 airplanes.

The utility of AWACS would be enhanced, according to the Department of Defense, by the deployment of a new long-range ground-based radar, the over-the-horizon backscatter (OTH-B) system. Over-the-horizon forward-scatter radar already forms a part of the U.S. early-warning network. Signals generated by transmitters in Japan, Taiwan and the Philippines are reflected by the ionosphere over the U.S.S.R. and are detected by receivers in western Europe. Activity in the intervening atmosphere, such as the launch of a volley of missiles or a fleet of bombers, produces an "OTH signature" on the signal that, when interpreted, gives a general picture of events in Russian airspace.

In the backscatter system signals returning to the vicinity of the transmitter would be affected in the same way by events over the U.S.S.R. and would be interpreted to provide the same information. Two units are planned, one looking east and the other west, according to a statement made in March by Secretary Richardson. OTH-B could give at least 30 minutes' warning of an impending attack, which would allow AWACS craft on ground alert to be launched and which thus would improve the chance that a radar and command station for antibomber defenses would survive. For further development of OTH-B $5.5 million has been requested this year.

Although the radar and command elements of the proposed air defenses seem adequate to their tasks, the weapons they would direct seem to offer no improvement over those now available in their ability to survive an interconti-

RUSSIAN LONG-RANGE BOMBERS could reach any targets in the U.S. and return to bases in the U.S.S.R. if they were launched from the Kola or Chukchi peninsulas. The polar routes illustrated are the shortest. Solid lines show maximum ranges for bombers flying round-trip missions without refueling; broken lines show the radii for round-trip flights in which the planes are refueled once.

nental-ballistic-missile attack. Among the aircraft being considered as interceptors are the F-111-X-7, the NR-349, the F-14 and the F-15; the last two have received the most attention. General John D. Ryan, the Air Force chief of staff, has said that he believes only the F-15, an Air Force plane with 50-mile radar and additional fuel tanks, should be considered for the air-defense mission. Others, in the Aerospace Defense Command, appear to favor the F-14, a Navy plane that has 100-mile radar and mounts the long-range Phoenix air-to-air missile. A recent study by the Director-ate of Defense Programs Analysis and Evaluation also concludes that the F-14 should be considered.

The cost of interceptors for bomber defense has not yet been estimated by the Air Force; the cost will depend on which aircraft is chosen and how many are procured. Published estimates of the strength of the proposed interceptor force have all set the number of aircraft at 200. For the F-15, 200 airplanes would cost about $2 billion, for the F-14 about $3 billion, based on prices now paid for other versions of the same craft.

Regardless of which airplane is chosen, its ability to survive a missile attack is doubtful. Interceptors are kept on the ground, and like ground-based radars they are probable targets for Russian ICBM's. Although several countermeasures are available to protect them, none seems entirely workable.

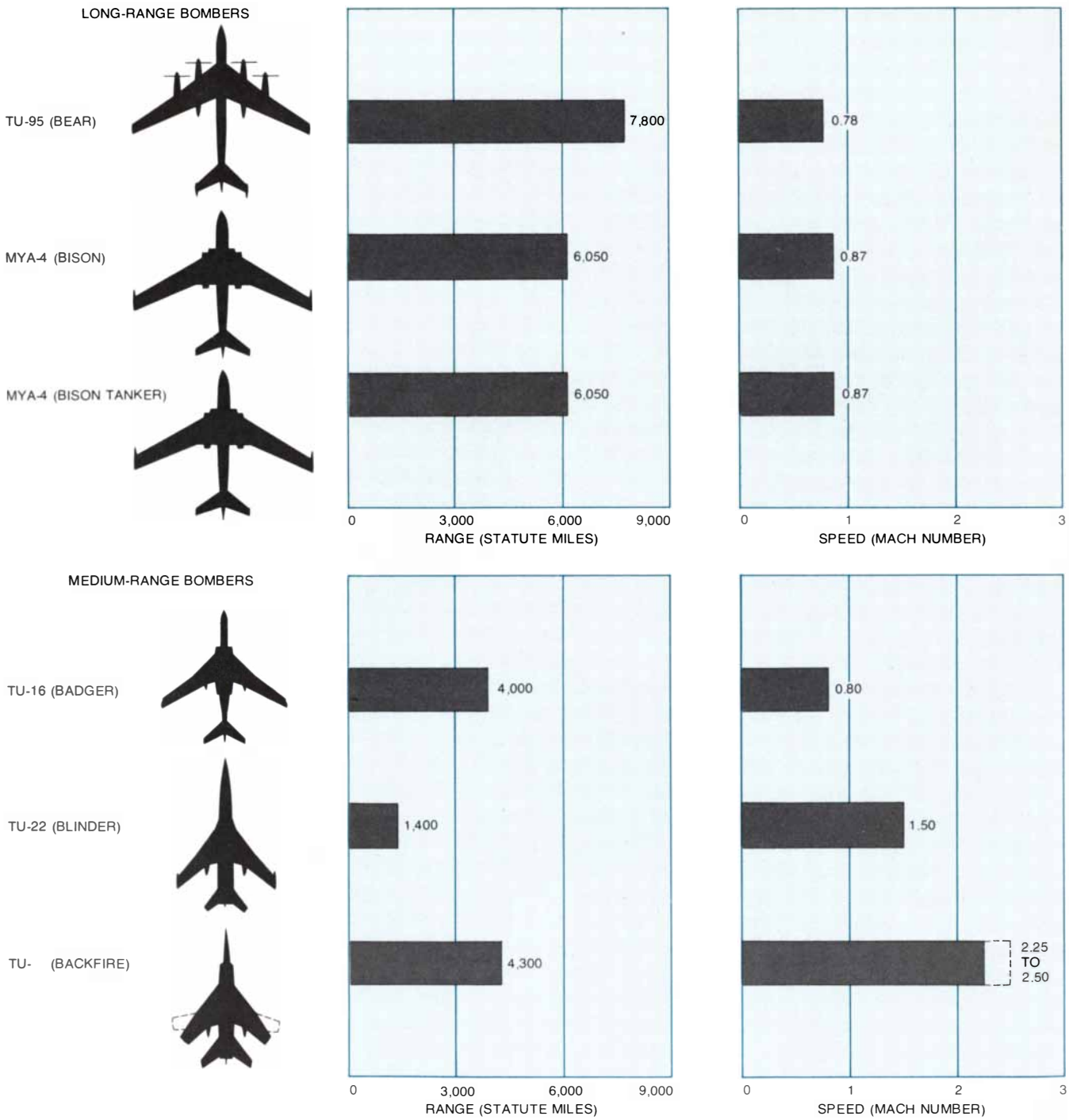
First, the interceptors could take off when the attack began and could wait out the missile barrage in the air. Neither the F-14 nor the F-15, however, could stay aloft long enough (seven to 12 hours) to meet the arriving enemy bombers. Aerial tankers could be used to

MEDIUM-RANGE BOMBERS could not attack the U.S. and return home (*solid lines*), although some might be capable of one-way missions (*broken lines*). The ranges indicated, however, are for flights without payload at optimum speed and altitude, and in combat would be reduced. Refueled ranges are not shown for medium bombers because U.S.S.R. has too few tankers to service them.

refuel them, but the present tanker force is entirely committed to the bomber fleet of the Strategic Air Command. New tankers would cost $40 million each, a price far beyond the scope of the projected bomber-defense budget.

The interceptors could also fly to alternate airfields on receiving an attack warning. Any fields with the facilities to service them, however, would also be likely targets. Indeed, the only airfields military planners may prudently assume would survive are those that have been abandoned, a situation that has led to the proposal of a "bare base" strategy. In this plan the interceptors would fly to abandoned fields, accompanied by transport aircraft carrying the men and equipment needed to set up ground control stations. Presumably the Russians would not know until after they had launched their attack which fields would be used.

Since the primary bomber threat is from the north, however, the bases would have to be in the northern U.S., which limits the potential dispersion of targets. In addition the bare-base strategy would require transports, and the Aerospace Defense Command has none. Borrowing them from other commands would be possible but risky. In a surprise attack they may be committed elsewhere and not available quickly enough.



LARGEST PORTION of Russian force is in medium-range bombers, which the authors maintain should not be considered a strategic threat to the U.S. The ranges shown were calculated for flights without payload, and under conditions that would promote the

The final element of the proposed air-defense system is the Army's new surface-to-air missile, called SAM-D for "SAM-Development." Four missiles, each in a protective canister that would also serve as a launching tube, would be carried on a large wheeled trailer. A similar trailer would carry phased-array radar, and a third communications and command facilities; combinations of these vehicles would make up a "fire unit." The Army estimates the cost of its planned missile force at $3.9 billion, in addition to research and development costs of $1.2 billion and operating expenses of about $1 billion a year.

The mobility of SAM-D has been mentioned as protection against destruction by ICBM's. With the missiles mounted on wheeled platforms and independent of fixed communications lines, they could be moved; they would not, however, be mobile enough to escape jeopardy.

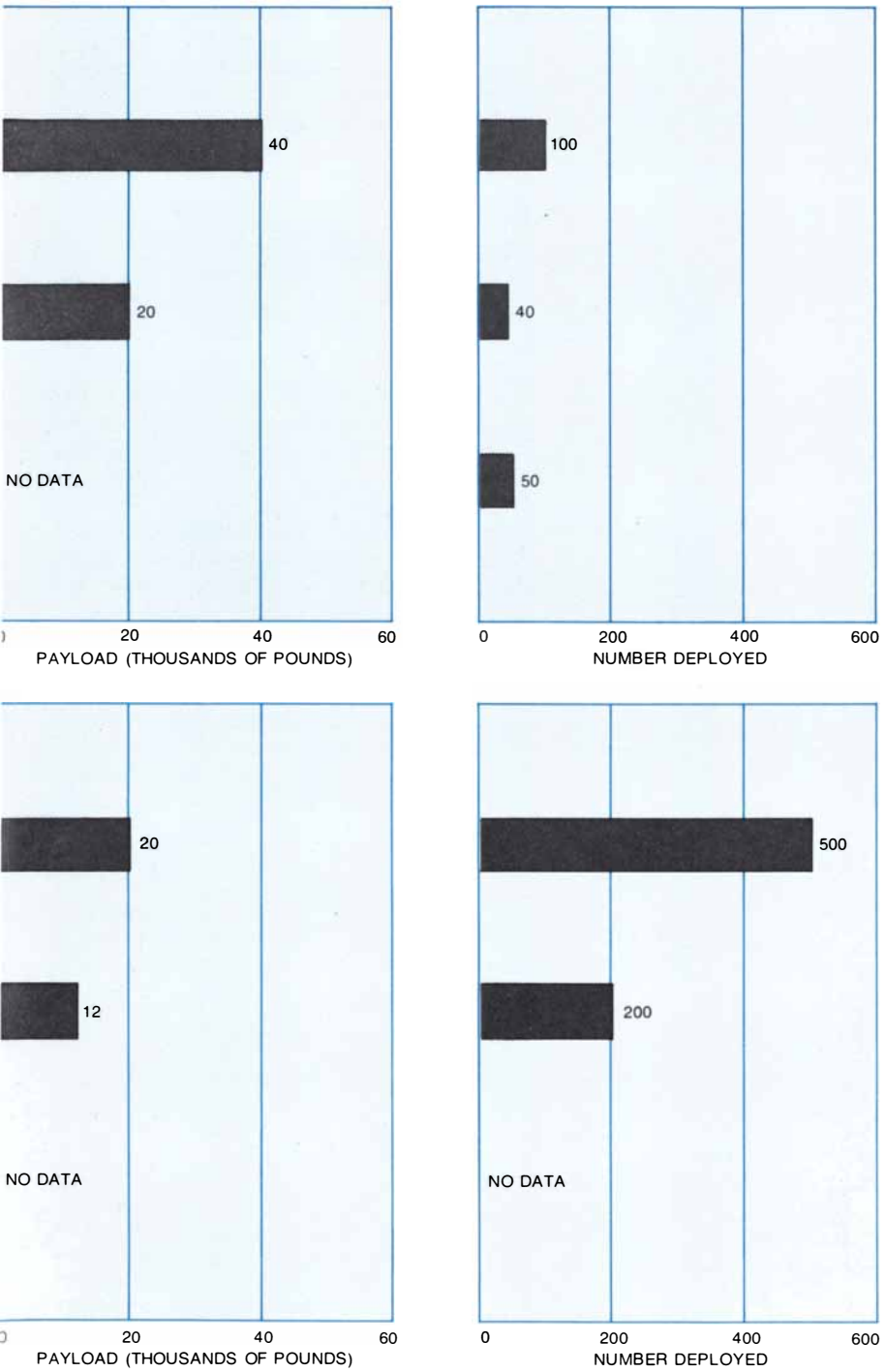
SAM-D would have a range of about 100 miles, which would restrict it to areas near the cities and military sites being defended. Moving such heavy and delicate vehicles on highways would be difficult, particularly in metropolitan areas, where air defense is presumably most needed. Moving them frequently and clandestinely, so that enemy missiles could not be again aimed on them, would seem almost impossible.

Thus SAM-D seems as vulnerable to missile attack as the proposed interceptor aircraft, and as unlikely to survive to repel a bomber attack. It is for this same reason that the present bomber defenses are considered obsolete. In a probable scenario for a nuclear war, over-the-horizon radar would give adequate warning, and AWACS aircraft would survive to coordinate the nation's defenses, but neither would be able to stop enemy bombers because the weapons they were intended to direct would have been destroyed.

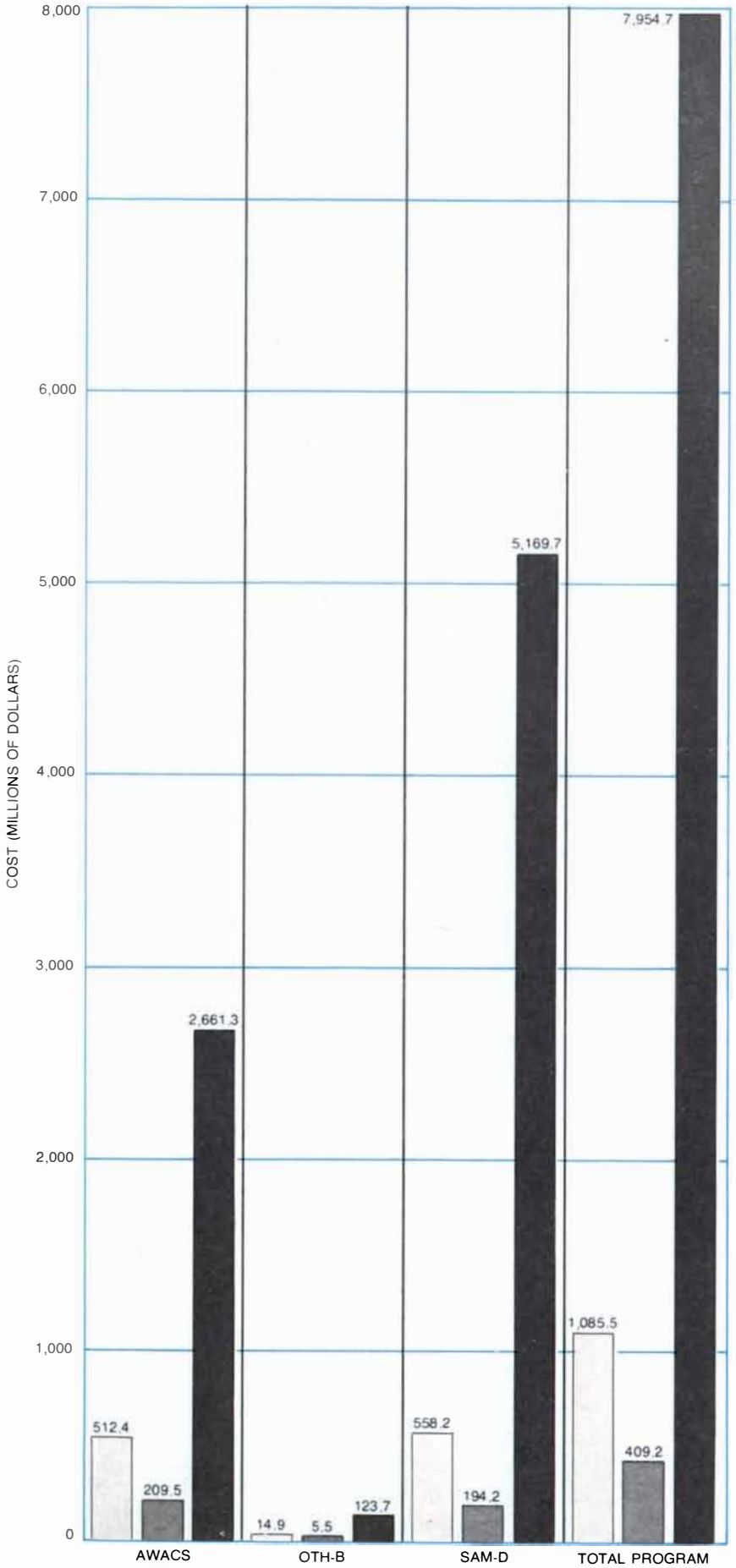
Both the AWACS aircraft and SAM-D have been proposed in Senate hearings as multipurpose weapons, with value both to the defense of the continental U.S. and to the Army and Air Force in field operations. The primary mission of SAM-D appears to be the battlefield defense of Army units, but the cost of its development is included in the continental air-defense section of the defense budget. AWACS is sought by both the Aerospace Defense Command and the Tactical Air Command. The latter would use it as a flying battlefield command post; instead of erecting ground-based radar units in a disputed territory, the Air Force could station an AWACS plane hundreds of miles from the battle, out of range of enemy weapons, and still control all air operations.

Although the 42 AWACS aircraft initially ordered would all apparently be assigned to the Aerospace Defense Command, the Air Force has said that, even if there were no bomber threat, it would still give high priority to the procurement of AWACS as a tactical weapon. Areas mentioned for its possible tactical deployment range from Southeast Asia to Western Europe, where the U.S. could act in support of a friendly government. It could be of greater utility in countries where the U.S. influence is not established, a possibility cited by Lt. Gen. George S. Boylan, Jr., deputy chief of staff of the Air Force. As an example, General Boylan described a hypothetical U.S. intervention in North Africa.

The political as well as the military



most efficient use of fuel. Several versions of the Tu-16 and the Tu-22 are known, and figures may vary somewhat in other-than-standard configurations. The Backfire is not yet deployed.



significance of such weapons ought to be evaluated. AWACS, together with remotely piloted vehicles and other automatic weapons developed in the course of the war in Southeast Asia, would make it possible to commit the U.S. to a Vietnam-like war more quickly, more quietly and more deeply than in the past. Ground-based radars would not be required. Because fewer troops would be engaged, both costs and casualties would be lower. Indeed, the ultimate goal is apparently to commit only air and naval power, and no ground troops at all.

Thus as a component of the Aerospace Defense Command, AWACS seems relatively harmless, if also relatively useless. In its tactical role, however, it could facilitate the prosecution of limited, nonnuclear wars. We believe the political questions raised by this capability should be answered before procurement of the aircraft is approved.

If one were to assume, contrary to our analysis, that Russian bombers do present a major threat to the U.S., and further that the proposed Aerospace Defense Command program could thwart an attack by those forces, would bomber defenses then be justified? We think not.

The main strategic-offensive forces of both the U.S. and the U.S.S.R. are not bombers but intercontinental ballistic missiles and sea-launched ballistic missiles. Because surprise would be important in any contemplated nuclear war, bombers are likely to remain weapons of secondary importance. They require several hours to reach their targets, whereas missiles need only about 30 minutes, and they may call for additional preliminary preparation that might be detected as much as a month before an attack.

Under the strategic-arms-limitation agreement the U.S.S.R. is allowed 2,358 missiles, and the U.S. is forbidden, except in two limited areas, to defend itself against them. Since this is a treaty, and intended to be permanent (as op-

FISCAL YEAR 1973 AND BEFORE
FISCAL YEAR 1974
ESTIMATED TOTAL

ESTIMATED COST of proposed bomber-defense system is based on information provided by the Department of Defense. Figures for fiscal year 1974 are defense budget requests; estimated totals include these amounts and the previous expenditures. Not included, however, is the cost of the manned interceptor aircraft; this would add $2 billion to $3 billion to the total, depending on what particular airplane is finally chosen.

posed to the five-year interim agreement on offensive weapons), planning for defense in the late 1970's or the 1980's should be expected to take its provisions into account. One of its consequences, we believe, is that in the absence of defenses against missiles any attempt to defend the nation against bombers would be futile.

While he was Secretary of Defense, Robert McNamara stated that the destruction of a fourth of a nation's population and half of its industry would be "unacceptable damage." That is, the nation could not survive this level of destruction as a functioning society; there would be no way to reestablish the network of production, transportation and communication necessary to the lives of the survivors.

The U.S.S.R. is capable of inflicting that much damage on the U.S. According to Department of Defense estimates, by 1975 Russian sea-launched missiles alone will be sufficient to attack all Strategic Air Command bomber bases and to "pin down" the Minuteman ICBM force. That would leave at least 1,000 ICBM's available to attack civilian areas. Assuming only 40 percent reliability, the lowest estimate we have seen, 400 warheads could be exploded over U.S. cities. Eight could be allotted to each of the 50 largest metropolitan areas. (The 50th-largest metropolitan area in the U.S. is Oklahoma City.) Such an attack would be more than enough to devastate those cities. Forty-two percent of the nation's population and 55 percent of its industry would be destroyed. Both figures exceed McNamara's criteria for "unacceptable damage," and doubtless the U.S.S.R. has more sophisticated means of choosing targets and calculating damage.

It has been argued that bomber defenses, however ineffective, would save American lives. Even if the Aerospace Defense Command failed entirely, some advocates of its programs have said, it would divert Russian missiles from civilian populations and force them to be aimed on military installations. These arguments are humane but specious. In military planning one must assume that one's enemy will use his forces in whatever way will cause the most destruction. That is, the U.S.S.R. would divert missiles from population centers only if that action would, by reducing the losses of its bombers, increase the damage overall; otherwise the missiles would be used directly on the cities. In any case the saving of lives, an issue of great magnitude in ordinary circumstances, becomes less important in a nuclear war. Once a threshold has been crossed, when a small fraction of an attacking force can demolish the opposing society, small increments in the number of deaths can have no effect on the outcome of the conflict.

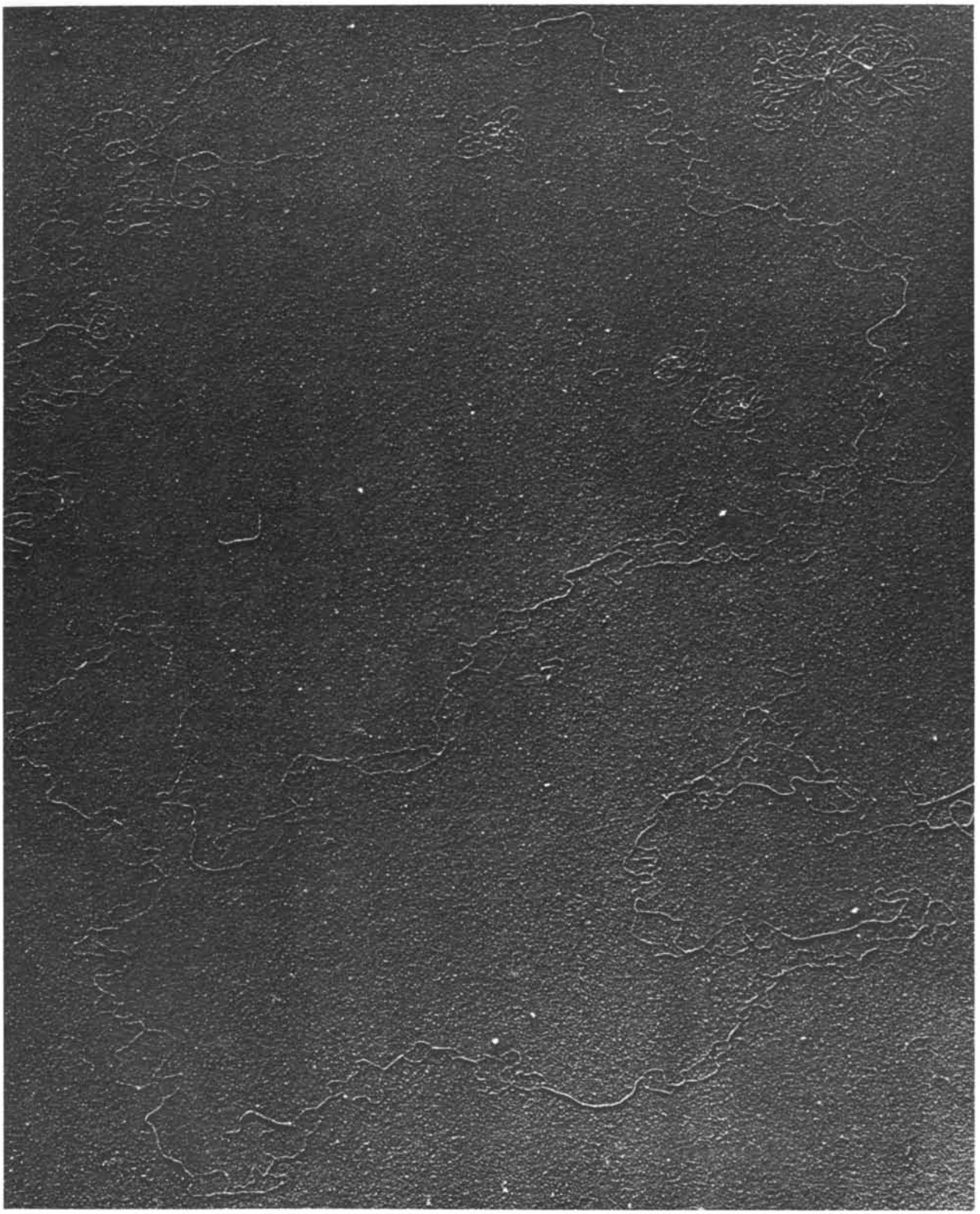
STYLIZED MAP of the eastern U.S. appears on a display console aboard an AWACS aircraft. Radar signals and identifying data are processed by computer, so that targets can be plotted against the fixed map rather than in the moving reference frame of the airplane.

Russian bombers, however small or large their number, could do grave damage to the U.S., but this potential for harm is not enough to justify the creation of expensive and complex weaponry to oppose them. It must be shown, rather, that bombers are a sufficiently large portion of the Russian force and account for so much potential destruction that thwarting their attack could ensure the survival of the nation. Given the preponderance of missiles in the Russian arsenal, we do not believe the bomber-defense system, as now planned, could save the U.S. from destruction. In any probable nuclear war, we feel, it would defend a society that no longer existed.

Even though the deployment of bomber defenses would probably not protect the nation, it would not be without consequences. First, it might foster the illusion that the U.S. is safe from nuclear attack, whereas even a perfect air-defense system would not meaningfully change the outcome of a nuclear war. Second, it might undermine the SALT agreements.

The U.S. and the U.S.S.R., in signing the strategic-arms-limitation agreements, recognized that the weapons programs forbidden or curtailed by SALT were too expensive for both countries and would improve the security of neither. The agreements are balanced on the razor's edge of what each nation considers an appropriate compromise between the needs of its security force and the limitations of its budget.

Should the U.S.S.R. conclude that a new U.S. bomber-defense system would threaten its security, the least it could do would be to respond in kind. The worst it might do, should the costs of bomber defenses seem too great, would be to abrogate the SALT agreements.



EIGHT GENES for ribosomal RNA are included in this molecule of DNA from the nucleus of an oöcyte of the African clawed toad *Xenopus mulleri*. DNA containing the genes, isolated by methods described in the article, was treated with alkali and thus partially denatured: the two strands of the DNA double helix were separated. Denaturing occurs first in parts of the molecule that have a relatively low "$G + C$ content," that is, a low proportion of two of the four nucleotides that combine to form DNA. The molecules are fixed, complexed with protein to make them thicker, placed on an electron-microscope grid and shadowed with platinum. The micrograph, in which the molecule is enlarged 25,000 diameters, was made by Pieter Wensink in the author's laboratory. It shows two kinds of regions: those in which the DNA is primarily denatured (*open loops*) and those in which its two strands remain unseparated (*thicker single lines*). Most of each denatured region is gene DNA; the remainder of those regions, together with the unseparated segments, is "spacer" DNA. Spacers alternate with genes and have a generally higher $G + C$ content (*see illustration on opposite page*).

The Isolation of Genes

Given a specific RNA product of a gene, it is now possible to find and purify the stretch of DNA that encoded the RNA. The first genes to have been isolated are those that make the RNA's in ribosomes

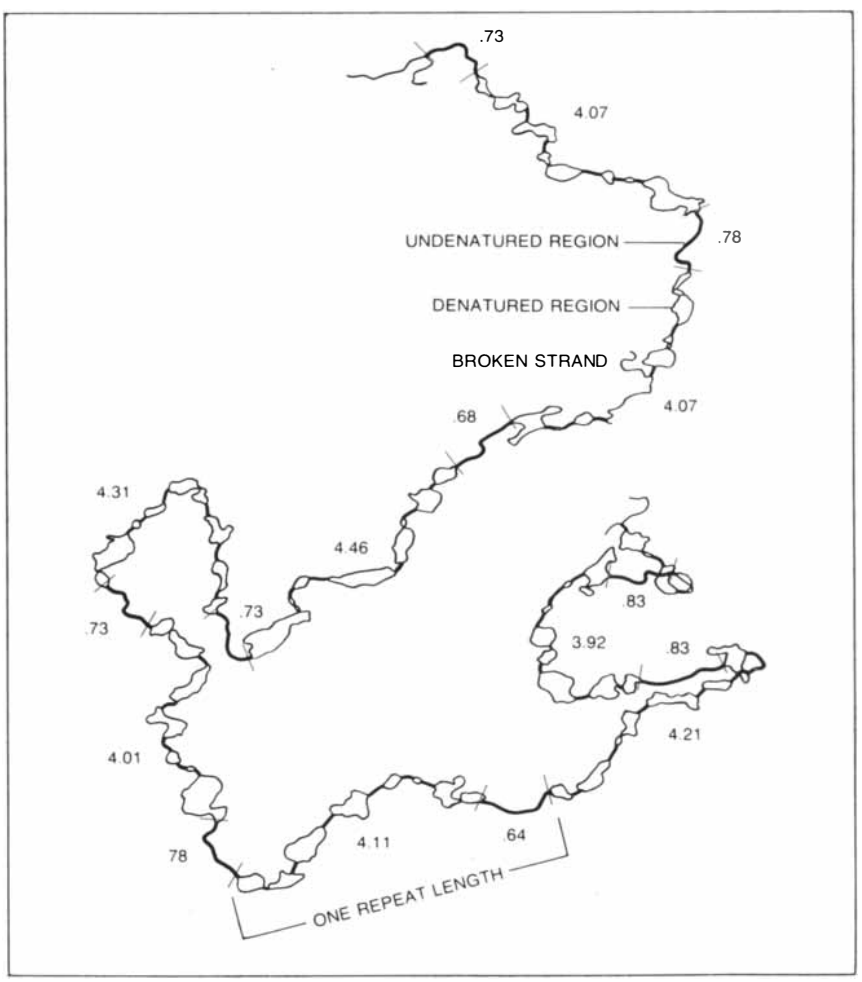
by Donald D. Brown

One of the scientific triumphs of this century has been the discovery that genetic information is encoded along the length of long molecules of deoxyribonucleic acid (DNA). The information is present in discrete segments of the DNA, the genes, each of which determines the structure of a complementary length of ribonucleic acid (RNA). Some of the RNA (messenger RNA) determines the structure of the proteins (primarily enzymes) that constitute or manufacture all the tissues of the organism; the remainder of the RNA (ribosomal RNA and transfer RNA) takes part in the synthesis of protein.

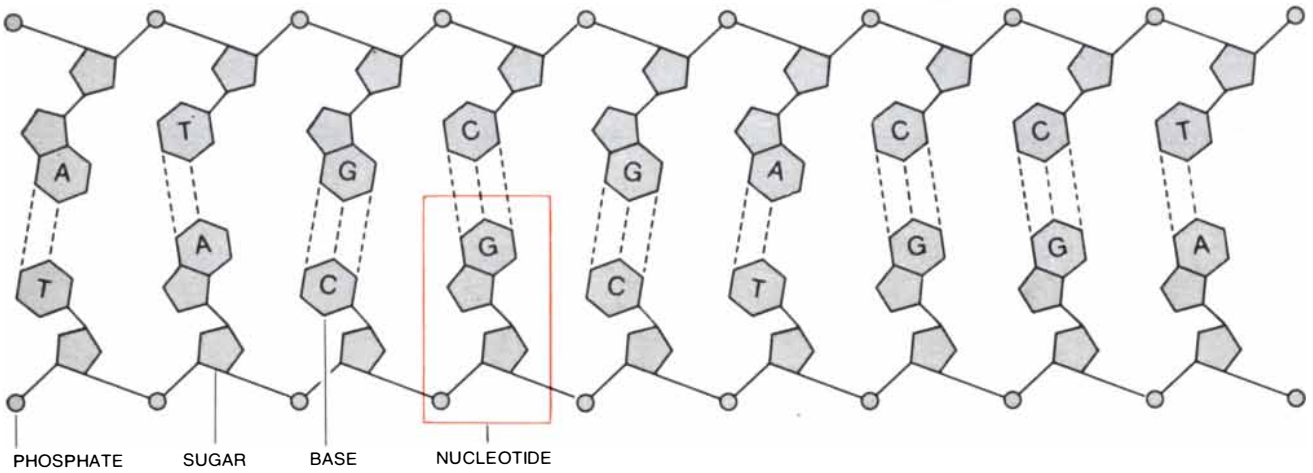
Although a great deal is now known about the nature of DNA and about the functioning of genes, most of this knowledge is based on genetic experiments with microorganisms or on biochemical studies with complex mixtures of DNA known to contain many genes. In the past 10 years a number of investigators have begun to develop methods that make it possible to purify individual genes of known function from the total DNA in an organism.

Techniques for gene purification are now at a stage comparable to those available 25 years ago for fractionating proteins. One difficulty is that, unlike proteins, which are composed of 20 different amino acids, the DNA molecules are long polymers composed of only the four building blocks called nucleotides. The result is that all DNA molecules have very similar chemical and physical properties, even though different arrangements of the four nucleotides can and do provide an almost infinite variety of different polymers. In spite of the chemical simplicity of DNA, techniques have been developed recently for fractionating complex DNA mixtures from animal cells, and so far two DNA components have been purified that contain genes of known function. The genes, which code for the RNA's of the ribosome, have been purified from the DNA of two related amphibians, the African clawed toads *Xenopus laevis* and *Xenopus mulleri*.

Gene purification became possible about 10 years ago with the development of an assay system that could detect the presence of a gene in a complex mixture of related DNA molecules. This assay system, called molecular hybridization, depends on the complementary structure of DNA, which is a long double-strand molecule of which each strand



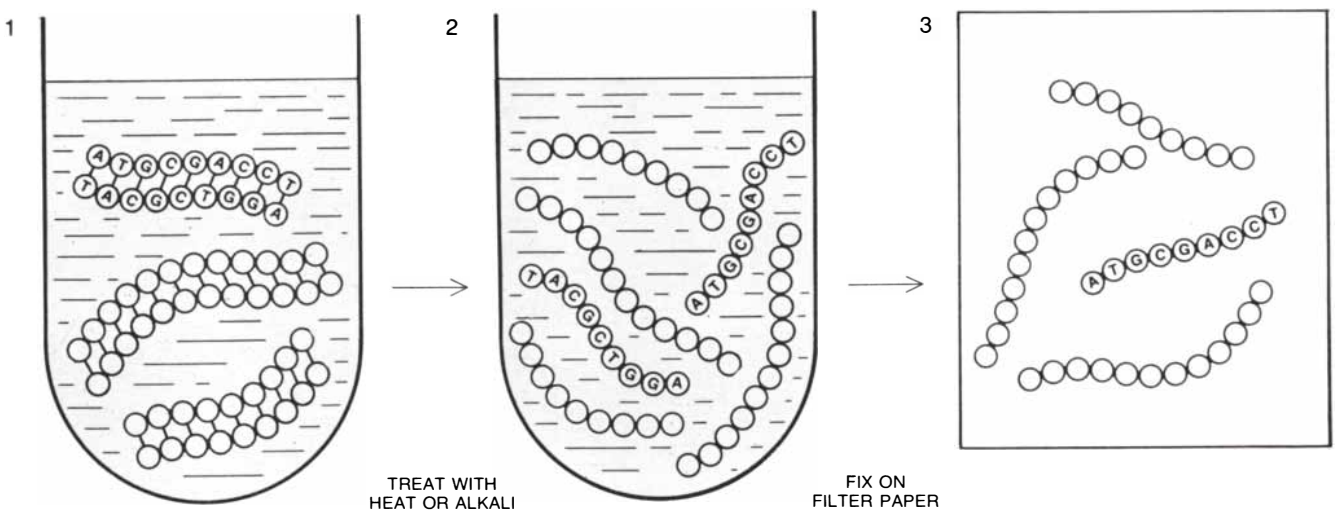
REPEATING PATTERN is evident in a tracing of the molecule on the opposite page. Each of the eight full repeats has one mainly denatured and one undenatured region. Lengths are in microns. Tracings such as this are analyzed to map genes (*see illustrations on page 25*).



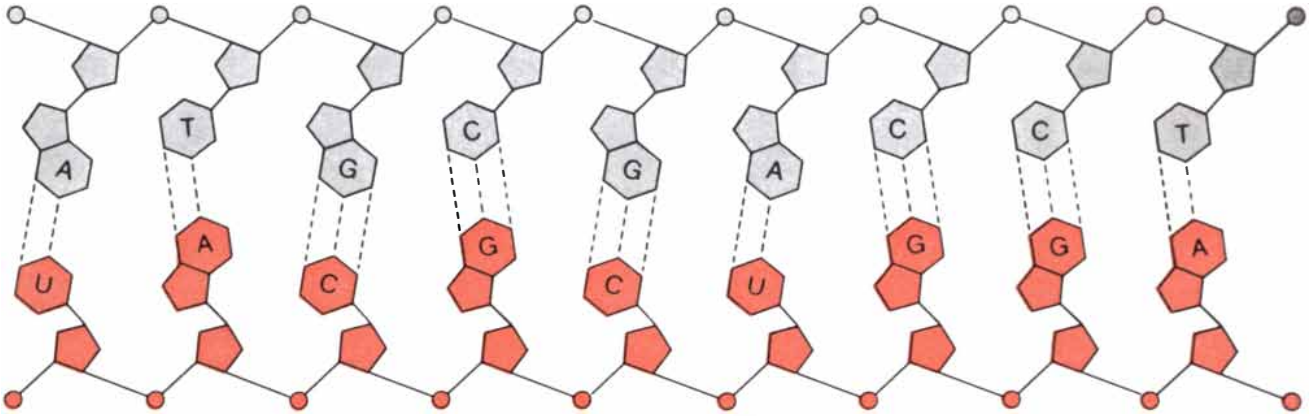
DNA MOLECULE (*gray*) **is a chain of nucleotides, each designated by the initial letter of the organic base it contains: adenine** (*A*), **guanine** (*G*), **cytosine** (*C*) **and thymine** (*T*). **In its usual double-helix form** (*left*) **DNA consists of two strands of nucleotides connected by hydrogen bonds** (*broken lines*) **between the bases. The nucleotides in one strand are always opposite the complementary**

is a chain of the four nucleotides [*see illustration above*]. Each nucleotide consists of an organic base (adenine, guanine, cytosine or thymine), a sugar (deoxyribose) and a phosphate group. All genes are different combinations of these four nucleotides, which are named for their bases: deoxyadenylic acid (*A*), deoxyguanylic acid (*G*), deoxycytidylic acid (*C*) and thymidylic acid (*T*). The sequence of the nucleotides on one strand of the DNA molecule precisely determines the nucleotide sequence on the other strand: *A* is always opposite *T* and *G* is always opposite *C*. (The complementarity has a well-documented physical basis in hydrogen bonding between the nucleotide pairs, since the bonds either cannot be formed or are formed weakly between other combinations of the nucleotides.) This specificity is what determines the exact nucleotide arrangement within the gene product RNA, a polymer that is chemically very similar to DNA. There are three differences: most RNA molecules have only one strand; RNA has three of the same bases as DNA (*A*, *G* and *C*) but the base thymine (*T*) is replaced by uracil (*U*); the sugar molecule (ribose) in each RNA nucleotide is slightly different from the deoxyribose of DNA. In the living cell one strand of the gene DNA serves as a template along which the RNA nucleotides are lined up according to the complementarity rules.

In the assay system for gene purification the investigator takes advantage of complementarity to determine just how much of a sample of DNA is the gene (or genes) for a specific RNA product [*see illustration below*]. First the DNA is denatured: its two strands are separated by treatment with heat or alkali. Then the DNA is fixed on a piece of filter paper, to which a solution of the RNA, which has been labeled with radioactive atoms, is added. The DNA and the labeled RNA are allowed to interact in the proper salt solution at the correct temperature long enough for the RNA molecules to find any complementary strands of DNA and pair with them, forming stable hybrid molecules in which one strand is DNA and one is RNA. The RNA that does not hybridize is washed away.



HYBRIDIZATION ASSAY pairs a gene with its product on the basis of complementarity. Double-strand DNA (*1*) **is denatured by treatment with heat or alkali and the separated strands** (*2*) **are fixed on filter paper** (*3*). **The filter paper is incubated in a solution of RNA molecules labeled by the incorporation of radioactive uracil** (*4*). **The RNA molecules diffuse over the paper and come in con-**

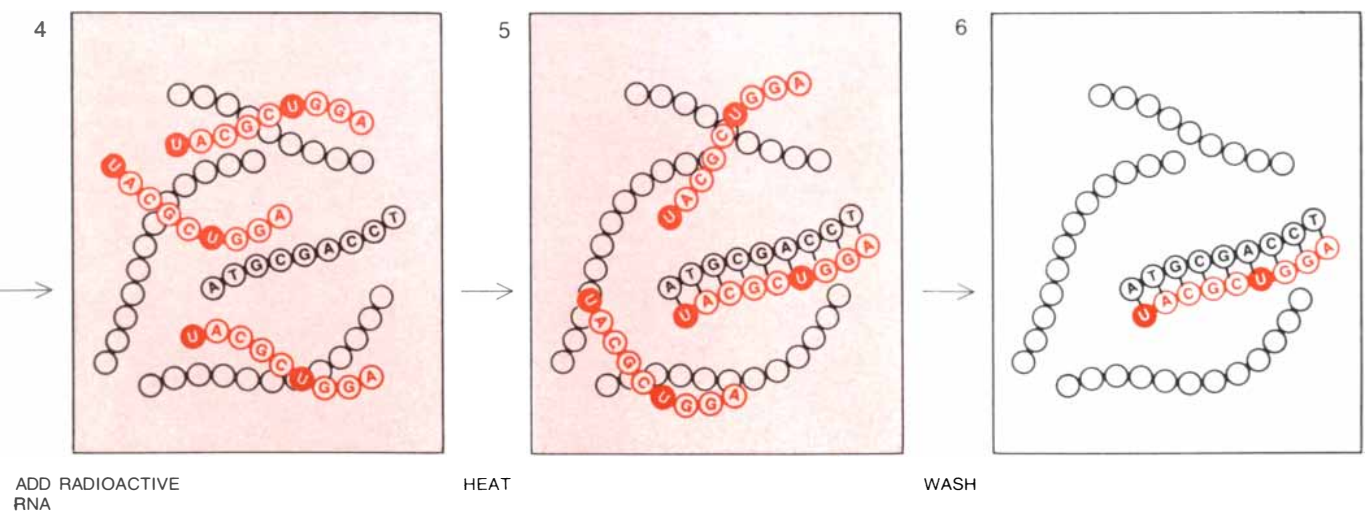


nucleotides in the other strands, with *A* opposite *T* and *G* opposite *C*. This complementarity determines the sequence of the RNA (*color*) that is transcribed from the DNA (*right*). In RNA the sugar is slightly different and uracil (*U*) replaces thymine. Complementarity is the basis of a number of techniques used in gene isolation, including the assay that locates the gene for a specific RNA product.

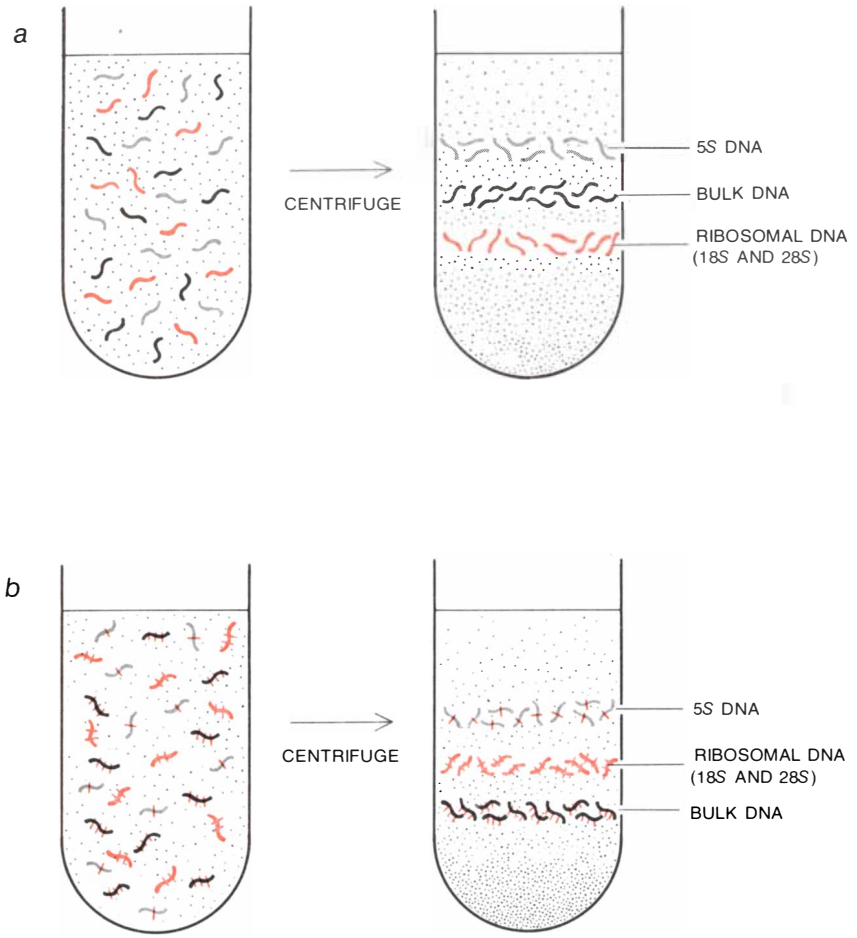
The next step is to measure the radioactivity on the filter paper. The amount of radioactivity shows how much of the RNA has bound to the DNA and therefore how much of the DNA has a nucleotide sequence that matches the RNA's sequence. That DNA is assumed to be the gene for the RNA. This assay system makes it possible to count the number of copies of a gene that exist in the complete DNA complement of a cell. In the course of gene isolation the assay tells one how close each successive DNA fraction is to containing copies of just a single gene.

Clearly the essential requirement for isolating a gene with such an assay is the availability of the RNA product of that gene in a pure form. About 80 percent of the RNA in a cell is in the ribosomes, particles composed of RNA and protein that serve as the site of protein synthesis. In addition to being abundant, ribosomal RNA is easily identified as such—much more easily, for example, than a messenger RNA that codes for a single protein can be identified. Each ribosome includes three kinds of RNA molecule of differing size, which are easily purified. In *Xenopus* they are called the 28S, 18S and 5S molecules. (S stands for Svedberg units, a measure of the speed with which a molecule sediments when subjected to a centrifugal force.) The three molecules respectively have a molecular weight of 1.5, .7 and .04 million daltons and contain 4,700, 2,200 and 120 nucleotides.

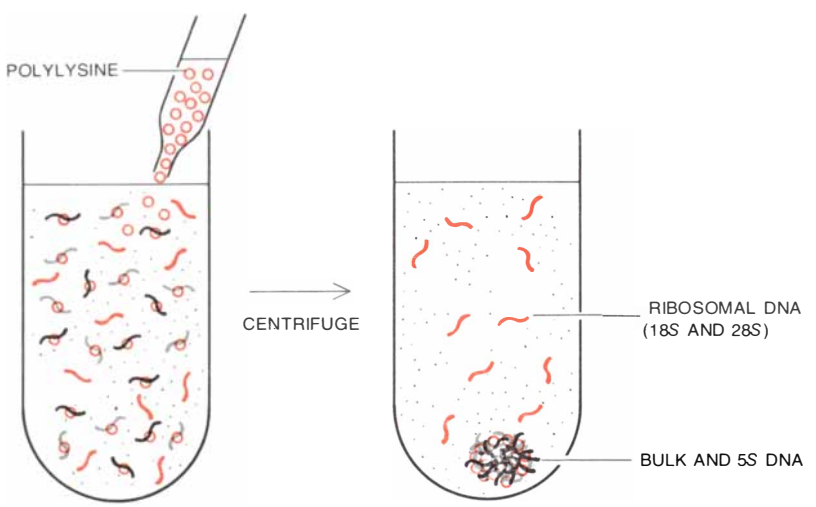
Another characteristic feature of ribosomal DNA that made its isolation possible is its unusual base composition compared with the composition of the rest of *Xenopus* DNA. About 40 percent of the DNA of most animals consists of guanine and cytosine; it is said to be "40 percent $G + C$" (and therefore 60 percent $A + T$). Some years ago, however, the ribosomal RNA's were found to be very rich in guanine and cytosine; their $G + C$ content ranges from 45 to 65 percent, depending on the animal. In the case of *Xenopus laevis* the 18S, 28S and 5S RNA's are respectively 53, 63 and 57 percent $G + C$, whereas the bulk DNA is only 40 percent $G + C$. Since the nucleotide composition of an RNA must be reflected in its gene, a high $G + C$ content



tact with any strands of DNA that are complementary to them, that is, with their gene. The radioactive RNA binds to this DNA, forming DNA-RNA hybrids (5). The remaining, unbound RNA is washed away, leaving only RNA that is paired with its gene (6). The amount of that RNA, determined by measuring the radioactivity, indicates how much of the DNA is the gene for the RNA.



DENSITY-GRADIENT CENTRIFUGATION separates different DNA molecules. In the basic method (*a*) the DNA is spun in a solution of cesium chloride. A density gradient is established by the salt molecules in the solution and each DNA molecule moves to the level at which its density matches that of the solution. The "ribosomal" DNA, containing the genes for 18S and 28S RNA, forms a band at a denser level than the bulk of the cell's DNA. The 5S DNA, on the other hand, is less dense than bulk DNA. It is best separated by a modified process (*b*): centrifugation in cesium sulfate after being complexed with heavy-metal ions (*colored hatching*), silver in the case of *X. laevis* and mercury for *X. mulleri*. The 5S DNA forms fewer complexes and is therefore lighter than the cell's other DNA.



POLYLYSINE METHOD gently separates large ribosomal (18S and 28S) DNA from other DNA. Polylysine, a synthetic polypeptide, reacts preferentially with DNA that has a high $A + T$ content, forming aggregates in a sodium chloride solution. Centrifugation precipitates the aggregates (bulk DNA and 5S DNA), leaving the ribosomal DNA in solution.

is one chemical characteristic that distinguishes the genes for ribosomal RNA from the bulk of the DNA.

That distinction, confirmed by Hugh Wallace and Max Birnstiel of the University of Edinburgh in 1966, is the basis of density-gradient centrifugation, a primary technique for the fractionation of DNA. DNA molecules with different $G + C$ contents have different densities: the greater the $G + C$ content, the higher the density of the DNA. The DNA to be fractionated is mixed with a solution of a cesium salt and spun in a centrifuge at a speed of from 30,000 to 40,000 revolutions per minute. The salt begins to sediment in response to the strong centrifugal force, forming a concentration gradient in the centrifuge tube. Each molecule of DNA moves to a level in the salt gradient at which the salt concentration exactly matches that molecule's density. In this way a mixture of DNA's with different base compositions separates into discrete bands of DNA's at different levels in the salt gradient.

What Wallace and Birnstiel did was to separate *Xenopus laevis* DNA in a centrifuge and test each of the bands by hybridization with radioactive ribosomal RNA. They found that the genes for 18S and 28S ribosomal RNA are denser than the bulk of the cell's DNA. In our laboratory in the Carnegie Institution of Washington we subsequently found that the DNA molecules containing the genes for 5S RNA are lighter than the main-band DNA. (That was surprising, since the 5S RNA gene has a high $G + C$ content. Because the genes themselves are very small we predicted, and later were able to demonstrate, that they are interspersed among stretches of DNA with a very different nucleotide composition, so that the molecule as a whole is low in $G + C$ content.)

The experiments established that two of the three kinds of RNA (18S and 28S) are encoded by genes that band together in cesium chloride and must therefore be on the same molecules (known as ribosomal DNA), whereas the genes for 5S RNA, having a different density, are on other DNA molecules. By measuring the amount of each type of RNA that could be hybridized to *Xenopus* DNA we could estimate the number of each kind of gene in a single set of chromosomes. There are about 450 copies each of the 18S and 28S genes and about 24,000 of the 5S genes in each set of *Xenopus* chromosomes. Since normal body cells have two sets of chromosomes, each cell must contain twice as many of the genes.

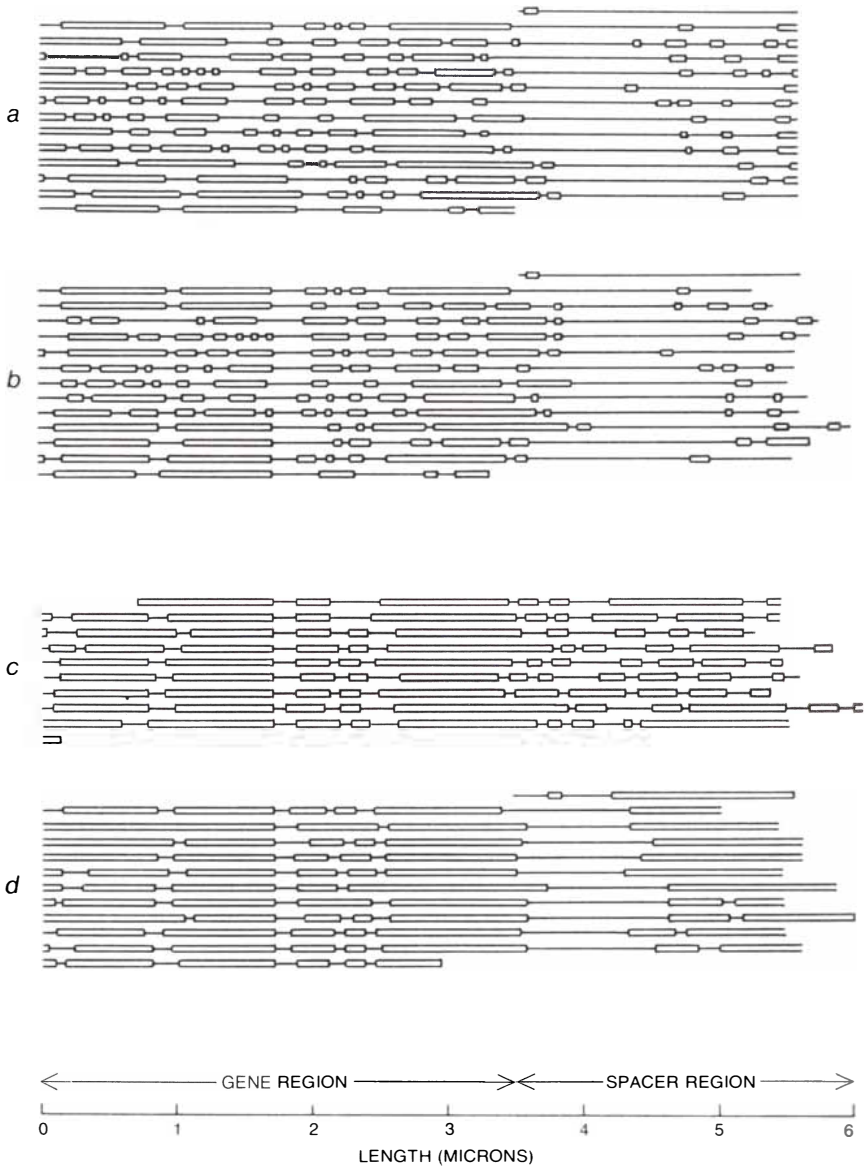
Three characteristics of the ribosomal genes made them good candidates for

isolation: the availability of homogeneous RNA's that could be radioactively labeled for the assay system during gene purification, the physical differences between these genes and the rest of the DNA, and the presence of the genes in multiple copies. The first small amount of *Xenopus laevis* ribosomal DNA, the DNA containing the genes for 18S and 28S RNA, was isolated in 1967 by Birnstiel and his colleagues. They did it by centrifuging *Xenopus* DNA repeatedly in a cesium chloride gradient, each time collecting only the densest fractions.

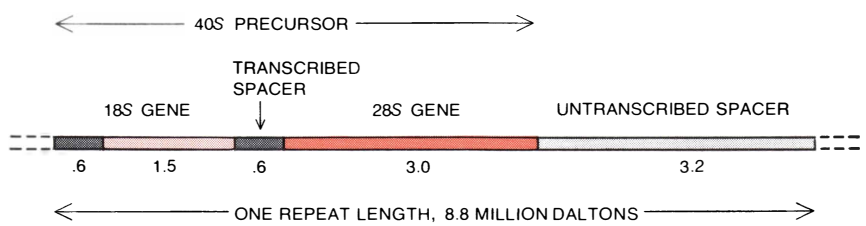
Since 1967 the centrifugation method has been improved and there have been several other advances in technique. For one thing, a number of methods other than centrifugation have been devised for purifying DNA that is high in $G + C$ content. One convenient method we sometimes use is based on an observation by Marc Leng and Gary Felsenfeld of the National Institutes of Health that polylysine, the synthetic polymer of the amino acid *L*-lysine, tends to bind to DNA molecules that are high in $A + T$; it precipitates such molecules selectively. Carefully applied, the polylysine technique can precipitate 99 percent of total *Xenopus* DNA, leaving in solution the ribosomal DNA (the DNA molecules that include the genes for 18S and 28S RNA). Another important development was the discovery that the amphibian oöcyte, the precursor of the egg cell, goes through a period of "gene amplification" in which a great deal of ribosomal DNA is synthesized. The ribosomal DNA in such an oöcyte accounts for 75 percent of the DNA in its nucleus, which amounts to a thousandfold amplification and simplifies the task of purification.

These advances in technique made possible the detailed examination of the structure of ribosomal DNA and the genes contained in it. Three laboratories have done most of the work: Birnstiel's at Edinburgh, O. L. Miller's at the Oak Ridge National Laboratory and our group's in the Carnegie Institution. The combined results of several kinds of experiment, including electron microscopy as well as chemical analysis, add up to a rather detailed picture of the DNA that governs the synthesis of the 18S and 28S RNA molecules [*see bottom illustration at right*].

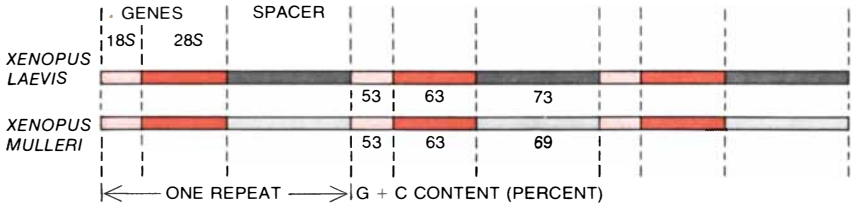
This ribosomal DNA molecule consists of a series of repeating units, each of which includes three major sequences: a gene for the 18S RNA, a gene for the 28S RNA and a "spacer" sequence that is not transcribed into RNA. Chemical evidence indicated that the genes for the



DENATURATION DIAGRAM is made from data in micrographs such as the one on page 20. An average repeat length is determined through computer analysis of measurements from the micrograph and is used to divide the molecule into successive repeats, with the denatured stretches of DNA diagrammed as double lines and the undenatured portions shown as single lines (*a*). It is clear already that there is a largely denatured region (*left*) and a largely unseparated region (*right*). The repeats are adjusted by eye for better alignment (*b*). Two such adjusted diagrams, for *X. laevis* (*c*) and *X. mulleri* (*d*) DNA at an advanced stage of denaturation, are compared. The gene regions are very similar, but the spacer regions are not. In *X. laevis* the spacer is mostly denatured; in *X. mulleri* about half of it is resistant, showing as a single line here and in the micrograph and tracing on pages 20 and 21.



RIBOSOMAL DNA of *Xenopus* is mapped in this diagram of one repeat length based on chemical and electron-micrograph data. The gene region codes for a 40S RNA molecule that is the precursor of the 18S and 28S RNA's, which are cleaved from it in the living cell by enzymes. Two small spacers in the gene region are transcribed into precursor but eliminated in the cleaving process. Then there is the main spacer region, which is not transcribed.



GENE AND SPACER REGIONS of ribosomal DNA from *X. laevis* and *X. mulleri* are compared. The schematic diagram shows three repeats. The gene regions have the same nucleotide compositions as well as the same lengths in the two species; the untranscribed spacers, on the other hand, have the same lengths but very different nucleotide compositions. The transcribed spacers (not diagrammed here) are also different in the two species.

two RNA molecules must be either adjacent or very close together. (Electron-microscope evidence and other experimental results later showed that they are in fact separated by a second very small spacer sequence.) This is consistent with an earlier finding that in the living cell the two large ribosomal RNA's are produced in two steps. First a 40S RNA precursor molecule is transcribed from about 60 percent of the length of one DNA repeat. Then the precursor is cleaved by enzymes; 80 percent of the precursor's length goes to make one 18S and one 28S molecule and the remaining 20 percent is discarded by cellular metabolism. The fraction of the 40S gene sequence that is occupied by the 18S gene and by the 28S gene can be deduced, of course, from the size of the RNA molecules they encode.

The spacer sequence was discovered independently in two laboratories by different methods. At Oak Ridge, Miller isolated functioning genes from living oöcyte nuclei and examined them in the electron microscope. In his micrographs he could see, in a regular repeating pattern along the length of the DNA, arrays of attached RNA molecules: the precursor molecules in the process of being synthesized. Between each of these RNA regions there were stretches of DNA with no RNA attached, and Miller gave the name "spacer" to these apparently inactive stretches [see "The Visualization of Genes in Action," by O. L. Miller, Jr.; SCIENTIFIC AMERICAN, March]. We had detected the spacer region too, first by chemical means. We found that some of the isolated ribosomal DNA would not form hybrids with either the 18S or the 28S RNA. By doing the hybridizing assay with pieces of DNA of various sizes we could determine that these noncomplementary regions were arranged so as to separate the repetitive gene regions. Further experiment showed that the inactive spacer DNA has an unusual sequence of nucleotides, which not only is different from that of the genes but also is not present anywhere else in the *Xenopus* DNA. The present evidence supports the view that no RNA is synthesized from this spacer DNA, and its function remains a mystery.

We went on to confirm the structure of ribosomal DNA by examining large molecules in the electron microscope. The method, developed by Ross Inman of the University of Wisconsin, depends again on the special nucleotide content of ribosomal DNA. The DNA to be studied is partially denatured by the application of heat or alkali and then fixed with formaldehyde, so that the complementary bases in the open loops where the strands have separated cannot snap back together. Inman had shown that stretches of DNA with a higher $A + T$ content denature first under certain conditions while stretches that are richer in $G + C$ still remain double-stranded. We had chemical evidence that much of the spacer region had an even higher $A + T$ content than the gene region. Pieter Wensink, a graduate student in my laboratory, made micrographs of partially denatured ribosomal DNA [*see illustration on page 20*]. Study of the micrographs disclosed a regular repeating pattern in which denatured loops alternated with solid double-strand DNA. Analysis of the pattern by several methods confirmed the earlier chemical data in detail.

Ribosomal DNA, then, consists of two alternating regions: the gene region, containing genes for the essential 18S and 28S RNA, and the spacer region, which has no known function. Now, if a gene defines an essential structure, it is usually found in at least closely related species and is often found throughout the animal kingdom. The genes for such generally essential components evolve more slowly than those for more specialized functions do, and the genes for ribosomal RNA are highly conserved in evolution. We have tested the nucleotide arrangement of *Xenopus* ribosomal RNA with the ribosomal genes in the cells of more than 50 organisms, including mammals, invertebrates, higher plants, protozoa, fungi and yeasts. In all cases the ribosomal genes have at least some nucleotide sequences in common with *Xenopus* ribosomal DNA.

Unlike the gene regions of ribosomal DNA, the spacers evolve rapidly. We have compared the ribosomal DNA's of *X. laevis* and *X. mulleri,* species so closely related as to produce viable hybrid progeny, some of which reach sexual maturity. The two species' ribosomal RNA's cannot be distinguished from each other by physical or chemical tests, which is to say that there has been no detectable change in the genes since the two species diverged. And yet their spacers are very different. We estimate that at least 10 percent of the bases within the spacers have changed during the period since the divergence of the *Xenopus* species. The evolutionary conservation of the gene regions is really extraordinary in contrast to this rapid divergence of the spacer regions. The genes that code for ribosomal RNA in higher plants are closer in sequence to the genes for that RNA in *Xenopus* than spacer sequences of *X. laevis* are to those of *X. mulleri!*

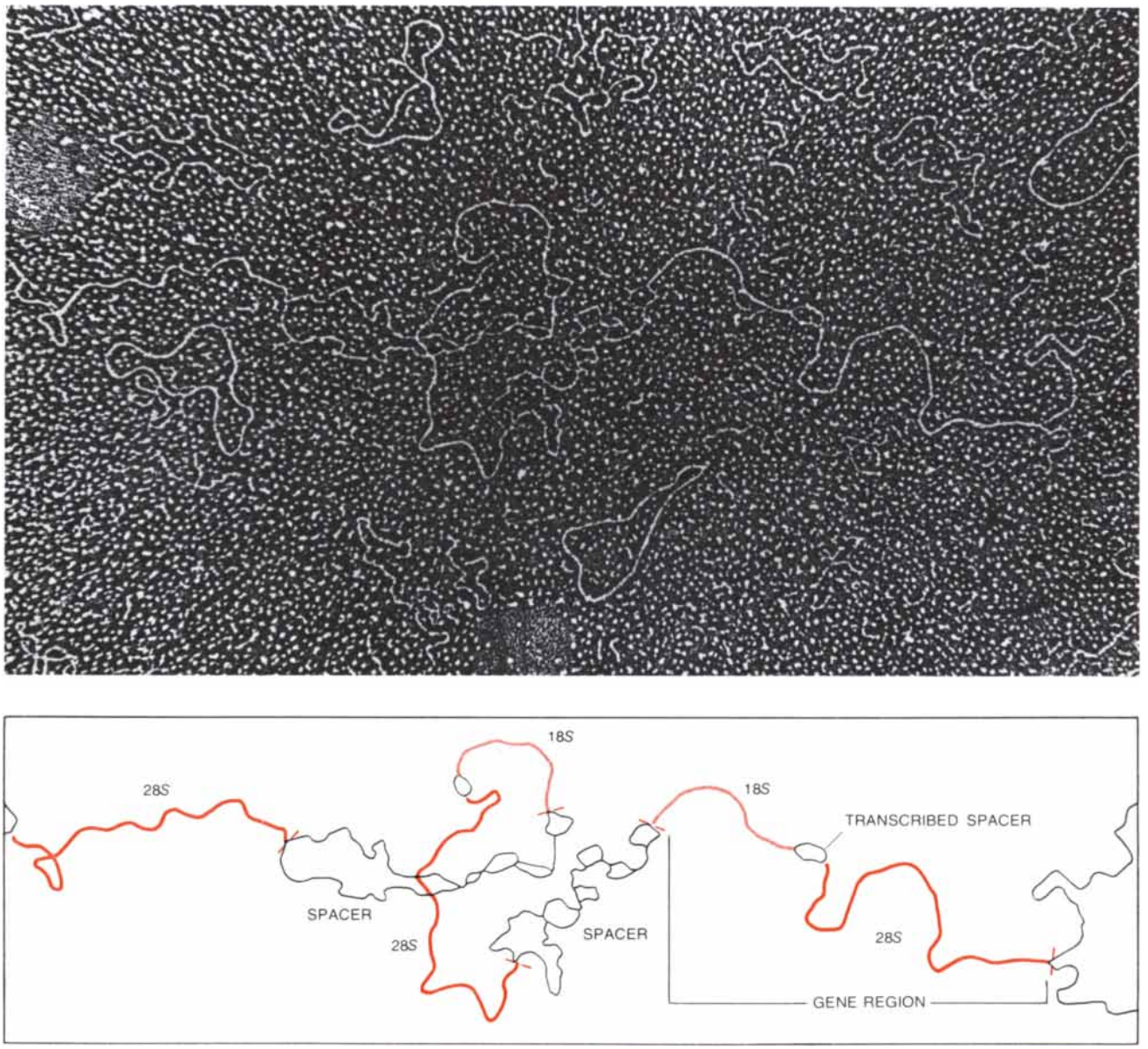
Since most of the spacer region is not conserved in evolution, we assume that the exact arrangement of bases within at least most of the region had no selective importance for the species and could therefore change rapidly. If this is true, however, we are faced with an extraordinary dilemma. Remember that there are about 450 spacer regions in each set of chromosomes. Our chemical studies show that the multiple spacers in the ribosomal DNA of *X. laevis,* although markedly different from those of *X. mulleri,* are nonetheless very similar to one another—perhaps identical. The same is true of the spacers in *X. mulleri.* How can 450 DNA regions evolve independently but almost identically? We call such simultaneous, parallel evolution of many sequences in the gene complement of a single species "horizontal" evolution; it contrasts sharply with the usual evolution of a single gene in two or more diverging species. Population genetics and evolutionary theory provide no guide for understanding horizontal evolution.

The 5S RNA component is small enough, and available in large enough amounts, so that the exact arrangement of its nucleotides has been determined. This feat was accomplished first by George Brownlee and Frederick

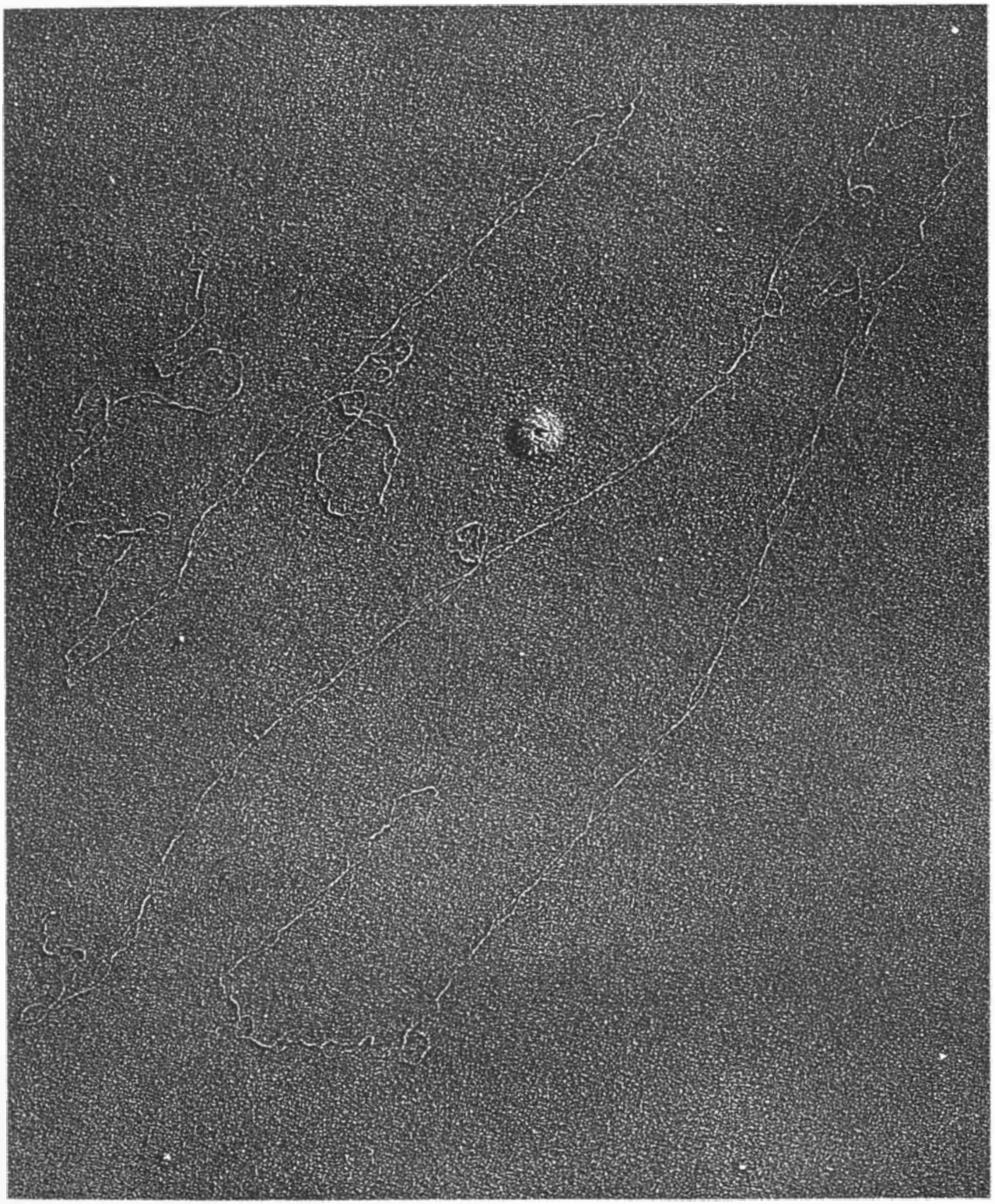
Sanger of the University of Cambridge in 1965 for bacterial 5S RNA; their technology, involving the repeated, selective cleaving of the molecule with different enzymes, has since been applied to the small ribosomal component from a number of animals, including *Xenopus*. The molecule contains 120 nucleotides, and like the 18S and 28S RNA it is highly conserved in evolution: the 5S molecules from all mammals studied to date, from marsupials to man, have exactly the same 120 nucleotides. In the amphibian *Xenopus* different tissues synthesize different kinds of 5S RNA. Body cells synthesize one kind, eight of whose 120 nucleotides are different from those in the mammalian molecule; the oöcytes synthesize a mixture of about four kinds of 5S RNA, of which the predominant form has seven bases that are different from those of the body-cell molecule. How many of the 24,000 gene sequences for 5S RNA code for each version of the molecule is not known.

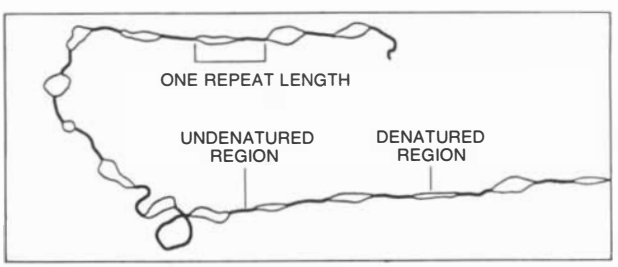
We have purified 5S DNA from both *X. laevis* and *X. mulleri* cells. Once again the unusual base composition of this DNA is what makes its isolation possible. In this case density-gradient centrifugation is carried out after the DNA has been allowed to react with heavy metal ions: silver for *X. laevis* and mercury for *X. mulleri*. The metal ions react with certain base pairs in the DNA, forming complexes that are denser than the original DNA. Because of its nucleotide content the 5S DNA forms fewer such complexes than the other DNA in the cell, and so it is lighter and can be isolated by centrifugation.

When we analyzed the structure of



HETERODUPLEX MOLECULE consisting of one strand of ribosomal DNA from *X. laevis* and one from *X. mulleri* was formed and pictured in an electron micrograph by Arleen Forsheit and Norman Davidson of the California Institute of Technology (*top*). The molecule has a structure much like the one formed by denaturing, on page 20, but the structure has a very different origin. DNA from the two species was mixed and completely denatured. Then the single strands of DNA were allowed to pair up. Some strands paired with their complementary strands from the same species, forming complete double-strand molecules. Other strands, such as the ones in the micrograph, came in contact with strands from the other species. Where the nucleotide sequences matched, the two strands became bound; where they did not match, the strands remained separated, forming looped regions, as shown in the tracing (*bottom*). The bound regions and the mismatched, looped regions are the gene and the spacer regions respectively.

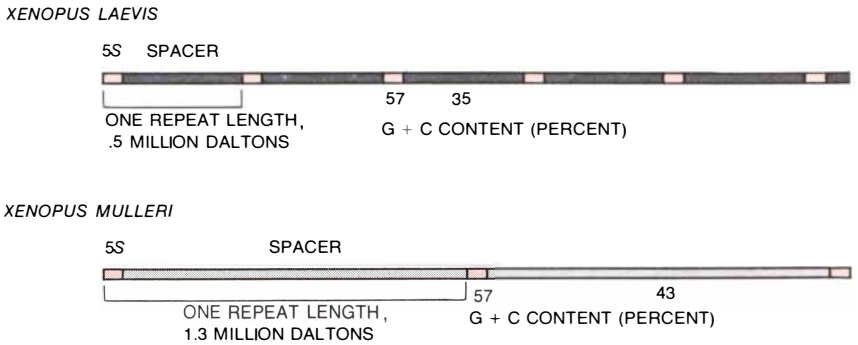




5S DNA from *X. laevis* is enlarged 50,000 diameters in an electron micrograph made by Wensink by the same technique as the one on page 20: the DNA was partially denatured so that the relatively high $A + T$ regions separated while those with the highest $G + C$ content remained unseparated. Again the denaturation pattern is regular along the entire length of the two molecules, as shown in the tracing of a part of one molecule (*right*). The high-$G + C$ gene for 5S RNA takes up about a third of the length of each undenatured region; the remaining six-sevenths of the repeat (all of the looped region and two-thirds of the undenatured region) is spacer.

5S DNA from *X. laevis* by many of the techniques described above for ribosomal DNA, it was apparent that other nucleotide sequences besides those of the genes for 5S RNA were present. When the DNA is hybridized with radioactive 5S RNA, only about a seventh of one of the DNA strands binds the RNA, indicating that the remaining six-sevenths of the DNA consists of spacer sequences. When Wensink did the electron-microscope experiment on purified 5S DNA, this structure was confirmed. The micrographs show clearly an arrangement in which loops of strand-separated DNA alternate with natural double-strand regions along each molecule [*see illustration on opposite page*]. One repeat length (the combined length of a loop and a double-strand stretch of DNA) is about seven times the predicted length of a gene for one 5S RNA molecule. Similar experiments with purified 5S DNA from *X. mulleri,* carried out with Kazunori Sugimoto, have shown that only about an eighteenth of the DNA molecule encodes the 5S RNA. Since the RNA molecule is the same size in both animals, the difference in repeat length must be due to a much longer spacer [*see top illustration at right*].

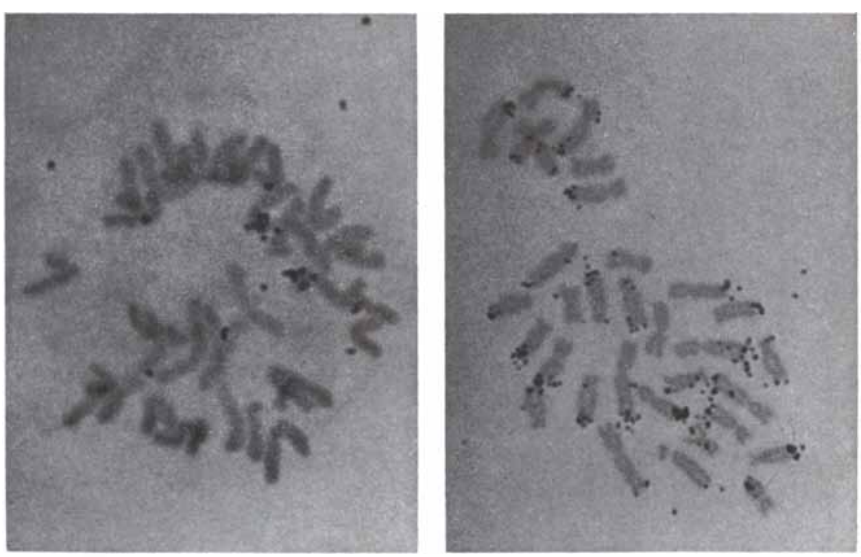
The 5S-DNA spacers are not only different in length in the two species but, like the spacers for the ribosomal genes, also very different in nucleotide sequence. Again, this interspecies difference contrasts with remarkable intraspecies similarity: the many thousands of spacer sequences in the 5S DNA of a single species are not quite identical (as they are in the ribosomal genes) but they are almost the same. The interspecies difference records very rapid evolution, implying that any function the spacers have must surely not depend on precise nucleotide sequence. The rapid evolution of the spacers contrasts, of course, with the 5S RNA genes themselves, which are highly conserved in evolution: we cannot distinguish *mulleri* 5S RNA from *laevis* 5S RNA, and so we assume that their genes are very similar and perhaps identical. Thus the repeating sequences of 5S DNA composed of spacer and gene regions behave as a family of closely related but not identical DNA regions.



SPACER REGIONS are very different in *X. laevis* and *X. mulleri*, in length as well as in nucleotide composition. The gene regions are the same length in both and appear to synthesize the same 5S RNA. The gene accounts for about a seventh of the total repeat length in *X. laevis* but only for about an eighteenth of the much longer repeat length in *X. mulleri*.

The chromosomal location of the 18S and 28S DNA was established first by genetic experiments, which showed that all this DNA was clustered at a single site on one of the 18 chromosomes of *Xenopus*. This has been confirmed by an elegant new mapping technique developed independently by Joseph Gall and Mary Lou Pardue at Yale University and by Kenneth Jones and his colleagues at the University of Edinburgh. In this method a complete set of discrete *Xenopus* chromosomes is fixed on a slide and treated with mild alkali, which denatures the DNA. A drop of radioactive RNA is placed over the slide and hybridization is allowed to occur. Then the slides are washed free of unbound RNA and dipped into a photographic emulsion; wherever there is radioactive RNA, silver grains are developed. When this experiment is done with a radioactive ribosomal RNA, the corresponding genes hybridize to the RNA and thus bind the radioactivity, and in this way the genes are easily located. Large ribosomal DNA turns out to be localized, as predicted, on only one chromosome per set: at the secondary constriction known as the nucleolar organizer, the site where the nucleolus is known to form in the actively synthesizing cell. On the other hand, most if not all of the *Xenopus* chromosomes contain 5S DNA, and it is always at the ends of the chromosomes [*see illustration below*].

In the future, with the development of more sophisticated methods for fractionating DNA, we should be able to purify many genes with a variety of other functions, in particular those that code for specific proteins. With these purified genes it will be possible to analyze their structure, chromosome location and evolution and the control mechanisms that regulate their action in the living cell.



CHROMOSOMAL SITES of large ribosomal DNA and 5S DNA in *X. laevis* are demonstrated in these autoradiographs made by Mary Lou Pardue of the Massachusetts Institute of Technology. Radioactive RNA's are allowed to hybridize to chromosomes fixed on slides and the unbound RNA is washed off. A photographic emulsion is applied and the bound RNA is revealed by the silver grains it develops. The autoradiographs locate the ribosomal DNA at a single site on one chromosome, two copies of which are present in the diploid cell (*left*). 5S DNA, on the other hand, is at the ends of most of the chromosomes (*right*).

Experiments with Neutrino Beams

The neutrino was first detected 17 years ago. Accelerators are now producing copious beams of these particles to probe the structure of other particles and the forces between them

by Barry C. Barish

In 1931 Wolfgang Pauli proposed the existence of a new particle to account for missing energy and momentum in the nuclear process known as beta decay. Named the neutrino, the particle would need to have no electric charge and almost no mass. With the advent of fission reactors, in which beta decay is torrential, there was a source intense enough to make it conceivable to detect such virtually impalpable entities, and in 1956 the first neutrino was recorded. In the 17 years since that pioneer effort the technology of particle accelerators has so advanced that it is now possible not merely to detect neutrinos but to produce beams of them for laboratory experiments. The most important instrument for neutrino physics is the powerful new accelerator at the National Accelerator Laboratory (NAL) near Batavia, Ill. This huge machine, in which protons are accelerated in a ring more than four miles in circumference, went into operation last summer. It can generate an intense beam of neutrinos at an energy in excess of 100 GeV (100 billion electron volts). Experiments are emplaced to observe the interactions of these high-energy neutrinos with other particles, and preliminary results are already available.

Beta decay is the process in which a neutron in an atomic nucleus emits an electron and becomes a proton, or a proton emits a positron (a positive electron) and becomes a neutron. The nucleus gains or loses positive charge as a result. It had been observed that the kinetic energy of the electron in this process was less than would be expected from the differences in the masses of the parent and daughter nuclei. The observation seemed to threaten two fundamental laws of physics: the conservation of energy and the conservation of momentum. Pauli suggested a solution by proposing the existence of the neutrino, which would be an additional product of beta decay and would presumably account for the missing energy and momentum.

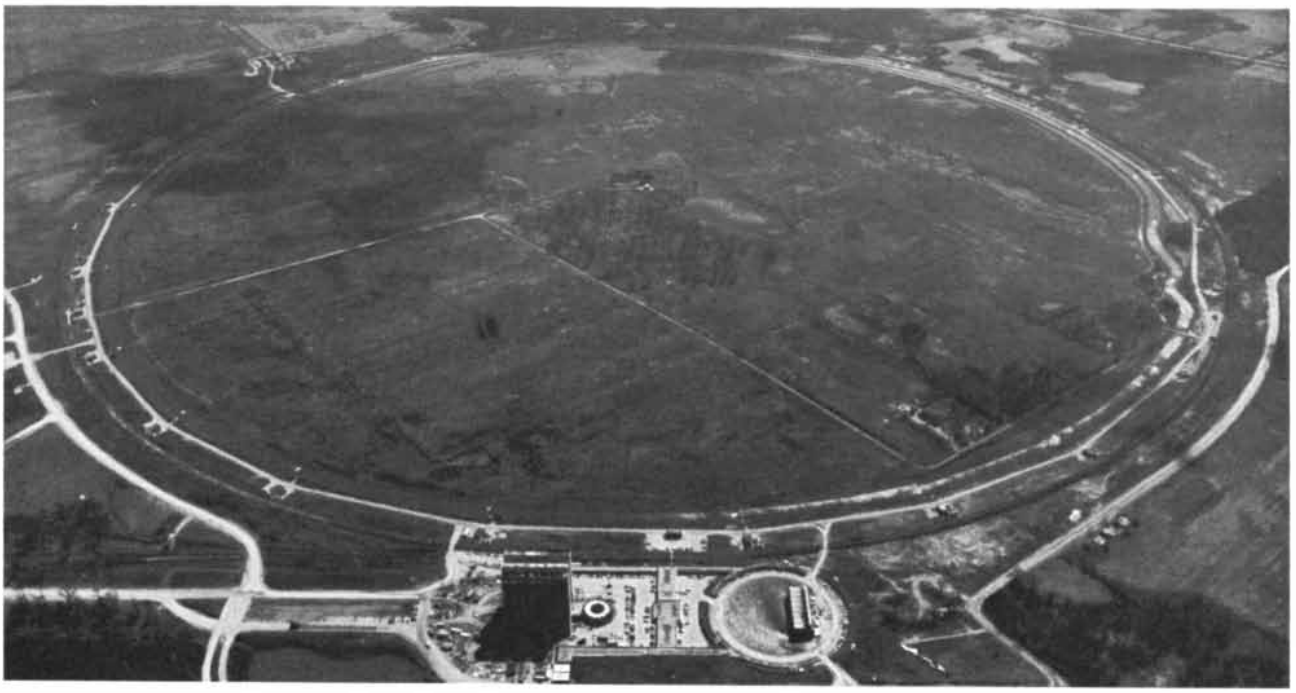
His proposal would save the conservation laws, but could it be accepted without verifying that such a particle truly existed? In subsequent years Enrico Fermi formulated a theory of weak interactions based on the existence of the elusive neutrino, a theory whose basic structure is still in use today. Weak interactions are the manifestation of the almost unimaginably weak forces associated with the spontaneous decay and transformation of certain nuclear particles, and beta decay is the most instructive and the best known of the weak interactions.

The neutrino had originally been deduced solely to account for the missing energy and momentum in beta decay. In order to confirm its existence as a real particle it would be necessary for a neutrino to somehow be observed as a free particle. In other words, one would have to observe an interaction of a neutrino and another particle away from the neutrino's point of origin. Such an observation represented a monumental experimental challenge. A single neutrino can go through the entire earth with scarcely any probability of its interacting with anything else. A fission reactor, however, manufactures large quantities of neutrons and artificially radioactive nuclei that undergo beta decay. Such a reactor would presumably liberate a flood of neutrinos, an almost infinitesimal fraction of which might then interact with other particles.

Frederick Reines and Clyde L. Cowan of the Los Alamos Scientific Laboratory took up the challenge and mounted a project that was ambitious even by today's standards. Choosing as their source of neutrinos a reactor in the Atomic Energy Commission's Savannah River plant in Georgia, they designed an experiment to detect the elusive particle through inverse beta decay in a tank of liquid adjacent to the reactor. The neutrino was postulated to be a product of the decay of the neutron; the reaction would be neutron $\rightarrow$ proton + electron + neutrino. In principle the process would be reversible. It is the reversed reaction that is inverse beta decay: neutrino + proton $\rightarrow$ neutron + positron. The positron participates in the reaction because charge must be conserved; it carries off the charge of the proton.

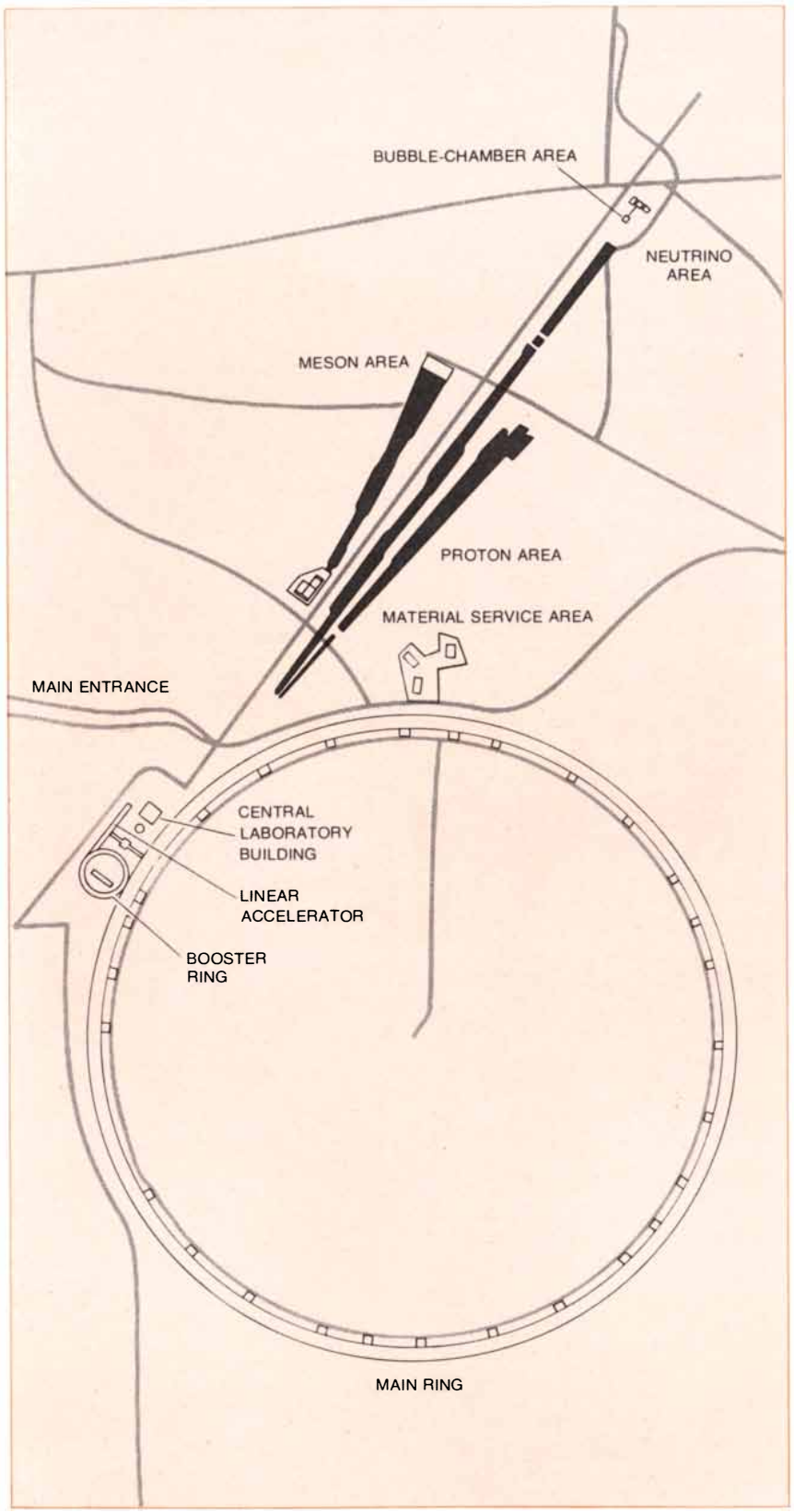
When the positron encounters an electron in the surrounding matter, the two particles annihilate each other. They give rise to two photons, which can then be caught by two detectors linked so that they respond only when photons pass through both of them simultaneously. Reines and Cowan's experiment was capable of identifying inverse beta decay by detecting such an annihilation in coincidence with the capture of a neutron. The two events together would point to an interaction involving a neutrino.

In order to identify these neutrino interactions Reines and Cowan had to screen out a formidable amount of background radiation from the reactor. Moreover, they had to be able to distinguish the neutrino interactions from those due to the large number of cosmic rays going through the apparatus. In spite of all these obstacles the experiment was a success, and the free neutrino was detected. The predictions made by Pauli and Fermi were at last confirmed, and the miss-



HUGE ACCELERATOR of the National Accelerator Laboratory (NAL) near Batavia, Ill., is one of those that is producing neutrino beams. This aerial photograph shows the main ring of the machine, which is 1.24 miles across and more than four miles in circumference. The NAL accelerator went into operation last year. It can accelerate protons up to 500 GeV (500 billion electron volts).

NEUTRINO AREA at NAL is visible as a long mound of dirt extending from the main accelerator ring near the horizon (*top left*). The mound filters out all particles in the beam from the accelerator except the neutrinos. The experiments are mounted in the various buildings along the line of the beam. The building with the geodesic dome (*lower right*) is the control center for the neutrino area.



MAP OF NAL MACHINE shows the three major accelerators in series that accelerate protons. Protons generated in an ion source are introduced into a linear accelerator some 500 feet long that accelerates them to an energy of 200 MeV (million electron volts). They are then injected into the booster ring accelerator, which carries them to an energy of 8 GeV. The booster is a rapid-cycling synchrotron 500 feet in diameter. From the booster the protons are injected into the main ring and are accelerated to full energy. They are extracted and directed to the experimental areas, which can work singly or in any combination.

ing energy and momentum of beta decay were accounted for.

Once neutrino interactions were actually observed it soon became apparent that a new field of particle physics had been born. What is this area called neutrino physics? Originally it was the study of the neutrino to determine its properties and its role in physical processes. Recently, however, neutrino physics has taken an interesting turn. New sources of neutrinos and new neutrino detectors are opening up a myriad of exciting possibilities. Neutrinos are now seen as an ideal probe for examining the weak interaction at high energies, where present theories break down. Moreover, neutrinos can be used as a "microscope" to study the structure of other nuclear particles.

In the proton accelerator at the National Accelerator Laboratory we have for the first time a source of neutrinos at such high energy that we might be able to perceive new phenomena in the weak interaction. The accelerator works in several stages. A linear accelerator provides protons at an energy of 200 MeV (million electron volts), which are injected into a booster ring. The booster ring accelerates the protons to 8 GeV and injects them into the four-mile main ring. The protons, held in orbit by magnets, are then accelerated to energies as high as 500 GeV. Finally, they are deflected out of the ring and are aimed into various experimental areas.

In neutrino experiments the energetic protons travel about 1,000 meters to a target area. The neutrino beam is formed by first directing the protons into a thick target. The protons interact strongly with the nuclei in the target, typically creating a large number of pions and kaons. A special focusing system captures these secondary particles and directs them down a 400-meter "decay tube." Since pions and kaons are unstable, some of them decay as they proceed down the tube, usually giving rise to a muon and a neutrino.

At the end of the tube is a long absorbing shield consisting of 1,000 meters of dirt fill, which serves to filter unwanted particles out of the neutrino beam. The pions, kaons and protons are absorbed in the shield quickly. Only the muons persist, and in the end even they are eliminated by the shield. The only particles that are oblivious to the 1,000 meters of dirt are the neutrinos. Therefore what emerges from the long shield is essentially a pure beam of neutrinos. A bubble chamber 15 feet in diameter

sits portentously at the end of the shield. When its piston pulls back, lowering the pressure on the liquid hydrogen or neon that fills the chamber, the products of the interactions of the neutrinos and the target nuclei are revealed as tracks of fine bubbles.

What kinds of event can we hope to observe with such powerful apparatus? To answer the question we need to say a bit more about the neutrino itself and the physical problems it presents.

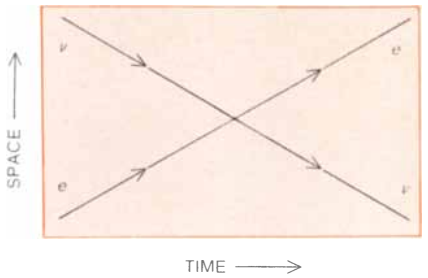
After the neutrino had first been observed, the main questions and experimental efforts revolved around determining the properties of this strange new particle. Several of its properties were already known from beta decay. For example, the conservation of charge told us that the neutrino is electrically neutral. The kinematics of the decay required that the mass of the particle be very small, perhaps even zero. Another property that could be determined from beta decay was the neutrino's intrinsic angular momentum, or spin on its axis.

It is a fundamental feature of quantum theory that a particle can have a spin of only integral or half-integral multiples of Planck's constant. (That constant, $\hbar$, relates the energy of a quantum of radiation to its wavelength: energy equals 2π times frequency times $\hbar$.) The neutron has a spin of $1/2\ \hbar$. It decays into two spin-$1/2\ \hbar$ particles (the electron and the proton) plus a neutrino. Hence in order for angular momentum to be conserved in the decay the neutrino must have a spin of $1/2\ \hbar$. Reines and Cowan were able to show, in an experiment conducted after they had first detected the neutrino, that the new particle had a preferred orientation for its spin with respect to its direction of motion.

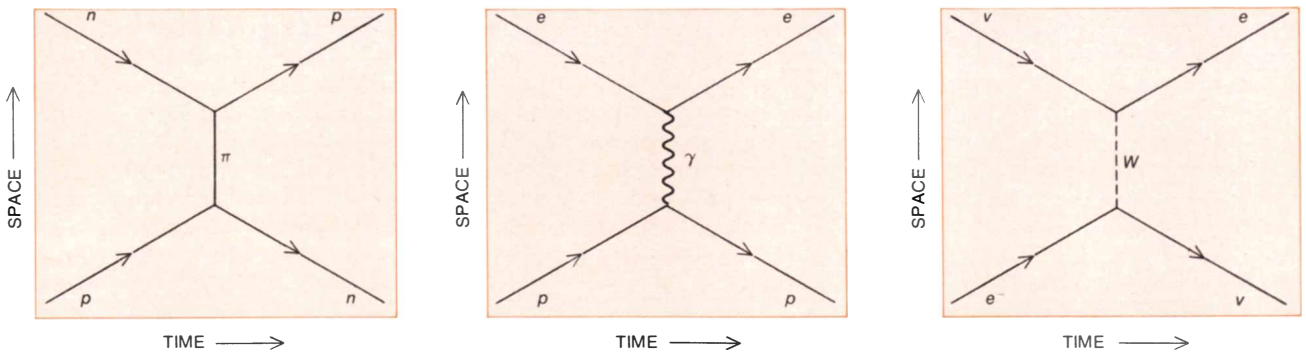
Was the neutrino that gave rise to inverse beta decay the same as the one emitted in "normal" beta decay? That is, was the particle associated with the positron different from the one associated with the electron? Raymond Davis, Jr., of the Brookhaven National Laboratory addressed this question in 1958. His source of neutrinos, like Reines and Cowan's, was a fission reactor at Savannah River. He discovered that the neutrino associated with the positron is the antiparticle of the one associated with the electron. The electron, the positron, the neutrino and the antineutrino are all weakly interacting light particles of the class designated leptons. The preference of the neutrino for the electron and of the antineutrino for the positron has been included in the theory of weak interactions by assigning a new quantum number, or defining quantity, called the lepton number to the various weakly interacting particles. The convention is that the electron and the neutrino have a lepton number of $+1$, whereas their antiparticles, the positron and the antineutrino, have a lepton number of -1. Lepton number is then assumed to be another quantity that is conserved in an interaction.

The next major step in neutrino physics came in the early 1960's, when intense sources of neutrinos were provided for the first time by the large proton accelerators at Brookhaven and at the European Organization for Nuclear Research (CERN) in Geneva. There were hints that there might be still more rules of selection governing the decay of weakly interacting particles. In particular there was still no explanation of why certain reactions were not observed, for example the decay of a positive muon into a positron and a photon. That problem was dramatically resolved at Brookhaven in a classic experiment performed by Leon M. Lederman and Melvin Schwartz of Columbia University [see "The Two-Neutrino Experiment," by Leon M. Lederman; SCIENTIFIC AMERICAN, March, 1963]. In the course of their experiment Lederman and Schwartz demonstrated still another remarkable fact about neutrinos. Not only are there neutrinos and antineutrinos; there are two kinds of each. Lederman and Schwartz discovered that the neutrino "remembered" that it came from a decay associated with a muon. In subsequent interactions these neutrinos made muons but never electrons.

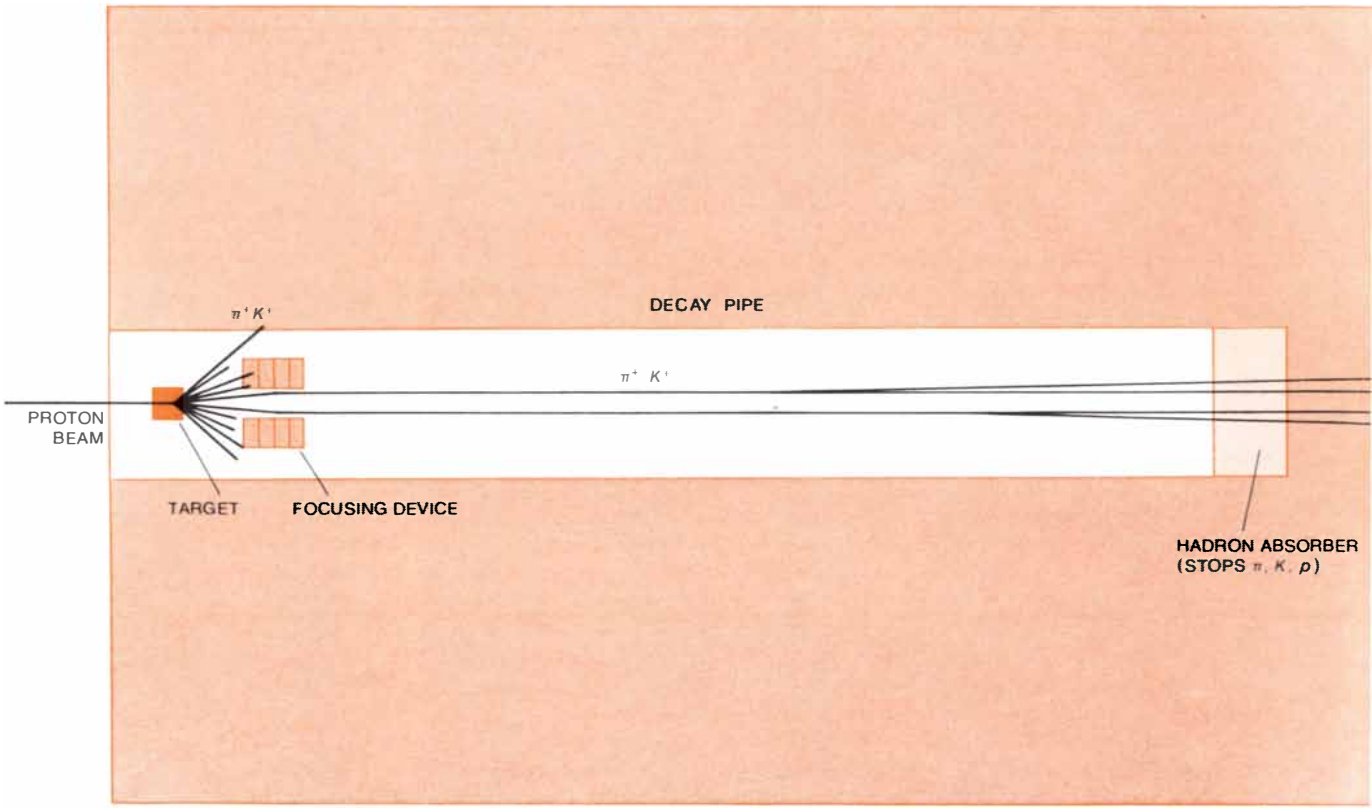
One of the greatest puzzles in particle physics is explaining nature's need for both muons and electrons in the first place. They seemingly have identical properties and roles, except that the muon has about 200 times the mass of the electron. Now, with the discovery that each is associated with its own neutrino, there is an even more puzzling situation. We have not only the unexplained need for both a muon and an electron but also an unexplained need



FERMI THEORY of weak interaction of two particles such as a neutrino (ν) and an electron (e) stated that particles would interact at single point in space and time.



***W*-BOSON HYPOTHESIS** modifies the Fermi theory of the weak interaction by introducing a "carrier" of the weak force called the W boson. The W boson is analogous to the carriers of the other forces in nature. In the strong interaction illustrated at left the attractive force between the neutron (n) and the proton (p) is carried by a pion (π). In the electromagnetic interaction (*middle*) the photon (γ) is the mediator between the proton and the electron. In the weak interaction (*right*) the W boson is exchanged. Here the weak interaction no longer takes place in a single point in space and time, and there is an analogy between the three forces.



BEAMS OF NEUTRINOS are formed at NAL through an elaborate series of particle interactions. High-energy protons from the main accelerator ring strike a target, producing a large number of pions and kaons (K). They are focused and directed down a pipe that is

for the fact that each has its own neutrino. There are many other outstanding questions that remain to be answered as well. Is the mass of the neutrino really zero or merely very small? Is the neutrino stable or does it eventually decay? Are such observations as the conservation of lepton number and the fact that there are two kinds of neutrino approximate rules or are they strictly true?

One deep mystery is related to another experiment conducted by Davis. He has observed that the number of neutrinos that reach the earth's surface from nuclear processes in the sun is well below the number expected on the basis of astrophysical calculations. Is it that we do not fully understand the processes within the sun, or could it be that we are not detecting the neutrinos because of some new property of the particles or some shortcoming of weak-interaction theory?

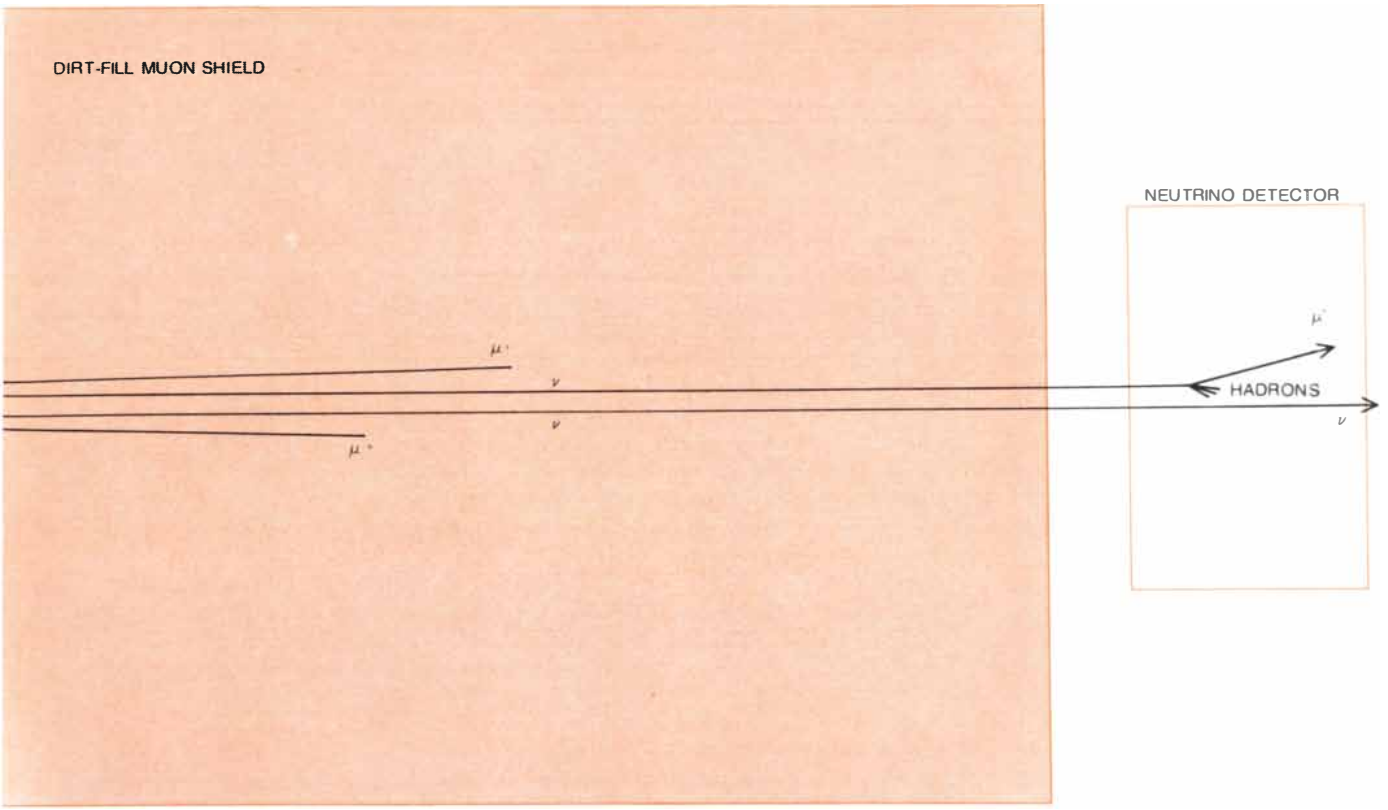
At the same time that work on these problems is continuing there has been a notable shift in the pursuit of neutrino physics. Although much remains to be learned about the properties of the neutrino, enough is known about it for it to be utilized as a powerful tool for studying some of the other fundamental problems in the physics of elementary particles. Neutrinos make an ideal instrument for studying strongly interacting particles at extremely short distances. Strongly interacting particles, which belong to the class designated hadrons, are particles such as the neutron, the proton and the pion that interact through the strong, short-range nuclear force that is responsible for the binding of these particles in atomic nuclei. All the particles discovered to date participate in strong interactions except the photon and the four weakly interacting leptons.

Hadrons have a diameter of about 10^{-13} centimeter. Neutrinos can be boosted to energies high enough for them to be used to examine structure within distances as short as 10^{-16} centimeter. Moreover, neutrinos are ideal for the purpose because their own structure is very simple. In other words, high-energy neutrinos provide a kind of hadronic microscope.

The fundamental question is whether or not there is still another layer of substructure underlying the elementary particles. The usefulness of lepton beams for this kind of study has been displayed at the Stanford Linear Accelerator Center (SLAC). There a beam of electrons was directed at targets of protons. The first general finding was that each hadron appears to consist of tiny pointlike bodies. It is as if the hadrons are made up of various combinations of other particles, which have been given the generic name parton by Richard P. Feynman of the California Institute of Technology.

Before Feynman had put forward the concept of partons, George Zweig and Murray Gell-Mann of Cal Tech had proposed a theoretical subhadronic entity called a quark. There is some evidence that partons and quarks are the same, although they have been postulated in different ways. In general the parton-quark models and some even more complex calculations are in agreement with present observations.

One of the most interesting predictions of such models concerns certain measurements of the scattering of neutrinos where the results will depend directly on the charge of the partons (or quarks). It will be possible to make such measurements soon. The prospect of probing the hadrons within such extraordinarily small dimensions to study their structure is exciting. Will the experiments confirm the predictions of the parton-quark models? Or will we find surprising new behavior? During our ini-



long enough for many of them to decay before they reach the end. When they decay, they create both neutrinos and muons (μ). The dirt-fill shield filters out unwanted particles, leaving an almost pure beam of neutrinos. Detectors include a 15-foot bubble chamber.

tial investigations within the past few months at the National Accelerator Laboratory my colleagues and I have found that neutrino scattering seems to at least qualitatively confirm the results from electron scattering at SLAC.

Neutrinos promise not only to be an effective probe of the hadrons but also to be a valuable means of testing Fermi's theory of the weak interaction. The theory is simple and is remarkably successful at describing both beta decay and the results of interactions between low-energy neutrinos and other particles. Yet we are certain that the theory is wrong!

The crux of the difficulty is that the Fermi theory is not finite. For example, if a neutrino is aimed at an electron, there is a certain area, called the effective cross section, where the two will interact. The cross section is a measure of the strength of the interaction. Physically there is a maximum allowable cross section that is related to the diameter of the target particle and to the wavelength that is associated with the projectile particle by virtue of its energy. The radius of the cross section can never be larger than one wavelength. The wavelength of a very slow projectile neutrino is long, and so the maximum effective cross section of its interaction with the electron is rather large. As the neutrino's energy is increased its wavelength gets shorter; thus the maximum cross section of the interaction gets smaller at higher energies. The Fermi theory, however, states that the particles will interact at a single point at low energy, and it predicts that the cross section will rise on a straight curve to infinity as the energy of the neutrino increases. At high energies (around 300 GeV in the center-of-mass system of the projectile and the target) these two statements conflict; the two particles would have to interact in a space that is larger than the wavelength of the neutrino would allow, and more would result from the interaction than was put into it.

The present theory has survived as long as it has mainly because almost all our experimental knowledge and measurements come from the beta decay of radioactive nuclei, where the energy is low and the theory is adequate. Now it can be subjected to the test of interactions at higher energies, a test to which physicists interested in neutrinos have looked forward for years. At high enough energies the interactions must be fundamentally different.

Exactly how might they differ? All we know is that the correct theory must include some mechanism for damping the weak interaction at high energies so that it is no longer quite linearly related to energy. At the same time the theory must retain the correct results at low energies. The simplest and most tempting modification is to include a carrier of the weak force that has a large mass. The picture is analogous to the conception of the photon being the carrier of the electromagnetic force. The hypothetical carrier particle of the weak force (which must have a charge) is usually referred to as the W boson. The modification that the carrier-particle concept introduces is that rather than having particles interacting at a point, as they do in the Fermi theory, they now act at a distance. The scale for that distance is related inversely to the mass of the W boson.

Unfortunately the introduction of the W boson will not completely solve the problems of weak-interaction theory. Although the divergences in the theory now become logarithmic instead of linear, they still exist. It requires more than the inclusion of a W boson to have a correct theory of the weak interaction. In other words, if only a W boson is included, the problem is not solved but is merely put off to still higher energies.

The correct theory might well include a W boson, but it will certainly need something in addition to the carrier-particle concept.

As we study particle physics at the new accelerators the pursuit of the W boson is of major importance. In spite of the fact that discovery of a W boson would not in itself save weak-interaction theory, it would surely take us a giant step toward understanding the problem. If the particle exists, it could be produced in collisions of two protons, of a muon and a proton or of a neutrino and a proton. Neutrinos are the most attractive means of directly producing W bosons, since we can reliably estimate their cross section. Moreover, once a W boson is produced, it will be quite unstable and will decay immediately, leaving unique products that will mark its passing.

So far there is no direct evidence for the existence of a W boson. Searches with neutrino beams at Brookhaven and at CERN have set lower limits on the probable mass of the particle at approximately twice the mass of the proton, which has 1,839 times the mass of the electron. My collaborators and I at the National Accelerator Laboratory have since shown that the W boson can be no lighter than five times the mass of the proton. This result is to be expected; since the range of the interaction is inversely proportional to the W boson's mass, the greater the mass of the W boson, the smaller and more pointlike its range, as in the Fermi theory.

Eventually W bosons with 10 times the mass of the proton will be detectable with the NAL accelerator. If the W boson is still heavier, it cannot be observed by direct production. By bombarding hadrons with ordinary high-energy neutrinos, however, it should be possible to detect the effects of a W boson with 25 times the mass of the proton. If such a particle existed, the number of interactions would deviate from the number predicted by the linearly increasing Fermi cross section.

The W boson is only one new weak-interaction phenomenon that might materialize as we study the interactions of neutrinos and other particles at high energies. For example, recently there have been a number of attempts to theoretically unify the weak interaction and the electromagnetic interaction. These theories predict other new particles, such as heavy leptons, in addition to the electron and the muon that might well be observable in high-energy neutrino collisions. Preliminary results at NAL have shown that there are no heavy muons with a mass less than twice the mass of the proton. That is, to be sure, a negative result, but it is enlightening nonetheless.

It should be obvious that our knowledge is so meager there is not much restriction on the possibilities. With high-energy neutrino beams, however, it appears that the nature of the weak interaction is now particularly vulnerable to experimental attack. Perhaps we shall soon discover some of the keys to understanding this mysterious force.

What kinds of instrument are required to fulfill our dreams of making detailed studies of neutrino interactions? Briefly, we need intense sources of neutrinos that can be finely focused, very large and sensitive detectors to explore the interactions and massive shields to protect the detectors from background radiation. Over the past few years there has been much progress in meeting these requirements.

Probably the most impressive single development has been the successful construction and operation of enormous bubble chambers specifically designed for neutrino physics. Bubble chambers long ago proved their value in elementary-particle physics. Until now, however, they have mainly been used to study strong-interaction phenomena. The primary problem is: How can a few neutrinos be induced to interact with other particles in the bubble chamber? The problem has been solved by developing chambers with a very large volume. The largest of these chambers is the one at the National Accelerator Laboratory that I have mentioned. This chamber has a volume of 30,000 liters and can be filled with hydrogen, deuterium (heavy hydrogen) or neon, depending on the requirements of the experiment. The chamber is 15 feet long in the direction of the beam and the liquid is surrounded by a large superconducting magnet.

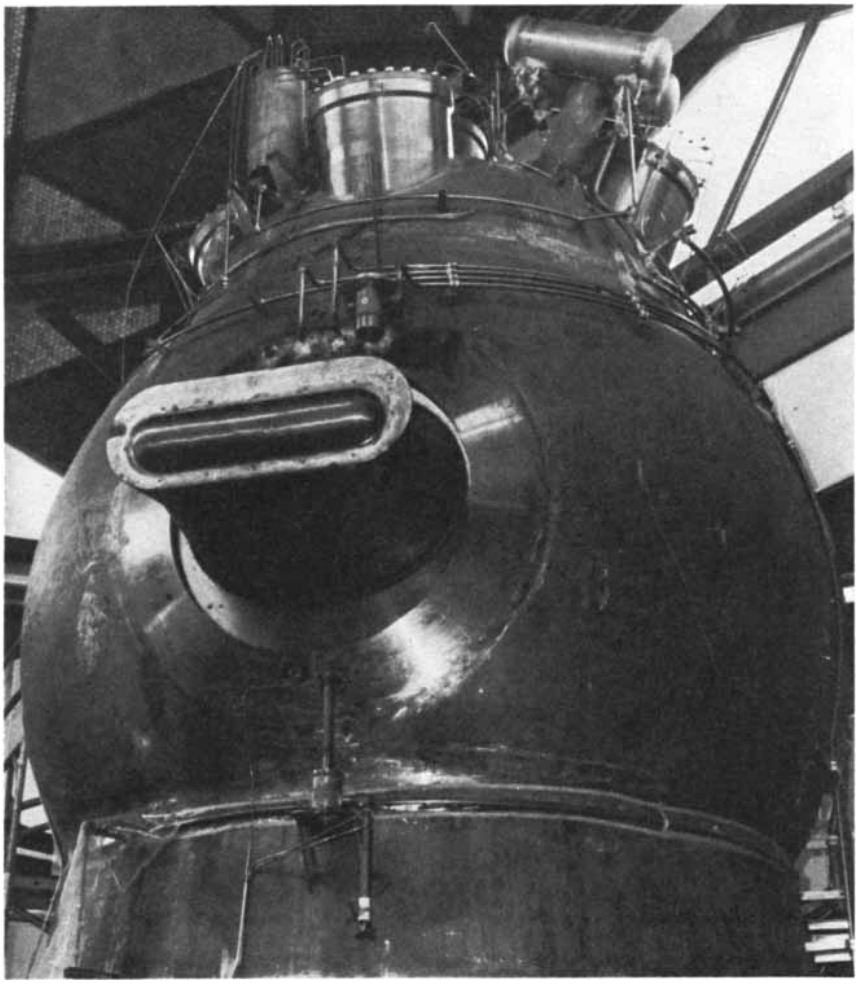
Another large bubble chamber, called "Gargamelle" after the mother of Gargantua in Rabelais's *Gargantua* and *Pantagruel,* is now in operation at CERN. A special neutrino-beam facility has been built for it at the 28-GeV CERN proton synchrotron. Much care has been taken to build an appropriate shield and to develop schemes for determining the neutrino flux and monitoring its intensity. The initial results emerging from Gargamelle are providing our first glimpse into the nature of neutrino interactions as we go to high energies.

Workers at CERN have found that the cross sections for both neutrinos and antineutrinos grow significantly as the neutrino energy is increased from 1 GeV to 10 GeV. Although accuracy is limited, the results are consistent with the linear increase predicted for the scattering of neutrinos by partonlike bodies within the target particles. The question that remains is whether or not the linear increase will continue up to 50 GeV, 100 GeV, 200 GeV or even higher energies. If and when we detect a deviation from this behavior, we shall know that we are starting to perceive new physical principles. The exciting experimental objective at that point will be to explore the reasons for the deviation.

Another striking result from the CERN experiment is that the cross section of the interactions involving antineutrinos is only about a third of the cross section of the interactions involving neutrinos. For all strongly interacting particles the cross sections for the particles and their antiparticles are approximately equal. The result for neutrinos and antineutrinos is consistent with a theory where these particles are pictured as being scattered mainly by pointlike particles within the proton that have a spin of $1/2\ \hbar$. The question is: Is this evidence for quarks or partons?

While these studies have been going on at CERN, a large bubble chamber filled with hydrogen or deuterium has been brought into operation at the Argonne National Laboratory. One consequence of developing intense neutrino beams and big bubble chambers is that we can now look forward to studying neutrino interactions in simple targets without possible complications from nuclear effects. One of the experiments at

NEUTRINO INTERACTION is visible in the photograph on the opposite page, which was made in the 12-foot bubble chamber at the Argonne National Laboratory. The neutrino, which leaves no track, enters the picture at the lower left. There it encounters a proton, giving rise to the dark three-pronged track in the bubble chamber. The short prong to the right was made by the proton. The longer prong to the left was made by a positive pion. The prong in the middle, which goes out of the picture, was made by a negative muon. Although the lines in the background give the picture the appearance of being a photograph mosaic, they mark segments in the sides and bottom of the bubble chamber. The spiral tracks were made by background electrons in the chamber.



FIFTEEN-FOOT BUBBLE CHAMBER at NAL is just being brought into operation this summer. It can be filled with 30,000 liters of hydrogen, deuterium or neon, depending on needs of the specific experiment. No photographs of neutrino events have yet been made.

Argonne is the observation of neutrinos scattered by neutrons and protons. Current theoretical ideas suggest that there is a fundamental connection between the electromagnetic structure of the proton, which has been studied for years by the scattering of electrons, and an analogous structure in weakly interacting particles. The new Argonne experiments will offer the first real information on such predictions.

With the emergence of neutrino physics at the large accelerators we should not overlook the fact that fundamental experiments are still being carried out at low energies. For example, at one of the Savannah River fission reactors a major effort has been mounted by Reines and his colleagues to detect the scattering of antineutrinos by electrons. Theoretical efforts have been made to relate this reaction to the known beta-decay processes. So far the experimental results are compatible with the theoretical expectations. Soon a meson-producing facility at the Los Alamos Scientific Laboratory will provide a very intense source of low-energy neutrinos associated with electrons. As the results from CERN, Argonne, Brookhaven, Savannah River, Los Alamos and the Serpukhov laboratory in the U.S.S.R. become available, we shall move into a new era in the study of the weak interaction where we shall no longer be solely dependent on information from the decay of weakly interacting particles.

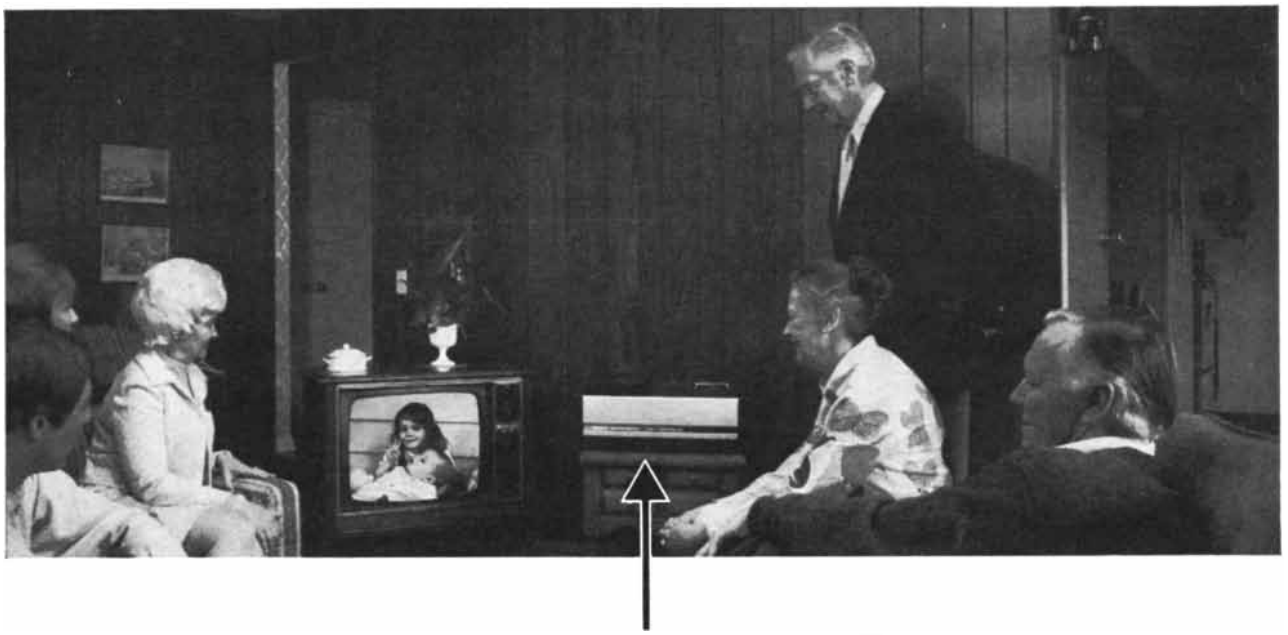
The most ambitious new instrument for neutrino physics is undoubtedly the NAL bubble chamber, with the neutrinos being supplied by the NAL accelerator. Bubble-chamber experiments, however, are only one way that neutrino physics is being approached at NAL. Two other neutrino experiments are in progress that in many ways are even more sensitive. One of these experiments has been mounted by our group from Cal Tech. The other is a collaboration among physicists from Harvard University, the University of Pennsylvania and the University of Wisconsin.

The Cal Tech experiment is somewhat unusual in that it involves the production of a neutrino beam in which the energy of the bombarding neutrinos is known. Usually neutrino experiments are done by measuring the energies of all the final products of an interaction and thereby determining the neutrino energy. In the Cal Tech experiment the neutrino interactions are observed in a 170-ton steel target with a system of scintillation detectors and spark chambers. The experiment is uniquely suited for measuring the total cross section of neutrino interactions.

A beam in which the energy of the neutrinos is known is also ideal for searching for exotic new particles such as heavy leptons and the *W* boson. Neutrino interactions were first detected in the Cal Tech apparatus last November. Now several hundred interactions of high-energy neutrinos have been recorded, and preliminary results are becoming available.

The Harvard-Pennsylvania-Wisconsin collaboration follows the more traditional "broad band" neutrino-beam concept, but it has emphasized the development of a very sensitive, high-resolution detector. The target-detector that has been installed at NAL consists of 100 tons of a liquid scintillator surrounded by a large number of image tubes. The energies of the products of an interaction are recorded when the particles streak through the liquid, generating flashes of light that are detected by the image tubes. Muons from the interaction are identified by penetration, and their momentum is measured by magnets. This experiment is particularly well suited for examining rare processes and the details of ordinary neutrino interactions. Neutrino interactions have also been observed in this apparatus.

In the 17 years since the neutrino was first observed as a free particle not only have many of its curious properties been determined but also it is rapidly becoming one of the most powerful tools in particle physics. Of course, the challenge for the future is for neutrinos to provide some of the answers to the questions and dilemmas I have mentioned. Will a *W* boson be found? Are there heavy leptons in addition to the electron and the muon? Will the neutrino cross section continue to rise linearly? The fundamental questions are seemingly endless. If the goal of answering them is achieved, it is more than likely that the information gained will present an entirely new set of questions and dilemmas for physicists to ponder. That is what physics is all about.



KODAK SUPERMATIC Film Videoplayer

Available late this year or early 1974. Show your own super 8 movies on your present TV set. Why bother setting up a screen and all that? We have a feeling that the output of movie cameras will be getting more use in the future.

Feeds into any receiver, color or black-and-white. Can also feed a whole roomful of receivers through a commercially available RF amplifier. Magnetic-striped sound or silent. Automatic threading. Accepts any of the four sizes of KODAK SUPERMATIC Cassette, or projection reels–50′, 100′, 200′, or 400′. Rewinds itself. Linger over a single frame if you like.

For a better illusion, the picture on the tube was pasted in after the cosy scene above was photographed. If we believed that every TV household will soon be buying a KODAK SUPERMATIC Film Videoplayer, we'd be illuding ourselves. But at around $1200, it's quite a bargain for institutions and business firms that want to educate and communicate through video cassettes. Now they will be able to draw on extant super 8 material and low-cost reporting of new developments and new visual viewpoints with any simple super 8 camera, up to and including the new ones that require no movie lights when loaded with the proper fast film. For the communicator—educator, scientist, salesman—the economy of super 8 stems from its interchangeability: home projection, study carrel projection, classroom projection, auditorium projection, and now TV viewing—private, closed circuit in industrial or school settings, CATV, or even broadcast.

It's because movie projectors can be expected to retain their popularity that our videoplayer for super 8 film makes economic sense.

Where we would look up the answers to your photographic questions

We at Kodak know quite a bit about the technical side of photography. At least many believe we do. So we receive a great many questions each day. It is very flattering. It is necessary to answer in terms appropriate to the questions.

"Technical" is one of the more flexible words in the language. It means one thing to a lady whose grandchild has presented her with a small camera for her very own and who now wishes to make a close-up portrait of her cat, Walter. It has another meaning for a specialist in solid-state theory who has heard, vaguely, that photographic manufacturers have ways to distribute lattice anomalies.

For one end of this semantic spectrum, extending roughly a third of the way, from our primer "How to Make Good Pictures" to the latest baring of deep mysteries in the preparation of four-color halftones, we offer a free guide to all the Kodak publications that can be purchased from photo dealers or from us. *(For a copy of this guide, request "L-5" from Kodak, Dept. 412-L, Rochester, N.Y. 14650.)*

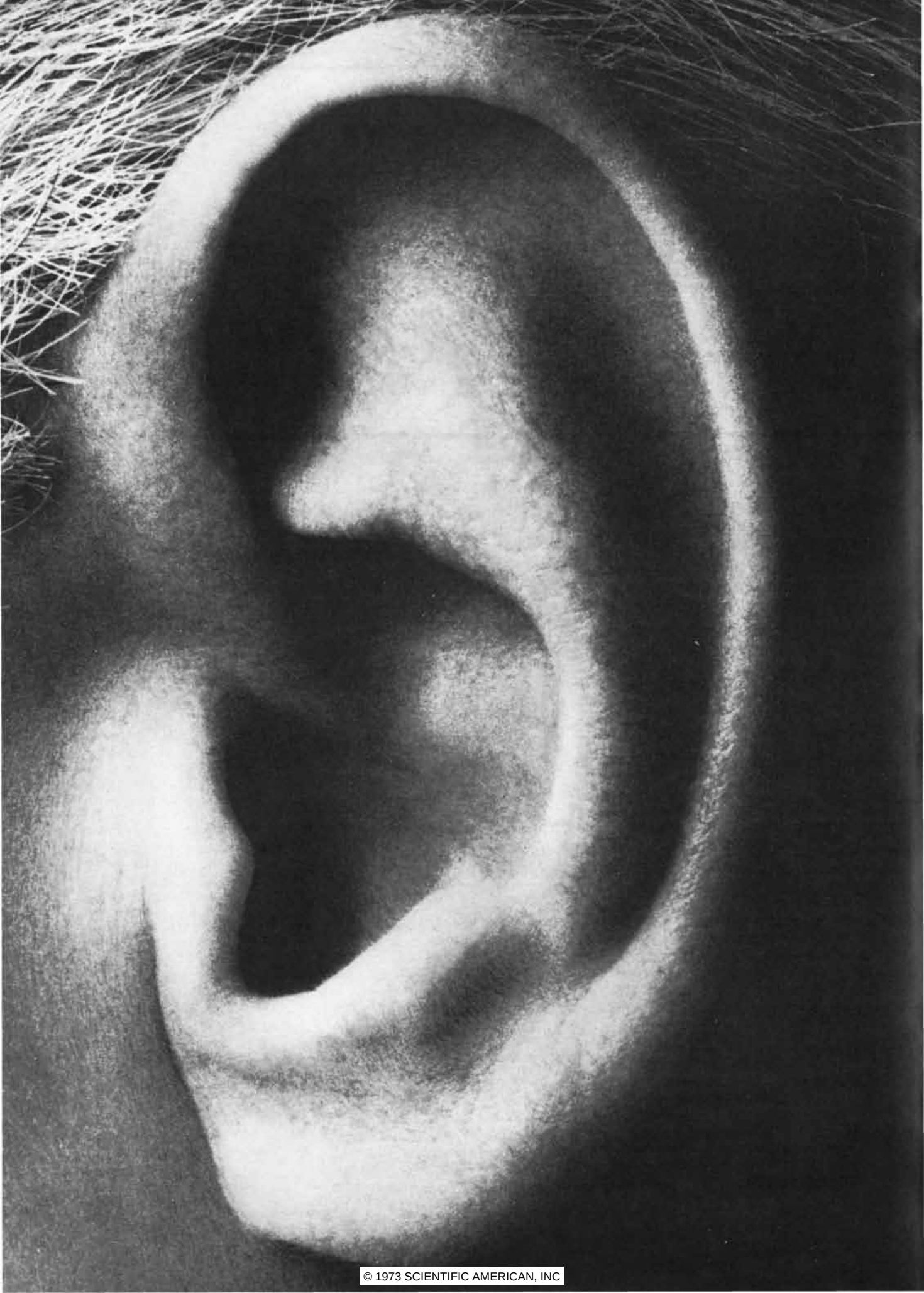
A classic work for the other end of the spectrum, born in the Kodak Research Laboratories but published by The Macmillan Company, is the 1966 Third Edition of "The Theory of the Photographic Process." It tells a great deal more than most people want to know, but there are some who can extract still more out of it by reading between the lines.

Now the gap in the middle has been filled by Wiley-Interscience for technical people (there's that word again) who want to do something photographic that may just possibly not have been done before and who need some background to appreciate the problem well enough to get down to the nitty-gritty with reasonable directness. Such is the purpose of a handbook. This one is entitled "SPSE [Society of Photographic Scientists and Engineers] Handbook of Photographic Science and Engineering." For this book, only about 50 of the hundred authors and editors were from Kodak. The others were from 24 organizations in the photographic community, not all commercial. The treatment of artificial radiation sources, for example, comes from a company better prepared than we to go into all the fine details, but as for fine details on the use of sun, moon, and sky for photographic lighting, there's a chap of our own who has spent 40 years gathering them. The Kodak man who made the book happen spent only eight years on the project.

1440 pages. Charts and tables in rich profusion. $37.50. Order from any bookseller. Not sold by Kodak.

Kodak

Channels clear



Sony introduces the latest technological advance in business communications.

The human potential.

The uniqueness of "me to you"

It's what's going to revolutionize the traditional, impersonal, typed or printed, mailed business communication.

With women's role in business changing rapidly, the stenographer is being phased out.

And sending business letters through the mails will one day be as archaic as the pony express.

Sony is helping that day arrive sooner.

Our technology is used to communicate through the mouth and ear, as well as the eye.

Our business machines are designed to pick up the subtlest shades of meaning in the human voice.

The nuances that are important because you hear exactly who said what (and can figure out why).

We developed a condenser microphone that'll catch even a whisper.

Automatic level control, so every voice sounds natural.

A wide range speed control, to help a secretary in transcribing.

And advanced circuitry (including hifi amplifiers) to give these machines their playback quality (the Sony sound).

In all our products, we have this desire to keep "the human touch."

Not only dictating/transcribing machines.

But in all the various and advanced word communications equipment we're working on.

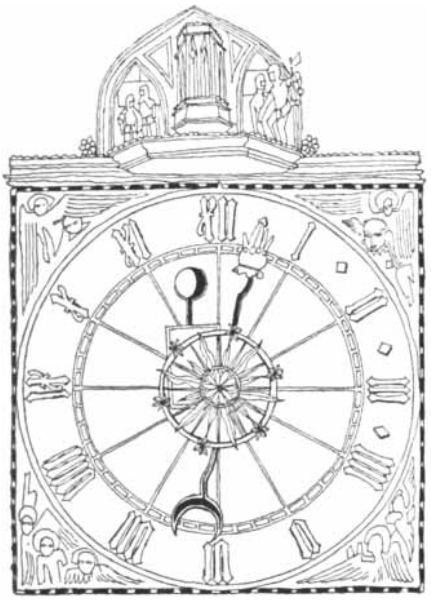
(Or even dreaming about.)

Maybe some day only machines will talk to people.

But we're still in favor of people talking to people.

SONY® keeps people in touch.

SCIENCE AND THE CITIZEN



Extending the Test Ban

The limited-test-ban treaty is 10 years old this month. In August, 1963, the U.S. and the U.S.S.R. signed a document prohibiting test explosions of nuclear weapons in the atmosphere, in space and under water. Within a few months it had been ratified by most other world powers. (The important exceptions are France and China.)

Excluded from that agreement were restrictions on the testing of weapons underground, and such tests have continued throughout the decade. Many ways of halting them have been proposed, but on this anniversary of the treaty the prospects for a comprehensive ban remain uncertain.

A resolution now before the Senate (S.R. 67) would call on the President to announce the immediate suspension of all U.S. underground tests, with the interruption to continue for as long as the U.S.S.R. also abstains. It would further urge that the President propose to the U.S.S.R. a new treaty that would permanently ban all nuclear testing. All other nations would be invited to join in the agreement.

Since the 1950's the obstacle in the way of a complete test ban has been the purported difficulty of detecting clandestine underground tests. The U.S.S.R. has been reluctant to allow on-site inspections, and the U.S. has been unwilling to trust any other means of verification. Advances in seismology and other methods of remote detection may have eliminated this obstruction.

W. Averell Harriman, in a *New York Times* article adapted from testimony given before the subcommittee on disarmament of the Senate Foreign Relations Committee, concludes that the limitations of technical means of verification should no longer hinder the effort to negotiate a test ban. "The risk of danger to U.S. security interests by clandestine Soviet underground tests is very limited," Harriman said. "Any test that might escape detection and identification would be quite small . . . and of relatively little importance in its possible effect on the strategic balance. Even with respect to tests of this size, there is sufficient uncertainty so that a potential evader could never be sure that any individual test would not be detected and identified." Harriman was the principal U.S. negotiator of the 1963 treaty.

A test ban enforced solely by technical means of verification would have a precedent in the 1972 strategic-arms-limitation agreements, which limit various categories of offensive and defensive weapons without a requirement for on-site inspections. The SALT accords further provide that neither side shall interfere with the other's detection methods, and that a commission shall be set up to consider any suspected violation.

If the verification issue were to be settled by negotiation, other impediments could arise. One might be the Plowshare program for the peaceful uses of nuclear explosives; another could be the difficulty of securing the compliance of nations that have so far refused to sign test-ban treaties.

Neither of these issues would impede a serious effort to stop underground testing, advocates of arms control believe. According to a report prepared by Senator Philip A. Hart and others, possible objections to the Plowshare program could be resolved by making special provisions for the inspection of Plowshare projects. If this is not possible, the report continues, the program should be abandoned before it is allowed to interfere with a possible test ban.

Regarding those nations that have refused to cooperate with the current partial test ban, Harriman and others have pointed out that the abstention of the two major powers from further testing would help, not hinder, efforts to limit the proliferation of nuclear arms.

A complete test ban is considered particularly important now because the SALT agreements regulate the number of weapons that may be deployed but not the nature of the weapons themselves. The testing of warheads thus remains uncontrolled. In his Senate statement Harriman noted: "A comprehensive test ban is the most immediate way to further reduce the dangerous and costly nuclear arms race. With the will of both sides it could be achieved promptly."

Comeback

Although there is still room for deep concern about the long-range prospects of the great whales, it appears that populations of certain species are now increasing. Some of these species have been protected against hunting but others are gaining in spite of it. Less than a decade ago it was estimated that of a prewhaling population of perhaps 425,000 fin whales (*Balaenoptera physalus*) only 36,000 still survived, and that of an original population of 210,000 sei whales (*B. borealis*) only 47,000 remained. Hunting of both species has continued in the years since these estimates were made, yet the fin and sei populations are now respectively estimated at 85,000 and 100,000.

The current estimates have been made by N. A. Mackintosh of the British National Institute of Oceanography. Writing in *Science Progress*, Mackintosh also gives estimates of the prewhaling and present populations of two protected baleen whales: the blue whale (*Balaenoptera musculus*) and the humpback whale (*Megaptera novae-angliae*), which have not been hunted since the mid-1960's. There were probably never more than 150,000 blue whales and 100,000 humpbacks; their present numbers are respectively about 6,000 and between 3,000 and 4,000. Mackintosh also estimates the prewhaling population of the sperm whale (*Physeter catodon*), the species killed today in greater numbers than any other. The original sperm population was 1.5 million; they thus once outnumbered all the large baleen-whale species combined and probably still do. Because sperm whales are taken in widely separated localities (by catcher vessels in the North Pacific and the Antarctic

and by shore-based whalemen off Japan, the Azores, Chile, Peru, Natal and Western Australia) no firm estimate of their present numbers is possible.

The severe reduction in the population of large baleen whales that followed the introduction of factory ships and fleets of catchers to Antarctic waters after World War II has affected the ocean ecosystem of the region. Mackintosh calculates that before intensive whaling began the large whales annually consumed some 150 million tons of krill, the shrimplike animals that are their principal food. As the biomass represented by the large whales dwindled from some 33 million tons to five million, the tonnage of "surplus" krill increased correspondingly. The beneficiaries of the increase, Mackintosh suggests, have been the krill-eating birds, fishes and squids of the region, along with two sea mammals: the crabeater seal (*Lobodon carcinophaga*) and a small baleen whale, the minke whale (*Balaenoptera acutorostrata*). Heretofore whalers in the Antarctic have rarely pursued the minke whale; up to 1971 the catch was fewer than 100 per year. In the 1971–1972 whaling season, however, they killed at least 3,000 of an estimated minke-whale population of some 200,000. This number is well below the maximum sustainable yield for such a large stock. Nonetheless, the danger exists that intensive hunting of sperm and minke whales could eventually reduce both species to the level of severe endangerment.

Enrichment Scheme

Since the earliest days of nuclear technology the U.S. has employed the gaseous-diffusion process for large-scale enrichment of natural uranium in the fissionable isotope uranium 235. Now, foreseeing that additional enrichment capacity will be needed by the early 1980's, the Atomic Energy Commission is exploring and urging private industry to explore the gas-centrifuge method as an alternative. Although much of the work remains classified, a general statement of where the matter stands was presented to the Joint Committee on Atomic Energy by F. P. Baranowski, director of production and materials management of the AEC.

For both the diffusion and the centrifuge processes the feed material is uranium hexafluoride. In diffusion the gas is blown through a series of thin, porous barriers; at each barrier the U-235 atoms, being lighter and therefore faster, get through more rapidly than the U-238 atoms, so that successively richer concentrations of U-235 are segregated. In centrifugation the isotopes are separated on the basis of the difference in their mass. A large plant based on the centrifuge process could be expected to show much lower costs for power than a gaseous-diffusion plant, but it remains to be seen whether various other operating costs would offset the advantage.

Baranowski told the committee that the AEC has set the fiscal year 1976 as the "target date for assessing the economic competitiveness of the gas-centrifuge process in comparison to the established gaseous-diffusion process." To that end the commission has intensified its own study of the centrifuge process and has accepted seven proposals from industry to do classified development work on the process. The commission's expectation is that private industry will provide the additional enriching capacity in privately owned plants.

Cross-eyed Tigers

There is a well-known family of genetic anomalies in which some form of albinism is combined with visual abnormalities. The Siamese cat is a prime exemplar. It is light-colored except for its paws and the tips of its tail, nose and ears (because the enzyme deficiency involved is temperature-sensitive, and the extremities are cooler). It is often cross-eyed, and even when it is not, the visual pathway is abnormal, with some optic-nerve fibers going to the wrong side of the brain. The pathway error is reflected in a derangement of the layers of cells in the lateral geniculate nucleus, a visual center in the brain. The same visual abnormality has been found to be associated with albinism in a number of other mammals—most recently in the so-called white tiger.

These great cats, which are seen from time to time in the wild in India, have blue eyes and an off-white stripe, instead of the normal yellow one, alternating with a gray-brown stripe. (The white tends to be darkened in cool environments, suggesting the same kind of temperature-dependence effect as the one found in the Siamese cat.) Some descendants of a captured white tiger have been bred at the National Zoological Park in Washington, D.C., and one of them, Rewati by name, is markedly cross-eyed. That interested R. W. Guillery and J. H. Kaas, who had been studying the Siamese cat's visual pathways at the University of Wisconsin School of Medicine. It seemed impractical to investigate optic-nerve endings with microelectrodes in a living tiger, as one does in small cats, but when a younger brother of Rewati died, Guillery and Kaas were able to study its brain. They found the same disruption of the layers of the lateral geniculate nucleus that they had seen in Siamese cats.

The white tiger not only is one more species exhibiting the albino abnormality, Guillery and Kaas point out in *Science,* but also is the most highly pigmented species in which it has been found. Is the visual abnormality directly related to pigment formation, and if so, what is the relation? Is it possible that in man too some forms of the cross-eye syndrome are produced by a similar misdirection of nerve fibers associated with reduced pigment formation, either in the body in general or in the eye itself?

The Greenhouse of Titan

Titan, the largest and brightest of Saturn's nine satellites, is the only moon in the solar system known to have an atmosphere. Analysis of the absorption bands in the spectrum of the sunlight reflected from it indicates the presence of methane and hydrogen, but these gases cannot account for the remarkable color of the object, which is as red as Mars. Recent measurements of the polarization of the reflected light made by J. Veverka of Cornell University and Ben Zellner of the University of Arizona suggest that Titan is completely covered by opaque clouds. The unusual red color could be due to the absorption of blue light and ultraviolet radiation within the cloud.

Such an atmosphere, according to the calculations of Carl Sagan of Cornell University, could produce a "greenhouse" effect that would trap heat and raise the surface temperature of the satellite. Assuming that the atmosphere of Titan is primarily hydrogen, Sagan calculates that the satellite's surface temperature could be as high as 200 degrees Kelvin (−100 degrees Fahrenheit). If methane and ammonia were also present, surface temperatures of up to 260 degrees K. (nine degrees F.) would be possible. The atmospheric pressure on Titan's surface is calculated to be between a tenth and nine-tenths of the earth's atmospheric pressure at sea level.

The density of Titan is low. To account for this fact it has been suggested that several dozen kilometers below the satellite's surface there exists a slush of methane, ammonia and water. If that is the case, volcanism and outgassing could be expected. It is well known, says Sagan, that an electrical discharge or ultraviolet radiation acts on a gaseous mixture of methane, ammonia and water to

produce hydrogen and complex organic molecules, some of which are bright red. Organic molecules resulting from such reactions are sugars, amino acids, purines and pyrimidines, all of which are basic building blocks of living organisms.

"Thus," writes Sagan in *Icarus,* "the picture which emerges of Titan is of a place with an atmospheric density approaching that of Earth, with surface temperatures of at least 150°K and possibly as high as 200°K, and with abundant organic compounds in the clouds, atmosphere and surface. Biology under these circumstances is by no means out of the question." He suggests that Titan is an ideal target for an unmanned or manned landing.

Mutation and Atherosclerosis

The main cellular component of the fibrous plaques that are the characteristic lesions in atherosclerosis is the smooth-muscle cells of the artery wall. It is their proliferation, together with the buildup of fatty debris, clotted blood and connective tissue, that can eventually close down an artery, causing disability or death. The big unanswered question is: What causes the plaque to form? Is it penetration of the artery wall by fatty acids and cholesterol, formation of a clot on the inner surface of the wall or some other kind of injury to the wall? A father-and-son team at the University of Washington School of Medicine suggests that mutation of the smooth-muscle cells may be implicated and that it might be due to some chemical agent or a virus.

Earl P. Benditt and John M. Benditt reasoned from the fact that the cells of atherosclerotic plaques differ in size and other characteristics from the cells in normal artery walls. They wondered why that should be. Are the plaque cells transformed cells, in which the genetic material has somehow become changed from that of the normal cell population? If so, the plaque cells could be expected to be monoclonal: a homogeneous population derived from a single cell, rather than a mixed population of cells as in normal human tissue. The tool for investigating the question was glucose-6-phosphate dehydrogenase, an enzyme that is encoded by a gene on the *X* chromosome. Every female cell has two *X* chromosomes, one from the father and one from the mother, but only one of them is expressed; the choice is made at random during embryonic development. Every woman is thus a "mosaic" of two cell types. The enzyme exists in two forms, called *A* and *B,* and normal tissue from a woman who is heterozygous for the enzyme (that is, whose two *X* chromosomes code for different forms) can be shown to contain a mixture of the *A* and *B* enzymes.

In *Proceedings of the National Academy of Sciences* the Benditts tell how they analyzed the glucose-6-phosphate dehydrogenase from normal tissue and from atherosclerotic plaques in arteries taken from four women at autopsy. Normal cells from even very small arterial samples regularly yielded a mixture of the two enzymes. The cells from large plaques, on the other hand, produced solely or predominantly one enzyme type.

If plaques were a simple response to an injury of some kind, as has been proposed, their cells should be polyclonal, the Benditts point out. The fact that they are monoclonal suggests that there is more to it than response to injury, that the plaques are largely growths that originate from a single cell, something like a wart or a tumor. The investigators propose that at some point a single cell somehow develops a selectional advantage through a change in its genetic material. Both the probability and the expression of such changes are enhanced by factors that stimulate cell proliferation—and increased levels of blood lipids and hypertension may stimulate proliferation in artery walls. The authors conclude that in the search for an initiating factor in atherosclerosis "consideration should be given to factors that transform cells, such as chemical mutagens or viruses, or both."

Baby Fat

In many sociocultural groups a plump, "healthy looking" baby has long been regarded as a sign of good mothering. Recent research indicates, however, that hypernutrition in the first year of life may lead to the onset of obesity in adolescence or early adulthood. Several investigations have shown that the adipose, or fat-storage, tissues of the body appear to have a critical period of development during the first year, in which the size and the number of fat cells in the body are established for life.

Genetic factors appear to play a role, since a high percentage of obese people have parents and siblings who are also obese. Studies of twins reared in different families, however, indicate that heredity is not the sole determinant. Indeed, one study shows that the rate of weight gain during the critical early months of life is a better indicator than parental size of obesity in later life.

Reviewing the problem in *American Journal of Diseases of Children,* S. L. Hammar suggests that the way to control obesity is prevention. Parents should be counseled not to overfeed their babies and not to introduce solid foods too early. Hammar also notes that if weight loss is used as the criterion, there is little evidence that current methods for the treatment of obesity in adolescents produce better long-term results than no therapy at all. For some, weight loss may be impossible.

"For these individuals," Hammar remarks, "obesity should be considered a handicapping condition like epilepsy or diabetes, with which the patient must learn to live. Efforts should be directed toward their social rehabilitation... [rather] than repeated unsuccessful attempts at weight reduction."

Feathered Lizards

It is commonly believed that the dinosaurs, the best-known of the reptiles that dominated the Mesozoic era, left no direct descendants when the last of their kind became extinct some 70 million years ago. Not so, says John H. Ostrom, a paleontologist at Yale University. In Ostrom's view every feathered animal in the world today, from the ptarmigans of the Arctic tundra to the penguins of the Antarctic ice, is a lineal descendant of the coelurosaurs: an infraorder of small, bipedal, carnivorous saurischian dinosaurs that flourished from late in Triassic times, some 200 million years ago, until the end of the Cretaceous.

Writing in *Nature,* Ostrom points out that a dinosaurian ancestry for birds has been proposed before but that some 50 years ago most scholars dismissed the notion because the coelurosaurs did not seem primitive enough to play the role of an ancestor. Over the past two years, however, Ostrom has closely scrutinized the only four existing fossils of *Archaeopteryx,* the earliest bird in the fossil record, in the course of preparing a publication on the ancestry of birds. His conclusion is that if it were not for the fact that clear imprints of feathers accompany the four skeletons, all four specimens would probably have been identified as coelurosaurs.

Ostrom suggests that the reason the birds descended from *Archaeopteryx* did not follow their dinosaur ancestors into extinction at the end of the Mesozoic was that their feathers gave them insulation from extreme heat and cold of a kind that was not available to their bare-skinned kin. Can it be a coincidence, Ostrom asks, that mammals evolved a comparable system of body-temperature

The stuff he's blasting out of could turn into a pickle jar overnight.

Probably no duffer has ever stopped to think about it, but every time he's been in a trap, he's also been in what glass containers are made from.

It's sand. The most common substance in the earth's crust. And the natural material that makes up about 73% of the jars and bottles for your favorite foods and beverages.

Because glass is a natural product, homemakers love it. Glass lets them see what's inside. They like the purity of glass, its variety of sizes, its re-sealability. And the fact that glass protects the flavor.

It's the naturalness of glass that also makes it an ideal environmental package. Glass can be returned to the earth in virtually its original form. It can also be recycled into new glass containers. And new uses for waste glass are continually being developed by the industry—uses such as road paving material, construction panels, bricks, insulation, terrazzo, and reflective paints.

Given its natural ingredients, its packaging benefits and its environmental virtues, it's no wonder that consumer preference for glass containers continues to grow. For a copy of "The glass container story," write: Glass Container Manufacturers Institute, Dept. D, 1800 K St., N.W., Washington, D.C. 20006.

Glass. It's a natural.

Stuck for a week-end in Frankfurt? Everybody should be so lucky.

We wouldn't worry about having to spend a week-end in Frankfurt if business demanded it. First of all, there are theatres to take business off your mind. And other entertainment. Then there's Heidelberg just around the corner, as are a number of Europe's most picturesque towns, each with its own attractions. From mediaeval castles and churches to more contemporary establishments.
And, if it should happen to you in the Fall, you'll be in time for wine-tasting festivities in 93 towns and villages. These you will definitely enjoy.
But don't say we didn't warn you.

Lufthansa
the more you fly

control that enabled them to outlive their reptilian ancestors at about the same time dinosaurs first grew feathers?

The Healthy Crowd

It has long been surmised that living in crowded conditions has a bad effect on human health and behavior. This concept received strong support from the work of John B. Calhoun of the National Institute of Mental Health, who observed such effects in extremely crowded colonies of laboratory rats [see "Population Density and Social Pathology," by John B. Calhoun; SCIENTIFIC AMERICAN, February, 1962]. Now Robert M. Factor and Ingrid Waldron of the University of Pennsylvania have conducted two surveys that lead them to the conclusion that "density . . . does not have large effects on human health." The surveys encompassed 10 neighborhoods in Chicago with comparable socioeconomic conditions but different population densities and several pairs of countries that were similarly matched.

Writing in *Nature*, the authors describe the 10 "internally homogeneous community areas" in Chicago as ranging in population density from 11,800 to 19,500 per square mile. In the second survey the pairs of countries were matched for per capita gross national product and for health-care facilities, and in each pair one country was more than twice as densely populated as the other and also more highly urbanized. In both surveys the authors examined mortality rates for the respective populations and found little correlation between density and high mortality. In the absence of such a correlation, they write, "it seems probable that density per se does not have a large effect on physical health."

Dada Processing

The influence of the digital computer has spread not only throughout commerce, technology and science but also to the generally unmechanized realm of the literary scholar. Concordances and bibliographies are now commonly compiled electronically. Word-frequency tables and other statistical studies are also made much more quickly by computer than by hand.

Rarely, however, have the data-handling capabilities of computers been applied to one of the central tasks of the critic: the divining of meaning in a work of literature. J. M. Coetzee, a lecturer in English at the University of Cape Town in South Africa, found a Univac 1106 computer useful in the interpretation of a work whose meaning on first reading may seem quite elusive: *Lessness*, by Samuel Beckett. Coetzee describes his machine-assisted investigation in the journal *Computers and the Humanities*.

Lessness is an English translation of *Sans*, a work published in French in 1969, the year Beckett was awarded a Nobel prize for literature. "The story line is pretty minimal," Coetzee says. It seems to involve a naked figure experiencing day and night and nothing else. The conventions, and even the grammar and punctuation, of narrative prose are absent.

Lessness consists of 120 sentences, the second 60 repeating the first 60 in a new arrangement. Coetzee found it suitable for mathematical treatment because of its "rule-governedness" and because it could be considered "as linguistic game rather than linguistic expression."

He first instructed the computer to search for any ordering principle in the rearrangement of sentences in the second half; none was found. The arrangement of paragraphs was also found to have no significance.

Turning to the finer structure of the work, Coetzee used a computer algorithm to break each sentence into "phrases": lexical units that within the work are indivisible. He found 106 phrases of from one to 12 words each. For example, the sentence "Flatness endless little body only upright same grey all sides earth sky body ruins" was divided into nine phrases.

Coetzee then investigated the distribution and ordering of the phrases. He derived two rules. First, there are no closed subsets of phrases, that is, there are no sets of phrases that always appear together in a sentence. The second rule states that, with a few exceptions, any sentence in *Lessness* can be constructed from the elements of no more than three other sentences, and most can be formed from just two others.

From these two rules and from other results of his analysis, Coetzee was able to reach certain conclusions about the nature of *Lessness*. "There are no determinate principles of ordering among phrases, sentences or paragraphs, yet . . . all are interdependent and connected," he says. Because the computer program demonstrated that, syntactically if not semantically, the structure of *Lessness* is random, he concludes: "The residue of the fiction is then *not* the final disposition of the fragments but the motions of the consciousness that disposes them according to the rules we have traced, and no doubt to others we have failed to trace."

METAL-OXIDE-SEMICONDUCTOR TECHNOLOGY

The MOS fabrication technique can put more than 10,000 electronic components on a silicon "chip" only a few millimeters across. Its economic impact can be seen in such devices as pocket calculators

by William C. Hittinger

Last December the electronics industry celebrated the 25th anniversary of the transistor, the solid-state successor to the vacuum tube. This revolutionary invention and its continuously evolving offspring have paced the penetration of electronics into almost every sphere of human activity. As late as 1960 transistors were still packaged in individual containers varying upward in size from that of the eraser on a pencil and costing an average of about $1 each. To construct a complete electronic circuit, whether for a portable radio or for a computer, individual transistors had to be linked together with various other components, such as resistors, capacitors and diodes. About 1960 methods were developed to combine most of the components of a circuit on a single-crystal wafer of silicon to form a miniature "integrated circuit." The first integrated circuits consisted of about a dozen components on a "chip" measuring a few millimeters on a side. Today many mass-produced integrated circuits consist of more than 3,000 components, chiefly transistors, on chips only slightly larger, and the most advanced circuits contain upward of 10,000 components. It is not unreasonable to expect that by 1980 it will be feasible to build integrated circuits made up of a million transistors and associated components.

The highest component densities are achieved by a fabrication technology known as MOS, which stands for metal-oxide-semiconductor. The MOS technology produces transistors of the unipolar type in contradistinction to earlier junction transistors, which are bipolar. Bipolar transistors use both electrons and "holes" (the absence of electrons) as carriers of electric charge; unipolar transistors use either electrons or holes, but not both. Fewer processing steps and a set of smaller minimum dimensions are needed for integrated circuits consisting of unipolar transistors than for circuits consisting of bipolar transistors. Most integrated circuits produced in the 1960's were of the bipolar type, but production of the newer unipolar type is growing rapidly.

The broad spectrum of electronic equipment and services available today, ranging from entertainment products to communication by satellite, is largely made possible by the decreasing cost per electronic function that is in turn made possible by integrated circuits. In circuits consisting of more than 200 transistors (once regarded as "large-scale integration," or LSI), the cost of each transistor is now no more than one cent. By the end of the decade the cost per transistor in integrated circuits could well drop by a factor of 30 or so, at which point a transistor will be about as cheap as a word printed on the page of a hardcover book [*see top illustration on page 50*]. Although small size and weight and increased reliability have contributed to the attractiveness of integrated circuits, they have been less important factors than the dramatic reduction in cost per function.

The key to low cost has been the economy achieved by batch fabrication. Integrated circuits are made by simultaneous processing of many wafers of single crystals of silicon, each bearing hundreds of integrated circuits, through a series of chemical and metallurgical steps. In spite of all efforts to achieve absolute uniformity in the crystal wafers and in their subsequent processing, there are inevitable variations in the properties of the structure being fabricated. Critical dimensions are commonly measured in micrometers and sometimes in angstroms, or tenths of a micrometer. Circuit malfunctions can result when these critical dimensions are not met, for example when there is a localized defect in the structure of the silicon crystal or particulate matter on the surface of the crystal. As a result not every circuit in a batch of silicon being processed will pass inspection. The yield of good circuits can be as low as a few percent in the early stages of manufacturing an advanced circuit, but the yield usually increases rapidly with experience. Over the years there has been a steady improvement in process control, achieved by new equipment and measuring techniques, which has made it possible to increase the density and complexity of the circuits. At any given time a practical limit is reached when the economy achieved by adding more functions per circuit is offset by the reduction in yield.

Transistor Origins

Since the history of the transistor is still not well known, let me recount it briefly. The building block of an MOS integrated circuit is the field-effect transistor, which is based on the control of lateral conductance within a silicon crystal by an electric field applied at right angles to the surface of the silicon. The effect was recognized as early as 1930 by Julius Edgar Lilienfeld, who was issued a patent for field-effect devices in 1935. Because the physical properties of surfaces and thin films were poorly understood at the time, little was done with the field-effect concept and interest waned in the face of the rapid development of vacuum tubes.

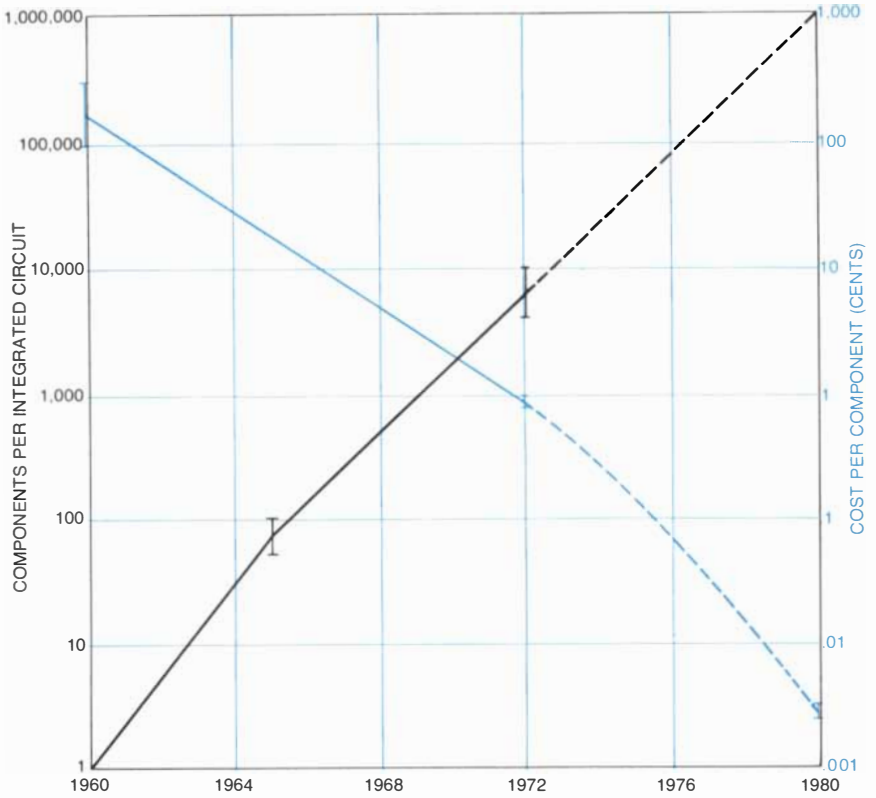
In the late 1930's a young physicist at Bell Laboratories, William B. Shockley, became interested in the possibility of developing a solid-state device of some kind to provide an electronic substitute for the electromechanical switches that were universally used in telephone exchanges to connect one telephone with

another. He and others believed that electronic telephone switching would eventually be needed and that vacuum tubes would probably be too costly and unreliable for the task.

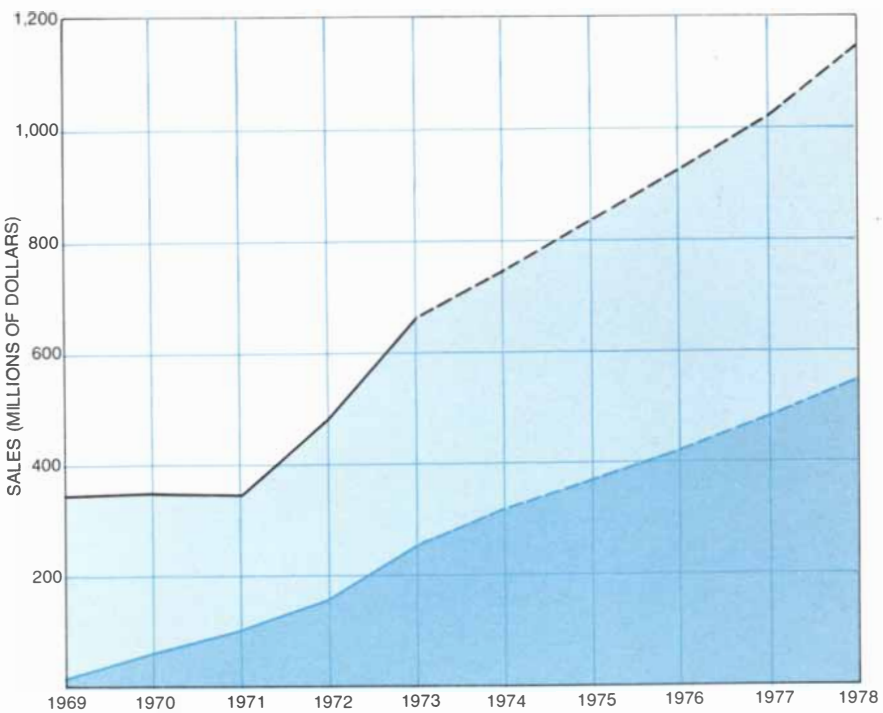
Shockley saw promise in a theory that had been proposed by Walter Schottky in studies of rectification (converting an alternating current into a direct current) at the interface between a metal and a semiconductor. Specifically Shockley foresaw the possibility of an amplifying action in the space-charge, or depletion, layer that would spread from the metal into the semiconductor. He believed that the spreading of this layer would constitute a kind of valve action or control of conductivity in the semiconductor at a substantial distance from the contact, thereby controlling the flow of current as a three-element vacuum tube does. In 1939 he tried building valvelike devices with a combination of copper and copper oxide, but he was unsuccessful.

After World War II Shockley returned to the problem at Bell Laboratories. He and John Bardeen and Walter H. Brattain began studying field-effect amplification in germanium, a semiconductor that by then was better understood than the combination of copper and copper oxide. Their studies of surface-contact potentials and the space-charge layer led in 1947 to the invention of the point-contact transistor. Although the point-contact transistor was difficult to produce in large volume, it was vital in establishing the new technology of semiconductors. In a few years the original device was succeeded by the junction transistor, which was easier to manufacture. Invented by Shockley in 1948, the

ELECTRONIC WATCH CIRCUIT CHIP with metal-oxide-semiconductor (MOS) transistors controls the light-emitting diodes on the face of the watch that give the time in numbers. The input to the circuit is from a quartz-crystal oscillator. The complementary-MOS, integrated-circuit chip shown here measures .149 by .150 inch and is made by RCA Solid State for HMW Industries, Inc.



DENSITY AND COST of integrated-circuit components are plotted for the period 1960–1980. The first rudimentary integrated circuits were made in 1960 and cost $1 or more each. As fabrication skills developed, the component content (mostly transistors) increased rapidly with a consequent decrease in cost per component. Up to 1970 the data represent bipolar integrated circuits. Since then unipolar MOS devices have led both in density and in cost reduction per component. By the early 1980's it is expected that circuits with a density of one million components will be produced at a cost of about .003 cent per component.



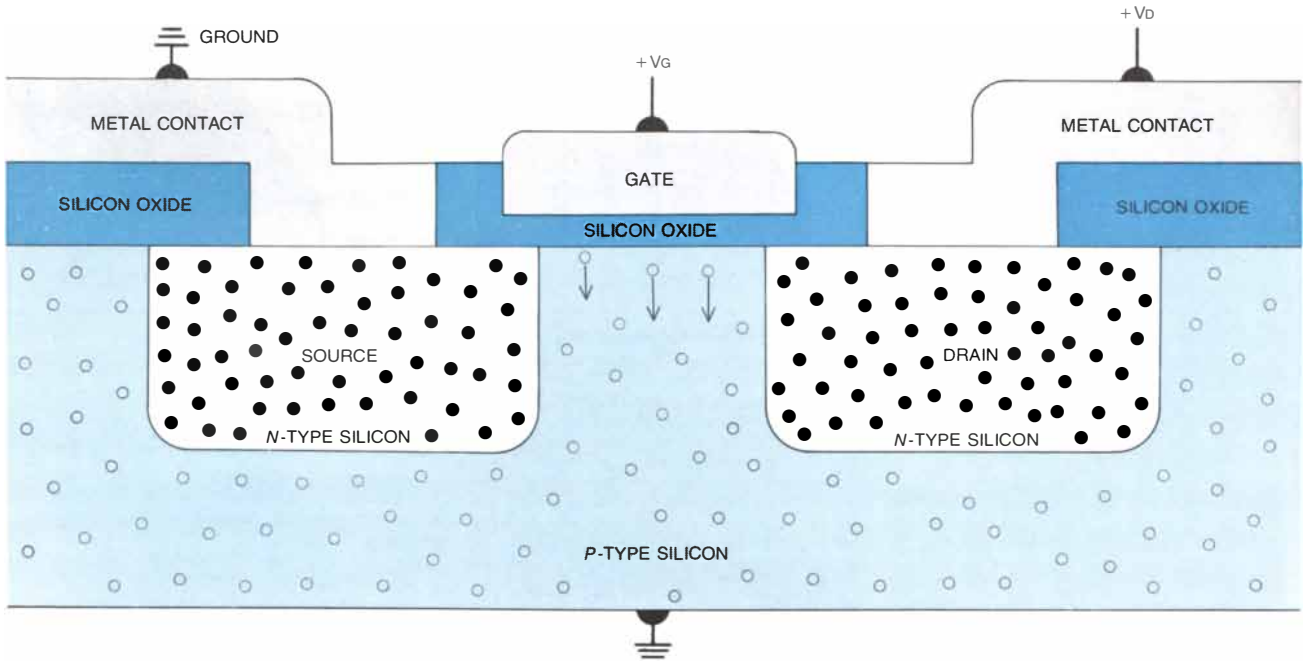
TOTAL FACTORY SALES of digital integrated circuits in the U.S. are shown from 1969 and projected to 1978. The dark colored area depicts the market of MOS circuits and the light colored area the share of bipolar devices. In 1973 sales of MOS devices will reach $250 million and of bipolar devices about $400 million. By 1975 the total market (outside of the socialist countries) for integrated circuits is expected to be twice as large as the U.S. market.

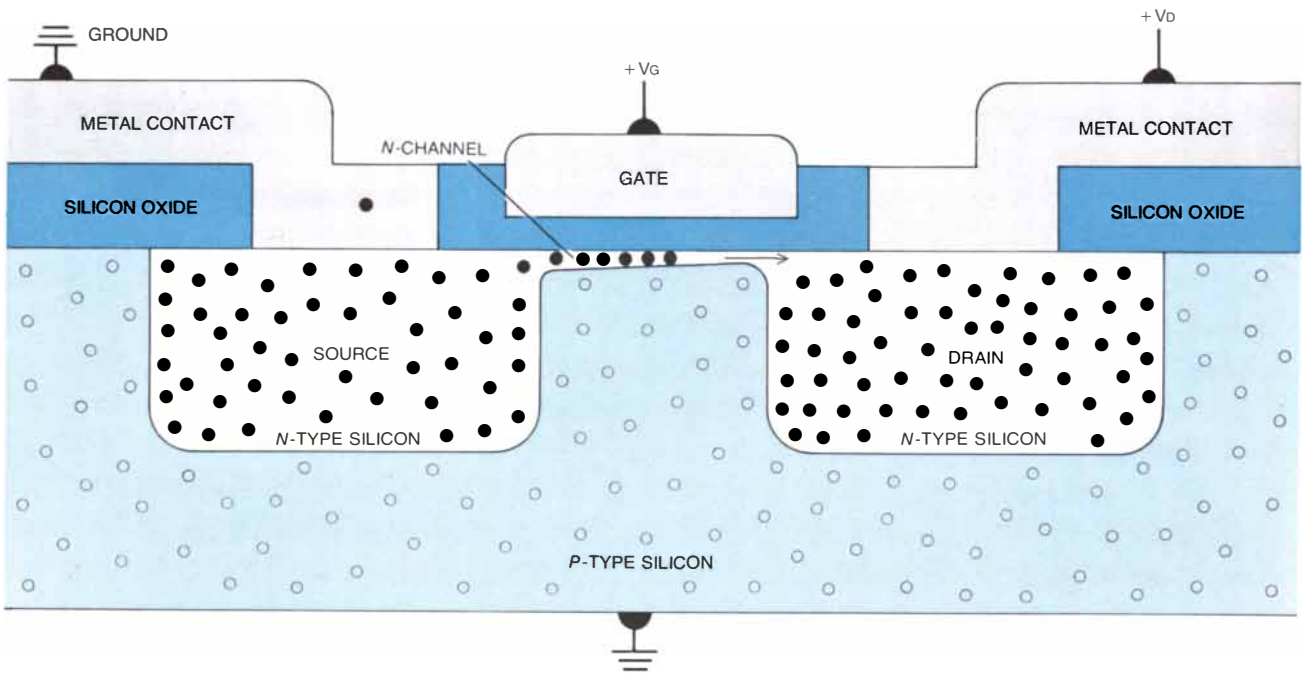
junction transistor, or bipolar transistor, consists of two junctions formed within the body of a single crystal of silicon. The two junctions separate three regions called the emitter, the base and the collector. The flow of current from emitter to collector is controlled by changes in a weak signal applied to the base.

Interest in the field effect remained high under Shockley's leadership. He and Gerald L. Pearson at Bell Laboratories observed field-effect behavior in 1948 with the structure in silicon called the *p-n* junction. In 1952 Shockley published the theory of the field-effect transistor, and a practical form of the device was built by George C. Dacey and Ian M. Ross in 1953. This device used an electric field to control conduction in a germanium semiconductor structure. Because it was costly to make and offered few advantages over the junction transistor, it was limited to a few specialized applications.

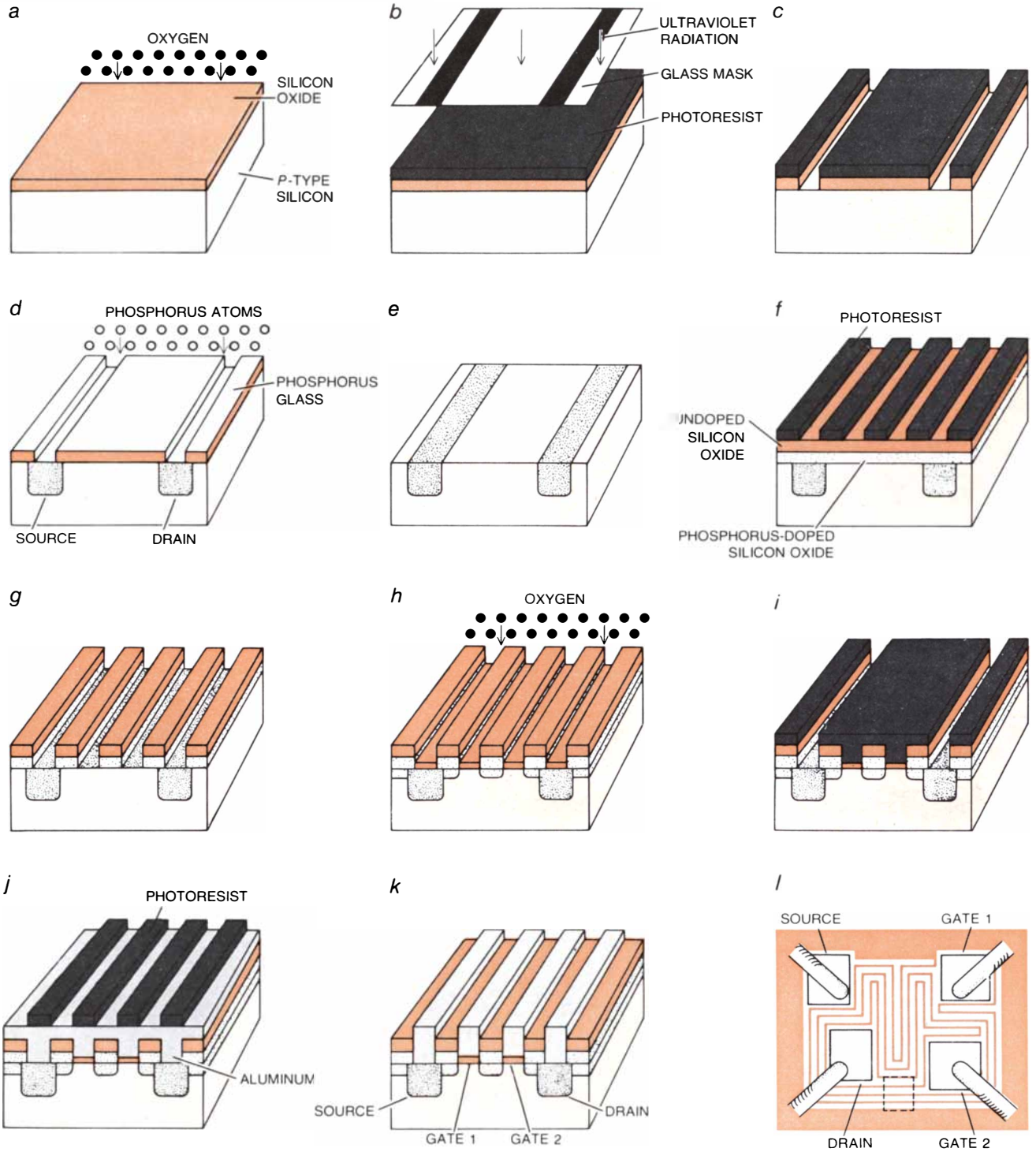
During the 1950's silicon began to supplant germanium as the preferred material for transistors because of its stability over a greater range of temperatures and because it promised better manufacturing control and hence lower cost. A body of knowledge about the surface properties of silicon developed rapidly, along with techniques for fabricating highly stable structures in which the interface properties between silicon and silicon dioxide were well understood and reproducible. In 1960 Dawon Kahng and John Atalla of Bell Laboratories proposed a silicon structure in which an insulated field plate, or gate, was used to induce a conducting surface channel between two surface *p-n* junctions. Two years later a silicon field-effect transistor successfully incorporating this concept was demonstrated by Stephen R. Hofstein and Frederick P. Heiman of the RCA Corporation. Their structure consists of two regions heavily "doped" with impurities, called a source and a drain, and a gate electrode of metal separated from the underlying silicon by an intermediate insulator, usually silicon dioxide. Hence the term metal-oxide-semiconductor (MOS) transistor [*see illustration on opposite page*].

When a negative potential is applied between a source and a drain, the conduction between them can be controlled by the modulation of the charge in a channel between them. The charge modulation is achieved by varying the voltage between the gate and the source. A charge placed on the gate electrode induces an equal but opposite charge in a shallow layer, or channel, in the silicon under the gate electrode. The charge in-





MOS TRANSISTOR derives its name from its construction materials: metal, oxide and semiconductor. The metal, aluminum, serves as an electrical conductor, the silicon oxide acts as an electrical insulator on the surface of the device and under the gate, and the semiconductor can be either *n*-type silicon or *p*-type silicon. *N*-type silicon has excess free electrons (*dots*) and *p*-type silicon has excess "holes," or vacancies in its electron shells (*circles*). Enhancement-type MOS transistors, such as that shown here, are normally off when the gate voltage (V_G) is zero. The source and the drain are strongly "doped" *n*-type silicon (*upper illustration*). When a positive voltage is gradually applied to the gate, holes are driven away from the interface of the oxide and silicon under the gate and a negative charge is induced. With a sufficiently high gate voltage an inversion occurs that changes a very thin region of the silicon under the gate from *p*-type (with holes) to *n*-type (with electrons). The inverted region is called an *n* channel and provides a conduction path for electrons (*lower illustration*). If a positive voltage is applied to the drain, current (electrons) will flow from the source to the drain through the channel. The gate voltage controls the amount of the current: the higher the voltage, the higher the current. Keeping the gate voltage constant (above the threshold for creating the channel) and increasing the drain voltage also increases the current from source to drain. The flow of the current, however, reduces the conductivity of the channel and the channel becomes "pinched off" at a certain drain voltage. The current then is said to be saturated and remains at a constant value as the drain voltage is increased. If the *p*-type and the *n*-type silicon are reversed, the result is a *p*-channel enhancement transistor, in which the charge carriers in the channel are holes. It is also possible to build a transistor in which a channel of the same polarity as the source and the drain exists without a charge on the gate. In this type of transistor, conduction is at a maximum at zero gate voltage and decreases as gate voltage is applied. The gate voltage depletes the channel charge. Such devices are called depletion-type transistors.



FABRICATION PROCESS for a discrete dual-gate MOS transistor is depicted. The process for making complex MOS integrated circuits is similar. Both are fabricated on wafers of doped silicon. From 100 to 1,000 chips, each with a complete device or integrated circuit, can be made from a wafer about two inches in diameter and less than a hundredth of an inch thick. To make an *n*-channel transistor a layer of silicon oxide is grown on *p*-type silicon (*a*). A photoresist layer is applied and exposed to ultraviolet radiation through a glass mask (*b*). Areas of photoresist exposed to radiation harden when developed. Unprotected silicon oxide is etched away by acid (*c*) and the remaining photoresist is removed. Phosphorus atoms are diffused into the unprotected silicon to create the source and drain areas (*d*). The glass that forms and the oxide layer are removed by acid stripping (*e*). Next a layer of silicon oxide doped with phosphorus is deposited, followed by a layer of undoped silicon oxide and a layer of photoresist. The photoresist is developed to form channels for the gates (*f*). An acid etch removes the oxide in the channels and the photoresist is removed (*g*). Next the wafer undergoes thermal oxidation to grow a thin insulating oxide layer in the channels. At the same time phosphorus atoms from the *p*-doped silicon oxide layer diffuse into the silicon substrate and extend the source and drain (*h*). Photoresist is added and developed so that the source and drain regions are open and an acid etch can remove the surface oxide (*i*). The photoresist is removed and a layer of aluminum is deposited by evaporation. Photoresist is added and developed to protect the contacts and channels (*j*). The unprotected aluminum is etched away, the photoresist is removed and the device is essentially complete (*k*). The broken lines in the top view of the entire dual-gate MOS transistor device (*l*) indicate the area shown by the cross sections in the fabrication steps.

duced in the channel can then be used to control the conduction between the source and the drain.

Conduction is very small for a negative gate voltage from zero to a value called the threshold voltage, then increases rapidly for voltages more negative than the threshold voltage. Above the conduction threshold the gate voltage enhances the channel charge between the source and the drain, thereby increasing the channel conductance. Such devices are called *p*-channel enhancement transistors, since the conducting channel consists of positive holes acting as charge carriers. If all polarities are reversed, including those of the silicon structure, the device is called an *n*-channel enhancement transistor.

It is also possible to build an MOS transistor in which a channel of the same polarity as the source and the drain exists without charge on the gate. In this mode conduction is at a maximum at a zero gate voltage and decreases as voltage is applied to deplete the channel of charge. This depletion type of transistor can also be made with either polarity. In practice the enhancement mode of operation provides greater flexibility in circuit design than the depletion mode, in particular where a number of transistors are directly coupled. Accordingly transistors of the enhancement type are used almost exclusively in MOS integrated circuits.

The MOS Transistor

An MOS transistor is fundamentally simpler than a junction transistor and therefore MOS integrated circuits are simpler than bipolar integrated circuits. For example, an important circuit building block known as an inverter requires four separate diffusion steps when it is designed as a bipolar, or junction, unit, whereas only one diffusion step is needed for a unipolar, or MOS, inverter. The bipolar inverter also calls for six masking steps, compared with five for a unipolar inverter.

It was this basic simplicity, with the promise of very low cost, that spurred work on MOS structures and processes in the 1960's. Several years were required, however, to solve stability problems associated with the oxide-silicon interface and the behavior of the oxide itself. Over the past six years the production of MOS integrated circuits has grown from almost nothing to an expected sales volume this year of about 48 million circuits for $250 million. For purposes of comparison, this year's sales of integrated circuits containing bipolar transistors are expected to reach about 720 million units for about $400 million.

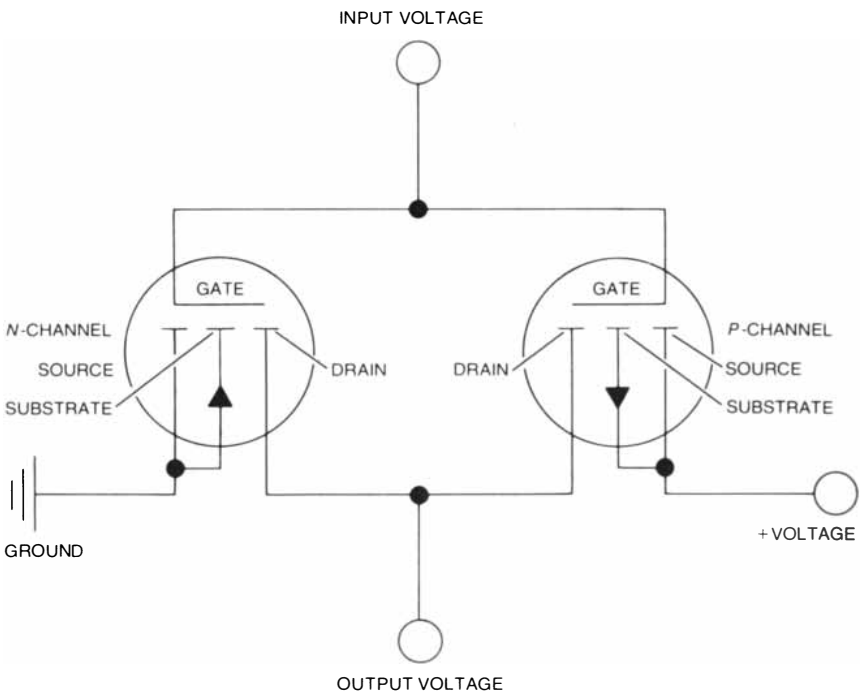
MOS transistors differ significantly from junction transistors in electrical properties stemming from fundamental differences in operating principles. MOS transistors can be used to advantage in many circuits because their properties are similar to those of vacuum tubes: their output response varies almost linearly with input and they have a high input impedance. (Impedance is a measure of the resistance an electronic device offers to the flow of an alternating current.) In contrast, bipolar transistors, which are current-operated rather than field-effect devices, are quite nonlinear and have a much lower input impedance. The high impedance of MOS transistors makes it simpler to drive a large number of successive circuit stages and thereby to design more complex integrated circuits than with bipolar transistors. In addition an MOS transistor consumes significantly less power than a bipolar transistor, both in the conducting and in the nonconducting modes.

Bipolar transistors now operate at much higher speeds than MOS transistors can, but this advantage can be attributed more to the comparative maturity of the two technologies than to an inherent limitation in the MOS principle. As things now stand, the two kinds of transistor tend to complement each other. The designer of new equipment is often faced with a difficult choice between the well-established bipolar technology and the rapidly improving performance and lower cost of MOS technology. My own belief is that MOS technology will dominate digital electronics before the end of the present decade.

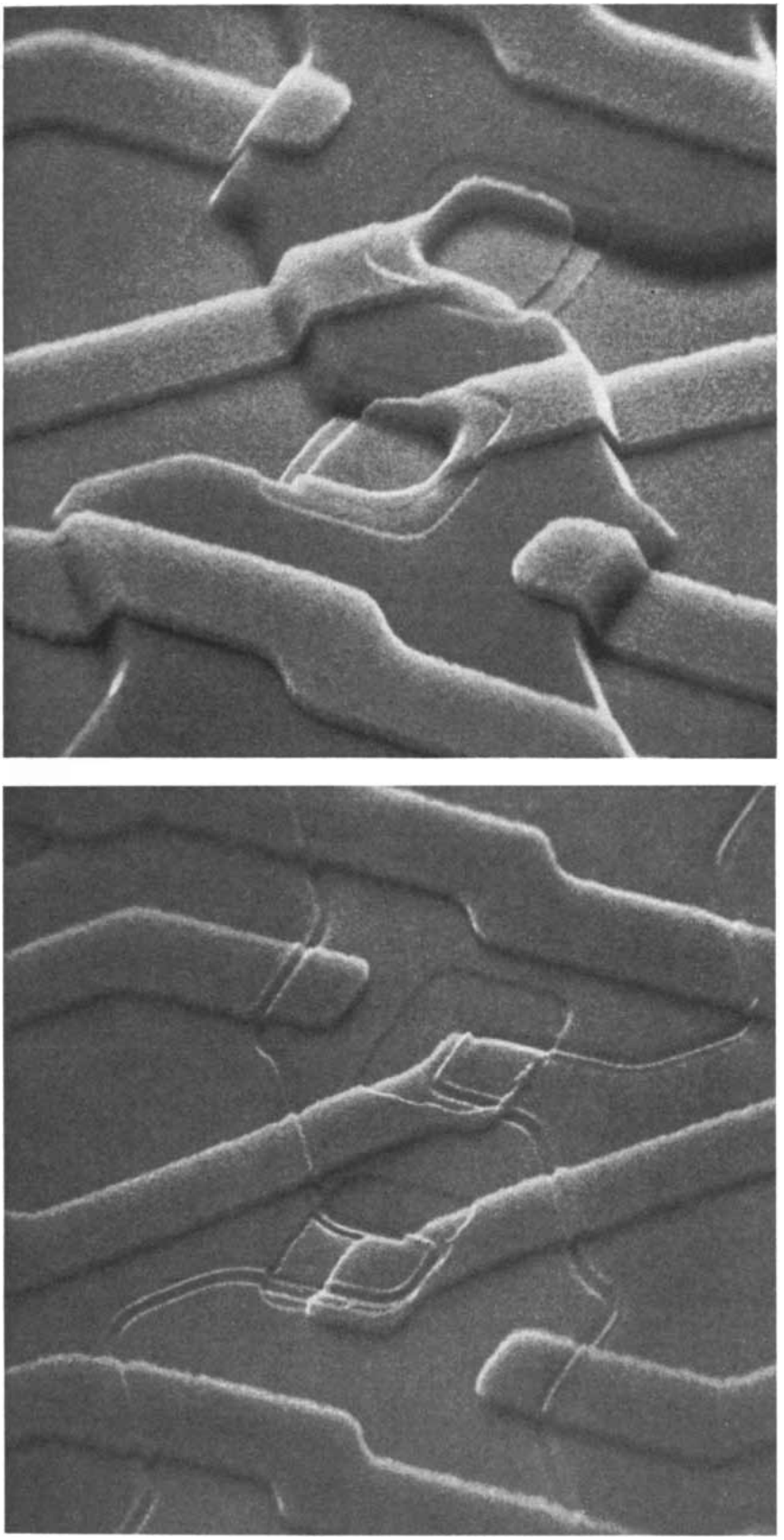
Many hundreds of MOS circuit types are being produced today, ranging from rather simple logic gates used in digital signal processing to custom designs combining both logic and memory functions on the same chip of silicon for service in such products as desk calculators and electronic organs. There is hardly an electronic device being designed today that does not incorporate some form of MOS circuitry, with the exception of high-speed computers.

The MOS Calculator

Probably the largest commercial application of MOS circuits during the past two years has been in desk and pocket calculators. Until the advent of complex MOS circuits most desk calculators were assembled from electromechanical com-



COMPLEMENTARY MOS (CMOS) INVERTER consists of a *p*-channel and an *n*-channel enhancement-type transistor with their gates connected together and their drains connected together. When the input voltage to the gates is at ground (logic 0), the *p*-channel transistor is on, the *n*-channel transistor is off and the output voltage is positive (logic 1). When the input voltage is positive (logic 1), the situation is reversed and the output voltage from the *n*-channel transistor is at ground (logic 0). In either logic state one transistor is on while the other is off, and the power consumption of the CMOS inverter in a quiescent state is low.



SCANNING ELECTRON MICROGRAPHS provided by Fairchild Semiconductor show a single bit in a 1,024-bit, four-transistor random-access memory circuit made by conventional diffusion and etching techniques (*top*) and by an isoplanar process that results in a flatter surface (*bottom*). Weak spots that can lead to defects tend to occur where the aluminum electrodes (*narrow strips*) climb a raised oxide surface. The isoplanar process extends the oxide into the silicon substrate rather than growing the oxide on top of the substrate.

ponents and cost between $500 and $1,000. Bipolar integrated circuits were used in some quantity, but their advantage in terms of performance with respect to cost was not striking. By 1969 MOS technology had reached the point where it became feasible to design calculators where all the computation functions could be accomplished by a few integrated circuits. Only three years later it became possible to incorporate essentially all the electronics contained in many bipolar integrated circuits into one complex MOS device. As a result high-performance calculators priced between $50 and $200 are now on the market, and still less expensive ones will soon be available. Today virtually all desk calculators and all pocket calculators are designed around MOS circuits.

Although MOS circuits still lack the speed needed for certain logic functions in the central processor of large computers, they are beginning to compete with magnetic cores in providing the computer's central memory. The ferrite magnetic-core memory has retained its supremacy for two decades in spite of threatened competition from such other magnetic devices as plated wire, ferrite sheet and the ribbonlike device called the twistor. Through continuous improvements in technology and intensive cost reduction makers of core memories have fought off competition by steadily lowering the cost per bit (binary digit) of memory until it is now around .8 cent.

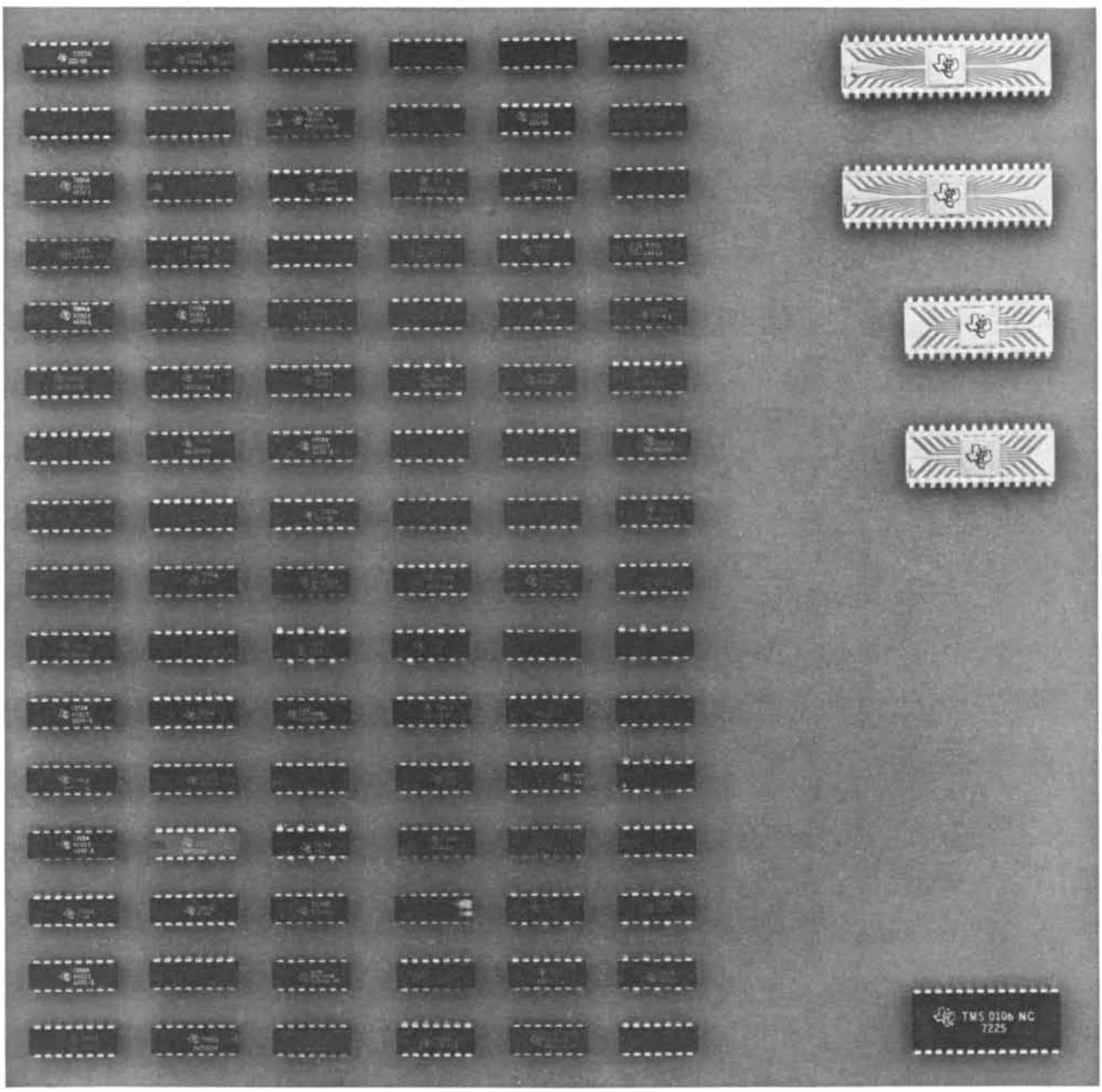
Recently, however, it has become possible to produce MOS random-access memory circuits for about the same cost as core memories, so that there is now intense competition between makers of the two products. MOS memories offer two distinct advantages over cores: lower power consumption and lower generation of heat, which will enable designers to increase the density of memory stores. In addition computers with core memories are designed with a single centralized memory, because cores require high-quality access circuitry that must be shared by all the cores in order to minimize cost. Individual MOS memory circuits, on the other hand, can be built economically with integrated-access electronics, so that the system designer now has the freedom to distribute memory functions throughout the computer without any penalty in terms of cost. This enhanced design flexibility should lead to improved performance in future computer systems. Although builders of core memories are striving vigorously to remain competitive, I believe it is only a matter of time until the price per bit of MOS memories will fall so low that MOS

will assume a dominant position in new mass-storage memory systems.

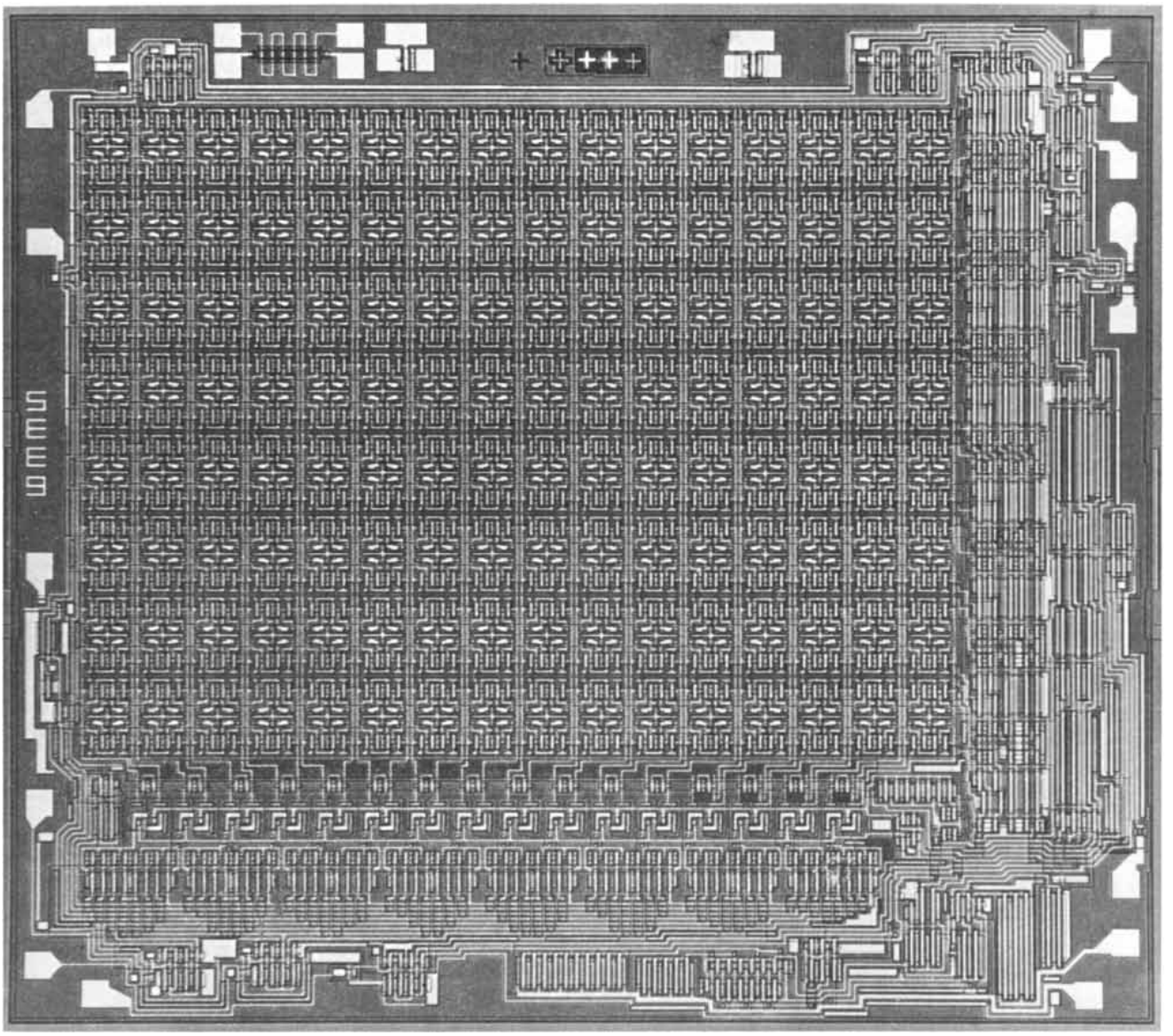
PMOS, NMOS and CMOS

Semiconductor technology has a remarkable history of rapid evolution. The high level of research and development in materials, structures and circuits has generated an impressive list of advances in cost effectiveness. In the case of MOS the field is already very broad and will continue to evolve in the foreseeable future. The first commercial MOS integrated circuits were of the *p*-channel enhancement type (PMOS), and today PMOS circuits constitute about 80 percent of the total MOS market. The PMOS process was the easiest to control and therefore emerged as the first in volume manufacture. Other types are now gaining in commercial importance and will occupy a larger share of the market over the next few years. Advances in processes and materials now make it possible to produce *n*-channel enhancement (NMOS) structures and to put NMOS and PMOS structures in the same circuit to form complementary (CMOS) devices, both of which have attractive advantages. NMOS circuits are two to three times faster than PMOS circuits because electrons are more mobile charge carriers (*n*-type conduction) than holes are (*p*-type conduction). Many producers are now manufacturing both types in order to optimize certain critical functions where speed is important.

The complementary, or CMOS, devices are rapidly gaining popularity and could very well become the most important of all. With both channel types com-



EVOLUTION OF THE ELECTRONIC CALCULATOR is depicted in this photograph. In the mid-1960's a typical electronic calculator contained from 90 to 150 bipolar integrated circuits. An array of 96 such circuits is at left. By 1969 four MOS integrated circuits could do the same job (*top right*). Then in 1971 the circuitry was reduced to a single MOS device (*bottom right*), which made possible the advent of the pocket calculator. All integrated circuits shown here were manufactured by Texas Instruments Incorporated.



MOS RANDOM-ACCESS MEMORY for use in computers requires lower power and generates less heat than the traditional ferrite magnetic-core memory. With MOS circuits the density of memory stores can be increased and memory functions can be distributed throughout the computer. Shown here is a 256-bit random-access memory array on a .146-by-.164-inch chip made by RCA Solid State.

bined in one circuit it is possible to gain performance advantages over all other varieties of integrated circuit currently being manufactured. A basic CMOS circuit is the inverter, which consists of one *p*-channel and one *n*-channel transistor connected in series [*see illustration on page 53*]. The inverter is a highly effective digital switch with the lowest power consumption in either the "on" or the "off" state of any semiconductor device. By interconnecting a number of inverter stages one can build a variety of useful circuits with extremely low power consumption. For example, a 14-stage binary counter, which can be used in a variety of timing applications, typically consumes a maximum of 2.5 microwatts at five volts, which is about 100,000 times less than comparable PMOS or bipolar circuits. This property alone makes CMOS circuits preferred for any application where the power supply is limited. Every battery-powered device is a candidate for their use.

A *p*-and-*n*-channel pair can also be connected in parallel with the power supply to form a transmission switch that can handle both digital signals and analogue signals in either direction. In principle it is possible to make such a switch with a complementary pair of bipolar transistors, but such a circuit is too costly to be commercially useful. Low-cost CMOS circuits also have the added feature of being highly immune to spurious input noise, which makes them particularly suitable for environments such as the automobile engine, where the electrical noise level is high. Circuit designers are now finding that CMOS inverters and transmission switches can achieve almost every desired switching and logic function with one basic technology in a very cost-competitive way.

A rapidly growing market for integrated circuits, CMOS circuits in particular, is in watches and clocks. Electronic timepieces are now being built that have an accuracy impossible to attain by mechanical means. The circuit consists of a quartz-crystal oscillator whose highly accurate fundamental frequency is divided down in many steps by counting circuitry to the speed required to drive clock hands or perhaps an electro-optic digital device such as a

liquid-crystal display or a light-emitting-diode array. The low price of integrated-circuit counters in combination with their accuracy and reliability has created a profound change in an industry that was based for generations on exquisite mechanical skills.

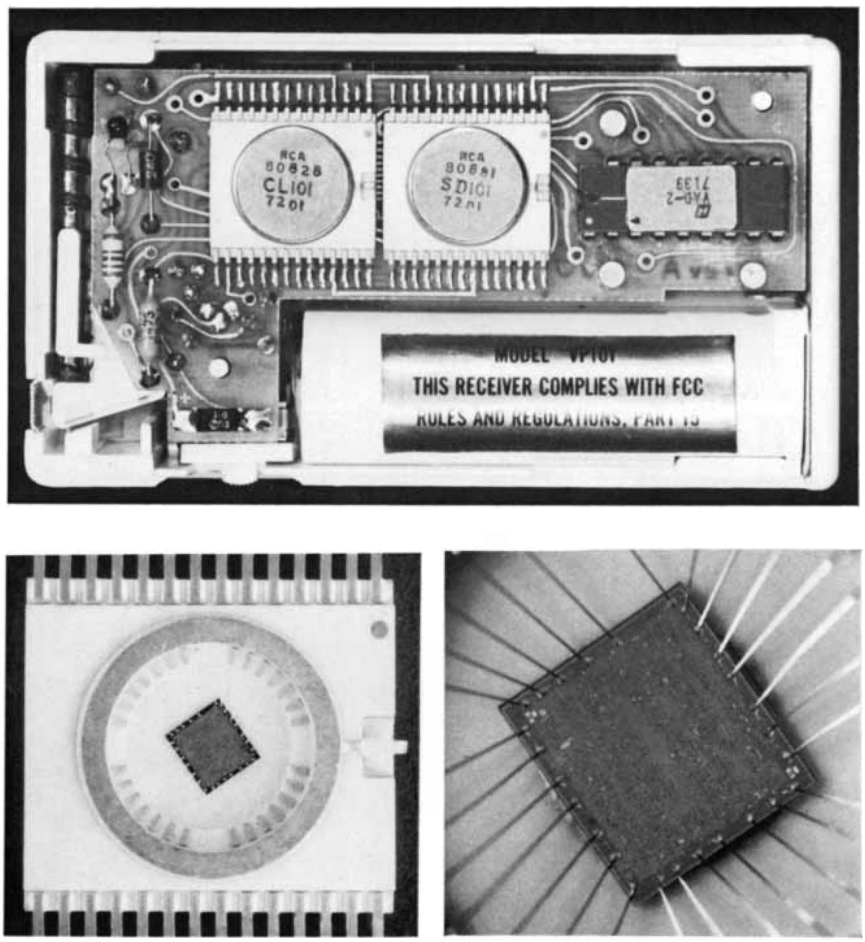
Although present-day MOS integrated circuits have lower operating speeds than the fastest bipolar circuits, even this limitation will be minimized in the near future. The relatively low MOS speeds are due primarily to the phenomenon known as parasitic or unwanted capacitance: a capacitance between the source and drain regions and the silicon substrate into which they are physically diffused. Capacitance refers to the ability of an insulator or a semiconductor to hold an electric charge; the effect of parasitic capacitance is to impede the flow of charge carriers through the field-effect structure, thereby slowing down its overall switching time. If these capacitances could be minimized, the inherent speed of the MOS transistor, which is defined by the mobility of the charge carriers and the dimensions of the structure, would be comparable to that of the fastest bipolar transistor.

An important step toward increasing the speed of MOS transistors has been the development of a technique for depositing a very thin layer of single-crystal silicon on an insulating substrate. It is now possible, for example, to deposit silicon layers only one micron thick on flat substrates of synthetically grown sapphire. As a result the capacitances of the source and the drain are reduced by a factor of roughly 20 over MOS transistors made in bulk silicon material, in which the source and the drain are diffused only partially into a relatively thick substrate. Individual transistors are easily isolated by chemically removing the silicon film between devices. The interconnecting wiring sits directly on the insulating sapphire and thereby does not contribute the additional unwanted capacitance that arises when the wiring is deposited on a bulk silicon substrate. When MOS circuits are made in this thin-film form, called SOS (silicon on sapphire), they combine the unique properties of MOS bulk devices with the high speed of bipolar ones. At the present time the SOS structures are more expensive than bulk devices and will be used only where the additional cost is justified. As the cost differential is narrowed with experience, SOS devices should have an important place in the integrated-circuit industry.

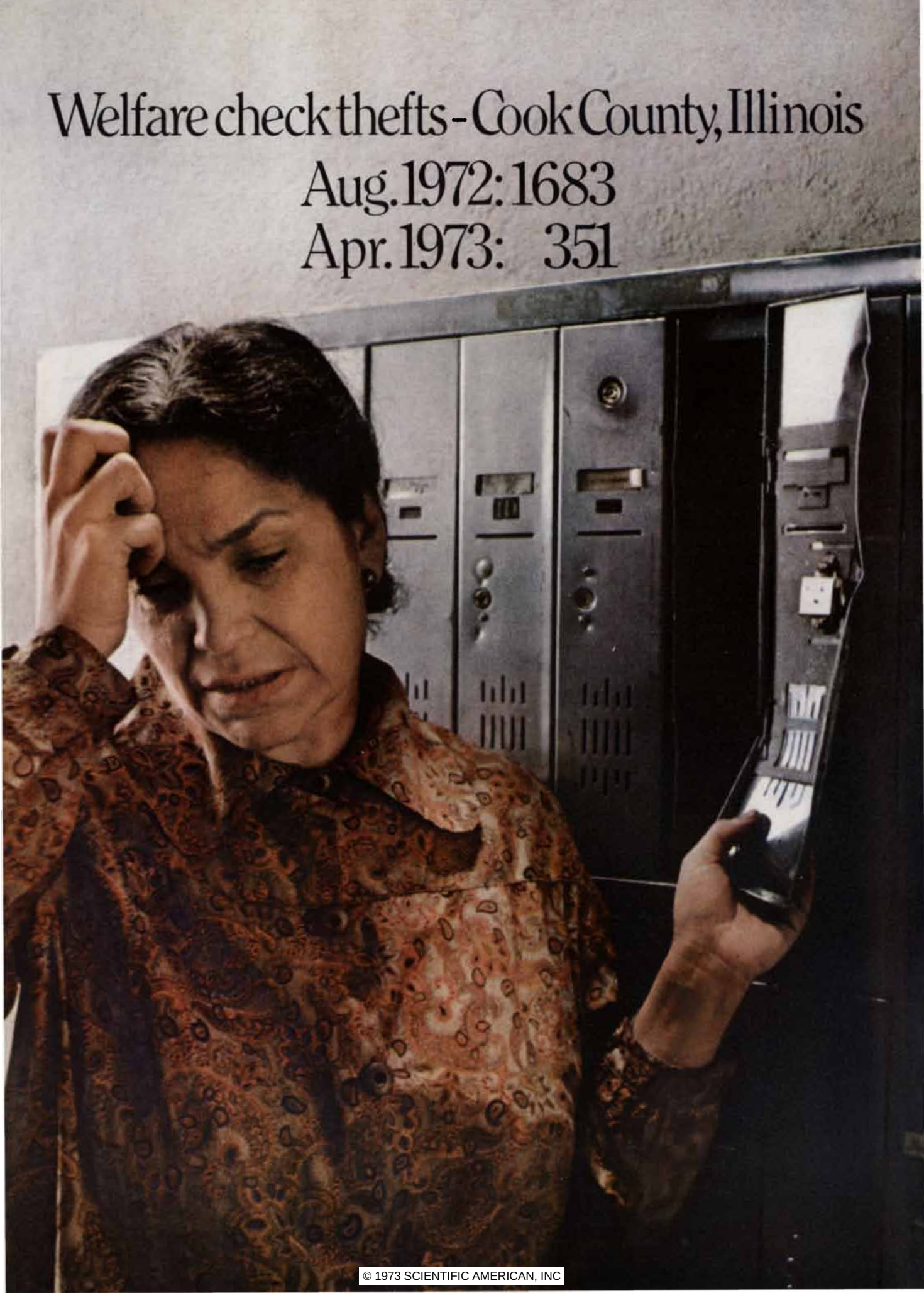
The ubiquitous MOS structure, now being exploited so successfully in field-effect transistor form, is also useful for storing and manipulating small quantities of electronic charge when it is built in capacitor form. The capacitor structure consists of the metal and silicon as end plates and the oxide as a dielectric. Electric charge can be stored in the capacitor in potential wells at the semiconductor surface. Connecting a number of capacitors in an array makes it possible to create charge in selected capacitors and to shift the charge through the array by an electrical clock pulse signal. These charge-coupled devices can be used as a shift register, a type of circuit widely used in information processing. Such registers can serve information storage, delay and the time compression of analogue signals.

In addition charge-coupled devices can be used as photosensitive elements in which minority carriers are generated and stored by the absorption of incident light. Thus an electrical signal can be generated in response to a light image and can be read out of the MOS array by applying a sequence of clock pulses. It is possible today to build small, rugged image sensors with a wide dynamic range and high light sensitivity using a single silicon chip with some 500 elements. A prototype television camera the size of a man's hand and weighing less than a pound has been made using a charge-coupled device as the image sensor. Such devices are suitable now for slow-scan television, facsimile transmission and other high-resolution uses. They have great potential for both consumer and industrial purposes.

The development of the MOS transistor is an instructive story of a device proposed many years ago that is only now being exploited commercially. It vividly demonstrates how a sound concept often needs a matching capability in materials technology before it can be brought into reality.



PAGING RECEIVER about the size of a pack of cigarettes contains 1,400 transistors in two CMOS integrated circuits made by RCA Solid State. The integrated circuits are in the square hermetically sealed packages that can be seen in the upper portion of the top photograph. The connections to the circuit chip in the package on the right are shown in the bottom photographs. The receiving pager, which is manufactured by Martin Marietta, has a range of 50 miles. It emits a tone when the person carrying it is wanted on the telephone.



When you're getting welfare assistance and your check is stolen, it means you have no money at all for groceries or the rent. It means inconvenient, time-consuming trips to fill out check replacement forms and to get an emergency food order. It means a physical and emotional crisis. And it happened to 1683 welfare clients in Cook County last August.

To make matters worse, the high incidence of fraudulent check cashers made many store managers refuse to cash bona fide checks, while others subjected welfare clients to mutually embarrassing questions for identification purposes.

Add the financial costs to those human costs, and you have the picture of how it used to be every month in Cook County. And then, last August, the Illinois Department of Public Aid began its voluntary Polaroid portrait identification card program. Here's how it works:

Each participating welfare client receives a laminated card with his portrait in color, his signature, and an account number that makes it easy to detect forgeries quickly.

The cards provide a fast, virtually tamperproof means of identification. And, since they *don't* identify the owner as a welfare recipient, they are readily accepted and used by over 90% of the welfare clients. (An added advantage is that no film negative or extra prints are produced. Since the card owner has the only photograph, his privacy is protected.)

In just eight months, Polaroid ID cards reduced thefts of welfare checks by almost 80% to only 351 this April.

And that's not all. Welfare clients find that many store managers readily cash their checks when the portrait card is shown. Which saves waits-in-line for half an hour, or more, at check cashing institutions. It also saves the check cashing fee. (And when check cashing services *are* used, the lowered risk results in lower fees.)

Last, but scarcely least, *the saving to the State of Illinois, and its financial and business organizations from decreased check thefts was well over $1,000,000.*

For more information about the Polaroid Portrait ID System, write: Polaroid Corporation, Department 57-263, Cambridge, Massachusetts 02139, or call collect (617) 864-6545.

The Polaroid Portrait ID System made the difference.

The Evolution of the Andes

The geology of the central Andes indicates that the history of the range can be understood in terms of the consumption of a plate of the earth's crust plunging under South America

by David E. James

A great series of mountain belts, the Andean cordillera, sweeps down the west coast of South America from Venezuela and Central America nearly to the southern tip of Chile. A long arc-shaped deep in the ocean floor, the Peru-Chile trench, is clearly allied with the mountain chain and runs roughly parallel to it from about four degrees north latitude to 40 degrees south. Some 15,000 meters (more than nine miles) of vertical relief separates the deepest part of the trench from the highest Andean peak. The Andean arc, comprised of both mountains and trench, is a living, evolving system. It is a part of the circumpacific "ring of fire," and the live volcanoes that dot the length of the cordillera and the devastating earthquakes that punctuate the flow of South American life are ever present reminders that the mountain-building processes that raised the Andean chain are still very active today.

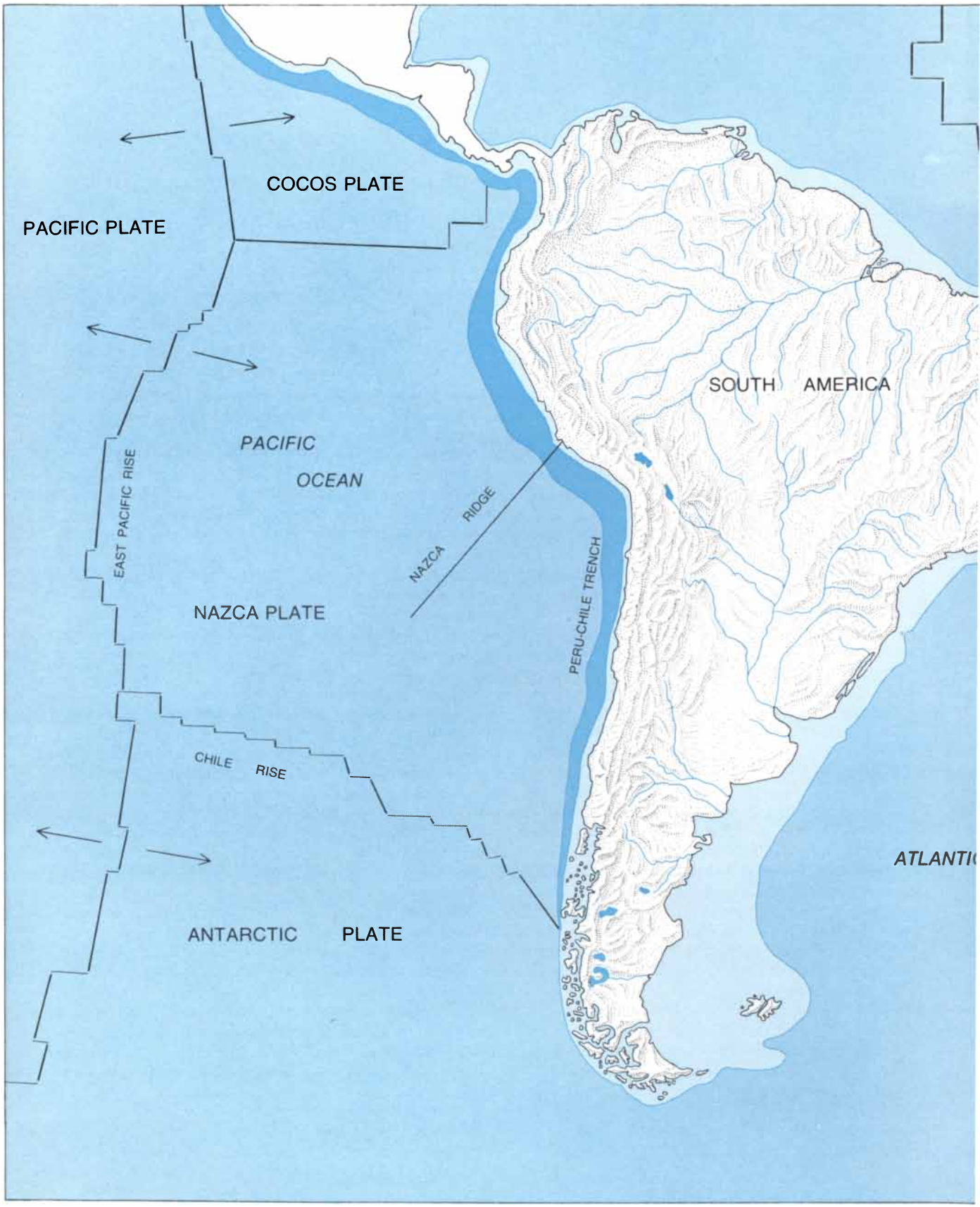
Many earth scientists, of whom I am one, believe that the Andean arc is a modern analogue of many older mountain belts: that if we had lived in the region of the Sierra Nevada 100 million years ago or the northern Appalachians 450 million years ago, we would have witnessed mountain-building activity of a similar kind. Until recently this view was not widely accepted. It was generally supposed that the processes through which older mountain belts evolved are not being duplicated today. That supposition is rapidly giving way to the concept that mountain-building activity has proceeded in accordance with consistent geologic patterns that have been repeated over and over again during at least the past two billion years of earth history, and that it continues to do so today. In this view the Andean chain is the foremost modern example of those mountain-building processes at work, and so, by understanding the Andes, we seek to learn how ancient mountain belts were born, matured and in some cases died long before man arrived to record their origin or their passage.

Concepts of orogeny, or mountain building, have recently been revolutionized by the theory of plate tectonics [see "Plate Tectonics," by John F. Dewey; SCIENTIFIC AMERICAN, May, 1972]. This theory, which has been developed with startling rapidity over the past decade, holds that the outer rind of the earth, the lithosphere, consists of a mosaic of rigid plates that are in motion with respect to one another. These plates, some 100 kilometers (60 miles) thick, include not only the earth's solid crust but also part of the denser upper mantle. Plate boundaries, or junctions, rarely coincide with continental margins, and so the relative movements of parts of the earth's surface are now viewed in terms of plate motions rather than "continental drift." The proud continents, which were once thought to plow through oceanic crust like great ships, are now reduced to the status of passive passengers on the lithospheric plates. Volcanic magma welling up from deep within the earth's mantle creates the lithospheric plates along oceanic ridges. Newly generated lithosphere moves away from the ridges to yield to constantly replenished volumes of magma injected along the axes of the ridge. These spreading plates are consumed at trenches, where they bend down and plunge into the earth's mantle.

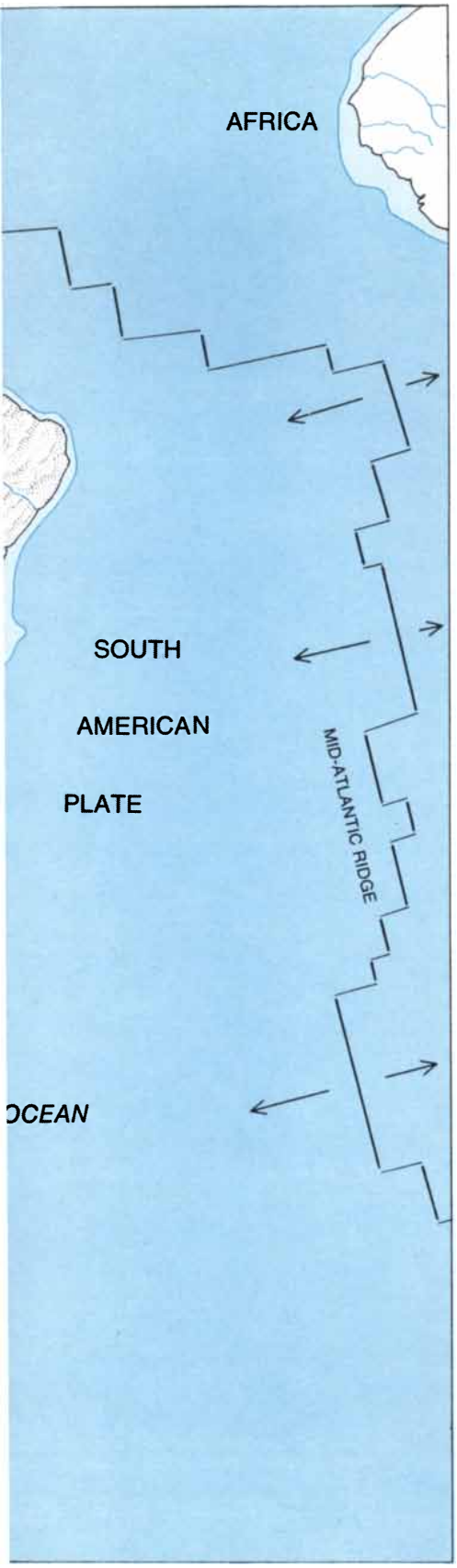
Most of the world's tectonic activity—earthquakes, volcanoes and mountain building—is concentrated along plate junctions. The west coast of South America is such a junction [*see illustration on next page*]. Here the oceanic Nazca plate, generated along the East Pacific Ridge, is consumed in the Peru-Chile trench, where it bends down and slides under the South American plate at a rate of about six centimeters per year. Examination of the structure and geology of the central Andes reveals that most of the present-day orogenic activity and the geologic evolution of the past 200 million years can be understood in terms of the subduction, or consumption, of the oceanic plate under South America. This interaction of the two plates accounts for the crumpling of the stable continental margin to form belts of fold mountains that now constitute the eastern ranges of the Andes, for the birth of the great Andean volcanic cordillera to the west and for the continental growth of western South America.

Lured by the sheer scale of the Andes mountains, the Carnegie Institution of Washington, in concert with a number of South American institutions, set out more than a decade ago to study the nature and evolution of the central Andes of southern Peru, Bolivia and northern Chile. Many years of work have led progressively to a rather complete

REGION OF THE ANDES near Lake Titicaca, on the boundary between Peru and Bolivia, is shown on the opposite page in a photograph made by the Earth Resources Technology Satellite. The lake is on the altiplano, or high plateau, nearly four kilometers above sea level. Rising above the altiplano to the west (*bottom*) are the volcanic mountains of the western cordillera; to the east (*top*) are the fold mountain belts of the eastern cordillera.



GEOPHYSICAL SETTING of the Andes is portrayed. According to the theory of plate tectonics, the lithosphere, or outer shell of the earth, consists of several rigid plates that are moving with respect to one another. The juncture between the Nazca plate and the South American plate represents the most important modern example of interaction of an oceanic and a continental plate. The Nazca plate is generated along the East Pacific Rise and consumed in the Peru-Chile trench. The interaction of the two plates at the trench



has caused deformation and growth of the continental margin of South America, thereby forming the Andean mountain system. Arrows indicate flow away from ridges.

understanding of the physical properties of the earth under the central Andes, but it is only recently that the conceptual tools have been available with which to understand the forces and sequences of orogenic events that produced the Andes and that continue to generate volcanism and seismicity along the Andean chain even today.

To understand the evolution of the central Andes it is necessary to draw on two distinct kinds of evidence, geophysical and geological. Important geophysical evidence includes the distribution of earthquakes and the distribution within the crust and upper mantle of certain physical properties: the velocity of seismic waves, the absorption of their energy (their attenuation) and the density of rock. From these properties one can deduce information on the kinds of rock in the earth's interior and whether they are rigid or mobile. Geological evidence derives from the study of rock types and their structures as observed on the earth's surface; it is only from geological evidence that one can read the history of past events. I shall consider the geophysical evidence first, since it provides the observational basis for most of our knowledge concerning the plate-tectonic framework of the juncture between the converging South American and Nazca plates.

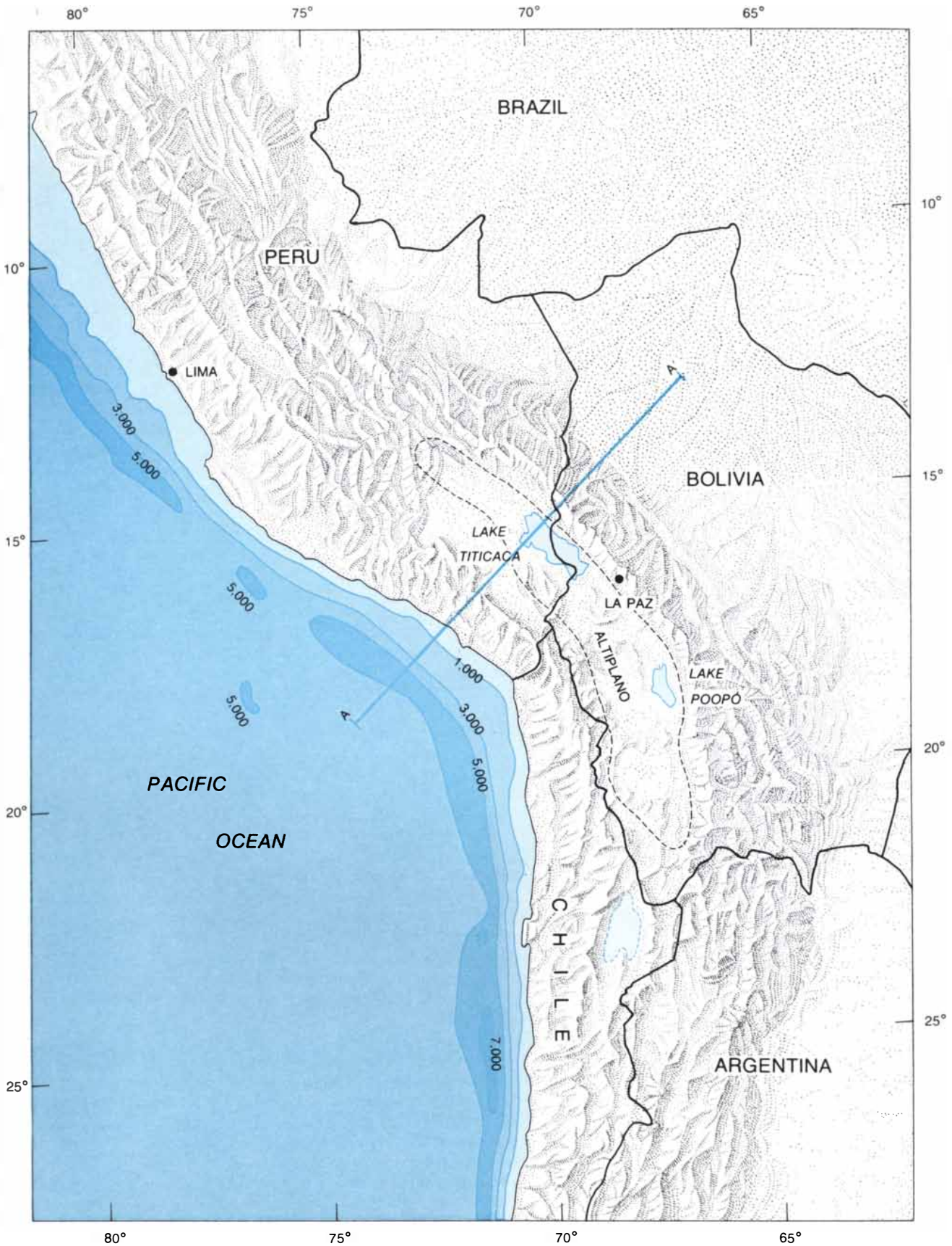
Earthquake distribution provides key evidence regarding plate interactions. It is a general tenet of plate theory that the inclined earthquake zones (the Benioff zones, named for Hugo Benioff) that bend down under trenches and volcanic arcs define the upper part of the descending oceanic plates. Descending lithospheric plates are cooler and hence more rigid than the asthenosphere (the hot mantle) through which they sink, and it is widely supposed that only within the lithosphere are rocks rigid enough to support brittle fracture, or earthquakes. This supposition is supported by the observation that in most island arcs earthquakes appear to be confined almost entirely to the descending plate.

Under the central Andes, however, earthquakes occur not only in the descending plate but also in a continuous wedge between the earth's surface and the top of the sinking oceanic plate to a depth of from 200 to 300 kilometers [*see top illustration on page 68*]. This observation leads one to suspect that the leading edge of the overriding South American lithosphere may be an abnormal 200 to 300 kilometers thick—much thicker than the usual 100 kilometers observed for most plates and in startling contrast to the 50-kilometer thickness measured seismically for the undersliding Nazca plate. More direct evidence that the Andean lithosphere is 200 to 300 kilometers thick derives from seismic results, which show that there is no low-velocity zone or zone of high seismic-wave attenuation in that region.

Both low-velocity zones and zones of high attenuation in the mantle are commonly interpreted as being regions of softer rock and are associated with the asthenosphere; conversely, the absence of these zones implies that the rock is comparatively rigid. An important consequence of the thick lithosphere under the Andes is that it precludes convective overturn in the mantle above the Benioff zone. It is this overturn that is believed to produce the secondary centers of seafloor spreading that develop behind island arcs but that are conspicuously absent behind the continental Andean arc.

A second important set of geophysical observations, on crustal thickness, provides an additional linkage between the present plate-tectonic regime of the Andes and the orogenic history of the Andes as it is read in the geologic record. Seismic data show that the crust of the central Andes is extremely variable in thickness. Over a distance of little more than 500 kilometers it varies from about 11 kilometers (including water) in the Pacific basin to 30 or 35 kilometers along the coast (a "normal" continental crust), more than 70 kilometers under the western-range volcanic crest, 50 to 55 kilometers under the eastern fold ranges and finally 35 kilometers under the Brazilian shield. Only in the Himalayas have crustal thicknesses of as much as 70 kilometers been observed elsewhere.

Yet one could have predicted these results obtained from seismic measurements simply on the basis of topography and the principle of isostasy. This principle states that mass excess above sea level—a mountain chain, for example—must be compensated by an equal mass deficiency at depth: the mountain chain's crustal root, which displaces mantle rock, must be less dense than the rock it displaces. It is evident, of course, that the reverse expression of the principle of isostasy is equally valid: Any mass deficiency at depth must be compensated by a mass excess at the surface; any growth or thickening of the crust, such as by the injection of magma into the crust from below, will result in surface uplift. From seismic studies of the crust it is known that areas that have long remained at sea level (and are thus presumed to be in isostatic balance) are underlain by "normal" continental crust 30 to 40 kilo-



CENTRAL ANDES is a region where mountain-building forces appear to be still at work. North of Lima the Andes form a single belt of closely spaced mountain chains running parallel to the coast. South of Lima, however, the mountains branch, with the easterly fold belt running hundreds of kilometers inland and the westerly volcanic chain continuing parallel to the coast. Between these ranges lies the altiplano, a broad flat plain underlain by an enormous wedge of sedimentary debris eroded from the adjacent cordilleras. The two ranges merge again in northern Chile. Line *A–A′* shows region portrayed in several succeeding illustrations.

meters thick. Areas below sea level are underlain by thinner crust, those above by thicker crust. This fact becomes central when considering crustal evolution, since geology indicates that many of the highest parts of the Andes were once near sea level.

One additional line of geophysical evidence, seismic velocity, can be used for guessing the composition of the rocks within the Andean crust and the adjacent oceanic crust. In laboratory experiments the velocity of seismic waves in common rock types has been measured over a large range of temperature and pressure. Seismic velocities of rock in the lower oceanic crust are appropriate to rock of basaltic composition at that temperature and pressure. Interestingly, even though the pressure and the temperature in the crustal root under the Andes are much greater than in the lower oceanic crust, the seismic velocities of the rocks are the same. Yet it is known that if basaltic rock were subjected to the pressures and temperatures found in the deep continental crust, it would undergo metamorphic transformation to a rock type with higher seismic velocities and densities than are observed.

When different kinds of rock are examined in detail, only lighter crustal rocks, similar in composition to those observed at the surface, exhibit the appropriate seismic velocities at the temperature and pressure of the lower crust. This leads to the inference that the Andean crust is rather homogeneous vertically, the rocks in the Andean root being similar in composition to the volcanic and plutonic rocks that form the upper levels of the crust. Exceptions to vertical crustal homogeneity occur in the altiplano (the high plateau between the two great volcanic cordilleras) and the eastern cordilleras. Variations in seismic velocity show that the altiplano could be underlain by up to 30 kilometers of sedimentary rock, perhaps mixed with volcanic deposits, and that the eastern cordillera is underlain by at least five to 10 kilometers of sedimentary rock.

In order to reconstruct the progressive development of the Andean system one must turn to the information that can be read from the geologic record. Paleozoic sedimentary rocks laid down some 250 to 450 million years ago are among the oldest rocks of the central Andes. Time has not treated them kindly, and we find them collapsed and crumpled to form the fold mountains of the eastern ranges. These rocks, some 10 kilometers of monotonously repetitive muddy, sandy sedimentary beds, are called geosynclinal rocks [see "Geosynclines, Mountains and Continent-building," by Robert S. Dietz; SCIENTIFIC AMERICAN, March, 1972]; once they formed the quiet Paleozoic continental margin of western South America. (A modern analogue to the western seaboard of Paleozoic South America is the inactive continental margin of the east coast of North America, where a great sedimentary wedge some 250 kilometers wide and up to about 10 kilometers thick has formed between the continental shelf and the abyssal plain of the ocean.)

During Permian and Triassic time, between about 200 and 250 million years ago, the quiet of the western coast of South America gradually gave way to rumblings brought on by the incipient breakup of the supercontinent Pangaea and the onset of the plate-tectonic cycle that is still under way. The continental edge became unstable and the geosynclinal strata were warped and buckled upward. Magma, possibly derived from partial melting of deeply buried geosynclinal strata, invaded and poked through the sedimentary pile to form batholiths (bodies of intrusive igneous rock) at depth and volcanoes at the surface. The old volcanoes have long since disappeared, but volcanic rocks from this period are found interlayered with sedimentary beds. The intrusive rocks, probably representing the deeper feeder levels of the volcanoes, are commonly found today exposed in the cores of the eastern-fold ranges, nestled among crumpled geosynclinal rocks.

About 190 million years ago, in earliest Jurassic time, the major axis of tectonic activity shifted several hundred kilometers oceanward, to the west. The lithosphere broke along the junction between the South American continent and the Pacific Ocean basin and the oceanic plate began its descent into the mantle under western South America. One can only speculate now on the reasons for the catastrophic rupture of the lithosphere, although there is some evidence that it may have been in response to the onset of sea-floor spreading along an ancestral East Pacific Rise. We do know that the South Atlantic Ocean had not yet begun to open and that South America and Africa were still one.

As the Pacific plate plunged under South America, andesitic magma welled to the surface, sweated out of the descending basaltic oceanic crust. Andesite is the characteristic volcanic rock of the Andes. It is richer in silica, and hence less dense, than basalt; its composition is what is termed intermediate, that is, between basalt and granite in chemical composition. These andesitic rocks of Jurassic age are well preserved all along the coastal regions of southern Peru and northern Chile. It is still somewhat problematical, however, whether the andesites were extruded on sialic (continental, shallow-water) crust or simatic (oceanic, deep-water) crust. The lavas themselves appear to have been extruded below water, since many have been highly altered by seawater.

One might infer from this that the volcanic arc of the central Andes began as an island arc in the ocean off the coast of ancestral southern Peru and northern Chile [*see illustration on next two pages*]. Yet the picture of the Jurassic arc is not simple, because Jurassic volcanic rocks in southern Peru are wedged in among crystalline metamorphic rocks at least 400 million years old. Just what these remnants of ancient sialic crust are doing some 300 kilometers west of the currently exposed geosynclinal rocks of the Paleozoic continental margin is not known. The presence of sialic rocks does not necessarily imply, however, that the Jurassic arc formed on continental crust. The rocks could be part of a Paleozoic microcontinent or peninsula that lay to the west of the South American coastline. Or they could be sialic flotsam swept into and plastered to the edge of South America, buoyant debris scraped from the top of the oceanic plate as it dived down at the trench.

These difficulties aside, formation of the Andean volcanic arc was well in progress during Jurassic time. The fact that Jurassic volcanic rocks are still widely preserved indicates that the Jurassic arc never stood much above sea level and consequently suffered little erosional destruction. Furthermore, it appears that the entire region to the east of the Jurassic arc, extending as far east as and encompassing the area of the present-day Andes, must have been near sea level, since marine sedimentary deposits are mapped throughout the area. This observation is important because it shows that the crust of the Andes was still thin, probably no more than 35 kilometers thick.

I have noted that the Jurassic arc formed during a period when South America and Africa were still a single continent. Dating of the magnetic pattern of the South Atlantic by Walter C. Pitman III and others at the Lamont-Doherty Geological Observatory of Columbia University indicates that it was only about 135 million years ago that the South Atlantic began to open by spreading along the Mid-Atlantic Ridge. For-

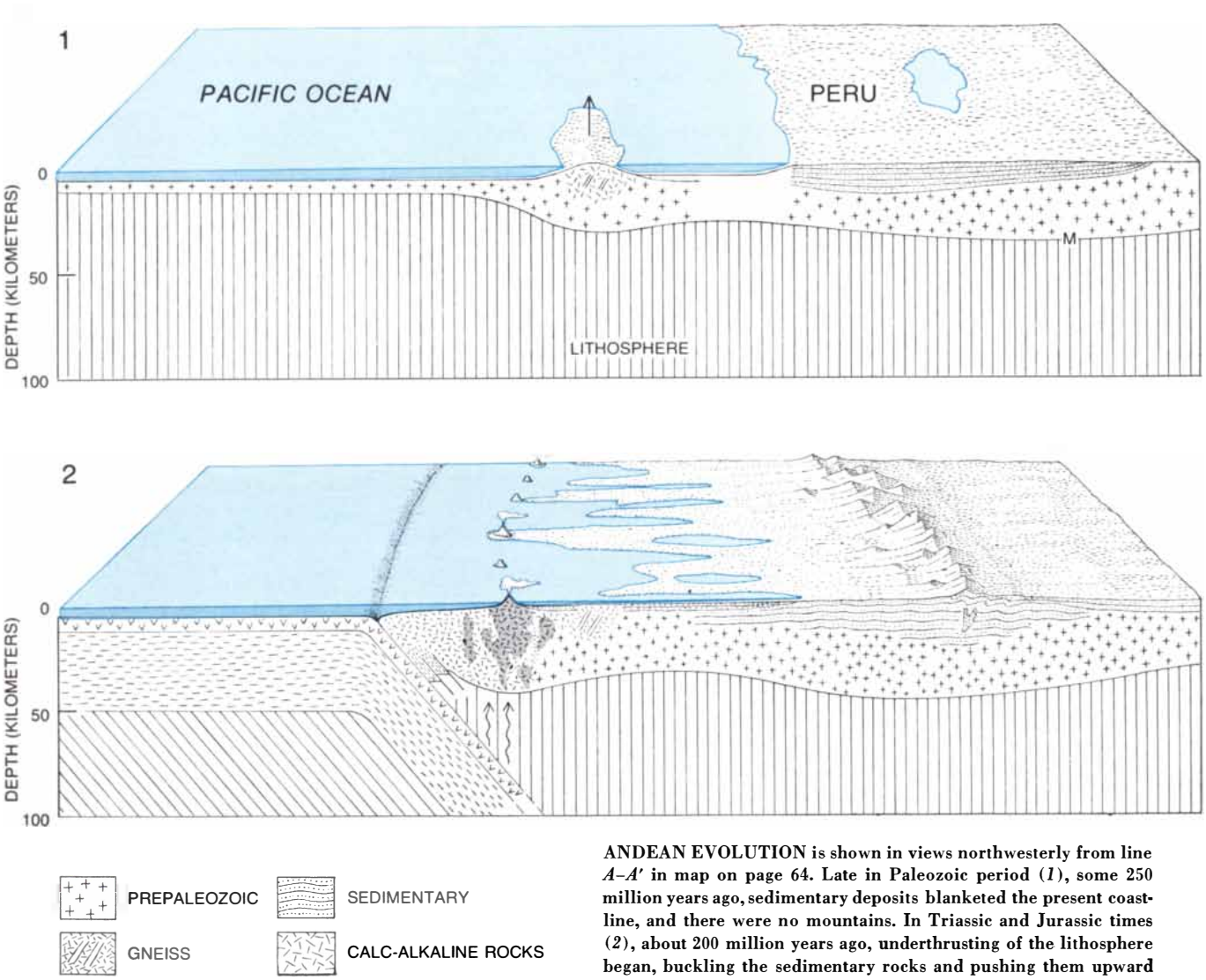
mation of the Jurassic arc inaugurated the impressive series of mountain-building episodes that produced the Andean belt, but it was not until sometime after the break-off of South America from Africa that Andean orogeny began in earnest.

About 100 million years ago, during Cretaceous time, a second major volcanic arc began to form parallel to and continentward of the Jurassic arc. (My treatment of Andean development as a series of a few distinct episodes is an oversimplification, but it serves to trace the growth of the mountain system; igneous activity has actually been rather continuous from about 200 million years ago to the present day.) The lavas of this volcanic chain were extruded above sea level—clear evidence that the volcanoes developed astride continental crust. Activity along the Cretaceous arc reached maximum intensity about 50 to 60 million years ago with the invasion of the crust by massive amounts of magma. The invading magmas crystallized to form enormous batholiths that now lie exposed all along the western flanks of the western ranges.

The contemporaneous volcanoes, which must have towered far above sea level, have long since been stripped off by erosion, exposing the intrusive underpinnings of the volcanic arc: the Andean batholith. In some areas, such as southern Peru, erosion has not eaten as deeply, so that small batholiths are mantled by volcanic debris of similar age and composition. These small intrusive bodies are probably parts of the feeder pipes that supplied the volcanoes. In time the volcanic roofs of the batholiths will be further stripped away, leaving only the deep-seated intrusive foundation of the vanished Cretaceous volcanic arc exposed to view.

The crustal swelling that accompanied the invasion of magma produced major traumas in the easily deformed geosynclinal rocks of the eastern ranges. Upwelling magmas dilated the crust along the line of the Cretaceous arc. The geosynclinal rocks were alternately sloughed off and pushed aside by the growing orogenic welt and were crumpled and thrown up into the fold mountains of the eastern ranges. Enormous volumes of sediments were eroded from the flanking mountain ranges and dumped into the intervening altiplano basin; some places in the altiplano received up to seven kilometers of sediment in Cretaceous time alone.

It is significant that the Cretaceous arc lies on the continent side of the older Jurassic arc because it has usually been supposed that, as debris from the oceanic crust is stuffed under volcanic arcs, the trench and arc will migrate oceanward. Quite the contrary is true in the Andean region: during the course of Andean evolution the axis of the volcanic arc has marched ever inland, away from the sea. Even on a fine scale one finds successive intrusions of the Andean batholith shifted continentward. One can only speculate on the reasons for the eastward mi-



ANDEAN EVOLUTION is shown in views northwesterly from line *A–A′* in map on page 64. Late in Paleozoic period (*1*), some 250 million years ago, sedimentary deposits blanketed the present coastline, and there were no mountains. In Triassic and Jurassic times (*2*), about 200 million years ago, underthrusting of the lithosphere began, buckling the sedimentary rocks and pushing them upward

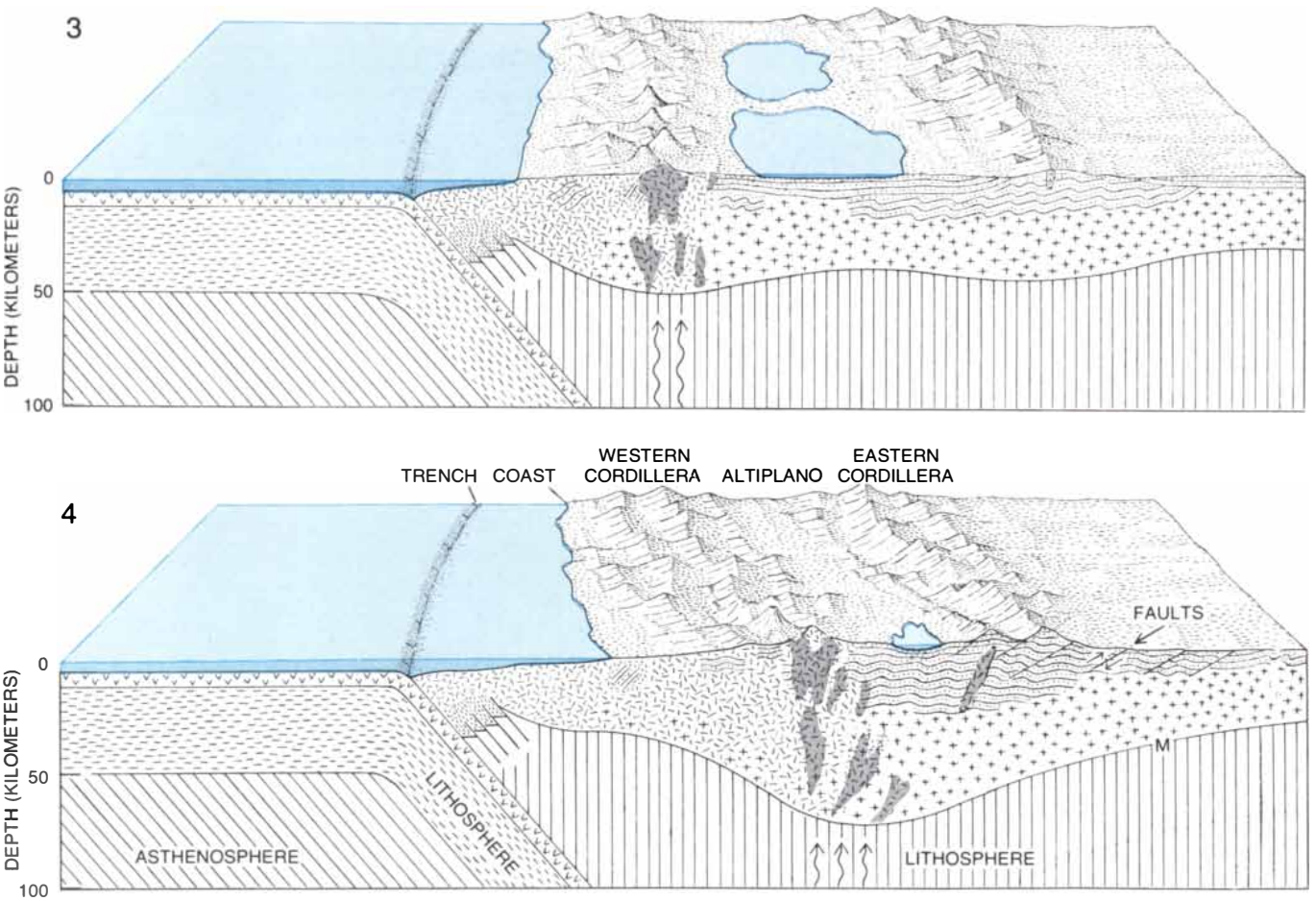
gration, but I shall return to this question when I consider the origins of the volumes of magma that have built the Andean crust.

The present-day Andean volcanic edifice began to emerge about 15 million years ago. In northern Chile and southern Peru huge volumes of silicic volcanic ash exploded out of fissures and flowed out from eruptive centers, eventually to blanket hundreds of thousands of square kilometers. Even now, when most of these ash deposits have long since been eroded away, it has been estimated that their remnants cover some 100,000 to 150,000 square kilometers to an average depth of 500 meters. The ash eruptions continued steadily until about four million years ago, when they ended abruptly and were followed closely by the outpouring of andesitic lavas from volcanic vents. These later lavas form the great stratovolcanoes, some still active, that dominate the Andean chain and rise to well over six kilometers in the central Andes.

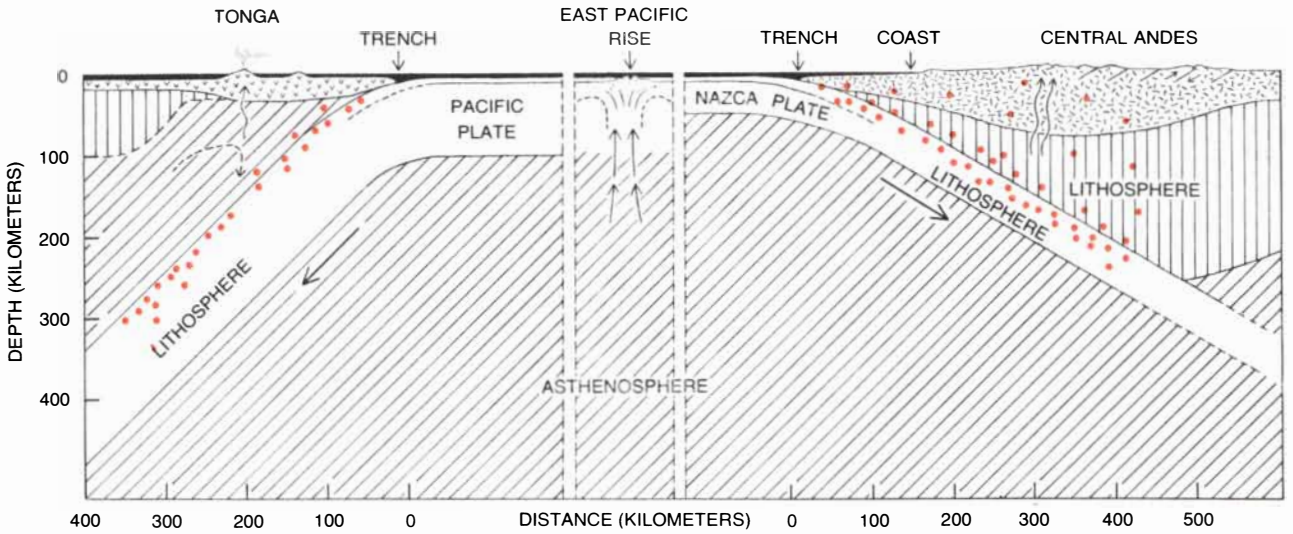
The massive influx of magma into the crust under the modern Andean volcanic crest swelled the crust and produced extensive folding and thrust-faulting in the altiplano and the eastern ranges. The geosynclinal sediments were further pushed aside by the expanding magmatic welt and the eastern ranges were jammed up to form narrow, high mountain chains. Sedimentary debris flooding in from both the eastern and the western range during Tertiary time piled up in the altiplano to thicknesses as great as 15 kilometers. We are now witnessing at least a temporary waning of activity, and through the relaxation of compressive stresses some extensional features have formed in the altiplano. Lake Titicaca is a notable site of one such feature: a graben, a long depression between fault lines.

So far I have considered Andean evolution from the point of view of surface geology, describing the inland march of the volcanic arc, the resulting progressive crumpling of the Paleozoic geosyncline and the gradual uplift of the Andean chain. I have not yet inquired in detail into the growth of the crust and the origin of the crustal rocks below the surface.

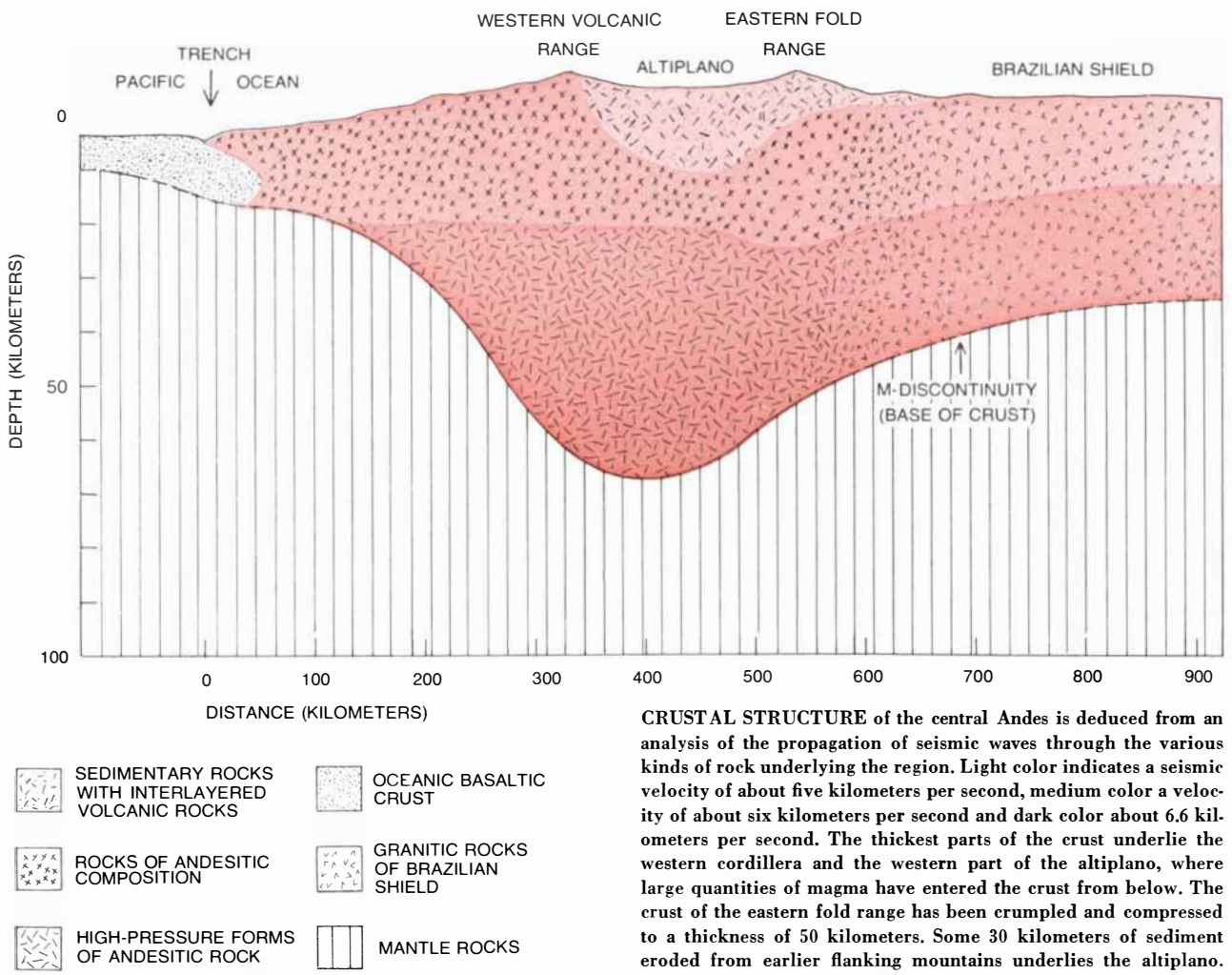
Classical concepts of mountain building, formulated long before the advent of plate tectonics, held that geosynclinal rocks formed in elongated basins of deep subsidence and that these basins were necessary precursors of later mountain belts, which were presumed to form through the deformation and melting of the thick geosynclinal sedimentary strata. The tectonic mechanics whereby these sedimentary rocks were melted and deformed was always a mystery, as was the origin of the subsiding basin itself. A number of problems are posed by classical concepts of geosynclines when they are applied to the central Andes. The crust under the volcanic crest is more than 70 kilometers thick, and yet no sedimentary basin can ever sit for long on crust more than about 35 or 40 kilo-



and eastward. Rising magma from the descending oceanic plate formed an arc of volcanoes in the coastal waters of western South America. Some batholiths, or bodies of intrusive igneous rock, formed in the sedimentary layers of the eastern ranges. In Cretaceous and early Cenozoic times (*3*), 100 to 60 million years ago, a second volcanic arc began to form eastward of the Jurassic arc. Upwelling magma swelled the crust, pushing aside the ancient sedimentary rocks, which crumpled to form the fold mountains of the eastern cordillera. Material eroded from these mountains poured into altiplano region. Formation of the present volcanic range began 10 to 15 million years ago, reaching by Pliocene or Pleistocene time (*4*), one or two million years ago, the present structure.



STRUCTURAL CONTRASTS between the Andean volcanic arc and the Tonga-Fiji island arc account for differences in physiographic settings and earthquake distributions. In the Tonga arc most earthquakes occur within the descending slab, which has mobile asthenosphere material above it and below it. There is also evidence of minor sea-floor spreading west of the arc. These observations suggest that drag from the sinking plate causes convective overturn (*arrows*) in the overlying asthenosphere, which in turn causes the spreading behind the arc. Earthquakes in the Andes are not confined to the descending plate, nor is there any evidence of crustal spreading east of the volcanic arc. Evidently the thick lithosphere under the leading edge of South America prevents convection and secondary spreading, and the forces generated by the descending slab produce earthquakes rather than convection above the slab. The intensity of earthquake activity under the Andes is indicated by the density of dots. Few dots show lighter activity.



CRUSTAL STRUCTURE of the central Andes is deduced from an analysis of the propagation of seismic waves through the various kinds of rock underlying the region. Light color indicates a seismic velocity of about five kilometers per second, medium color a velocity of about six kilometers per second and dark color about 6.6 kilometers per second. The thickest parts of the crust underlie the western cordillera and the western part of the altiplano, where large quantities of magma have entered the crust from below. The crust of the eastern fold range has been crumpled and compressed to a thickness of 50 kilometers. Some 30 kilometers of sediment eroded from earlier flanking mountains underlies the altiplano.

meters thick without bobbing above sea level and thus ending sedimentation.

If remobilized geosynclinal rocks are to make up the crust of the western cordillera, they must therefore have undergone twofold foreshortening to produce a doubling in crustal thickness. Yet in the volcanic chain we find few tectonic patterns such as thrust faults and tightly compressed folds that would suggest crustal shortening. On the contrary, studies of seismic earthquake mechanisms and modes of faulting indicate that the deformational style is extensional: the earth under the volcanic arc is swelling rather than contracting. Clearly knowledge of the Andean crust could not be easily mated with classical concepts of geosynclinal mountain building.

Only the advent of plate-tectonic theory has adjusted this unhappy mismatch between mountains and mountain-building theory, and it has done so by providing mountain-building mechanisms and processes that bear a clear relation to the Andean orogenic belt. It now appears to be evident (although some rearguard scientific battles still rage) that the geosynclinal sedimentary rocks are simply victims trapped at the continental margins and caught up in the tectonic processes that accompany subduction. The geosynclinal wedge must participate in, but can never be a prime mover of, the mountain-building process itself. If this analysis is correct, it follows that the magmas that built the central Andean volcanic arc were derived from below the crust and that the great volumes of rock that solidified from these magmas have produced crustal thickening and massive uplift of the arc.

We are left, then, with the question of the sources of magma below the crust. There are two principal possibilities: partial melting of the dense mantle rocks in the wedge between the crust and the top of the underthrust plate, or partial melting of the basalts and sediments of the oceanic crust riding down into the mantle on the lithospheric conveyor belt. Most observations argue against a mantle source for the andesitic rocks of the Andes. If the thickness of the crust in the volcanic arc had been doubled by the addition of magma from below, then some 20 percent partial melting of mantle rock between the crust and the subduction zone would be required to provide the necessary volume of additions to the crust. Experimental studies of the partial melting of probable mantle materials do not readily support the extraction of this quantity of andesitic magma from dense mantle rocks. A source in the mantle for the Andean rocks seems possible only if mantle flow provides a constant supply of rock undepleted in its low-melting-point fraction. Yet we have already seen that the upper mantle under the Andes is rigid lithosphere to a depth of from 200 to 300 kilometers, and therefore probably quite immobile.

An alternate proposition, that the magmas are derived from the descending oceanic crust, is currently most in favor and is a theory that I believe provides a viable explanation of Andean crustal genesis. Rocks of the oceanic crust have melting temperatures several hundred degrees Celsius below those of mantle rocks. As the cold oceanic crust descends into the mantle it is heated by the hot mantle in which it is enveloped, and at depths of about 100 to 150 kilometers the basaltic crust may well reach its melting point. The partly melted fraction of oceanic crust would be less dense than either the parent rock or the surrounding mantle rocks and would migrate upward into the crust. Most of the magma would be trapped and would solidify at the lower and intermediate levels of the crust; only a small fraction would ever make its way to the surface to come out through volcanoes, and so the volcanic pile on the earth's surface represents only a small part of the total increase in crustal volume.

A phenomenon related to the origins of the crustal rocks is the eastward migration of igneous activity. Assuming that the rocks originate through partial melting of the descending plate, there are a number of ways to explain the migration. If, for example, the depth in the earth at which melting occurs remains constant, then either a progressive decrease in the dip of the descending plate or a progressive continentward migration of the trench could cause an eastward migration of igneous activity. Or, if the position of the descending plate remains constant, the depth at which melting occurs on the slab may increase with time, thus pushing the source of magma deeper and eastward. A critical consideration here is the position and dip of the descending plate since the onset of subduction. Has the dip of the Benioff zone changed or has the trench migrated over the past 200 million years?

Examination of the chemistry of the rocks themselves provides a partial answer to this question. William R. Dickinson of Stanford University and Trevor Hatherton of the New Zealand Department of Scientific and Industrial Research have demonstrated a clear positive correlation between the amount of potash (K_2O) in modern volcanic rocks and the depth to the underlying Benioff zone. For particular rock types such as andesite collected from island arcs all over the world, the potash content is a more or less linear function of the depth to the underlying Benioff zone. By the inverse corollary of this relation one should be able to estimate the depth to defunct Benioff zones by measuring the potash content in older rocks.

We have just begun to apply this technique in the Andes. The data are by no means complete, but as we proceed from rocks of the Jurassic arc to those of the Cretaceous arc and on to those of the modern arc we find a progressive increase in potash content. This progressive increase, of course, is also correlated with increasing depth to the present Benioff zone. We find similar variations for strontium content, another chemical variable that has been shown to exhibit a positive correlation with the depth to the Benioff zone.

It is therefore possible to speculate with some basis in fact that the position of the Benioff zone has not changed greatly with time. If that is so, the eastward migration of the arc over the past 200 million years implies that the position in the mantle at which the oceanic crust melts has been pushed to progressively greater depths. It may well be that in the Andes the continued underthrusting of the cold oceanic slab into the mantle under the continent has gradually cooled the surrounding mantle rocks, so that the temperature at which melting occurs in the descending plate is attained only at greater and greater depths. If such a process accounts for the migration of igneous activity ever inland from the trench, then as long as subduction continues under the central Andes the volcanic chains of the future will impinge ever more on the old geosynclinal rocks, driving them even farther inland.

One day, however, Andean orogeny will end. The volcanic chain will stop growing and the inevitable destruction by erosion will set in. To reduce the Andes to sea level, some 35 kilometers of crust will have to be stripped away, exposing the center of the earth's crust. The sedimentary debris removed from the mountain edifice will be dumped on the continental shelf, the continental slope and the oceanic abyssal plain of the Pacific. There the sediments will pile up to await the next tectonic cycle, in which they will be crumpled, buckled upward and swept aside to give way to the next great volcanic mountain belt.

DUET-SINGING BIRDS

The male and female of certain tropical species join each other in remarkably precise song. The primary function of this behavior is to maintain close communication between the birds in dense foliage

by W. H. Thorpe

Newcomers to East Africa soon become familiar with a striking birdsong, heard sometimes in large gardens or parks but more frequently in open forest and bushy savanna country. The song is brief, about a couple of seconds in duration, but it is often repeated with great regularity over long periods. What makes the song particularly outstanding is that its few notes possess a flutelike or bell-like quality, exceedingly pleasant to the ear, that strikes almost every listener as being in some curious sense "musical." The singing bird is a shrike (*Laniarius aethiopicus*), known in some parts of its range as the bellbird or bell shrike. Obvious and conspicuous though its song is, many quite observant people have lived in East Africa for years without realizing that the performance comes from two singers rather than one. Not until the listener happens to get between the two birds does he realize that the first few notes of the song come from one direction and the rest from another, yet with an almost incredible precision of timing.

It is usual to speak of such a performance, with one member of a pair starting the song and the other completing it, as antiphonal song. Antiphonal song, however, is only one particular kind of what may in general be called duetting, a term that also includes polyphonic performances, when the two birds sing at the same time, each coordinating its individual song pattern with the other. Occasionally each bird may sing exactly the same pattern of notes at the same moment, that is, in unison. Alternatively, the two contributions may be different in pattern but overlap each other, again with precise coordination [*see bottom illustration on page* 72].

Shrikes of the genus *Laniarius,* which are found only in Africa, provide the most striking examples of duetting. It appears that all 15 species of the genus, ranging throughout tropical Africa, exhibit the behavior. This mode of mutual singing between a paired male and female is not, however, confined to shrikes. It is well developed in eight or nine other families of birds in different parts of the world, nearly all of them tropical in range. Such behavior seems in fact to be most characteristic of species that live in very dense tropical vegetation, where it must be difficult for the two members of the pair to keep in sight of each other. In such a setting the birds would find a vocal pair bond useful, if not essential. Duetting also appears to be characteristic of species in which the male and female are of identical appearance, remain paired for life and maintain their territories for the greater part of the year, perhaps for their entire lives.

My colleagues and I first studied duetting in the field in Kenya and Uganda during the years from 1962 to 1967. We were able to extend the work greatly with studies of captive birds in large tropical aviaries in England from 1964 to 1970. In our field studies we concentrated on three populations of bell shrikes in widely separated areas: on the shores of Lake Nakuru in Kenya, in southwestern Uganda about 400 miles to the west of the lake, and near Kapenguria in Kenya, 150 miles northwest of the lake. Our aim was to record as far as possible all the main duet patterns in the area, to plot territories where possible, and again where possible to identify individual birds by marking them with colored rings.

In the Lake Nakuru area we recorded 102 different duet patterns, in the Uganda area we had 22 examples and in the Kapenguria region 24. The tape recordings were analyzed by sound spectrograph, but because of the very pure tonal quality of the bell shrike's notes we found it much more satisfactory to represent the songs by simple musical notation rather than by sound spectrograms. The songs can be almost completely specified by the pitch of the notes, by the intervals (the difference in pitch between any two notes), by whether the sounds come in harmonic intervals (simultaneously) or in melodic intervals (successively) and by the duration of the notes, their timing and their overall pattern.

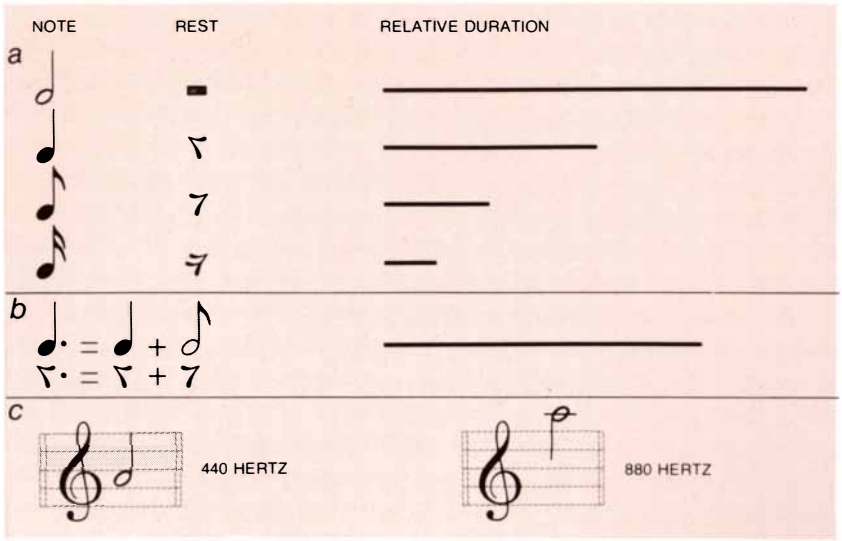
Both absolute and relative pitch are of great importance in the recognition of the duets, and these features can only be roughly assumed from sound-spectrographic analyses. Fortunately standard musical notation has been developed over the centuries for the specific purpose of communicating details of pitch and time. It is an elegant and foolproof method for the purpose. On the other hand, sound-spectrographic and other electronic methods of analysis give much information about acoustic structure (tonal quality), relative intensity (loudness) and the minutiae of timing (which are particularly needed in assessing response times). Hence for these purposes vocalizations are better portrayed by such methods.

In the Lake Nakuru and Uganda field areas the vegetation was so dense that it was often extremely difficult, if not impossible, to map the paired birds' territorial boundaries. Birds in the Kapenguria area were living under more open conditions, so that here the mapping of the territories was much easier. In one of the Kapenguria areas four pairs of birds were holding territories, and characteristic song patterns were recorded for each [*see illustration on page* 74]. It will be seen that certain series of notes and certain patterns are fairly general to the species. Indeed, one quickly repeated series of low notes (either G or G-sharp)

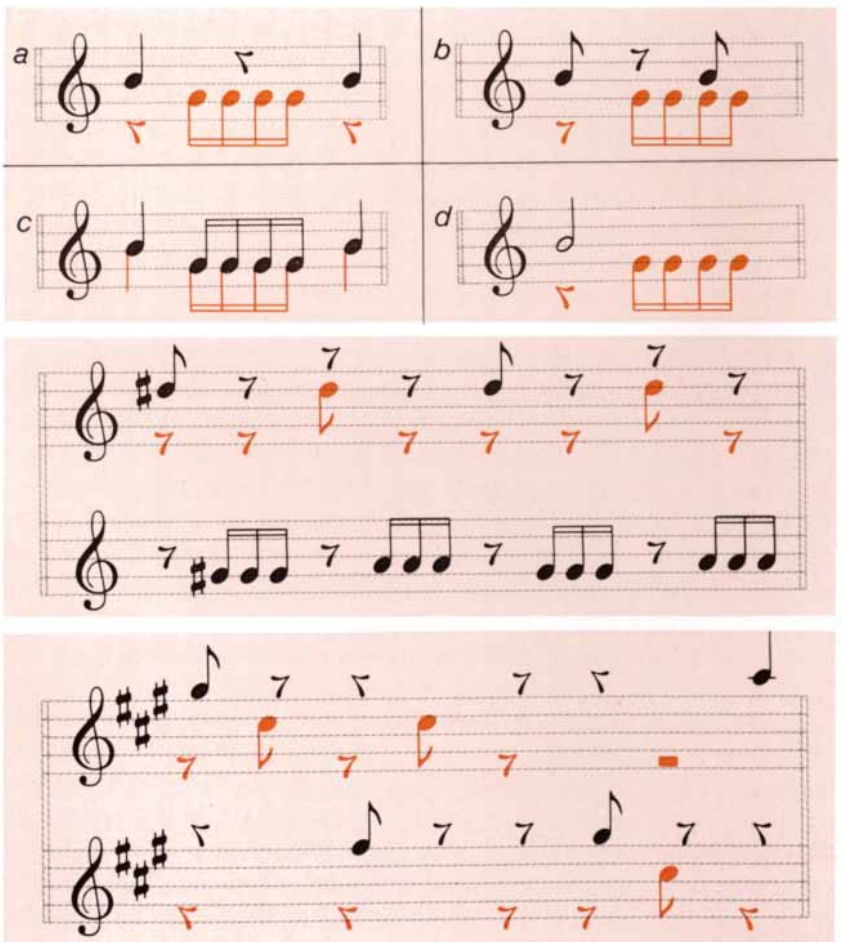
DUET SINGERS, a paired male and female bell shrike, are shown in an untypically open environment. Also known as the bellbird, the bell shrike is an African species belonging to the genus *Laniarius*. The paintings on this page were made by David Bygott.

POLYPHONIC SINGER that also maintains contact with its mate by singing duets, the white-browed robin chat (*Cossypha heuglini*) is found in the thick forest undergrowth and scrub of East Africa.

UNMUSICAL SINGER, the black-headed gonolek is another bird of the genus *Laniarius*. Of the 15 African shrike species in the genus few are as musical as the bell shrike, but they all sing duets.



KEY TO STANDARD NOTATION indicates (*a*) the relative duration (*right*) of a half note and half rest compared with successively briefer intervals of sound or silence: quarter, eighth and 16th notes and rests respectively. The presence of a dot after either a note or a rest (*b*) increases the value of the symbol by one-half. Two staffs with G clefs are shown (*c*). The note on the staff at the left is a "tuning" A (440 hertz). The note on the staff at the right is one octave higher (880 hertz). The birdsongs are written one octave below true pitch.



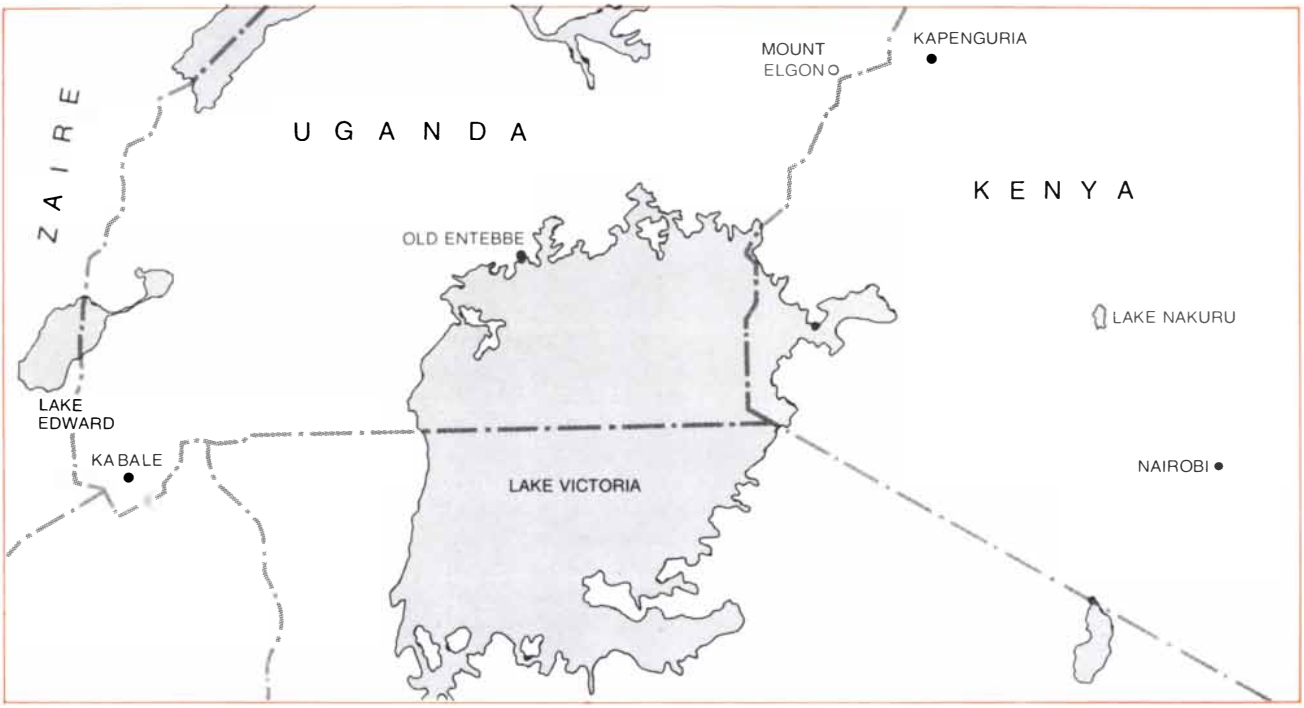
VARIETIES OF SONG include duets (*top*) that may be antiphonal (*a*) or polyphonic. The notes and rests of one singer appear in black and those of the other singer in color. Polyphonic duets (*b through d*) can include phrases sung in unison (*c*) or phrases that overlap (*d*). Not all bell shrike songs are duets. Shown here is a trio (*middle*) recorded at Lake Nakuru in Kenya and a quartet (*bottom*) by captive shrikes at the University of Cambridge.

has a mellow sound that has given rise to another common name for the bell shrike: the tropical boubou shrike. Another clearly characteristic feature of the species is its tendency to produce intervals between successive notes of about an octave, either an octave below or an octave above. The species can be recognized by these features wherever it is found throughout the thousands of square miles of its range. Yet the more individual pairs are studied, the clearer it becomes that each pair of bell shrikes works out its own particular repertoire of duet patterns, so that the repertoire tends to be unique to that pair. Once one knows it well enough, it provides a means for individual identification.

Quite early in our studies we came on examples of trio singing and sometimes even quartet and quintet singing. One way that trio singing arises is through the intervention of one bird of a pair in a neighboring territory. (We were seldom able to determine whether it was the male or the female.) The "outside" singer would interpose its notes between those of the pair in "home" territory in an ordered manner and with extremely precise timing. The home pair in turn would sometimes minutely adjust the pattern and timing of its own duet to allow the neighbor to participate [*see bottom illustration at left*]. On one or two occasions the mate of the outside bird was seen standing by but not taking part. Trios, and perhaps quartets, can also be formed when one or two grown offspring of a pair, still residing within the parental territory or on its margins, join in with their parents' songs.

Apart from certain snarling or buzzing sounds used as alarm notes, vocalizations by bell shrikes in the wild under normal conditions consist of antiphonal singing between the members of a mated pair. The male bird is usually, but not necessarily, the leader. When the female starts the duet, she usually does so with a particular snarling note that may then be incorporated into the overall pattern of musical notes. In other species of shrike (and in bell shrikes in captivity) it may be much more usual for the female to start the duet. We have good evidence that the vocal repertoire is worked out and developed between the two members of a mated pair. Indeed, we have noted that when a bird is isolated by some mischance in the wild (as in a Nairobi garden where one bell shrike was kept under observation for nearly six months), it appears unable to produce any complex pattern of vocalization.

In this connection we found that



FIELD STUDIES of duetting birds were conducted at two areas in Kenya: Lake Nakuru, northwest of Nairobi, and Kapenguria, east of Mount Elgon. The third area was near Kabale, southeast of Lake Edward in Uganda. In all 148 different duet patterns were recorded.

where the birds seemed to be crowded in the wild the song patterns tend to be more elaborate than in areas where the birds have more space. This field observation was confirmed by studies of birds in captivity. When such birds, on arrival at our aviaries at the University of Cambridge, were all in cages in one room but were unable to see one another, trios and quartets of remarkable elaboration would quickly be built up, would be sung for a few hours or a day or two and then would be abandoned. When the birds were put out as pairs in the aviaries and therefore were no longer in close contact with other pairs, the duet patterns usually regressed to rather simple forms.

On the basis of both the field data and our aviary studies we can list the function of duetting under four headings in decreasing order of importance. The most important is the singer's recognition of, location of and maintenance of contact with its mate. The next is mutual stimulation between the two birds that form a pair, as a part of (or a substitute for) the ordinary methods of visual display. The third is an aggressive maintenance of the pair's territorial integrity. Last and least is mutual reassurance after some disturbance.

With respect to the first two functions it is important to realize that in the earth's temperate zones seasonal changes in the length of the day provide the most important of the cues that initiate the secretion of birds' sex hormones and so bring potential pairs into the breeding condition. In the Tropics the day-length cue is very slight, if not completely absent. Moreover, other possible "seasonal" cues (such as variations in humidity, rainfall, degree of cloud cover and the like) tend to be unpredictable and to give little advance warning. It follows that if birds in the Tropics are to take full advantage of the time when breeding conditions are at an optimum, paired males and females must be in constant contact so that their behavior and their reproductive cycles are fully coordinated.

Our aviary experiments yielded many interesting facts about duetting. To mention a few, studies in the aviaries confirmed our field observation that, if one member of a pair is absent, the bird remaining in the territory tends to sing the whole duet pattern by itself: both its own contribution and its mate's. When the mate returns, however, a period of unison singing is not uncommon. The duet in unison will last for a few seconds; thereafter the pair resumes its antiphonal song as before. These shrike studies evidently confirm an argument put forward by Konrad Lorenz to the effect that "whole duet" vocalization in a partner's absence may be intended to secure the partner's return. It is almost as if one bird were using the characteristic vocal performance of the other as a "name" that might serve as a recall.

A series of separation experiments in the aviaries yielded these conclusions. First, separation leads to an increase in the amount of vocalization by the deserted bird. In contrast, a bird that is moved to a new "territory" tends to decrease its vocalization. Second, if the bird remaining in the territory is a male, he employs all his usual vocalizations and in addition some of those of his missing partner. Third, if two members of a pair are separated for a long period, there is sooner or later a tendency to show a regression of vocalizations that can reduce the performance to what amounts to a juvenile condition.

If an isolated male hears a playback of his mate's voice, he will reply with the appropriate item in his own repertoire. The male is much less likely to respond to a playback of his own voice, however, and does so only by repeating the song that has just been played. This finding during our aviary studies confirmed still another conclusion based on fieldwork: even though notes of the bell shrike are very uniform in acoustic structure, they may be sufficiently different to be recognized on the basis of their vocal quality as well as by the pattern they form.

Our aviary experiments provided further evidence of this kind of recognizability. We found that a female bell shrike in the aviary will answer such

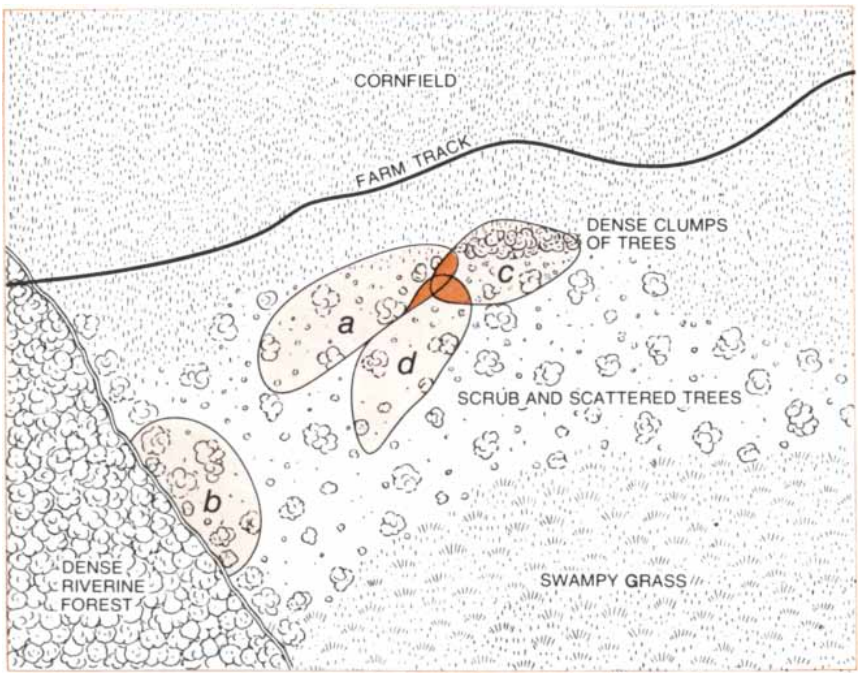
notes of her own male as she may be able to hear but will not answer the notes of any neighboring pairs. In one set of experiments an extra male was kept in association with a duetting pair. The outside male was never heard to vocalize until the resident male was removed. When the removal took place, however, it became clear that the outside male had at least in part learned not only his rival male's repertoire but also that of the female. The resident female would respond to the educated outside male's song although she would not respond to the songs of strangers.

A remarkable result emerged from our difficulty in identifying the sex of captive birds. Because the male and female are exactly alike in appearance we made occasional mistakes in our caging, inadvertently constructing male-male and female-female "pairs." It sometimes happened that for weeks or months our mistakes had surprisingly little effect on the birds' behavior and vocalization. And so it became clear to us that two bell shrikes of the same sex, either male or female, can behave in a way that is indistinguishable from the behavior of a true pair. They may engage in mutual duetting, picking up and carrying nesting material, mutual preening, begging and even some forms of display.

The precision in timing in the bell shrike's duets is excellent. It was easier to study this, however, in another species of the same genus: the black-headed gonolek (*Laniarius erythrogaster*). The general build and pattern of the gonolek is very similar to the bell shrike's except that its underparts, instead of being white suffused with pink, are a brilliant crimson [*see bottom illustration at right on page 71*]. Throughout its range the gonolek is a bird of the thickest bush and undergrowth. We studied it both in Kenya and Uganda and in captivity. Its duet is unmusical and extremely simple. The initiating bird, in this species almost always a male, has a "yoick"-like note. This note is immediately followed by a tearing hiss, sometimes suggesting the ripping of cloth and sometimes being more like a sharp sneeze, from the female. Here, as with the bell shrike, the timing is so perfect as to make the duet sound like the song of one bird.

The timing of the gonolek is easy to investigate because the onset of the second bird's note is extremely sharp, giving the investigator an exact point of measurement. We recorded a consecutive series of eight duets of a pair of these birds at Old Entebbe in Uganda. The second bird, which was completely out of sight of the first, took its time cue with extraordinary precision from the start of the first bird's note. The mean response time of the female in this series of duets was little more than 144 milliseconds, with a standard deviation of 12.6 milliseconds [*see illustration on page 76*]. In another series of seven consecutive duets the response time was much longer (425 milliseconds) but the standard deviation was even less (4.9 milliseconds). It is obvious that the species must have an extremely precise time sense; the accuracy does not decrease even if the response time is extended by a factor of four. I am not aware of any auditory reaction time in humans that has a standard deviation of less than 20 milliseconds.

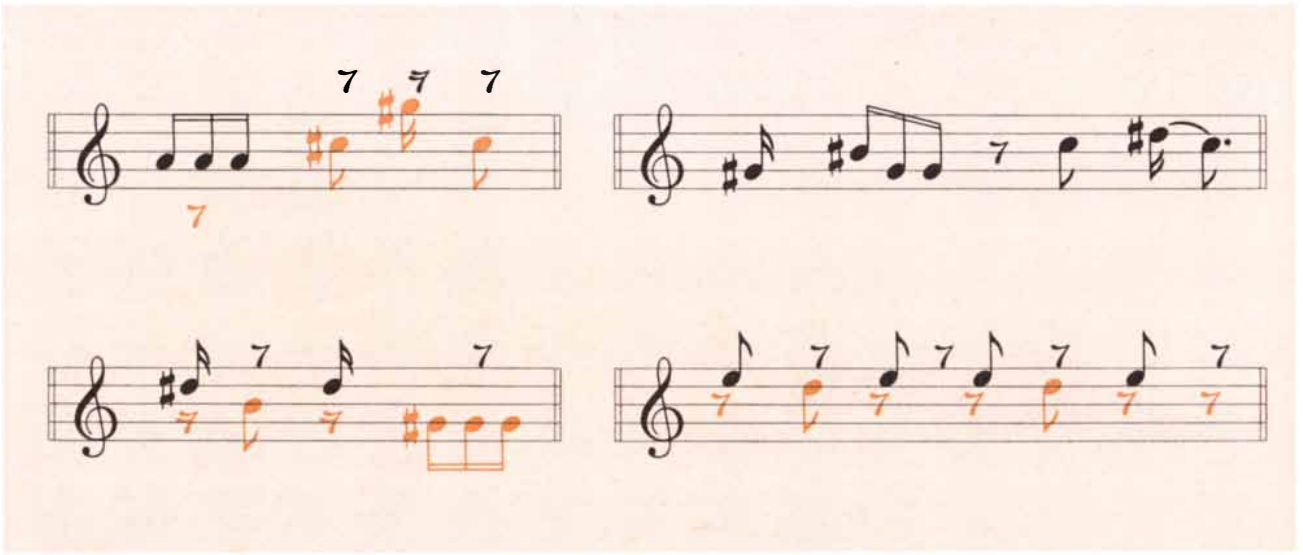
Even more unusual is the performance of another African duetting species, a member of the genus *Cisticola* (the grass warblers) known as Chubb's cisticola (*Cisticola chubbi*). Chubb's cisticola is a small streaky-brown bird that inhabits long grass in bush-clad clearings at altitudes of between 5,000 and 8,000 feet. A series of six consecutive duets that I recorded, the two birds in this instance being in sight of each other, had a mean response time of 396 milliseconds and the remarkably small standard deviation of 2.9 milliseconds. That is about an eighth of the error a man would make under similar circumstances.



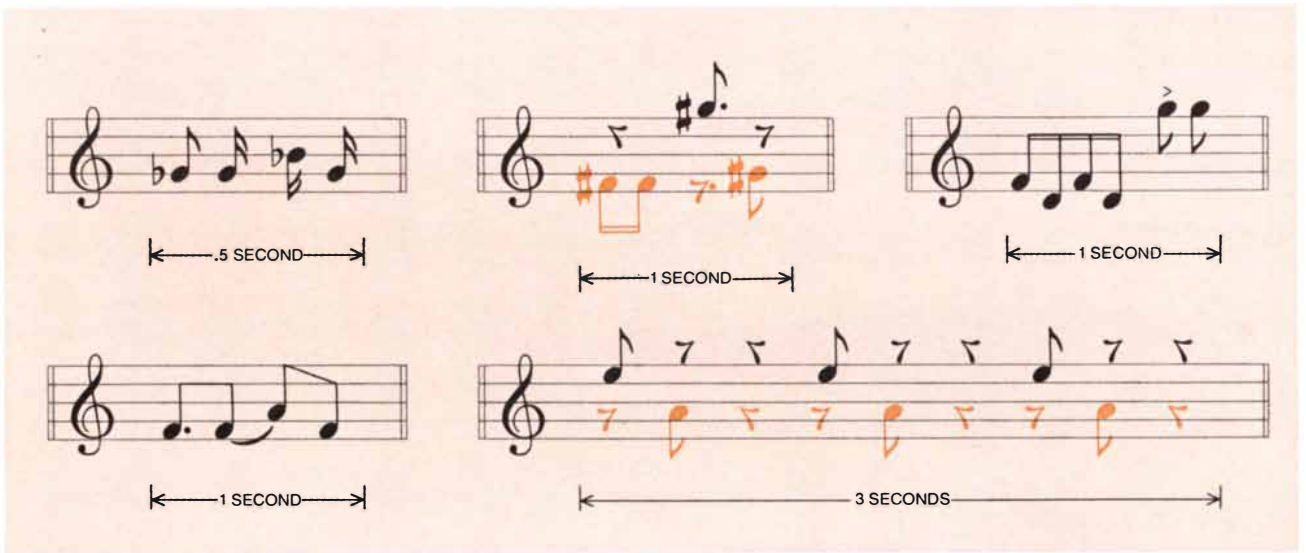
DENSE GROWTH along the fringe of a Kenya cornfield, backed by scrub, scattered trees and a thickly forested stream, contained four bell shrike breeding territories, three of them overlapping (*solid color*). Territories *a* and *b* contained paired adults only; the adult pairs in *c* and *d* shared the territory with juvenile birds. Both the territory overlap and the number of birds that were present led to frequent singing in trios, quartets and even quintets.

Is there any scientific basis for the impression of "musicality" the duets of the bell shrike so strongly suggest? The main investigator of this topic is Joan Hall-Craggs. As she has stated, it is useful to begin any such inquiry with the reasonable assumption that, since a bird's ear is similar to the human ear in its essential structure, it displays many of the same characteristics. At the same time we should remember that the perception of pitch probably begins at higher frequencies among most birds than among humans. The shortest time for the identification of pitch by man is approximately .05 second, assuming a signal in the middle range of frequency, but for birds it is quite likely that this time may be much reduced. It has also been calculated that the pitch of a tone is detectable when 70 percent of the energy in the spectrum lies within ±5 percent of the principal frequency. Such energy concentration is clearly discernible in the sound spectrograms of bell shrike songs, and most of the notes are of sufficient duration and concentration of sound energy to enable us to assess their pitch by ear.

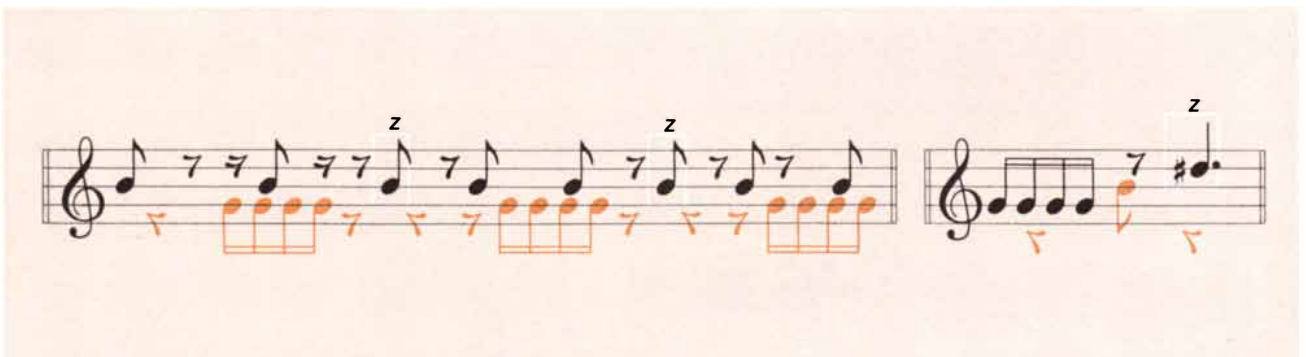
Musical form, at its simplest, consists



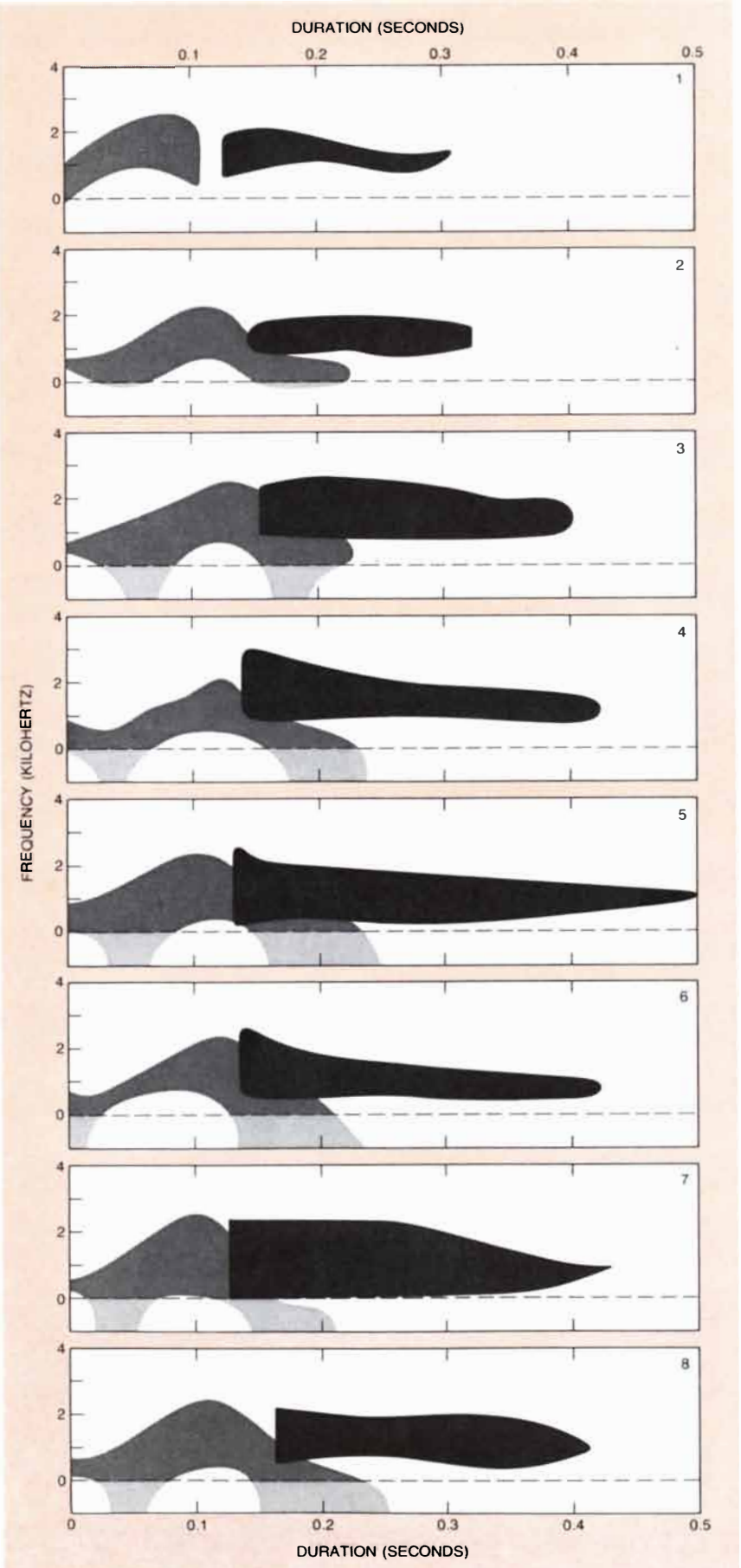
FOUR DUETS recorded in the Lake Nakuru area exhibit distinctly different patterns. In instances when the observer could not tell which notes were sung by which bird the notes and rests appear in black only. Otherwise the second singer's part is shown in color.



FIVE DUETS recorded in the Uganda study area provide a further example of differing bell shrike song patterns. In the third song, where the individual singers' contributions are unknown, an accent over the first G indicates that the singer put stress on the note.



TWO TRIOS were recorded in the Kapenguria study area shown on the facing page, where such singing was a frequent event. The notes of the first two singers appear in black and in color respectively; the notes of *z*, the third singer, are outlined in white.



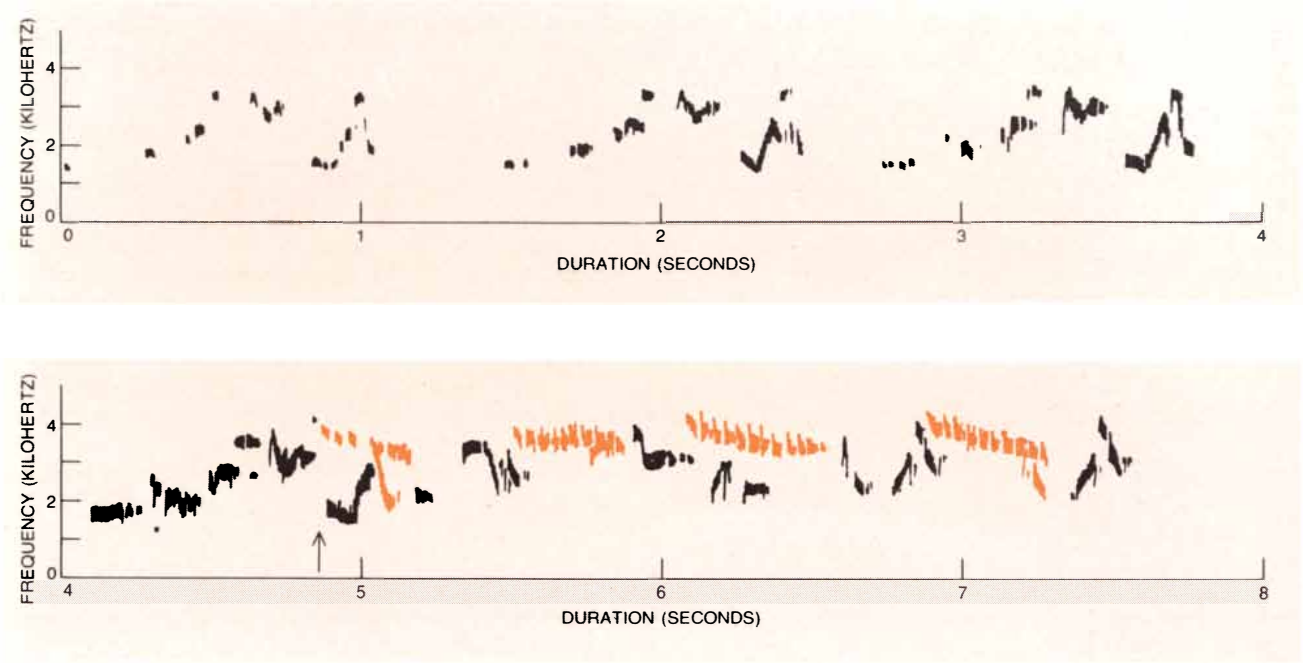
in the grouping together of units of sound energy in repeatable and consequently recognizable patterns of time and pitch. As Hermann von Helmholtz put it more than a century ago: "Melodic motion is change of pitch in time." It is obvious at a glance that the shrike songs epitomize melodic motion and therefore have musical form. To analyze that form one must first describe the intervals: the difference in pitch between any two tones.

We undertook such a study, using the duets of 10 pairs of one population of bell shrikes. The tones and intervals of the duets were subjected to physical analysis by taking single sounds from the field tapes and recording them on a sound spectrograph, which was then set to play the sound repeatedly at 2.5-second intervals. That signal was then matched, initially by ear, to a pure tone from a signal generator. The two tones were next matched precisely on a cathode ray oscilloscope by tuning the generated tone until a clear Lissajous figure with a ratio of 1 : 1 appeared repeatedly on the screen. (More recently we have employed a new instrument for this work. It is the "Melograph," designed and built at the University of Uppsala.) A frequency counter then gave a reading in hertz (cycles per second).

The absolute frequencies, as distinct from the intervals between them, were measured. It was found that the duets of all these birds fell within the range of 656 to 2,064 hertz. That is about one octave plus a sixth. The range of an individual pair varied from as little as a minor seventh (787 to 1,405 hertz) to a little less than a perfect 11th, that is, an octave plus a perfect fourth (798 to 2,064 hertz). The paired birds normally use only portions of the available frequency spectrum. In pairs where this uniformity is most marked a low band of sung frequencies is centered around 800 hertz, a middle band around 1,000 hertz and a smaller high band just below 1,200 to 1,220 hertz.

It was clear in almost every instance that the gaps of unused frequencies span

RAPID RESPONSE of female black-headed gonolek in eight successive duets is shown in the series of sound spectrograms at left. The song of the male is shown in gray; the female's response is in black. The series of duets was completed in some 30 seconds; the average response delay was .144 second. The apparent subzero frequencies are due to distortion and interference below 50 hertz.



CRESCENDO DUET of the white-browed robin chat, some 7.5 seconds in duration, is seen in a sound spectrogram. During the first part of the duet the male utters four five-syllable phrases that progressively increase in amplitude. The female joins in (*arrow*), continuing the crescendo, whereupon the male mutes his higher notes but continues to accompany the female in a low-pitched song.

wider ranges in the available spectrum than the bands of sung notes do, giving the impression that the birds are using a series of notes resembling a gapped scale. There is not sufficient consistency, however, to postulate such a scale, nor is there good reason to suppose that the birds are unable to sing notes in the unused section of the spectrum. The distribution of the used frequencies suggests that the syrinx, the sound-producing organ of the bird, may be such that these notes are produced more easily and/or to greater effect than such notes as might have been expected to be present in the bands of unused frequency. The tendency in most pairs of birds for the notes to cluster around 800, 1,000 and 1,200 hertz, however, suggests that the birds are producing most readily the fourth, fifth and sixth harmonics of a fundamental at about 200 hertz. These tones, derived from the lower reaches of the harmonic series, give rise to the ordinary major triad approximately on G in octaves five to six. If indeed the birds are utilizing the harmonics in this manner, it goes part of the way toward explaining their decided proclivity for singing duets that incorporate, or are based on, a major or minor triad.

In music if two tones or frequencies are adjacent, then the interval is described as "melodic." When, as often happens in bell shrike duets, two sounds occur synchronously, or when one tone is sustained while another tone is sounded above or below it, the interval is termed harmonic. All 15 species of the genus *Laniarius* sing antiphonally, but the bell shrike and one other species I shall not discuss here appear to be unique among them in singing melodic lines that are sometimes antiphonal and sometimes polyphonic and sometimes even sung in unison.

According to musical theory, the sounding together of two notes may lead to a predominant use of consonant intervals in two-part or multipart singing. Helmholtz, who investigated the properties of consonance and dissonance from the physical point of view in the 1860's, showed that the "roughness" of dissonance that we (and presumably birds as well) experience is based on a physical phenomenon, namely "beats": periodic variations in the amplitude of the sound pressure due to the interference of two sound waves of different frequencies. According to Helmholtz, beats are maximally disturbing to man at about 33 per second. It might be expected, however, that with birds this critical figure would be substantially increased because of their faster identification of pitch.

Now, the original physiological explanation of beats was that two tones sounding together forced into vibration overlapping regions of the basilar membrane of the ear. This explanation, which was once much in doubt, has now been largely rehabilitated, although no completely satisfactory physiological description of the phenomenon has yet been devised. This does not, however, affect the reality of the phenomena of dissonance and beats. Helmholtz drew a curve illustrating the degree of consonance and dissonance of intervals within the octave. If we draw this curve to the same scale as a curve that shows the incidence of these intervals as they are used by the bell shrikes, we find close agreement between them.

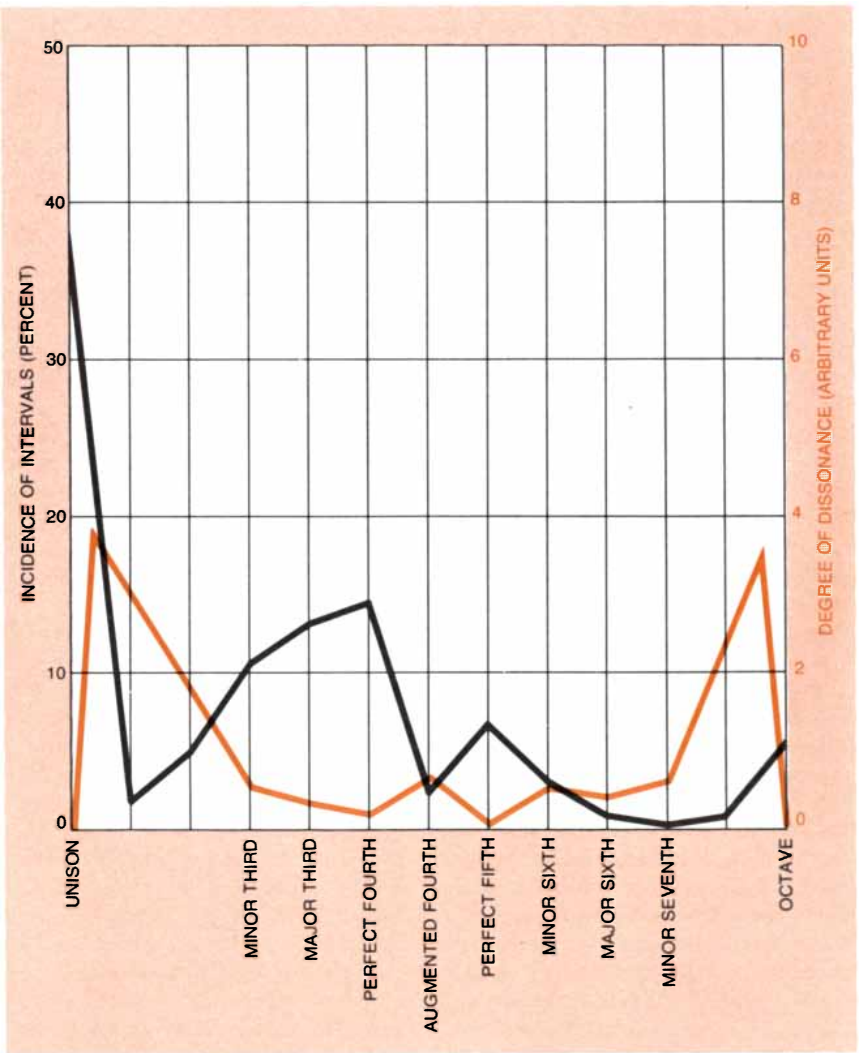
The shrikes' predisposition to sing consonant intervals is fully demonstrated by both aural and physical analyses. But since there are dissonant intervals as well, although in much smaller numbers, it cannot be argued that these birds are compelled by the structure of their syrinx to sing consonant intervals. It may be, however, that once such intervals are learned they are found to be functionally the best. It can also be argued that good, steady consonances might well assist in the effectiveness of the duets as a contact-maintenance system, particularly if it is important for the distance to be judged accurately.

Hence we can conclude the topic of bell shrike aesthetics by saying that the

apparent musicality of the songs of these species depends primarily on the birds' having a hearing apparatus that responds in the same way to the roughness of dissonance as our own does. To quote Joan Hall-Craggs: "To the musical listener these songs may seem overharmonious; nevertheless, it is the kind of harmony to which man aspired and which probably reached its peak in Mozart. No musical listener could call the songs 'unmusical' or 'displeasing'; their only fault from our point of view rests in their brevity and simplicity."

Two other examples of antiphonal singing among East African birds are particularly interesting when they are compared with the songs of the shrikes because they illustrate still other aspects of duetting. One of the birds is a member of the genus *Cossypha* (family Turdidae): the white-browed robin chat *Cossypha heuglini*. It is the sole known duettist among the 15 species of the genus.

Here again the sexes look alike. The birds are found in dense vegetation, in riverine forest and secondary scrub whether in farmland or garden, and from the East African coast up to an altitude of 6,000 feet. They are excessively shy, spending most of the time hidden in the undergrowth, feeding on the ground and usually singing at dawn and dusk from a perch in a low bush. The song is a long one, lasting five seconds or more, and starts with male solo phrases showing a



CONSONANT INTERVALS are characteristic of the bell shrike's songs. The black line indicates the intervals most frequently sung among selected semitones within the octave, expressed as percentages of the total number of intervals in the sample. The colored line is a modification of Hermann von Helmholtz' curve of consonance and dissonance between a note of fixed frequency and another note that changes smoothly from the same frequency to a frequency one octave higher. The closer the curve is to the base line, the greater the consonance is. Most of the intervals in the songs coincide with minimally dissonant frequencies.

gradual crescendo. When the loudness reaches a critical level, it provides the signal for the female to join in. The birds then proceed together, continuing the crescendo and with the pitch steadily rising, although after the female takes part the male tends to cut out his higher notes, giving a lower-pitched accompaniment to the female's downward glissandos [*see illustration on page 77*].

The grass warblers (genus *Cisticola*) are a huge assemblage: 40 species and 153 races. Of these it seems likely that only four species are duettists, yet they are duettists of the highest precision. As I have mentioned, I have recorded the duets of Chubb's cisticola. One other species, *Cisticola nigriloris,* which is found in the highlands of northern Malawi and southern Tanzania, is a singer of special distinction. It is a persistent duettist, and the pair's usual theme is a very high-pitched four-note whistle: G (at 3,240 hertz), E-flat (near 2,568), E-flat again (but at 5,000) and B (at 4,064), all with a continuous squeaking or croaking accompaniment. In two separate recordings of this species, made in areas of Tanzania more than 100 miles apart, the entire four-note phrase is suddenly transposed to another key after the first half-dozen or so bouts. This suggests the interesting possibility that the musical transposition of the song is a species-specific character! If that is true, it is, as far as I am aware, unique.

To sum up, we can say that duetting clearly plays a very important part in the signal system between male and female in a large number of bird species, in particular species that inhabit tropical regions. These elaborate song patterns show many interesting features, of which only a few have been discussed here. Perhaps the most interesting result of our investigations of duetting is the light cast on the heretofore little appreciated precision and synthesizing power of avian aural perception, the great precision of response time and the equally great exactness of control of the vocal organs. The use of these vocal powers for individual recognition is in line with observations made over the past decade by Beat Tschanz of the University of Bern, by C. G. Beer of Rutgers University and by other investigators. Their work has shown that in many colonial nesting birds (for example auks, terns, gulls, gannets and penguins) brief calls of a half-second duration or less can have enough acoustic detail not only to serve as labels identifying the calling species but also to label the individual caller.

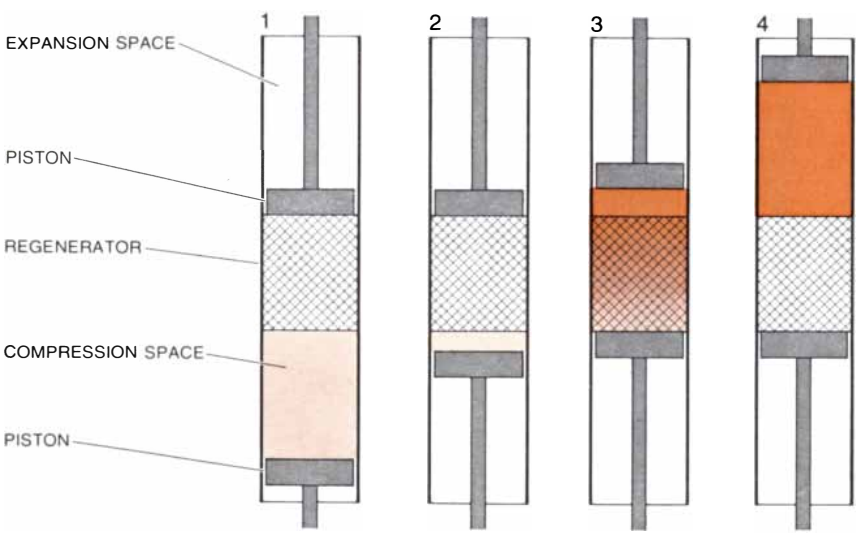
THE STIRLING ENGINE

This external-combustion engine contributes little or nothing to air pollution and is exceptionally quiet. Its virtues make it a possible alternative to the internal-combustion engine

by Graham Walker

The purpose of any engine or motor is to convert some form of energy into mechanical work. Since the beginning of this century the internal-combustion engine has played a major role in such energy conversion because it can be mass-produced inexpensively in a large range of sizes and serves well in a wide variety of applications. Now, however, it is coming under criticism because of its substantial contribution to air pollution and, to a lesser extent, to noise pollution. Conditions appear to be more favorable than they once were for the development of alternative types of engine.

One of several possible alternatives is the Stirling engine, which was invented in 1816 by the Reverend Robert Stirling of the Church of Scotland, attained a measure of commercial standing late in the 19th century and then was sent into limbo by the internal-combustion engine. A notable advantage of the Stirling engine is that it runs on external combustion, which can be designed to proceed with high efficiency and therefore to contribute little or nothing to air pollution. In a time of rising concern over shortages of petroleum fuels it is also noteworthy that the Stirling engine will operate on any fuel, including sunlight. Moreover, it is an exceptionally quiet engine. In addition to these advantages the potential of the engine has been increased in recent years by improvements in materials and by the developmental efforts of several companies. Now the engine can be envisioned as a prime mover for land and marine vehicles, as a means of low-temperature refrigeration and (in smaller forms) as a means of generating power at remote unattended sites and as the source of energy for an artificial heart.



IDEAL STIRLING CYCLE proceeds in four overlapping steps in a closed cylinder containing two pistons and a regenerator, which is a heat exchanger. The compression space is kept at a low temperature by an external cooling system, and the expansion space is kept at a high temperature by an external source of heat. The working fluid is air (*color*), which rises in temperature when it is compressed and cools when it expands. As a cycle begins (*1 and 2*) the air is compressed by the compression piston and moves into the expansion space (*2 and 3*), absorbing heat that has been stored in the regenerator during the preceding cycle. Work is done during expansion (*3 and 4*). The movement of the pistons and the gas is then reversed (*4 and 1*) with the aid of heat supplied from the external combustion.

The physical principles involved in the operation of a Stirling engine can be seen in an idealized way by visualizing a closed cylinder containing two opposed pistons with a regenerator between them [*see illustration on this page*]. The regenerator is a matrix of finely divided metal wires or strips; it can be regarded as a thermodynamic sponge, alternately accepting and releasing heat. One of the two volumes between the regenerator and the pistons is called the expansion space and is maintained at a high temperature by an external source of heat. The other volume is called the compression space and is maintained at a low temperature by a cooling system. The working fluid in the cylinder is air or some other gas. A full cycle of operation consists of four overlapping parts.

A convenient way to follow the cycle is to assume that the piston in the compression space is at its outer dead point (the outer end of its travel back and forth in the cylinder) and that the piston in the expansion space is at its inner dead point, close to the regenerator. All the working fluid is in the cold compression space; the gas is at its maximum volume and therefore at its lowest pressure. During compression (parts one and two of the cycle) the compression piston moves toward the regenerator, compressing the working fluid. Since the temperature of a gas rises as the gas is compressed, heat must be abstracted from the working fluid in order to maintain the temperature in the compression space at a con-

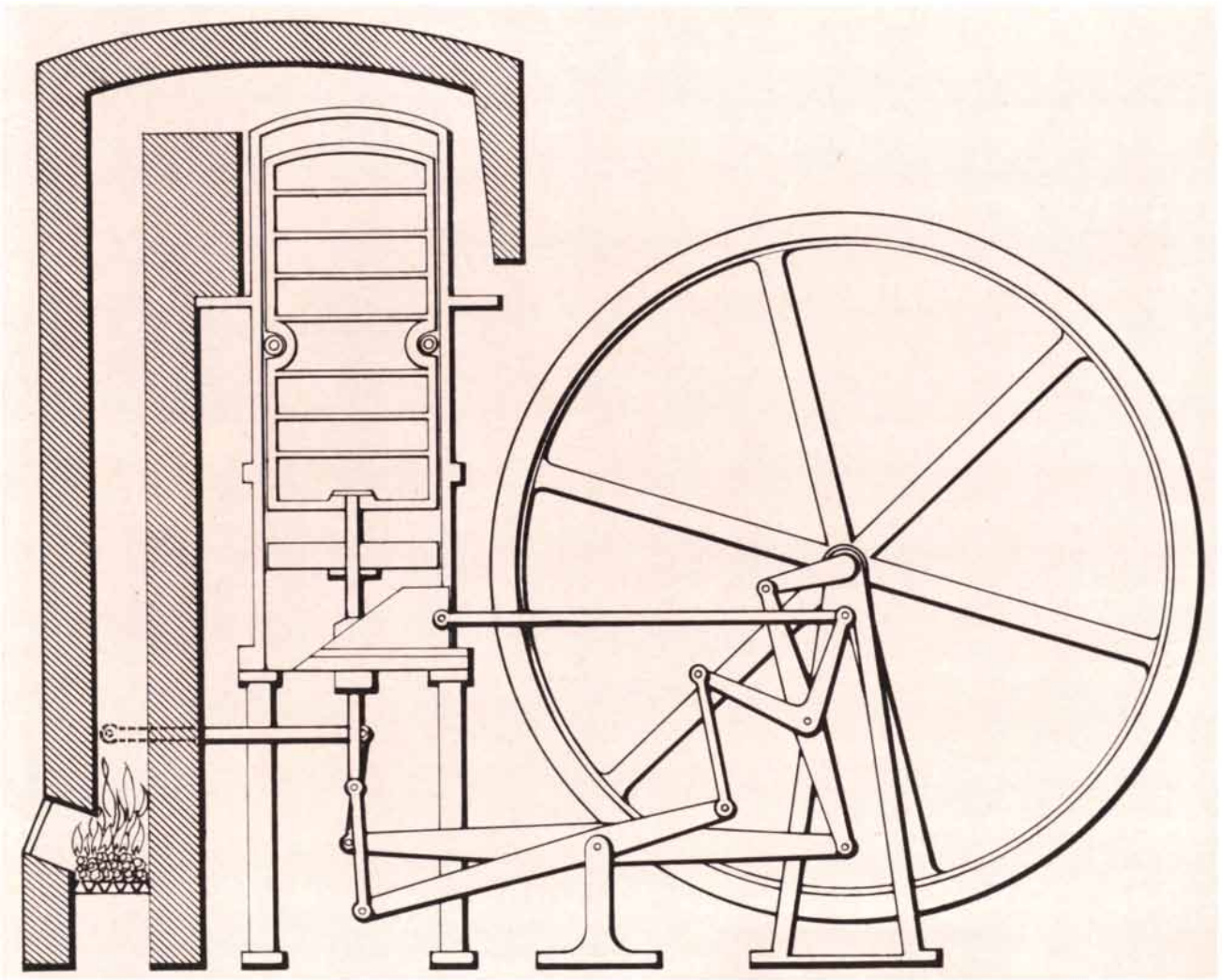
stant level. The cooling is done by an external cooling system analogous to the radiator of an automobile. In parts two and three of the cycle, called the constant-volume regenerative transfer, both pistons move simultaneously, and the working fluid passes through the porous regenerator and into the expansion space. In going through the regenerator the working fluid absorbs heat. The temperature of the fluid rises, resulting in an increase in the pressure.

During expansion (parts three and four of the cycle) the expansion piston moves to its outer dead point. The pressure of the gas falls. Since the temperature of a gas falls as the gas expands, heat must be supplied to the working fluid in order to maintain the temperature in the expansion space at a constant level. The heat is supplied by the external heat source. Finally, in a second constant-volume regenerative transfer process (parts four and one of the cycle), both pistons move simultaneously to their original positions, causing the working fluid to pass back through the regenerator and into the compression space. In going through the regenerator the working fluid gives up heat, which is stored in the matrix for use in the next cycle. The temperature of the working fluid falls from the high level to the low level, resulting in a decrease of the fluid's pressure to the original level.

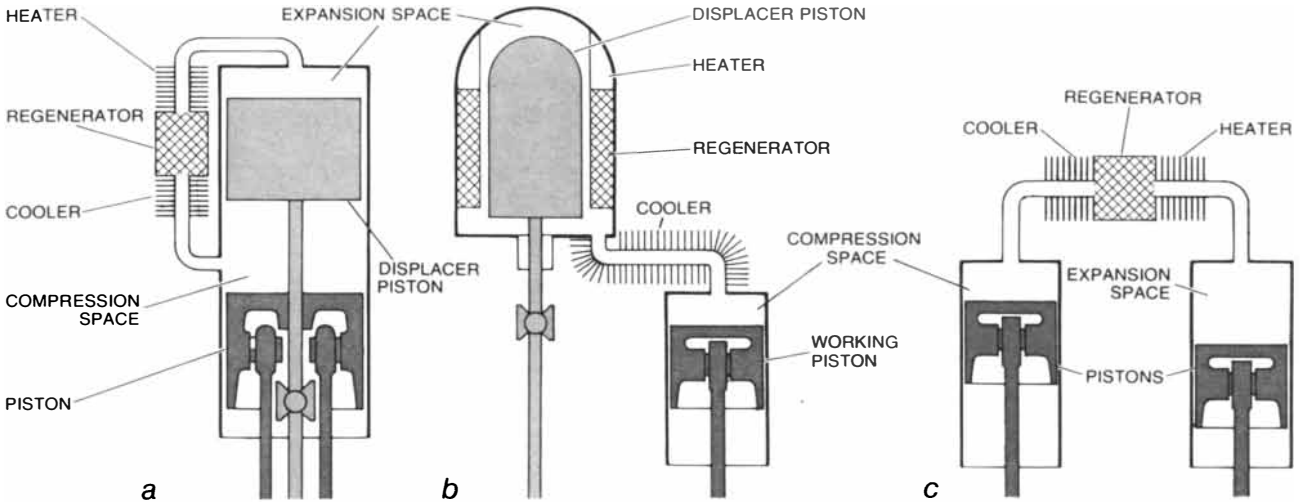
In practice Stirling engines do not work on the Stirling cycle. It is not possible to have isothermal (constant temperature) compression and expansion processes, although in a practical engine various heat exchangers are usually provided to enhance the heating and cooling processes and bring about the closest possible approximation of isothermal conditions. Practical engines also need to have pistons that operate with a continuous harmonic motion, so that the working fluid is never wholly in the expansion space or in the compression space. Another problem is that the regenerator and the other heat exchangers have internal void volumes that detract significantly from the engine's idealized performance. The regenerative process is less than perfect, and the effects of aerodynamic and mechanical friction are appreciable. For all these reasons the output and the efficiency of practical Stirling engines are only fractions of the theoretical values.

The elements of a practical Stirling engine include two spaces, which are coupled through a regenerator and auxiliary heat exchangers. In one space during operation the temperature is high and in the other one it is low. The volumes of the two spaces can be varied cyclically.

These simple elements can be combined in a remarkable variety of mechanical arrangements. For single-cycle machines the arrangements can be broadly classified into two groups: two-piston machines and machines combin-



ORIGINAL STIRLING ENGINE is portrayed in an adaptation of a drawing in the British patent specifications of 1816. The inventor was the Reverend Robert Stirling of the Church of Scotland. Such an engine was used in 1818 for pumping water from a quarry.



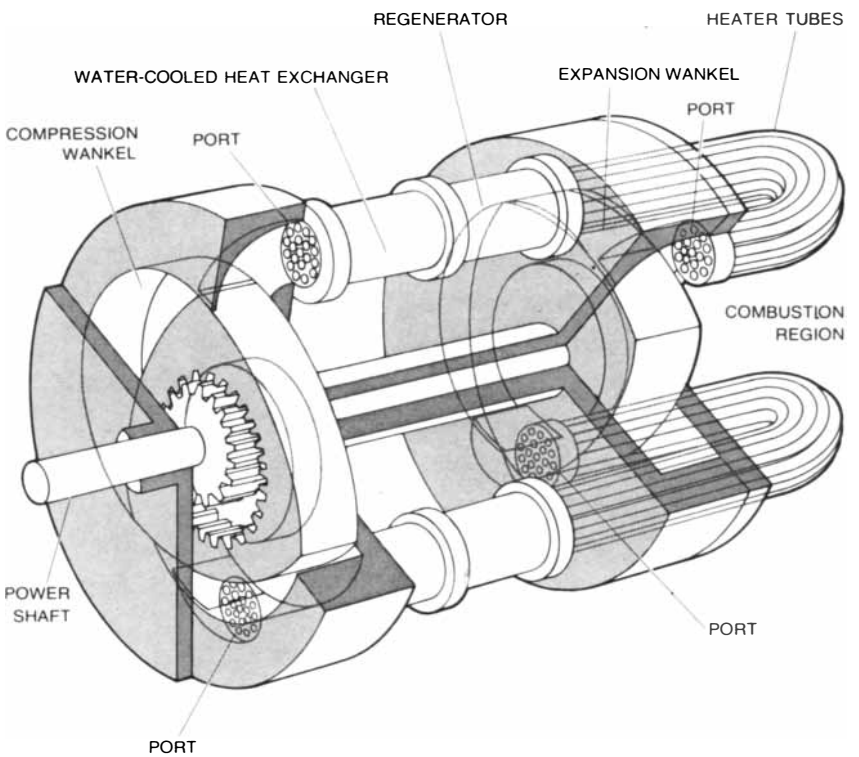
BASIC MECHANICAL ARRANGEMENTS for single-cycle Stirling engines are depicted. In machines combining a working piston and a displacer piston, which helps to move the gas back and forth between the compression space and the expansion space, the two pistons can be in the same cylinder (*a*) or in separate cylinders (*b*). The third basic arrangement (*c*) is a two-piston Stirling engine.

ing a working piston and a displacer piston, which helps to move the working fluid back and forth between the two spaces. The second group can be further subdivided into machines that have the working piston and the displacer in a single cylinder and machines with separate cylinders for the working piston and the displacer [*see illustration above*].

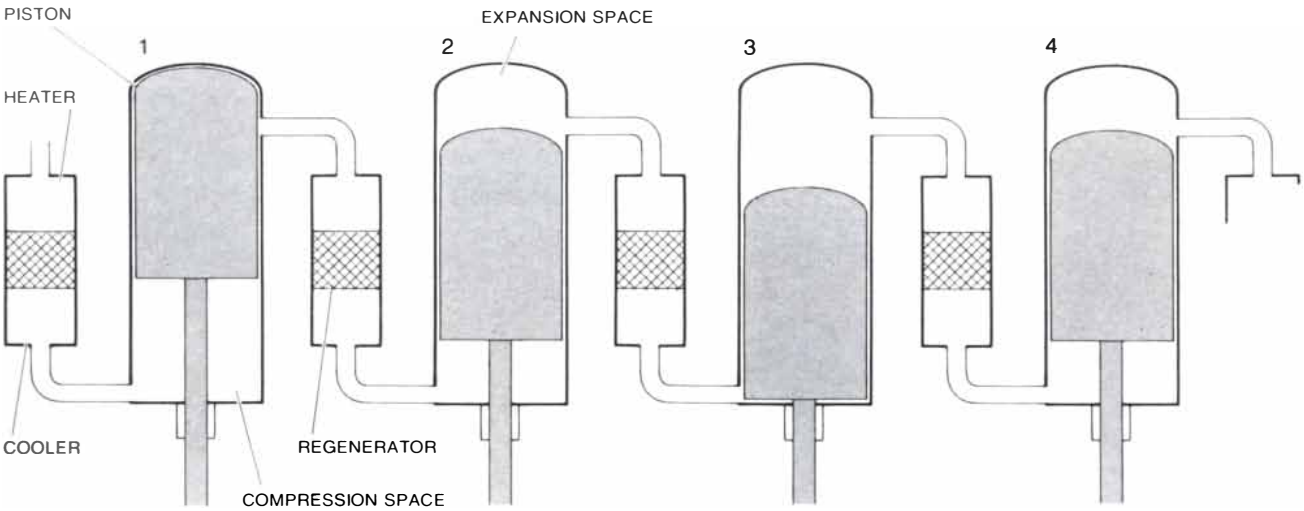
In engines with multiple cylinders a Rinia arrangement can be adopted [*see top illustration on opposite page*]. In this arrangement, which was devised by Herre Rinia of the Philips Research Laboratories in the Netherlands, adjacent cylinders are interconnected through a regenerator, and the number of reciprocating elements is halved to one per cycle. In *V* form or with a swash-plate drive the Rinia arrangement is well suited for a compact automotive-traction engine of high performance. (A swash plate is a disk mounted at a slant on a revolving shaft, so that as the shaft rotates a rod in contact with the edge of the disk's face is given a reciprocating motion.)

Development of prototype Stirling engines for automotive purposes is being carried out at the Philips Research Laboratories and by licensees of Philips in Sweden, West Germany and the U.S. For several years the focus of the effort was on a four-cylinder engine of 200 horsepower that was characterized by a rhombic drive, which achieves a relatively efficient out-of-phase operation of the two pistons in a Stirling cylinder by mounting the cylinders at approximately a right angle to each other. The engines were designed for service in trucks, buses and boats. Prototype units installed in buses and boats in the Netherlands and Sweden performed well, but it became clear that the cost of the engine could not be reduced enough to meet the competition of other engines.

As a result work on the rhombic drive for automotive engines has ceased, and the focus of efforts is now on engines that have the Rinia arrangement. Philips and its U.S. licensee, the Ford Motor Company, are working on a four-cylinder, 180-horsepower engine with a swash-plate drive for automobiles. The Swedish licensee, United Stirling, is concentrating on a four-cylinder, 65-horsepower *V* engine.



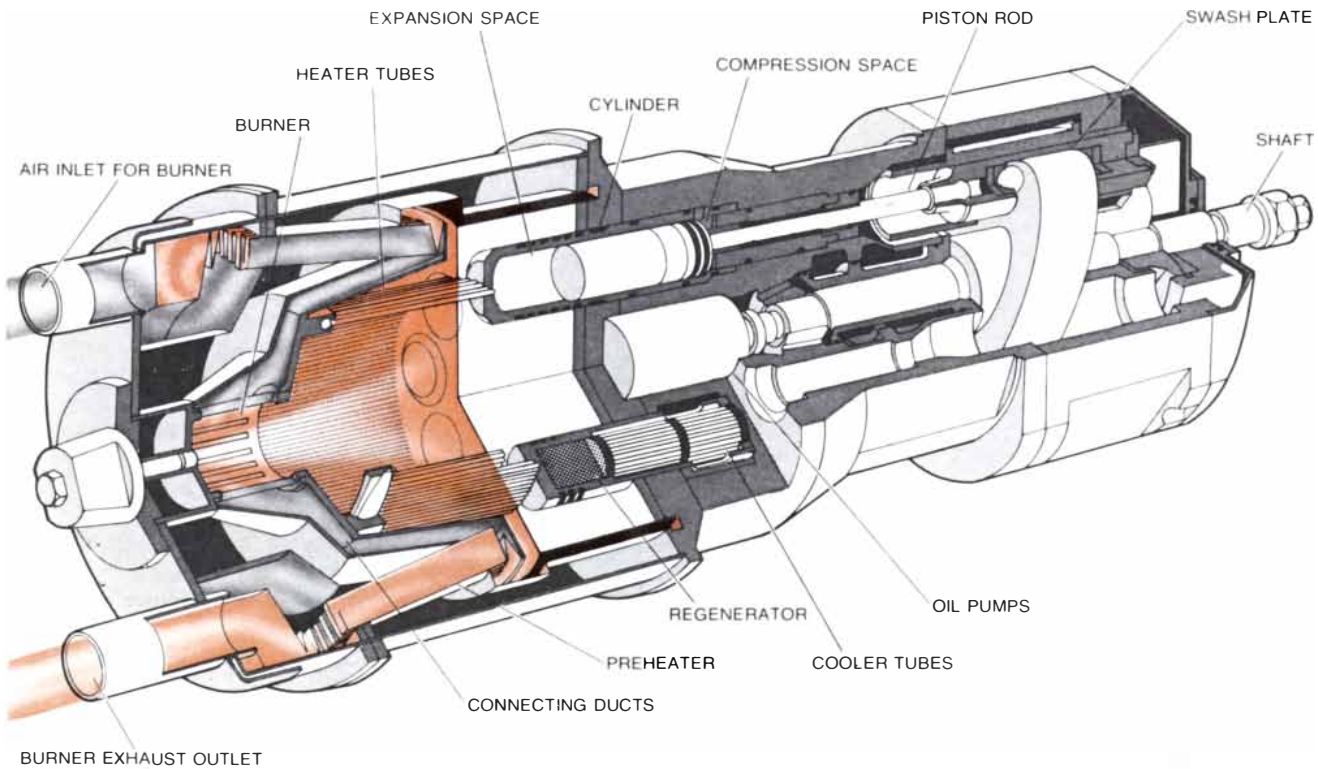
WANKEL-STIRLING SYSTEM was devised at the University of Calgary by the author and his colleagues. The Wankel, or rotary, engine employs a rotor and three chambers of variable size in place of the piston-and-cylinder arrangement of the conventional internal-combustion engine. In the Wankel-Stirling combination one Wankel engine functions as the compression part of a Stirling cycle and the other one functions as the expansion part.



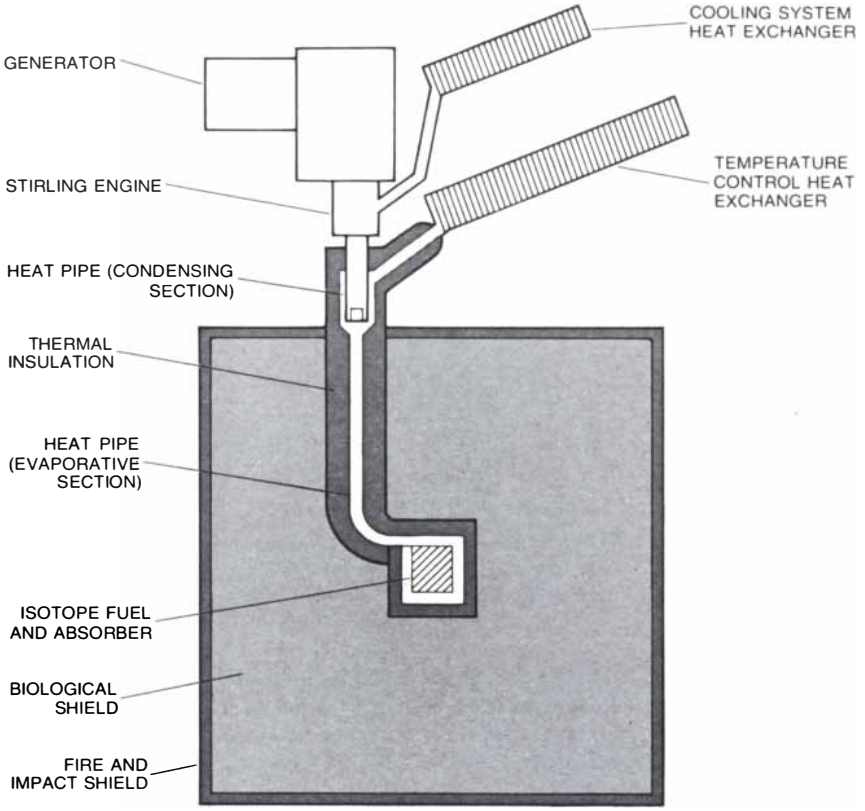
RINIA ARRANGEMENT is often employed for Stirling engines with more than one cylinder. The arrangement was devised by Herre Rinia of the Philips Research Laboratories in the Netherlands. Adjacent cylinders are interconnected through a regenerator. With this arrangement there is only one reciprocating element per cycle, rather than two elements as in single-cylinder engines.

Whatever the form, the strong virtue of the engine in the present state of affairs is its cleanness of operation. The external combustion takes place continuously in a hot-walled chamber, and there is no limit on the amount of air that can be supplied. As a result the unburned residual gases characteristic of internal-combustion engines are eliminated. Moreover, with recirculation of the exhaust gas the production of nitrogen oxides by Stirling engines operated on diesel fuel is already lower than the standard set for 1976 by the Federal law on air pollution in the U.S.

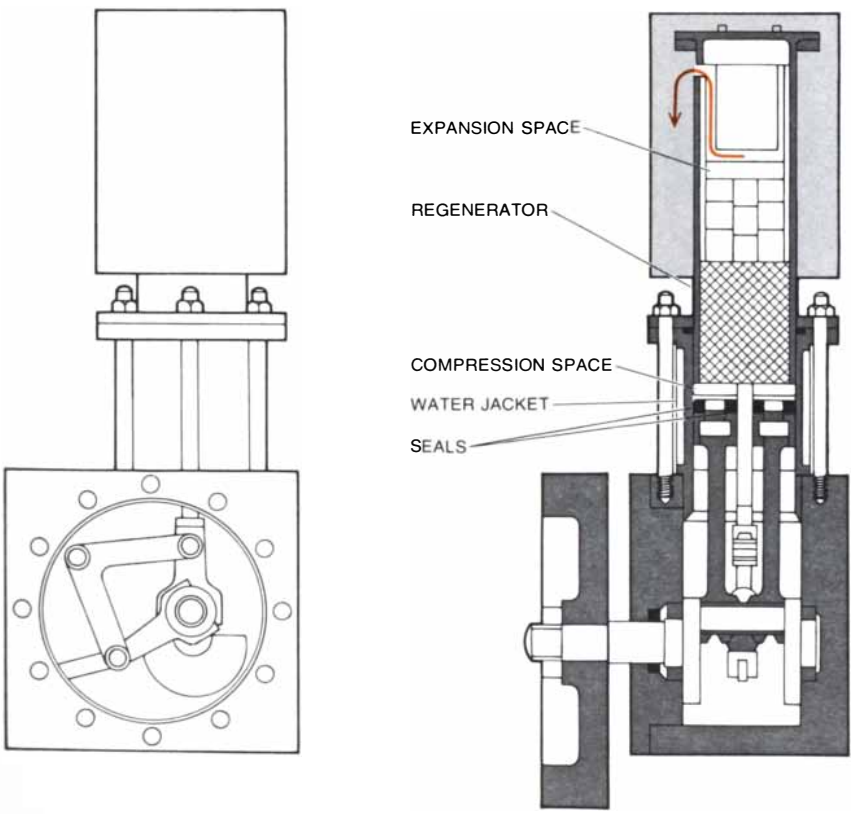
The Stirling engine has several other favorable characteristics as a prime mover for land vehicles. It runs without vibration and is very quiet, because there are no valves or periodic explosions, and the engine can therefore be balanced perfectly. The efficiency and the specific output of the engine (the power produced for a given size) are comparable



AUTOMOBILE ENGINE incorporating the Stirling principle is under development by Philips and its U.S. licensee, the Ford Motor Company. The four cylinders are in a Rinia arrangement, and the drive is by swash plate. A swash plate is a disk mounted at a slant on a revolving shaft, so that as the shaft rotates a rod attached at a point on the face of the disk is given a reciprocating motion. In the current design this is a 180-horsepower engine. Because the combustion is external to the cylinders it takes place continuously in a hot-walled chamber with an unlimited supply of air. Hence the engine produces no unburned gases to pollute the environment.



SMALL GENERATOR fueled by the radioactive decay of cobalt 60 has been designed by the Calgary group. It would employ a Stirling engine running on the heat produced in the radioactive decay. Elements of the generator are depicted schematically. As now designed the engine could be expected to run steadily for more than two years with little attention.



DETAILS OF STIRLING ENGINE for the cobalt-60 generator are depicted in end (*left*) and side (*right*) views. Engine converts heat to electricity at an efficiency of 20 percent.

to those of a diesel engine, and the economy of operation at partial load is good. Engine braking, with up to 80 percent of the driving force, is possible. Since the engine has a wide range of speeds and favorable torque characteristics, a simple transmission system can be employed. Oil consumption is close to zero, because the combustion products are not in contact with moving parts, and oil changes are required infrequently. The engine is capable of rapid acceleration and a quick response to sudden changes in load. It is reliable and has a long service life. Finally, it is not susceptible to contamination or damage from dust or salt in the environment.

Advanced Stirling engines with the specific output of gasoline engines and the thermal efficiency of diesel engines are in prospect. They are highly pressurized engines of compact design and are distinguished in particular by the fact that they embody a system of indirect heating of the Stirling cylinders. With this arrangement the fuel-combustion system and the engine-heating system can be designed independently, with highly beneficial results for both. Moreover, with indirect heating the designer can resort to schemes for storing heat, as a battery stores electricity. One possibility is that a thermal storage battery could be charged overnight with heat from electricity or natural gas and discharged during the day by means of a Stirling engine. The Philips laboratories have investigated the possibility of storing heat in lithium fluoride. They found the system suitable for such vehicles as buses, taxis and delivery trucks. A bus, for example, could be operated all day on the heat stored in a cubic meter of lithium fluoride.

Another concept that is under study entails burning hydrogen to heat a Stirling engine. A recent discovery by the Philips workers is that certain compounds consisting of a rare-earth metal and nickel or cobalt can absorb and release large amounts of hydrogen. For example, at room temperature and a pressure of 2.5 atmospheres the density of hydrogen in a compound of lanthanum and nickel ($LaNi_5$) is nearly twice the density of liquid hydrogen. If the discovery can be brought to commercial application, it would mean that gaseous hydrogen could be employed as a fuel in vehicle engines. In present systems only liquid hydrogen will serve, and it must be stored at an extremely low temperature (22 degrees Kelvin, or 22 degrees Celsius above absolute zero). Indeed, the development could have a significant ef-

fect on the problem of air pollution generally, since gaseous hydrogen would serve as a clean fuel in a conventional internal-combustion engine.

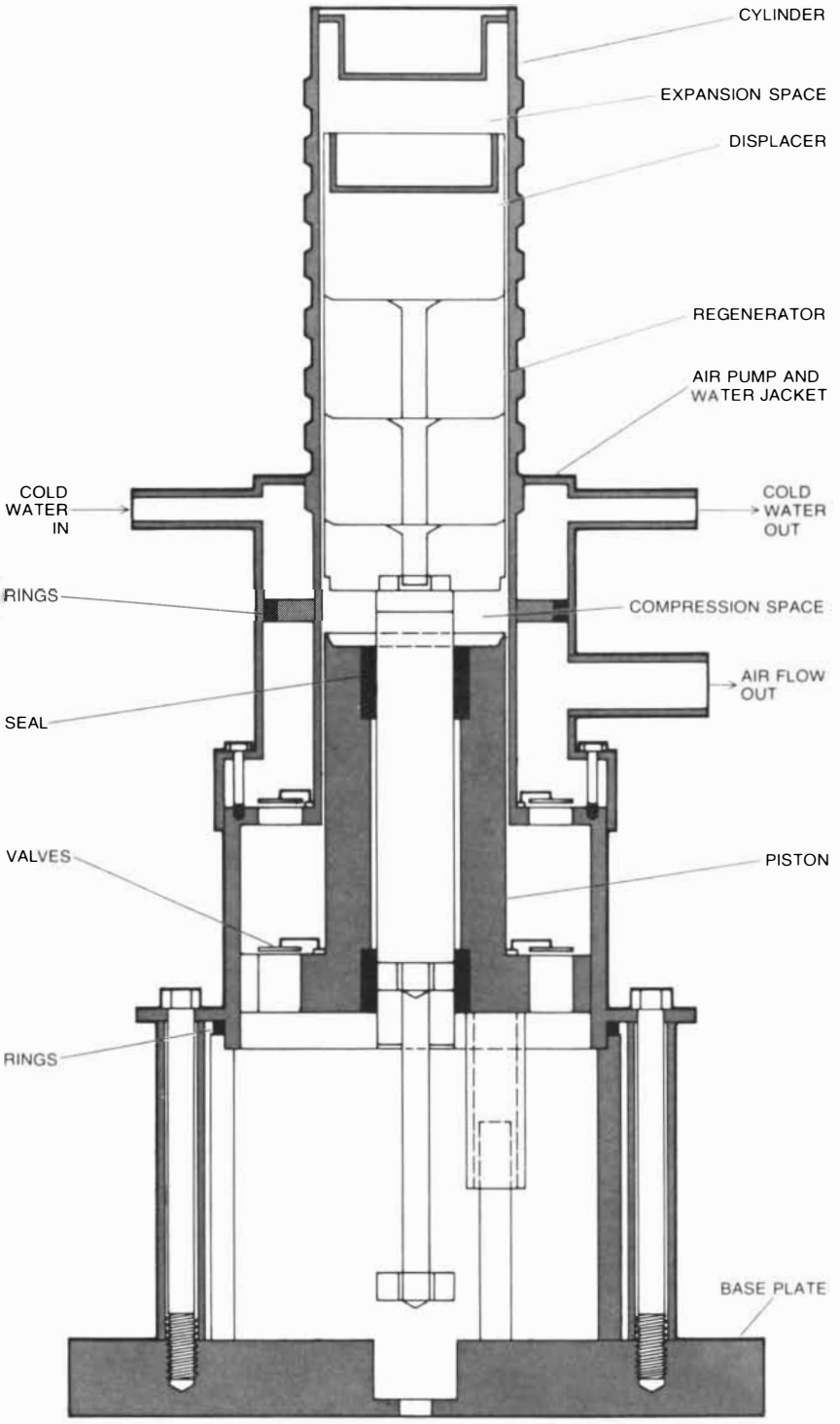
An arrangement that my colleagues and I at the University of Calgary have devised combines two Wankel engines in a Stirling-engine system [*see bottom illustration on page 82*]. One engine functions as the expansion part of a Stirling cycle and the other functions as the compression part. The system has three separate expansion and compression spaces, connected in our arrangement by two ducts containing the regenerative heat exchangers. Each of the three spaces experiences two operating cycles per revolution. Given concentrated research efforts, this engine could provide the basis for a compact prime mover.

The Stirling engine has achieved its greatest commercial success as a means of refrigeration [see "The Stirling Refrigeration Cycle," by J. W. L. Köhler; SCIENTIFIC AMERICAN, April, 1965]. The procedure involves something of a reversal of the hot-air system, since the source of heat for the Stirling cycle is not a burner but rather whatever it is that one wants to cool. Until now Stirling-cycle cooling engines have been employed almost exclusively in the cryogenic temperature range, that is, for achieving extremely low temperatures. Many opportunities for commercial applications at higher temperatures are in sight. They include air-conditioning units for vehicles such as automobiles and railroad cars. A Stirling engine in such an application would probably not be any better at cooling than a conventional refrigerator is, but the Stirling concept holds the promise of substantial reductions in size and weight.

Another promising area for Stirling engines is in the field of small generators of electric power, particularly in situations that call for a machine capable of operating unattended in remote sites for long periods of time. The power level of interest ranges from five watts to five kilowatts but is concentrated mainly in the range from 200 to 500 watts. Examples of installations where a small Stirling engine would be an efficient and reliable source of power include lighthouses, navigation buoys, automatic weather stations and booster stations for various communication systems.

In most of the applications the main virtue of the engine would be its reliability. It is also important that the engine can have almost any size, weight and speed. Starting and stopping would usually not be a matter of concern, since in most cases it would be possible to include a system of batteries storing electricity, so that power could be drained off at a high rate at certain times (such as at night in a lighthouse) while the engine generated power continuously at a steady rate. The same feature reduces the importance of being able to regulate the engine, since in most applications it is preferable and easier to adjust the system electrically than to control the engine.

Thermal efficiency is an important



BEALE ENGINE, invented by William Beale of Ohio University, is a variant of the Stirling engine. It consists of a heavy working piston and a light displacer piston, which are moved within a cylinder by the fluid mechanics generated by the working fluid. There is no crank mechanism coupling the working piston to a shaft, and the displacer piston is not mechanically coupled to the working piston. This engine was designed to serve as a gas compressor.

factor, since it governs how much fossil fuel or radioisotope fuel must be provided to supply heat. In this respect the Stirling engine offers a ratio of conversion of heat to electricity of about 20 percent. The only other systems available are thermoelectric units (converting heat to electricity directly by means of a thermocouple), with conversion ratios of 8 percent or less. Neither diesel engines nor gasoline engines with sufficient reliability are available for the kind of applications I have been discussing.

My colleagues and I have done research on a small Stirling engine that would be fueled by the radioactive decay of cobalt 60 [*see illustrations on page 84*]. Heat is produced in the radioactive decay, and the engine is designed to convert the heat to electricity with an efficiency of 20 percent. The half-life of cobalt 60 is five years; we envision the generator as capable of operating steadily and with little or no attention for two and a half years.

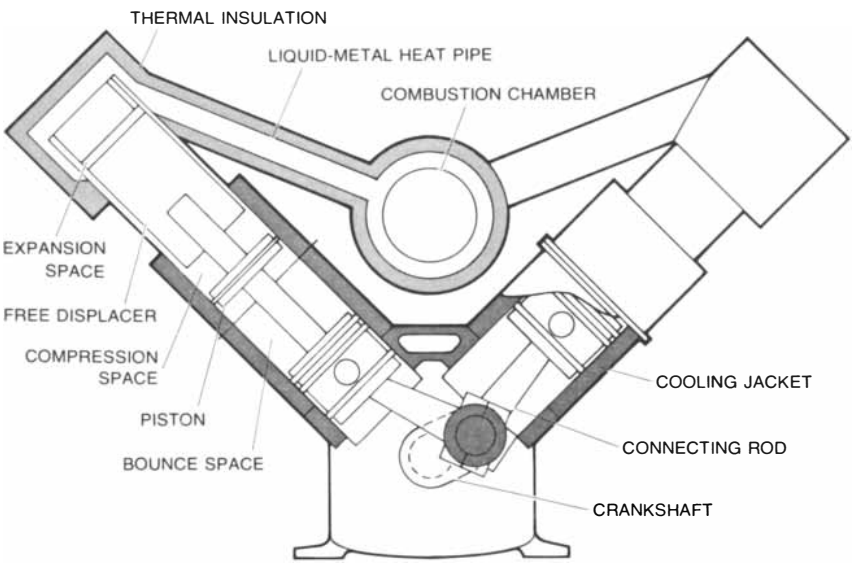
There is a large demand for small, inexpensive engines that would run on sun power to produce electricity in developing countries in the Tropics, particularly for irrigation and small lighting installations. The Stirling engine could fill such a role well. The problem with it, as with any solar engine, is that in addition to the engine one must have a device that collects and concentrates the sunlight; for the best results one must also have a mechanism that keeps the collector oriented to the sun. Usually these devices are fairly elaborate installations external to the engine. Theodore Finkelstein of Trans-Computer Associates has devised a cylinder head made of quartz, so that the focus of the collector is within the cylinder. With this arrangement the working fluid in a Stirling-cycle cylinder can be heated to a high level by sunlight alone. William Beale of Ohio University has done a good deal of work on sun-powered Stirling engines, as a result of which a few engines of this type are in limited production.

Beale has also invented a novel type of Stirling engine in which a heavy working piston and a light displacer piston oscillate in the cylinder solely on the fluid mechanics provided by the movement of the working fluid [*see illustration on preceding page*]. The cylinder is closed at one end and can be either open or closed at the other end. The engine has no crank mechanism coupling the working piston to a shaft, and the displacer piston is not mechanically coupled to the working piston but moves in relation to it by fluid forces generated by the motion of the working piston. Beale engines are self-starting, which is a characteristic not shared by any single-cylinder engine controlled by a crank. Moreover, they can be made without an external seal for the gas that is the working fluid. In this form they can be pressurized and hermetically sealed, thereby providing for a relatively high specific output and eliminating the possibility of contamination of the moving parts by dirt from outside. Such an engine is well suited for applications where maintenance might be a problem, as in domestic appliances. Beale engines perform in any orientation—vertical, horizontal, inclined or upside down. They are amazingly simple in construction and do not depend on springs, valves or any kind of mechanical gadget.

Work can be extracted from the Beale engine by attaching a load to the oscillating shaft of the working piston. Such an engine can serve as a gas compressor, a water pump or the drive for a linear generator of electricity. A particularly valuable application for a self-starting, thermally active drive system of this kind is in providing a small supply of electricity to run a fan or a water pump associated with a large furnace. In the normal installation a failure of the commercial electric supply system immobilizes the furnace, even though the gas or oil contributing 99.9 percent of the energy needed by the furnace is still available.

Our group at the University of Calgary admired the Beale engine but wanted to find a way of coupling it to a crankshaft, inasmuch as most modern machines (such as the automobile) function not by means of reciprocating shafts but by the power derived from rotating shafts. The result of our efforts is a hybrid machine with a free displacer piston and a working piston that is controlled by a crank [*see illustration on this page*]. An important advantage of the hybrid is that the working piston and the connecting-rod assembly for the crankshaft can all be standard parts utilized in the internal-combustion engine, thereby opening the possibility that the cost of the hybrid on a commercial scale could be comparable to the cost of other piston engines.

Because of its reliability and its ability to run for a long time, the Stirling engine seems well suited to serve as the source of power for an artificial heart, both the relatively large type that operates outside the body (taking over the function of the heart temporarily during medical procedures) and the small type that may someday be implanted to replace a defective heart. In the implanted type a Stirling engine fueled with a radioactive isotope could be expected to function effectively for at least three years before needing replacement. The main problem at present is that only isotopes emitting alpha particles would be suitable, since other kinds of emission would require elaborate shielding to protect the patient and people near him from radiation. Alpha-emitting isotopes are rare and expensive, and at present no adequate source of supply exists. Nonetheless, the National Heart and Lung Institute sees



HYBRID ENGINE designed by the Calgary group employs the Beale free-piston principle but couples the engine to a crankshaft, so that the power can be taken off by way of a rotating shaft rather than a reciprocating shaft. In this *V* design the working piston and the connecting-rod assembly are intended to be standard internal-combustion-engine parts.

the radioactive-isotope-fueled artificial heart as promising and is supporting the exploration of several concepts.

At the other end of the size spectrum is a concept advanced by W. J. Bradley of the Chalk River Nuclear Laboratories in Canada to replace the steam turbine in a nuclear power station with a large Stirling engine. In present installations the heat from the nuclear fuel boils water, producing steam that drives the turbines that drive the generators. Because of the problem of containing the nuclear fuel, the steam temperature must be kept relatively low, and so the efficiency of conversion of input energy to electricity is low. Bradley proposes to incorporate the cooling channel of the nuclear reactor as the heater of the Stirling engine, with helium at high pressure acting simultaneously as the working fluid of the engine and the coolant fluid of the reactor. The system would be much simpler and cheaper than a steam turbine and could be expected to increase the efficiency of the plant by at least 20 percent.

In the light of the many advantages of the Stirling engine, the reader may wonder why the engine is such a rarity in commercial applications. The main reason is the vast amount of research that has gone into the internal-combustion engine and the enormous technological investment in building that engine. It is also true that developmental efforts are required to bring the Stirling engine to an optimum level. Stirling engines of advanced form are complex and expensive. Simpler versions are not so costly, but they lack the efficiency and the specific output of internal-combustion engines. Among the particular problems needing work are the heat exchangers for the hot part of the cycle, which must be made of material that will operate continuously at high temperature, and the cooling arrangements for the cool part of the cycle. Sealing is another crucial problem, because the air that constitutes the working fluid cannot be replenished readily and so must not be allowed to leak.

It seems probable that sociological pressures will soon force the addition of a number of expensive and complex pollution-control devices for gasoline and diesel engines. In that case the economics of the Stirling engine would begin to look better than they have up to the present time. Conditions therefore appear to be appropriate for the intensive developmental efforts that would make the Stirling engine commercially competitive with the internal-combustion engine.

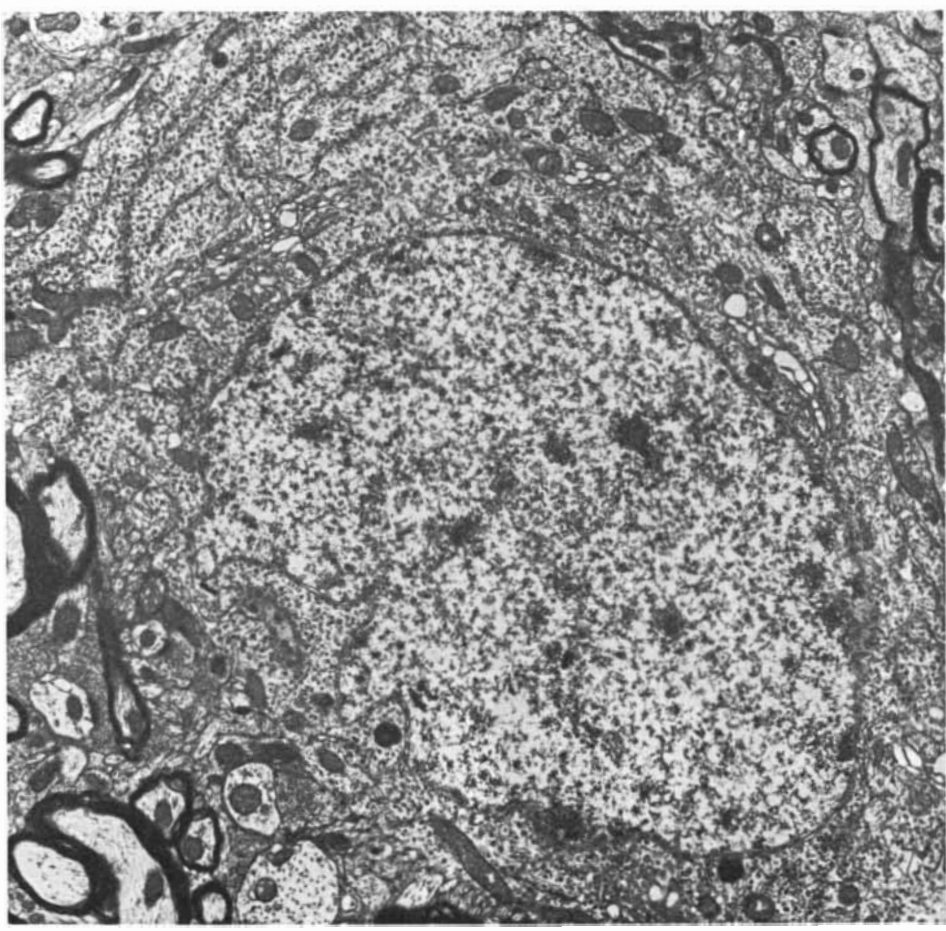
HEREDITARY FAT-METABOLISM DISEASES

New tests for detecting adult carriers and for prenatal genetic diagnosis make it possible to control the incidence of 10 usually fatal lipid-storage diseases caused by inherited enzyme deficiencies

by Roscoe O. Brady

Lipids are fatty materials found in the bloodstream and in all tissues of the body. They may be simple molecules of neutral fat (which the body stores as an energy reserve) or they may be combined with other substances (often a sugar) in the membranes of cells. Aged cells or cells that have otherwise lost their usefulness are degraded into simpler materials that can be used for supplying energy to other cells or for forming new cells. A normal person has a large variety of enzymes that catalyze the breakdown of the lipids in such cells. When the enzyme that catalyzes a particular lipid reaction is inactive or absent, excessive amounts of that lipid begin to accumulate in certain tissues. In human beings lipid-storage disorders often result in mental retardation and enlargement of the spleen and the liver. Most of the known lipid-storage disorders are fatal. Here I shall describe how the discovery of the enzyme deficiencies responsible for lipid-storage disorders came about and describe some of the practical developments that followed, including the possibility of treating some of the disorders by administering the missing enzyme.

The investigation of lipid-storage diseases has passed through four overlapping phases in the past 90 years. The first phase was the description of symptoms of patients with disorders caused by excessive accumulation of lipid. It began in 1881 with a report by Warren Tay, an English ophthalmologist. He described a brownish-red spot surrounded by a contrasting white patch on the retina of a one-year-old child suffering from muscular weakness. In 1887 Bernard Sachs, an American neurologist, noted the same characteristic feature in a child with muscular weakness and severe, progressive mental retardation. Sachs went on to discover that the disease was hereditary and was caused by the excessive accumulation of lipids in nerve cells. In time the disorder became known as Tay-Sachs disease. The disease occurs predominantly in children of Ashkenazic (middle and northern European) Jewish descent. Although the disease is much less frequent in non-Jews, it has been reported in most races. In children with Tay-Sachs disease the nerve cells in the brain are swollen, and round membranous bodies are found in the protoplasm outside the nucleus of the cell. The



NERVE CELL from a normal brain (*left*) and a nerve cell from the brain of a child with Tay-Sachs disease (*middle*) are shown enlarged about 10,000 diameters. In each case the cell nucleus is the membrane-enclosed area at the center of the electron micrographs. A prominent nucleolus can be seen at the center of the nucleus of the cell from the diseased child (*middle micrograph*). The protoplasm outside the diseased cell's nucleus is crowded

symptoms of the disease progress from muscular weakness to mental retardation, seizures and blindness, with death usually coming at three years of age.

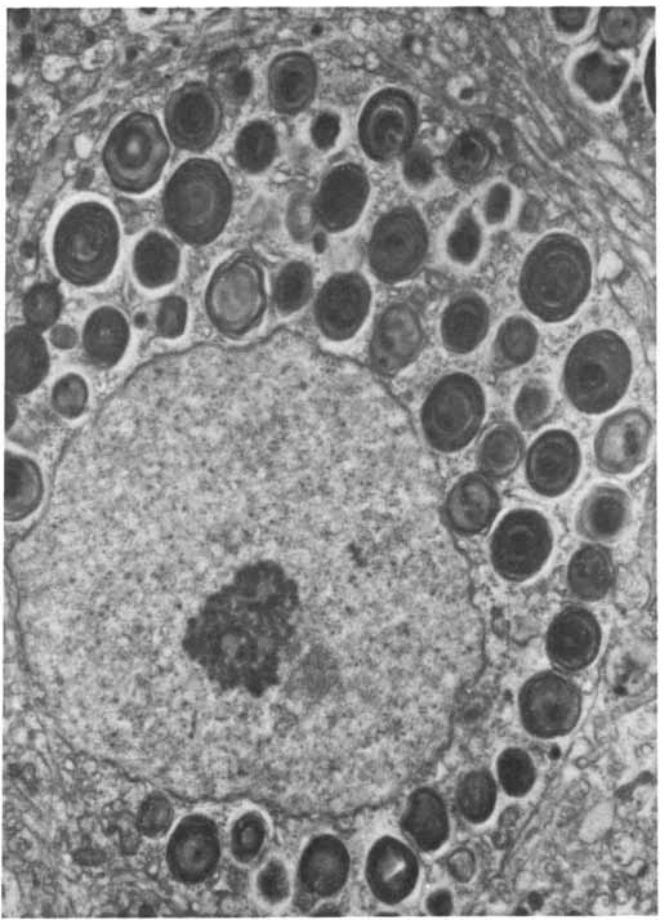
Another lipid-storage disease involving enlargement of the spleen and the liver was described by Philippe C. E. Gaucher of France in 1882. Here the accumulation of lipid was primarily in the cells of the spleen. Then in 1898 Johannes Fabry of Germany and William Anderson of England reported a particular type of skin rash in patients who also showed kidney damage. This disorder is called Fabry's disease.

Today there are 10 well-characterized human diseases that are known to be caused by the excessive accumulation of lipids in tissue cells [*see illustration on pages 92 and 93*]. The nomenclature of the diseases leaves much to be desired. Some are named for the physicians who first reported them, some according to the term for the accumulating lipid and some according to the characteristic structure of the affected tissues.

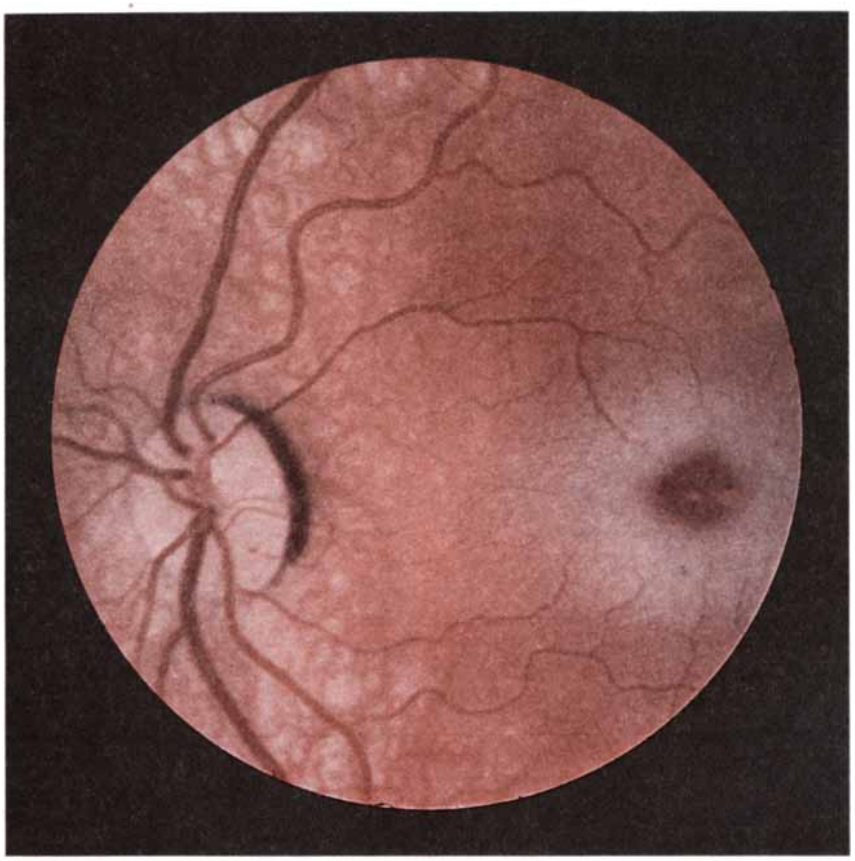
The second phase of investigation into lipid-storage diseases dealt with their hereditary aspect. Of the 10 known diseases all except Fabry's disease are transmitted as autosomal recessive genetic defects. This term means that the trait is transmitted by a chromosome other than the sex chromosomes (the *X* and *Y* chromosomes) and that both parents must be carriers of the defective chromosome. Statistically a fourth of the children of couples who are both carriers will be affected by the disease, half of the children will not be affected but will be carriers and a fourth will neither be affected nor be carriers. In reality for each couple the birth of affected children may be more frequent because the game of chance is played with the same odds for each pregnancy. On the other hand, with the aid of prenatal genetic diagnosis and genetic counseling parents who are both carriers now have a much better hope for having only normal children.

The exception to autosomal recessive inheritance, Fabry's disease, is transmitted as a recessive characteristic by the *X* chromosome. Since males have only one *X* chromosome, which is inherited from the mother, only the mother need be a carrier to have a male child who is affected by Fabry's disease. Statistically half of the sons of a carrier mother will have the disorder and half of her daughters will be carriers. Some of the female carriers show manifestations of the disease, although they are usually not severe.

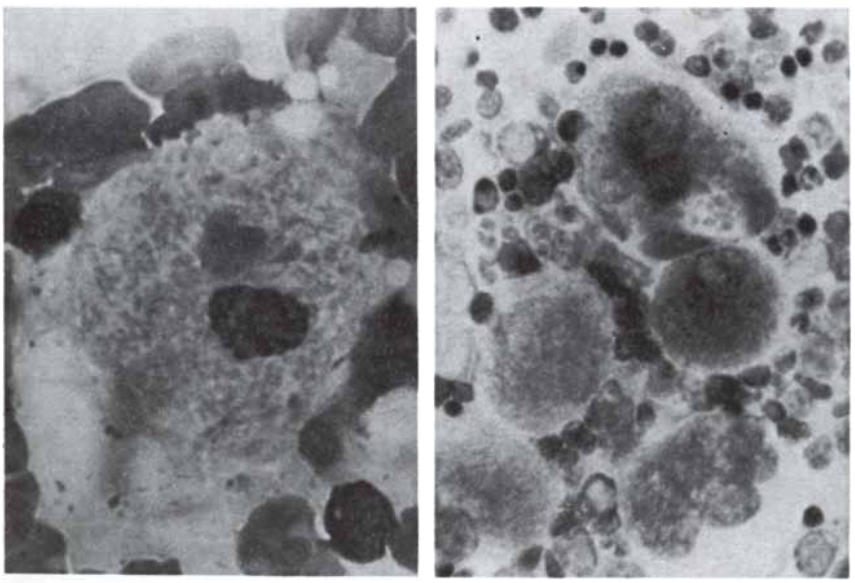
The third phase in the study of lipid-storage diseases was the identification of the chemical nature of the lipids that accumulate in the tissues. This development began in 1934 with the demonstration by A. Aghion of France that the lipid that accumulates in the spleen of a person with Gaucher's disease is a compound called glucocerebroside. The molecule of this lipid has three parts: sphingosine, a long-chain fatty acid and a sugar [*see illustration on page 91*]. The combination of sphingosine and a fatty acid is called a ceramide, and it is



with ovoid membranous bodies that have displaced most of the other tissue elements. Few mitochondria, which are numerous in the outer protoplasm of the normal cell, remain between the membranous bodies. The membranous bodies are produced by the accumulation of lipids in layers. The micrograph at right shows the concentric structure of a membranous body, enlarged about 67,000 diameters. The electron micrograph of the normal cell was made by Henryk M. Wisniewski. The micrographs of the Tay-Sachs nerve cell were made by Robert D. Terry and Martin Weiss. All are at the Albert Einstein College of Medicine in New York.



CHERRY RED SPOT in the macular region of the retina (*right*) is characteristic of several lipid-storage diseases. This view, made through a fundoscope by George N. Wise of the Albert Einstein College of Medicine in New York, is of the eye of a child with Tay-Sachs disease. The lipid ganglioside GM_2 accumulates in the ganglion cells adjacent to the fovea of the retina and forms a whitish halo. The red spot on the retina is not in itself abnormal; it is simply the normal color of the fovea produced by blood in underlying vessels.



ABNORMAL BONE MARROW CELL from a patient with Gaucher's disease is shown enlarged about 2,200 diameters in the photograph at left made by Neal Weinreb of the Mount Sinai Hospital and School of Medicine. The Gaucher cell has a wrinkled cytoplasm and an eccentric condensed nucleus. Normal multinucleated giant cells of the bone marrow, from which blood platelets are derived, are shown enlarged 1,000 diameters in the photograph at right by Robert S. Hillman of the University of Washington's Harborview Medical Center.

the common portion of all the substances that accumulate in lipid-storage diseases. In the normal human spleen the concentration of glucocerebroside is low; in Gaucher's disease the concentration may be as much as 4 percent of the dry weight of the spleen.

Another disorder involving enlargement of the liver and the spleen and mental retardation, Niemann-Pick disease, was shown by Ernst Klenk of the University of Cologne to involve the accumulation of the phospholipid sphingomyelin. Klenk also reported a massive accumulation of gangliosides in the brain tissues of infants affected with Tay-Sachs disease, and in 1962 Lars Svennerholm of the University of Göteborg identified the lipid as ganglioside GM_2.

The fourth phase, and from a practical standpoint the most important at present, was the investigation of why the lipid accumulated. That phase began with the study of Gaucher's disease, which involves the accumulation of glucocerebroside in the spleen. Initially there were three possible explanations. The first was that the accumulation was the result of an abnormality in the metabolism of carbohydrate, specifically the improper utilization of galactose, a sugar closely related to the glucose found in glucocerebroside. This concept was derived from the fact that individuals with the hereditary disorder called galactosemia cannot metabolize galactose in a normal manner. When galactose was infused into the blood of people with Gaucher's disease, however, no metabolic intolerance was found.

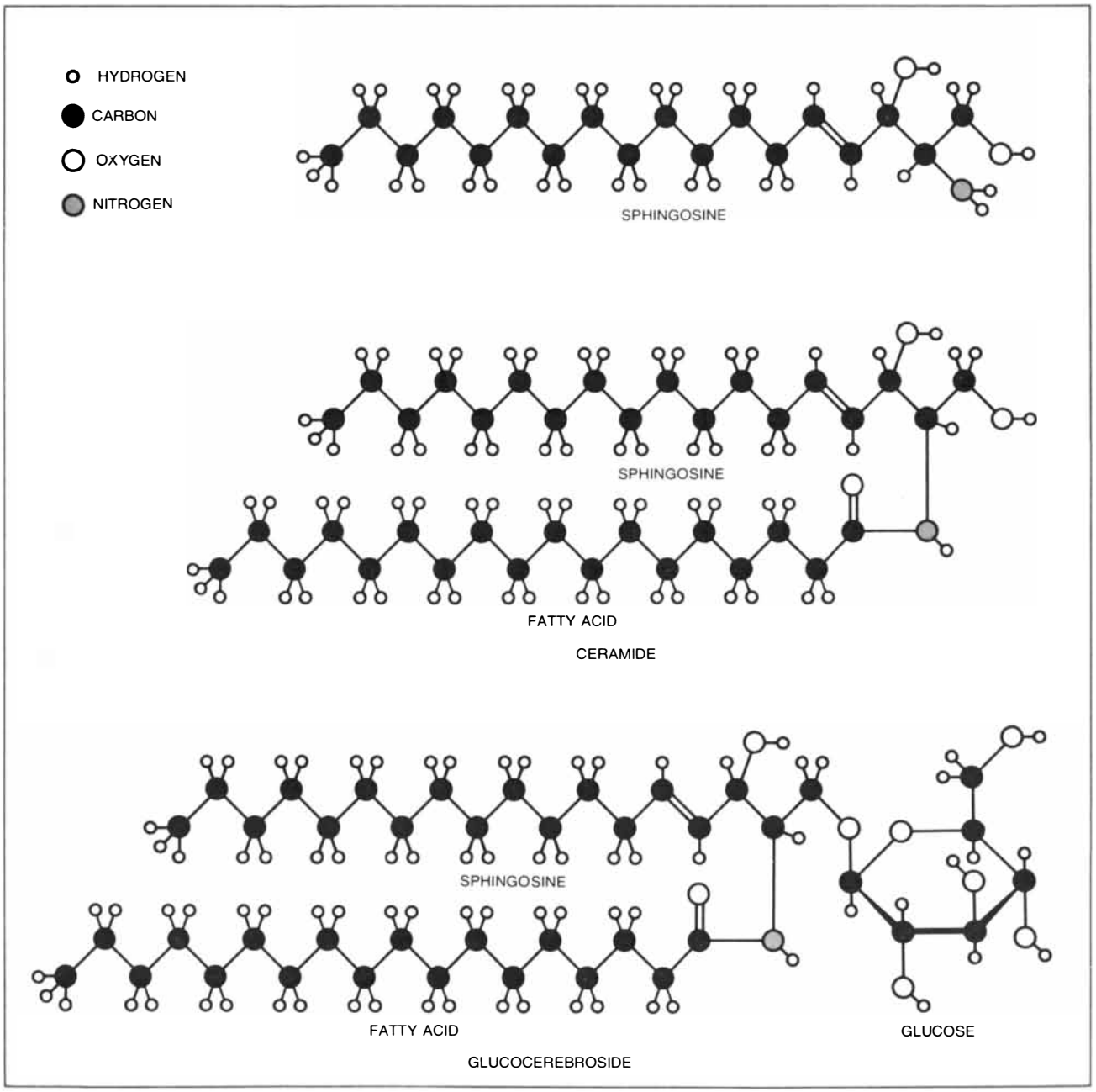
In 1957 Eberhard G. Trams and I, working at the National Institute of Neurological Diseases and Stroke, investigated the problem in a more direct fashion. Some patients with Gaucher's disease must have their spleen removed because of the anemia that frequently attends damage to the spleen. We obtained specimens of spleen from such patients and incubated slices of the tissue with radioactively labeled glucose or radioactively labeled galactose. Spleen cells continue to perform their normal functions when incubated in suitable nutrients. We found that the spleen cells from patients with Gaucher's disease could synthesize both glucocerebroside and galactocerebroside from either glucose or galactose. This finding conclusively ruled out the possibility that a defect in carbohydrate metabolism was involved.

The second possibility was that the lipid accumulated because of rapid or

uncontrolled synthesis. We obtained spleen slices from patients with Gaucher's disease and from patients who had had their spleen removed for other reasons. We found that the rate of glucocerebroside synthesis was essentially the same in both spleen samples. This eliminated the possibility that excessively rapid synthesis was responsible for the accumulation of the lipid.

The third possibility was that diminished activity or lack of the enzyme involved in the degradation of the lipid was the responsible factor. We felt that the best way to investigate the catabolism of the lipid glucocerebroside was to introduce a radioactive tracer into a portion of the molecule. This, however, was not an easy task. We attempted biosynthesis of the lipid with a radioactive glucose as a precursor. The end product was not radioactive enough for our studies. Finally we decided it would be best to include a radioactive glucose molecule in glucocerebroside by chemically synthesizing the molecule.

David Shapiro of the Weizmann Institute of Science in Israel joined Julian N. Kanfer and me in synthesizing the labeled glucocerebroside in 1964. With this labeled compound we were able to show that all normal mammalian tissues contain an enzyme that cleaves the glucose molecule from the ceramide molecule. We were then able to purify some of the enzyme from normal human



SPHINGOLIPIDS involved in lipid-storage disease are all based on a long-chain amino alcohol called sphingosine (*top*), which is combined with a fatty acid (such as stearic acid) to produce lipids called ceramides (*middle*). When a single sugar is added to a ceramide, a cerebroside is the result. If the sugar is glucose, the lipid is called glucocerebroside (*bottom*). When a polysaccharide (complex sugar) is added to the ceramide along with one molecule or more of N-acetyl neuraminic acid, the result is a ganglioside.

spleen tissue. We were also able to show that the activity of the enzyme in spleen tissue from patients with Gaucher's disease was drastically reduced. We concluded that Gaucher's disease was due to a deficiency of the specific enzyme required for the cleavage of glucocerebroside into glucose and ceramide. The following year we showed that Niemann-Pick disease, which involved the accumulation of the lipid sphingomyelin, was due to a deficiency of the enzyme that catalyzes the cleavage of the phosphorylcholine portion of the molecule from the ceramide portion.

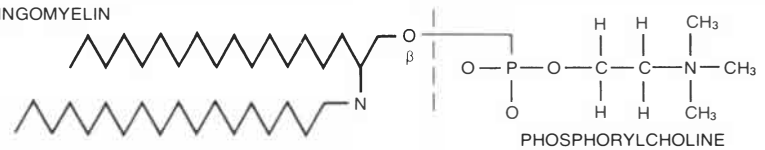
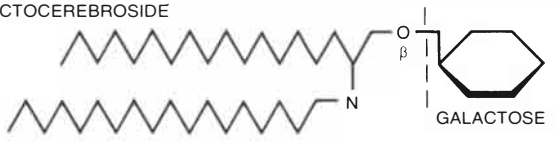
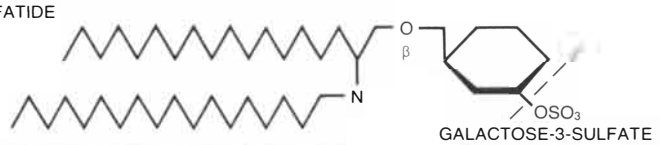
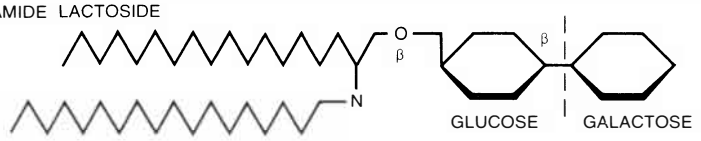
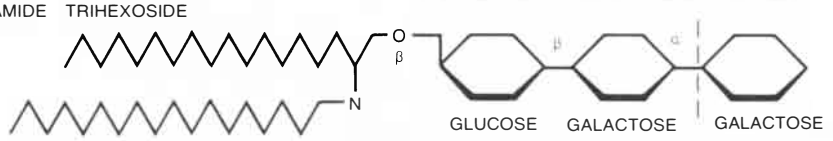
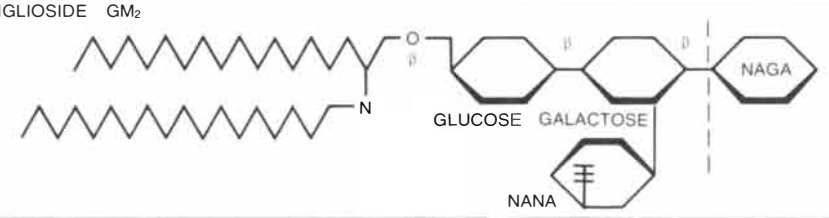
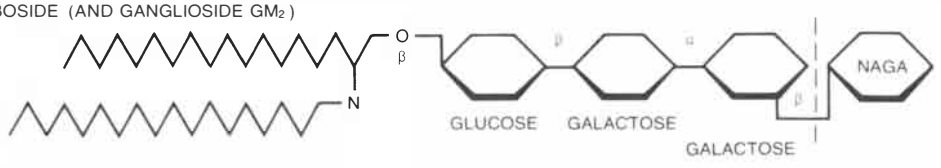
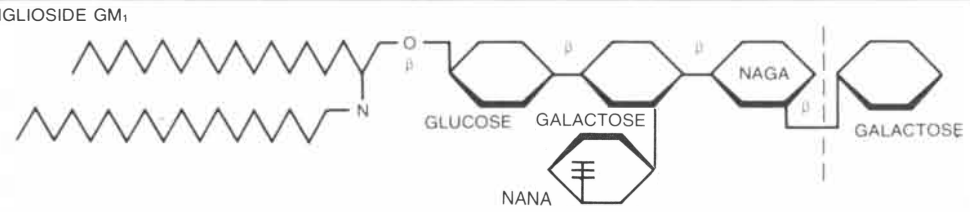
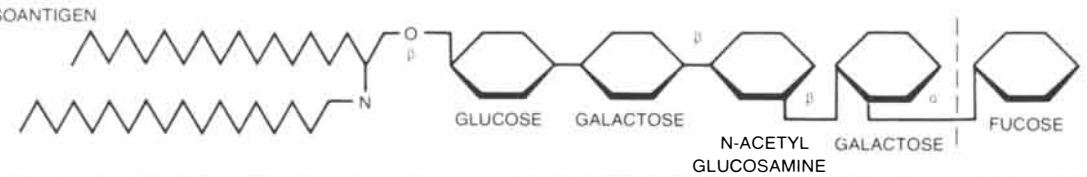
We now had the key that opened Pandora's box. We predicted in 1966 that Fabry's disease, Tay-Sachs disease and generalized gangliosidosis were due to deficiencies of specific enzymes that normally catalyze the cleavage of the accumulating lipid molecule in these diseases. Subsequent investigations have proved the predictions to be correct. In fact, it has now been shown that all the lipid-storage diseases are caused by the decreased activity of specific enzymes required to degrade the various accumulating lipids.

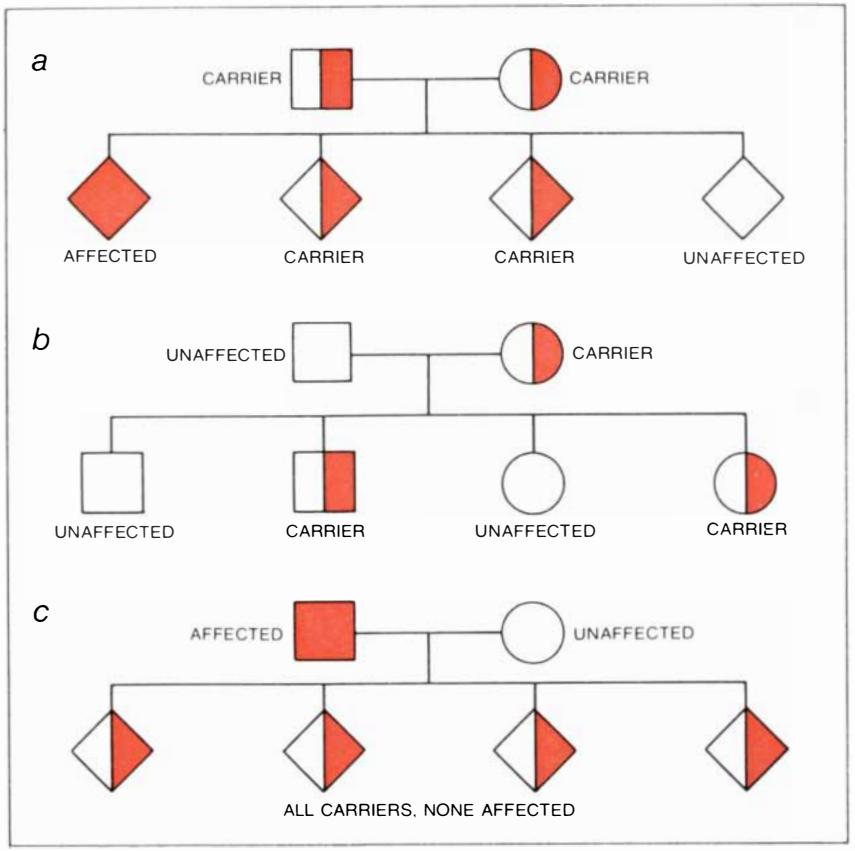
The lipid that accumulates in people with Fabry's disease is ceramide trihexoside (ceramide-glucose-galactose-galactose). Because chemical synthesis of the lipid was not possible, we labeled it by exposing it to gaseous tritium, the radioactive isotope of hydrogen, in a sealed ampoule for one week. A small fraction of the ordinary hydrogen atoms in the lipid were replaced by tritium atoms, but a complex mixture of other substances was also produced. Andrew E. Gal of my laboratory was able to purify the labeled lipid enough for it to be usable for enzyme assays. We found that in normal people the galactose at the end of the lipid molecule was removed by a specific enzyme, and that this enzyme was virtually absent in the tissue of patients with Fabry's disease.

The lipid that accumulates in children

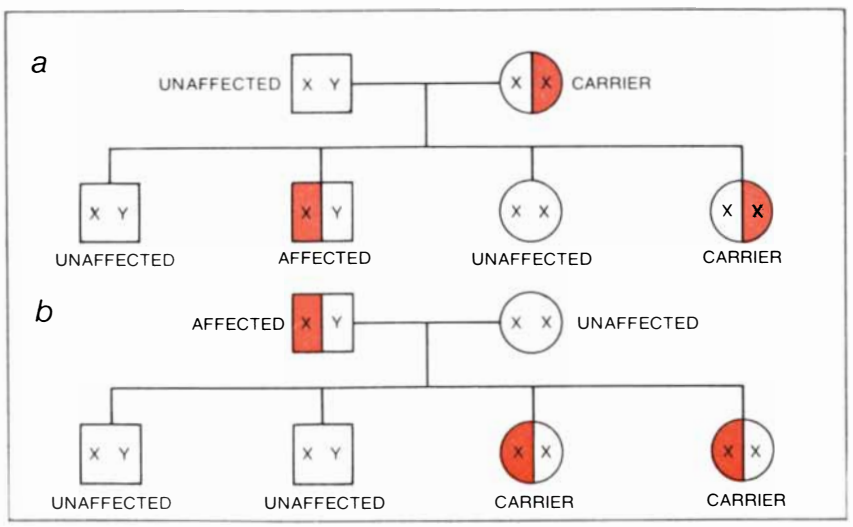
DISEASE	SIGNS AND SYMPTOMS
GAUCHER'S DISEASE	SPLEEN AND LIVER ENLARGEMENT EROSION OF LONG BONES AND PELVIS, MENTAL RETARDATION ONLY IN INFANTILE FORM
NIEMANN-PICK DISEASE	LIVER AND SPLEEN ENLARGEMENT MENTAL RETARDATION, ABOUT 30 PERCENT WITH RED SPOT IN RETINA
KRABBE'S DISEASE (GLOBOID LEUKODYSTROPHY)	MENTAL RETARDATION, ALMOST TOTAL ABSENCE OF MYELIN, GLOBOID BODIES IN WHITE MATTER OF BRAIN
METACHROMATIC LEUKODYSTROPHY	MENTAL RETARDATION, PSYCHOLOGICAL DISTURBANCES IN ADULT FORM, NERVES STAIN YELLOW-BROWN WITH CRESYL VIOLET DYE
CERAMIDE LACTOSIDE LIPIDOSIS	SLOWLY PROGRESSING BRAIN DAMAGE, LIVER AND SPLEEN ENLARGEMENT
FABRY'S DISEASE	REDDISH-PURPLE SKIN RASH, KIDNEY FAILURE, PAIN IN LOWER EXTREMITIES
TAY-SACHS DISEASE	MENTAL RETARDATION, RED SPOT IN RETINA, BLINDNESS, MUSCULAR WEAKNESS
TAY-SACHS VARIANT (SANDHOFF'S DISEASE)	SAME AS TAY-SACHS DISEASE BUT PROGRESSING MORE RAPIDLY
GENERALIZED GANGLIOSIDOSIS	MENTAL RETARDATION, LIVER ENLARGEMENT, SKELETAL DEFORMITIES, ABOUT 50 PERCENT WITH RED SPOT IN RETINA
FUCOSIDOSIS	CEREBRAL DEGENERATION, MUSCLE SPASTICITY, THICK SKIN

TEN LIPID-STORAGE DISEASES caused by excessive accumulation of sphingolipids in tissue cells are listed, along with the principal signs and symptoms of the disease, the major lipid that accumulates and the defective enzyme. The lipids accumulate because of a deficiency of the specific enzymes that normally catalyze the cleavage of the lipid molecules. The broken lines indicate the site of the cleavage. The alpha and beta signs indicate the type of bond. NANA is an abbreviation for N-acetyl neuraminic acid and NAGA stands for N-acetyl galactosamine.

MAJOR LIPID ACCUMULATION	ENZYME DEFECT
GLUCOCEREBROSIDE 	BETA-GLUCOSIDASE
SPHINGOMYELIN 	SPHINGOMYELINASE
GALACTOCEREBROSIDE 	BETA-GALACTOSIDASE
SULFATIDE 	SULFATIDASE
CERAMIDE LACTOSIDE 	BETA-GALACTOSIDASE
CERAMIDE TRIHEXOSIDE 	ALPHA-GALACTOSIDASE
GANGLIOSIDE GM_2 	HEXOSAMINIDASE A
GLOBOSIDE (AND GANGLIOSIDE GM_2) 	HEXOSAMINIDASE A AND B
GANGLIOSIDE GM_1 	BETA-GALACTOSIDASE
H-ISOANTIGEN 	ALPHA FUCOSIDASE



AUTOSOMAL RECESSIVE TRAITS are transmitted by chromosomes other than the *X* and *Y* sex chromosomes. Nine lipid-storage diseases are known to be transmitted by this mode of inheritance. A carrier has one normal and one abnormal chromosome affecting a trait, an unaffected person has two normal chromosomes and an affected person has two abnormal chromosomes. Because the traits are recessive, to be affected by a disease an offspring must receive a defective chromosome (*color*) from each parent. If both parents are carriers, statistically one in four offspring will be affected (*a*). If only one parent is a carrier (*b*) or affected (*c*), some offspring will be carriers but none will be affected. Males are indicated by squares, females by circles. Diamonds indicate offspring can be of either sex.

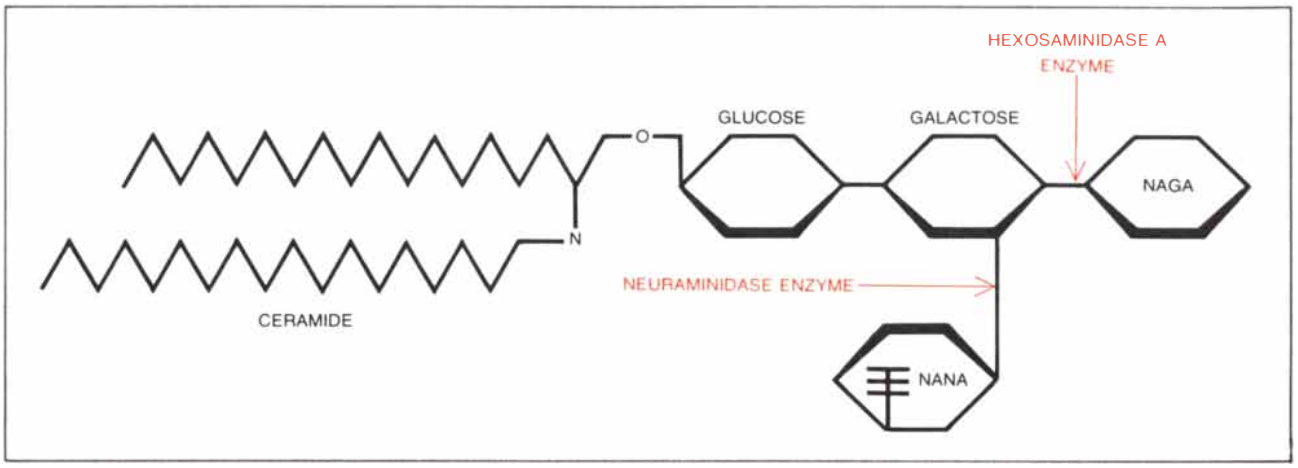


SEX-LINKED RECESSIVE TRAITS transmitted by the *X* chromosome will cause all male offspring who receive defective chromosome (*color*) to be affected by the disorder. Statistically half of the sons of a carrier mother will be affected and half of the daughters will be carriers (*a*). If a father has an *X*-linked recessive trait, he will be affected by the disorder but none of his sons will be. All his daughters will be carriers (*b*). Of the known lipid-storage diseases, only Fabry's disease is a recessive characteristic of the *X* chromosome.

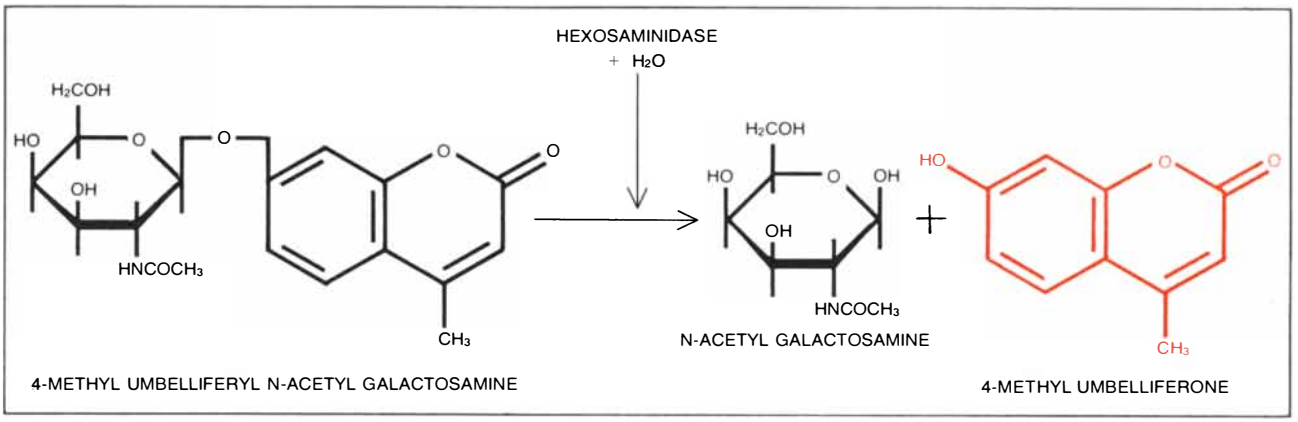
with Tay-Sachs disease is branched in the terminal portion of the molecule. Conceivably the enzymatic cleavage could occur at either branch [*see top illustration on opposite page*]. In one case N-acetyl neuraminic acid would be cleaved from the lipid; in the other case N-acetyl galactosamine would be split off. After several years of intensive work Edwin Kolodny, who was working in my laboratory at the time, was able to synthesize this lipid with a radioactive label in the N-acetyl neuraminic acid portion of the molecule. We found that the activity of the enzyme that cleaves the neuraminic acid portion from the lipid is perfectly normal in the tissue of people with Tay-Sachs disease. We then undertook the labeling of the sugar portion of the lipid and eventually showed that the activity of the enzyme that splits the sugar N-acetyl galactosamine from the lipid is extremely low in the tissues of children with Tay-Sachs disease.

Thus, after this long search, we finally identified the metabolic defect responsible for Tay-Sachs disease. While our investigation was in progress a most important contribution was made by Shintaro Okada and John S. O'Brien of the University of California at San Diego School of Medicine. Following a report by British workers in 1968 that human spleen contains two enzymes called hexosaminidase *A* and hexosaminidase *B*, Okada and O'Brien demonstrated that hexosaminidase *A* activity is almost totally lacking in the tissues of people with Tay-Sachs disease. They also established that both enzymes are active in cells taken from the amniotic fluid surrounding a normal fetus. This opened the way for prenatal diagnosis of fetuses for Tay-Sachs disease. The first diagnosis of a fetus with the disease was reported in 1970 by Larry Schneck of Kingsbrook Jewish Medical Center in New York.

Prenatal diagnosis involves obtaining a sample of the amniotic fluid surrounding the fetus. The fluid contains viable cells from the fetus that can be grown in a tissue culture and then tested [see "Prenatal Diagnosis of Genetic Disease," by Theodore Friedmann; SCIENTIFIC AMERICAN, November, 1971]. In 1971 prenatal diagnoses of Fabry's disease, Niemann-Pick disease and Gaucher's disease were made in my laboratory. In the latter part of the same year K. Suzuki of the University of Pennsylvania School of Medicine made a prenatal diagnosis of a fetus with the lipid-storage disorder known as Krabbe's disease. When prenatal diagnosis confirms that a fetus is affected with the disease, the



GANGLIOSIDE GM_2, the lipid that accumulates in nerve cells of children with Tay-Sachs disease, is branched at the terminal portion of the molecule. Either the terminal N-acetyl neuraminic acid (NANA) or the terminal N-acetyl galactosamine (NAGA) can be cleaved from the molecule at the site indicated by the arrows. Tests show that the activity of the enzyme that cleaves NANA is perfectly normal in the tissues of people with Tay-Sachs disease but that the enzyme that splits off NAGA is almost totally lacking. Thus the accumulation of the lipid ganglioside GM_2 in Tay-Sachs disease has been shown to be caused by a lack of the enzyme hexosaminidase *A*.



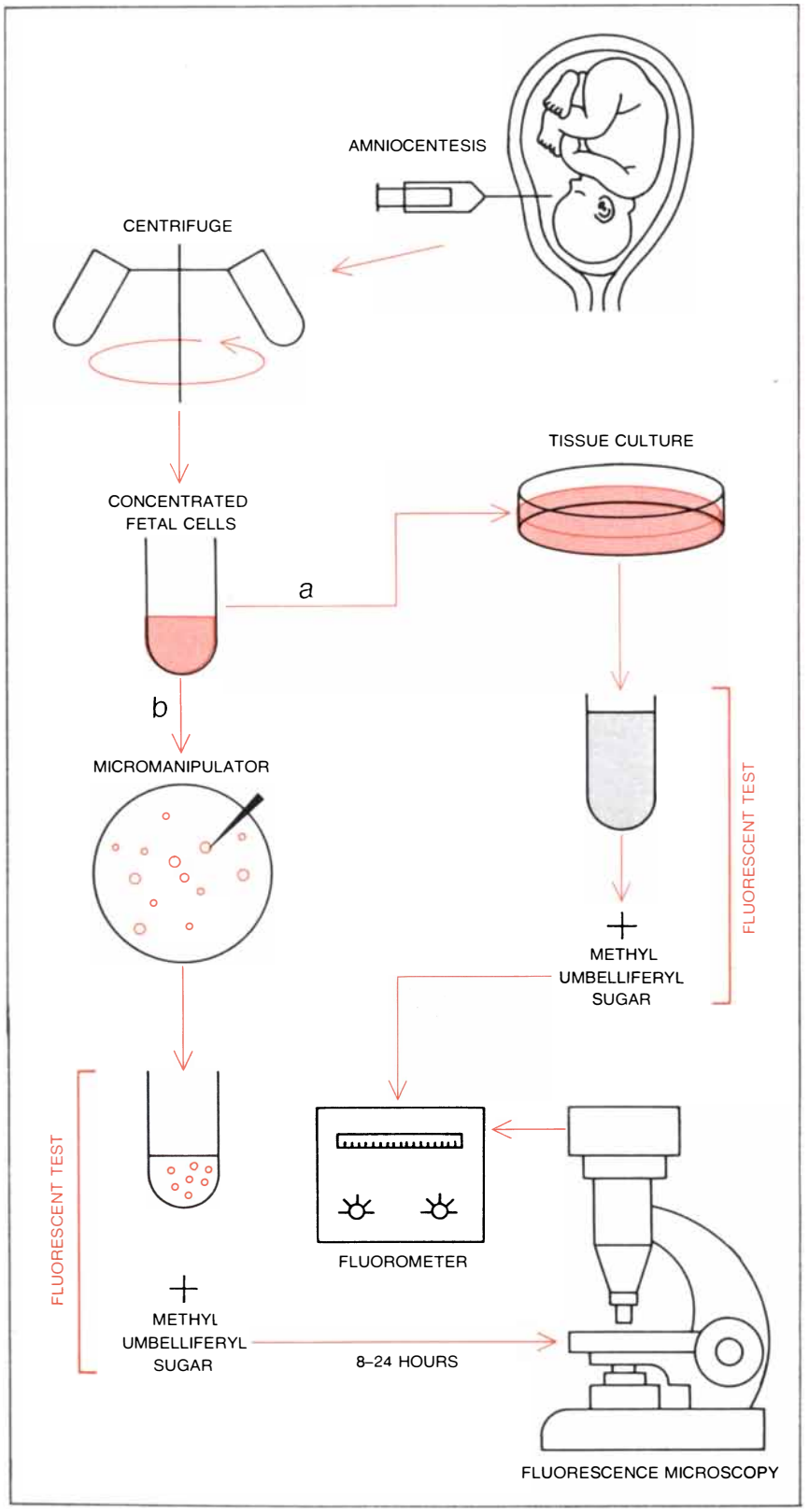
ASSAY OF HEXOSAMINIDASE ACTIVITY is carried out by adding a homogenate of tissue to a solution containing the synthetic sugar, 4-methyl umbelliferyl N-acetyl galactosamine or N-acetyl glucosamine. The enzyme hexosaminidase splits the amino sugar from the molecule. The other product of the reaction, 4-methyl umbelliferone, fluoresces under ultraviolet radiation. Intensity of the fluorescence is a measure of the activity of the enzyme. Sugar derivative can be varied depending on what enzyme is to be assayed.

parents have the choice of continuing the pregnancy or having it terminated by abortion.

Other benefits stemming from fundamental investigations into the biochemistry of lipid-storage diseases include the development of tests for determining the activity of enzymes that are so sensitive that even carriers, who have only somewhat less than normal activity of the particular enzyme involved, can be identified. It was quickly established that the degree of enzyme deficiency in people with lipid-storage disorders is similar in most tissues of the body. A very sensitive test using white blood cells has been developed for detecting people who have Gaucher's disease and Niemann-Pick disease or are carriers. Only a small sample of blood is required for the diagnosis. The enzymes involved in these two diseases are also found in skin cells. A small sample of skin can be taken and grown in tissue culture; the activity of a particular enzyme is then measured. The cell-culture technique can also detect carriers.

There has been an important new development in the prenatal diagnosis of lipid-storage diseases. M. L. Moss and Warren W. Harris of the Oak Ridge National Laboratory have reported a technique that greatly speeds up the process. They take a small number of viable cells from the amniotic fluid and place them in a solution containing a synthetic sugar derivative, such as 4-methyl umbelliferyl glycoside. The terminal sugar on the derivative molecule can be varied, depending on what enzyme is to be assayed. The enzyme cleaves the sugar from the molecule. The other product of the reaction, 4-methyl umbelliferone, fluoresces when it is exposed to ultraviolet radiation. Therefore the extent of the reaction and the activity of the enzyme can be determined by measuring the intensity of the fluorescence by fluorescent microscopy. The fast fluorescent assay can be carried out within 24 hours, whereas the assay of cells grown in tissue culture takes from 10 days to four weeks.

The next step in the research in lipid-storage diseases is to find a way to treat affected individuals by supplying the appropriate enzyme. To achieve this result sufficient quantities of enzyme are



PRENATAL DIAGNOSIS of lipid-storage diseases involves removing a sample of amniotic fluid containing fetal cells, then centrifuging the sample to separate fluid and cells. The standard technique (*a*) is to grow the fetal cells in a tissue culture, which may take from 10 days to four weeks. Cells from the culture are added to a synthetic sugar solution to assay the activity of a sugar-cleaving enzyme. The enzyme's action releases a chemical that fluoresces under ultraviolet radiation. The intensity of the fluorescence depends on the amount of enzyme present. A new, faster technique (*b*) involves taking a small number of viable cells directly from the centrifuged sample with a micromanipulator and adding the cells to the synthetic sugar solution. Activity of enzyme is measured by fluorescent microscopy.

required. The problems involved in purifying the enzymes are enormous, but promising techniques are currently under development in several medical centers.

In 1971 we intravenously injected a small amount of purified hexosaminidase A into an infant with Tay-Sachs disease to see if the enzyme would break down the lipids stored in the nerve cells of the brain. The enzyme was rapidly removed from the infant's bloodstream and taken up by the liver. The enzyme degraded a lipid that is normally found in the blood and the enzyme was rapidly withdrawn from the circulation. However, the enzyme did not appear to cross the blood-brain barrier, the mechanism by which the blood vessels of the brain exclude certain substances that interfere with the metabolism of brain cells. This result indicates that simple intravenous injection of the missing enzyme will probably not help patients with a defect of lipid storage in the brain.

One possibility being explored is the modification of the enzyme so that it can pass the blood-brain barrier and still retain its activity. Or the blood-brain barrier itself could be opened temporarily so that the enzyme can enter the brain. Stanley I. Rapoport and his associates at the National Institute of Mental Health have been able to open the barrier in animals by injecting a concentrated solution of sucrose into an artery leading to the brain. Although the idea is attractive, such procedures could give rise to serious complications if they are attempted on infants during their first three years of life.

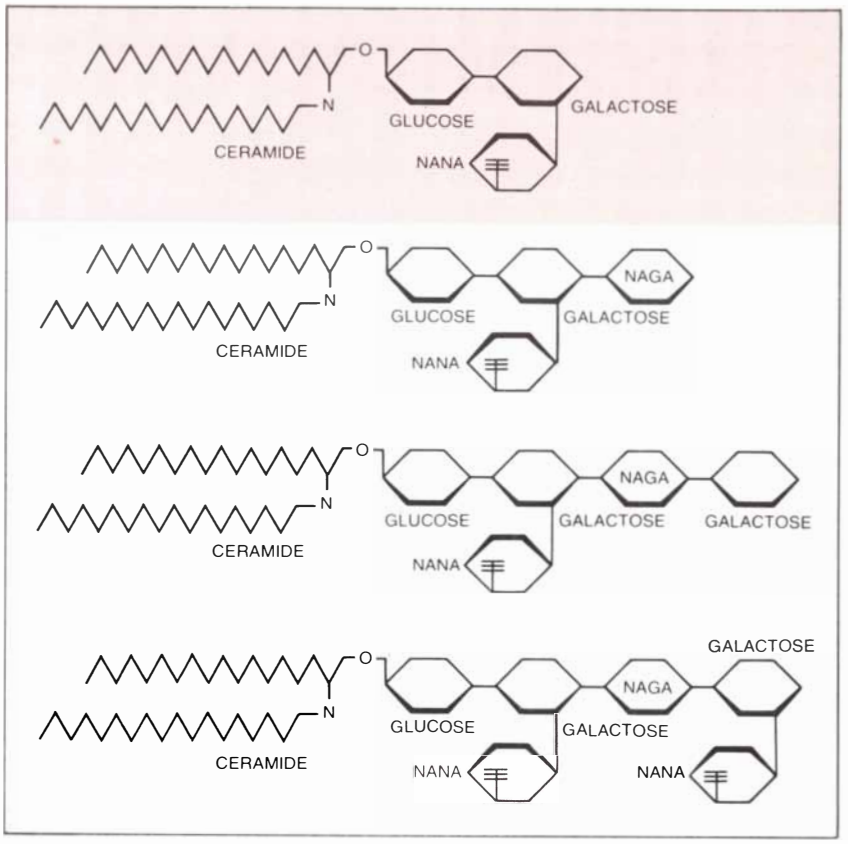
Some investigators are looking into the possibility of introducing the genetic message for producing a missing enzyme by means of a nonpathogenic virus. At present the prospect of developing such a procedure for the treatment of lipid-storage diseases seems remote.

The most promising development is the treatment of people with lipid-storage diseases that do not affect brain cells. Here again the treatment is the injection of suitable enzymes. Some children with Gaucher's disease have a sufficient activity of the enzyme glucocerebrosidase to prevent the accumulation of lipids in the nervous system but show an accumulation of lipids in the spleen and the liver. Although it has been difficult to produce pure glucocerebrosidase, enzyme-replacement trials have recently been undertaken by my colleagues and me at the National Institutes of Health.

We have also been able to purify a small amount of the enzyme involved in

Fabry's disease from human placental tissue. We injected it into two males with the disease and obtained encouraging results. We appear to be at a stage comparable to the first injection of insulin for diabetes. If the therapy proves successful, new methods for producing large quantities of pure enzyme will have to be developed, as they were for insulin. The study of lipid-storage diseases has yielded some important clues for investigating other disorders, specifically other forms of mental retardation and cancer. In the course of purifying enzymes from human placental tissue for injection into patients with lipid-storage diseases, William Johnson of my laboratory found the enzyme alpha-hexosaminidase, which splits off derivatives of N-acetyl galactosamine that are attached to another molecule by what is known as an alpha bond. Hexosaminidase *A*, the deficient enzyme in patients with Tay-Sachs disease, cleaves N-acetyl galactosamines that are attached by a beta bond.

By analogy with Tay-Sachs disease we can predict that individuals with a deficiency of the enzyme alpha-hexosaminidase will be found. Such affected individuals will probably have a considerable degree of mental retardation but will probably not show the extensive accumulation of lipid in nerve cells found in Tay-Sachs disease. The logic of the analogy can be extended. We know that other sugars are linked to proteins. It may be that the inability to cleave these sugars from protein is the cause of still undiagnosed forms of mental retardation. The metabolic derangements would be very much like the one found in lipid-storage diseases: the lack of a particular enzyme for splitting the sugar from the remainder of the molecule.



FAMILY OF LIPID-SUGAR COMPLEXES called gangliosides are found in high concentrations on the surfaces of most mammalian cells. The surfaces of normal cells contain the entire family of these gangliosides but most strains of cultured mouse cells, when made cancerous by certain viruses, have only the simplest member of the family (*top*). Synthesis of the next-largest ganglioside apparently is blocked by the lack of the enzyme that catalyzes the addition of the sugar N-acetyl galactosamine. The deficiency is exactly the opposite of the one in Tay-Sachs disease, where the lack of an enzyme prevents cleavage of the same sugar. The lack of the larger gangliosides on the surface of cancerous cells may play a part in the loss of contact inhibition characteristic of cancer. Normal cells will stop growing when they contact other cells but cancerous cells lose this inhibition and continue to grow.

The cells of a cancer continue to grow after they come in contact with other cells. Normal cells will stop growing when they touch another cell, a phenomenon known as contact inhibition. There is a family of lipid-sugar complexes called gangliosides that are highly concentrated on the surfaces of most mammalian cells. Normal cells contain the entire family of these lipid-sugar complexes, but some cancerous cells have only the simplest member of the family [*see illustration on this page*]. For example, the only ganglioside in cells transformed into tumor cells by certain viruses has two molecules of sugar and one molecule of N-acetyl neuraminic acid. There appears to be a blockage in the synthesis of the next-largest ganglioside, which is the same ganglioside involved in Tay-Sachs disease. The change in the virus-transformed cells is caused by a deficiency of the enzyme that catalyzes the addition of a sugar molecule. The situation is exactly the opposite of the one in Tay-Sachs disease, where the lack of an enzyme prevents the cleavage of the same sugar molecule.

The deficiency of enzyme activity in genetic diseases is very likely due to an alteration in the sequence of the amino acids that comprise the molecules of various enzymes. The change is a result of a mutation in the genetic code that specifies the sequence of amino acids in the enzyme. It is now well recognized that the growth and division of cells are regulated by feedback systems within the cell, and there is some evidence that the feedback substances are proteins. It could be that a mutation of the genetic code for the formation of the feedback protein results in a protein that cannot perform its function effectively. Possibly such a modified protein would fail to exert control over the growth of the cells. Uncontrolled growth, or cancer, would be the result.

It is known that some strains of mice are much more susceptible than other strains to cancer-causing agents such as certain viruses and chemicals. The trait is probably inherited. A good candidate for the cause might be an inherited mutant protein on the cell's surface. The altered protein might allow cancer-causing agents to penetrate into the cell more readily than the normal protein would. It seems likely that this possibility can be tested with procedures that are currently available. Hence many of the concepts established in the course of investigating lipid-storage diseases may be helpful in gaining a better understanding of other biochemical disorders.

MATHEMATICAL GAMES

An astounding self-test of clairvoyance by Dr. Matrix

by Martin Gardner

An institute in Los Angeles (I must leave it nameless for the present) claims to be able to train anyone in the power of clairvoyance: the ability to perceive hidden or distant objects by extrasensory perception. Students are given tests before and after an intensive six-week training course—for which they pay a fee of $500. Their scores on the final test are, I am told, invariably high.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

The number test

A	B	D	H	P
C	C	E	I	Q
E	F	F	J	R
G	G	G	K	S
I	J	L	L	T
K	K	M	M	U
M	N	N	N	V
O	O	O	O	W
Q	R	T	X	X
S	S	U	Y	Y
U	V	V	Z	Z
W	W	W		
Y	Z			

The alphabet test

A young lady who is a file clerk at the institute wrote me recently that the school's owner and founder looks suspiciously like Dr. Irving Joshua Matrix, the famous numerologist about whom I have written many columns. And his chief assistant, she adds, closely resembles Matrix' attractive Eurasian daughter, Iva. My informant managed to photocopy the clairvoyance test that students are given at the end of their training. The test is given below and the "targets" for each of the 25 problems are given upside down on page 101. Readers are urged not to look at the targets until they have completed the test.

Before starting the test you must have on hand the following material: a pencil and several sheets of paper, an unused matchbook, a deck of playing cards, a pair of dice, a ruler, scissors, a box of raisins, an eight-cent stamp, a penny, a nickel, a dime, a quarter and a King James Bible. Answer each question as quickly as you can, without thinking too long or hard about what the target might be. Write your answers on a sheet of paper, and when you have finished the test, compare your answers with the given targets to see how many hits you have made. A score of more than 15 hits indicates, according to the literature of the California school, an extremely high degree of clairvoyant ability.

1. Draw a circle around any one of the 16 numbers in the upper illustration at the left. Cross out all cells in the same row and column. Draw a circle around one of the remaining nine numbers. Cross out the cells in its row and column. Circle one of the four remaining numbers. Cross out its row and column. Circle the single number that is left. Add the four circled numbers and write down the sum.

2. From an unused matchbook, which should have 20 matches, tear out any number of matches up to nine and discard them. Count the matches that remain. Add the two digits of this number, then tear that many matches from the book and put them aside. Tear out two more matches. Write down the number of matches still in the matchbook.

3. From a deck of 52 playing cards remove the red queens, the black aces, the red fours, the six of clubs and the jack of diamonds. Shuffle the deck, hold it face down and take the cards in pairs from the top. If the first pair contains a red and a black card, turn it face down and discard it. If both cards are red, put them face up on the table to start a pile of red pairs. If both cards are black, put them face up at another spot on the table to form a pile of black pairs. Continue through the deck in this fashion, discarding all red-black pairs and building up the piles of red-red and black-black pairs. When you finish, count the cards in each pile. Subtract the smaller number from the larger and write down the difference.

4. Draw a simple geometric figure. Inscribe within it another and different simple geometric figure.

5. Write the name of a wild beast.

6. Place two dice, *A* and *B*, on the table, any side up. Add the number on the top of *A* to the number on the bottom of *B*, then find the chapter of Genesis (in a King James Bible) that corresponds to the sum. Locate the verse indicated by the sum of the top of *B* and the bottom of *A*. Write the first word of that verse.

7. Think of any number, *k*, between 10 and 50. Place your finger on the bottom ESP symbol in the illustration on the opposite page. Say "One." Tap the next symbol above it, saying "Two," and continue upward, counting aloud with each tap. When you come to the colored star, turn right and proceed counterclockwise around the circle, tapping and counting until you reach *k*. This may take you more than once around the circle. If it does, ignore the tail portion of the illustration.

After you tap the symbol on the count of *k*, stop and reverse direction, then tap and count from one to *k* as before, but this time go around the circle clockwise. The symbol you tapped for *k* should get the count of one. (Do not make the mistake of starting the count on the symbol next to it.) Ignore the tail portion. Stop when you reach *k* again and write down the symbol just tapped.

8. Write a two-digit number between 10 and 50 that meets the following provisos: both digits must be odd and the digits must not be alike. (For example, 11 is ruled out because the same digit is repeated.)

9. Take any 20 cards from the deck and hold them face down. Turn face up the top pair of cards, leaving them on top

of the packet, and cut the packet at any spot. Again reverse the top two cards and cut. Continue turning pairs and cutting for as long as you like. In reversing top pairs you will, of course, sometimes turn reversed cards face down again, but it does not matter. The procedure is designed to randomize the number of reversed cards in the packet.

Deal the randomized cards in a row on the table, taking care not to reverse any cards as you do so. Now turn over all the cards in even positions (2, 4, 8, . . . , 16). Count the number of cards that are face up in the row and write that number.

10. Form three equal piles of raisins on the table. There must be at least four raisins in each pile, and the number in each pile must be the same. Call the piles *A*, *B* and *C*.

Take two raisins from *A* and put them in *B*.

Take three raisins from *C* and put them in *B*.

Count the raisins in *A*, then take the same number from *B* and put them in either of the other two piles.

Take one raisin from either *A* or *C* and put it in *B*.

Write the number of raisins in *B*.

11. Think of any letter of the alphabet. Inspect the five columns in the lower illustration on the opposite page and circle that letter wherever it appears. Jot down the letter that is at the top of each column containing the letter you thought of. Convert these letters to numbers by using the cipher $A = 1$, $B = 2$, $C = 3$ and so on. Add the numbers. Using the same cipher key, convert the sum back to a letter. Write that letter.

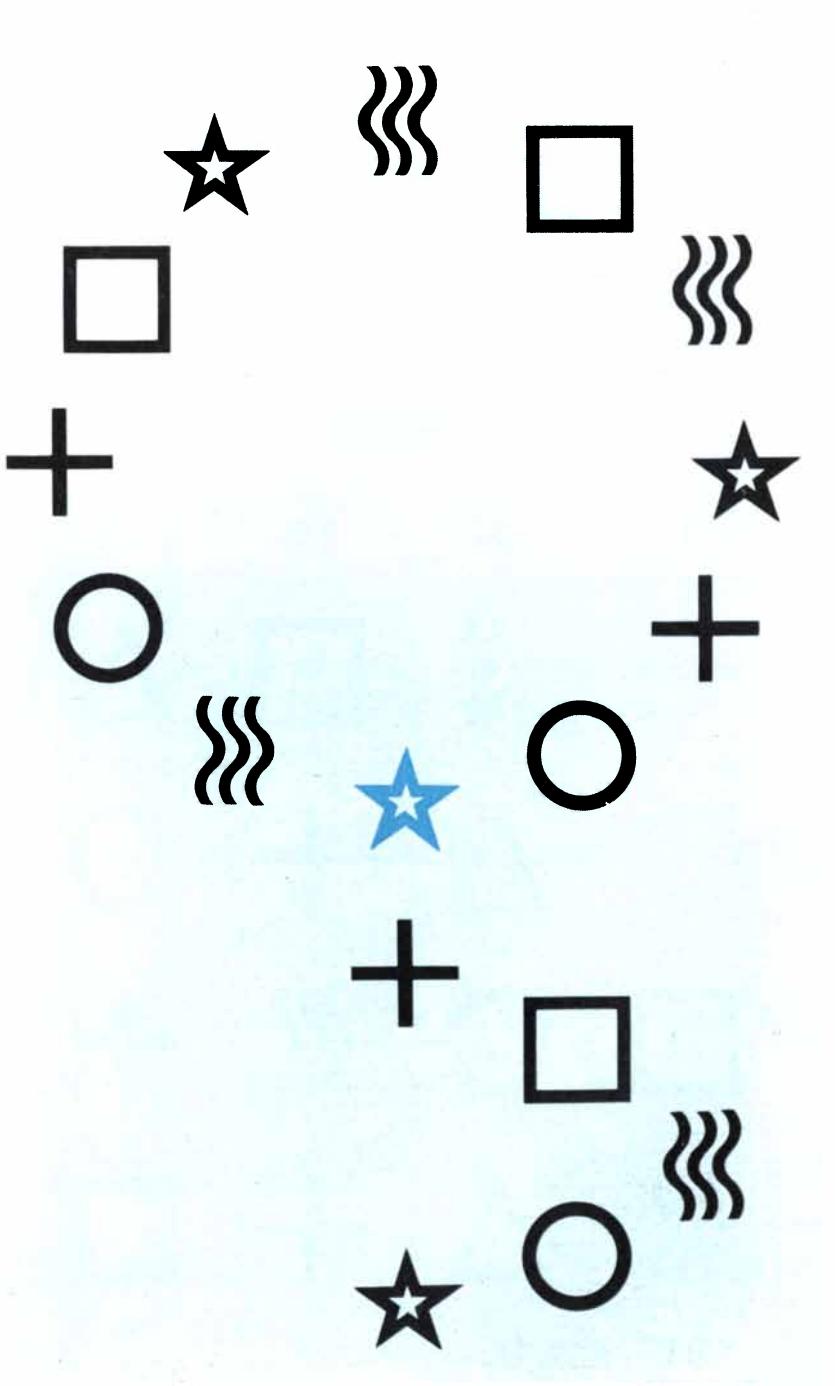
12. Place one die on the left side of the table, the other on the right, each with any side up. Now obtain four products as follows: multiply the top two numbers, multiply the bottom two numbers, multiply the top of the left die by the bottom of the right, and multiply the top of the right by the bottom of the left. Add the four products and write down the sum.

13. Take a square sheet of paper, about eight inches on a side, and fold it in half four times so that the creases will mark a four-by-four matrix of small squares. Fold each crease both forward and back so that the paper will fold easily either way along every crease. Number the cells from 1 to 16 as shown in the upper illustration on the opposite page.

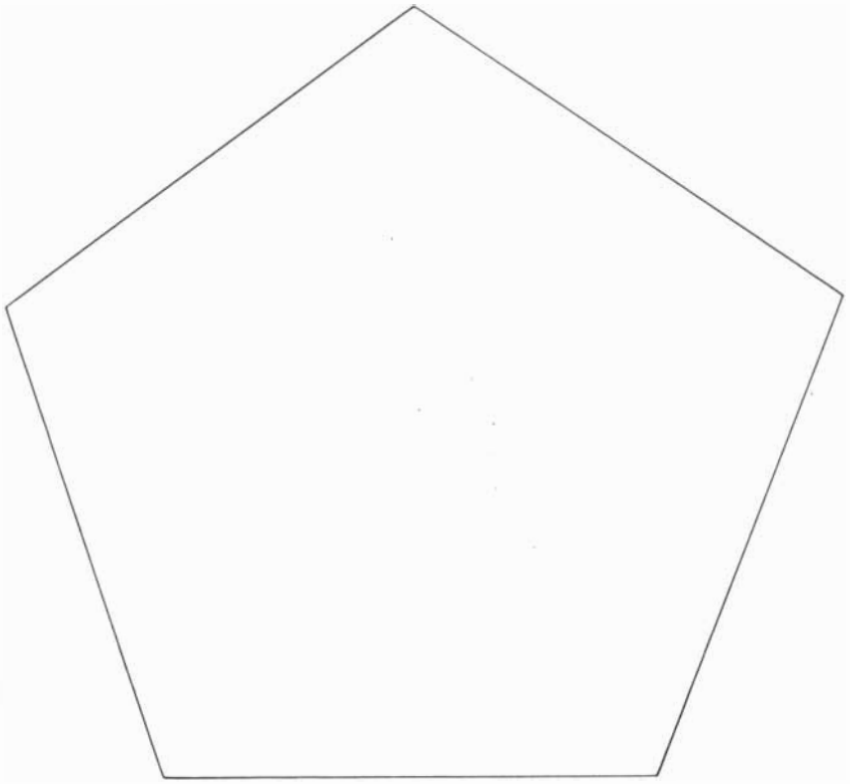
Fold the square into a one-by-one packet, making each fold in any manner. Indeed, your folding may be as tricky as you please, including the tucking of folds between folds if you like. With a pair of scissors, trim away the four edges of the final packet so that it will contain 16 separate squares. Spread the squares on a table. Some will be number side up, others number side down. Add all the face-up numbers and write the sum.

14. Write any number, provided only that its digits are not all alike. Form a new number by rearranging those digits any way you wish. Subtract the smaller number from the larger. Add the digits of the answer. If that gives a number of more than one digit, add the digits again, and continue in this way until a single digit remains. Increase that digit by 4 and write down the sum.

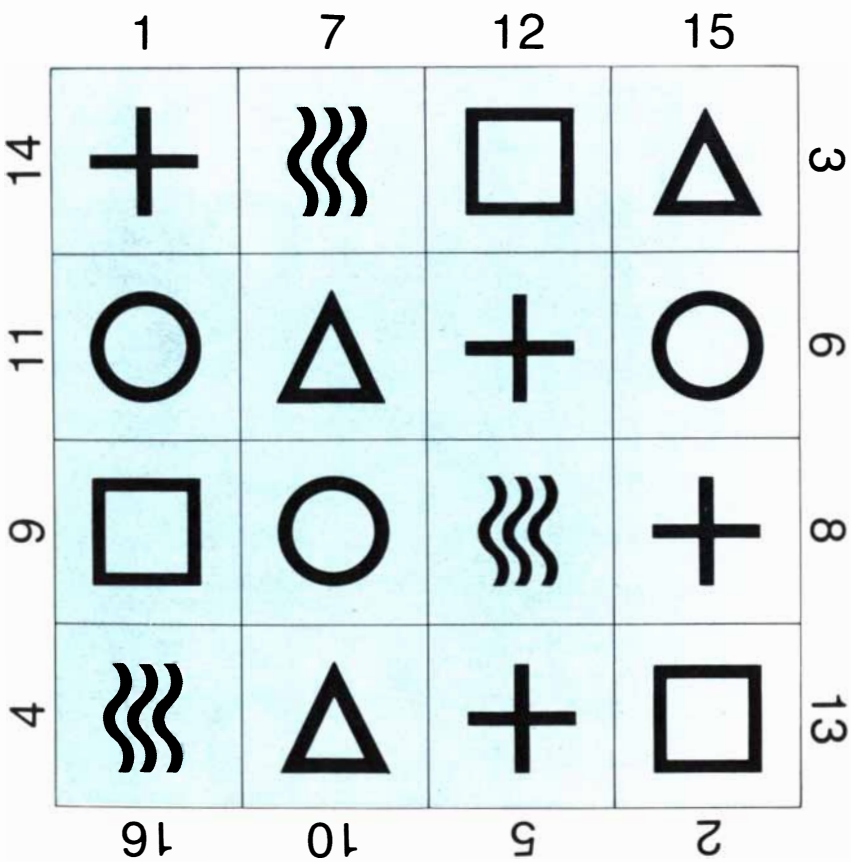
15. A pentagram, the mysterious occult symbol of medieval witchcraft and the ancient Pythagoreans, is shown in the top illustration on the next page. With a pencil put a dot anywhere you like inside the pentagonal border or on the border. Draw perpendiculars from the dot to each of the pentagon's five sides, extending the sides with a ruler if necesssary. The perpendiculars are easily



The "Q" ESP test



The pentagram test



The rotating-matrix test

drawn by using the corner of a sheet of paper to provide a right angle. Add the lengths of the five perpendicular line segments by marking them along the edge of a sheet of paper. Measure the total length carefully with the ruler and write down its length to the nearest half-inch.

16. Write the name of a city that is the capital of a large foreign country. (Not London—that is too obvious.)

17. Place in a row (in order from left to right) a penny, a dime, a nickel, an eight-cent stamp and a quarter. Put a matchbook on any of the five objects. A move consists in transferring the matchbook to an adjacent object, either left or right. Of course if the matchbook is at either end of the row, the next move will be limited to one direction. Move the matchbook randomly left and right as many times as indicated by the value in cents of the coin (or stamp) on which you first placed it. When you finish, if the matchbook is not on the penny, remove the penny from the row. Again move the matchbook as many times as indicated by the value of the object it is on. If the matchbook is not on the quarter when you finish, take away the quarter. Move the matchbook once. Write down the value of the coin (or stamp) on which it now rests.

18. A row of five face-up playing cards is formed on the table. From left to right the cards must be the nine of diamonds, the four of hearts, the queen of hearts, the ace of diamonds and the seven of clubs. As you can see, there is one picture card, one ace and one black card. Look the five cards over carefully, focus your attention on one of them and then write down its name.

19. Think of a number from 1 to 16 inclusive. Find that number on the border of the lower illustration at the left and turn the page so that the number is right side up above the matrix. With the page still turned count the cells of the matrix beginning with the cell in the upper left corner until you reach your chosen number. Write down the ESP symbol inside that cell.

20. Write the name of a flower.

21. Shuffle a deck of cards. Assign to the face cards any value you wish from one through 10. (For example, you may decide to give each face card a value of 3.) Start dealing the cards face up from the top of the deck to form a pile on the table. Say "Ten" when you deal the first card, then continue with "Nine, eight, seven...," counting backward as each card is dealt. If you put down a card that happens to coincide in value with the number named, stop dealing on that pile and start another face-up pile

next to it. If you fail to hit a coincidence by the time you deal and say "One," cover the last card dealt with a face-down card from the cards in your hand and begin another pile. Form four piles in this manner.

To recapitulate: As each face-up pile is being formed, count backward from 10 until you "hit" or until you count "One" without having made a hit. The "failure" piles are covered with a face-down card. While you deal, be sure to keep in mind the value you assigned to the face cards. In our example it was 3. Consequently, if you called "Three" when you dealt a face card, it is counted as a hit and you start another pile.

After four piles have been formed add the values of your hits—that is, the face-up cards on the tops of piles. A face card has, of course, whatever value you assigned to it. Call the sum of your hits *X*. Discard *X* cards from those that have not been dealt. Count the cards left and write down the number.

22. Write a two-digit number between 50 and 100. Both digits must be even and not alike.

23. Roll a die on the table. Think of a number from 1 through 6 and put a second die on top of the first one so that the number you selected is on top of the stack. To your thought-of number add the sum of the two touching faces of the dice. Think of another number from 1 through 6 and add it to the previous total.

Remove the top die and turn it so that your second chosen number is on top. Place it alongside the other die. Lift up both dice and add the sum of their bottom faces to the previous total. Add 3 to the last result and write down the final sum.

24. Write the name of a color.

25. Put 10 cards on the table. Turn them so that five are face up and five face down. Shuffle them around the tabletop, mixing them thoroughly, then separate them into two sets, *A* and *B*. Reverse all the cards in set *B*. Count the face-up cards in each set and write down the difference between the two numbers.

26. Select any digit from 1 through 5. Call it *k*. Look at the *k*th chapter of Revelation, in a King James Bible, and count to its *k*th word. Write the word.

The short problems given in May prompted many informative letters. Edward N. Peters, professor of pediatrics at the University of Rochester, found a general procedure for constructing maximum-length rook tours on square boards of any order, but I must postpone giving the lengthy details. Mitchell Taibleson of St. Louis wrote that the retrograde-chess problem given in May had been devised about 20 years ago by the mathematician Raymond Smullyan and that it is one of a large collection of unpublished chess problems invented by Smullyan when he was a student at the University of Chicago.

So many letters were received on the polypowers of Aristid V. Grosse, and so many are still coming in, that I cannot attempt a summary now. Perhaps in a later column I can report some of the highlights.

Several readers "proved" that an 11-sided polygon could not be formed with squares and equilateral triangles of unit sides. The flaw, of course, was failing to realize that a side could be more than one unit long.

Alan Brown of Urbana, Ill., pointed out that if the word "exactly" is removed from each of the 10 statements in the logic problem, there is a different and unique solution: the first five statements are true, the last five false.

Targets for the Clairvoyance Test in "Mathematical Games"

1. 34.
2. 7.
3. 2.
4. Triangle with inscribed circle.
5. Lion.
6. And.
7. Star.
8. 37.
9. 10.
10. 8.
11. It is the letter you thought of.
12. 49.
13. 68.
14. 13.
15. 9¾ inches.
16. Paris.
17. Five cents.
18. Four of hearts.
19. Cross.
20. Rose.
21. 8.
22. 68.
23. 24.
24. Blue.
25. 0.
26. The.

For some of the test problems I am indebted to the following people: Fitch Cheney (Problem 12), Henry Christ (21), Frederic DeMuth (2), Victor Eigen (6, 13, 15, 23, 26), Uri Geller (4, 16), Walter B. Gibson (19), Robert Hummer (9, 25), Stewart James (3), David L. Silverman (11), Dai Vernon (18), Jack Yates (17).

These calculators have such special significance for scientists and engineers that we devoted this space to describe them in some detail. Other new instruments and systems stemming from our measurement/computation technologies are being developed.

Three alternatives (two of them brand new) to the drudgery of paper-and-pencil mathematics.

Almost 300 years ago, Gottfried Wilhelm Leibniz wrote, "It is unworthy of excellent men to lose hours like slaves in the labor of calculation." He was right, but how could anyone avoid it then?

Today, in any field of science or engineering, tedious manual calculation is neither wise nor necessary because things have changed significantly for the better, even in the last few months. We're convinced you should never again labor with slide rule, tables, scratch pads and adding machines . . . wasting your creative time getting answers that aren't as accurate as you'd like.

Any of the economical calculators that we describe in this month's message is as easy to operate as an adding machine but incomparably more powerful. All are pre-programmed to perform not only the basic arithmetic operations but also transcendental and statistical functions. All calculate positive and negative numbers throughout a 200-decade range. All automatically keep track of the decimal point and can display answers to the tenth significant digit.

One of the traits that sets them apart from the recent flood of electronic calculators is a four-register operational stack that is solidly based on computer theory. The stack automatically stores intermediate results obtained during your calculations — whether they be serial, chain or mixed chain — and brings them back to the working register when they are needed to complete the calculations. In plain English, the stack relieves you of the necessity to make scratch notes and re-enter intermediate values: it does it for you, automatically and without error.

HP-35.
The electronic slide rule

Small enough so that you can easily carry it around in your shirt pocket — it weighs only 9 ounces including rechargeable battery — the HP-35 is the original electronic slide rule introduced a little over a year ago. It has since become the constant companion of more than 75,000 scientists and engineers around the world.

Due to the economies realized in this long production run, the price of the popular HP-35 has been reduced to $295.*

HP-35

The HP-35 is easier to use, 10 times faster and significantly more accurate than the slide rule. With a single keystroke and in less than a second, it performs trigonometric (sin, cos, tan), logarithmic (log x, ln x, e^x) and other commonly used functions (x^y, $1/x$, $\sqrt{x}$, π) as well as the four arithmetic operations. It also calculates inverse trig functions.

In addition to its computer-like operational stack, the HP-35 has a constant storage register which lets you store any number and recall it as often as you want for repeat operations, without ever having to re-enter it.

The HP-35 comes with owner's handbook, battery pack, AC adapter/recharger, carrying case and travel case.

HP-45.
The scientific pocket calculator

A direct descendant of the "electronic slide rule," the new HP-45 packs nearly twice the computational power into the same package. The trick is that it has a unique gold-colored "shift" key that doubles the function of 24 of its 35 keys. Hence it does all that the HP-35 does . . . and then some.

The HP-45 is the first pocket calculator with *nine* addressable memory registers besides its operational stack. You can store data in each one — any number that appears on the display — and recall it to the working register whenever you want. (Let your imagination picture the calculating horsepower of this feature for register arithmetic, conversions, continued products, payrolls . . .)

The HP-45 also has a 14th register, called "Last X" in which the last input argument is auto-

HP-45

HP-46

matically stored. You can recall this number by pushing the "Last X" key . . . then proceed to correct it or to perform calculations with it.

There's more. The HP-45 lets you do trig calculations in any of three angular modes (degrees, radians or grads) and converts angles in any mode to degrees/minutes/seconds instantly, and vice-versa. It lets you convert polar coordinates to rectangular and vice-versa, at a single keystroke. Add or subtract vector components in polar or rectangular coordinates. Perform two dimensional accumulations for vector calculations. And convert U.S. units of length, weight or volume to metric, and vice-versa . . . to 10-digit accuracy.

The HP-45 costs $395* including owner's handbook, quick reference guide, battery pack, AC adapter/recharger, carrying case and travel case.

HP-46.
The scientific printing calculator

If you need a permanent record of your calculations and don't insist on the size and portability of our two shirt-pocket wizards, the HP-46 was designed especially for you.

It has all the computational ability of the HP-45 with the important addition of a printer. The printer's extended set of alphanumeric symbols lets it "talk back" to you as you perform your calculations. The HP-46 prints easy-to-read symbols with all operations you perform, and clearly labels the results. Its printout thus constitutes a complete permanent record.

For example, when you calculate the mean and standard deviation of a series of numbers, the printer lists each entry with a $\Sigma+$ symbol. Then, after you push the $\bar{x}$, s key, it prints the results, in order: the number of entries, the calculated standard deviation and the mean . . . each clearly labeled.

On command, the printer will also record the contents of the operational stack or the 9 addressable memory registers, each also clearly labeled. Should you make a logical error in data entry or call for an improper calculation, the printer will make an error note with a reference to an explanation contained in the operating manual.

A 15-digit LED display is available as an option.

The HP-46 costs $695* including owner's handbook, printer paper, and carrying case.

*Domestic USA prices only.

00310

Mail to: Hewlett-Packard,
1505 Page Mill Road, Palo Alto, CA 94304

I want an alternative. Send me complete information on:
( ) HP-35 Electronic Slide Rule
( ) HP-45 Scientific Pocket Calculator
( ) HP-46 Scientific Printing Calculator

Name_____

Title_____

Company_____

Address_____

City_____ State_____ Zip_____

HEWLETT hp PACKARD

Sales, service and support in 172 centers in 65 countries.
Palo Alto, California 94304. Offices in principal cities throughout the U S

THE AMATEUR SCIENTIST

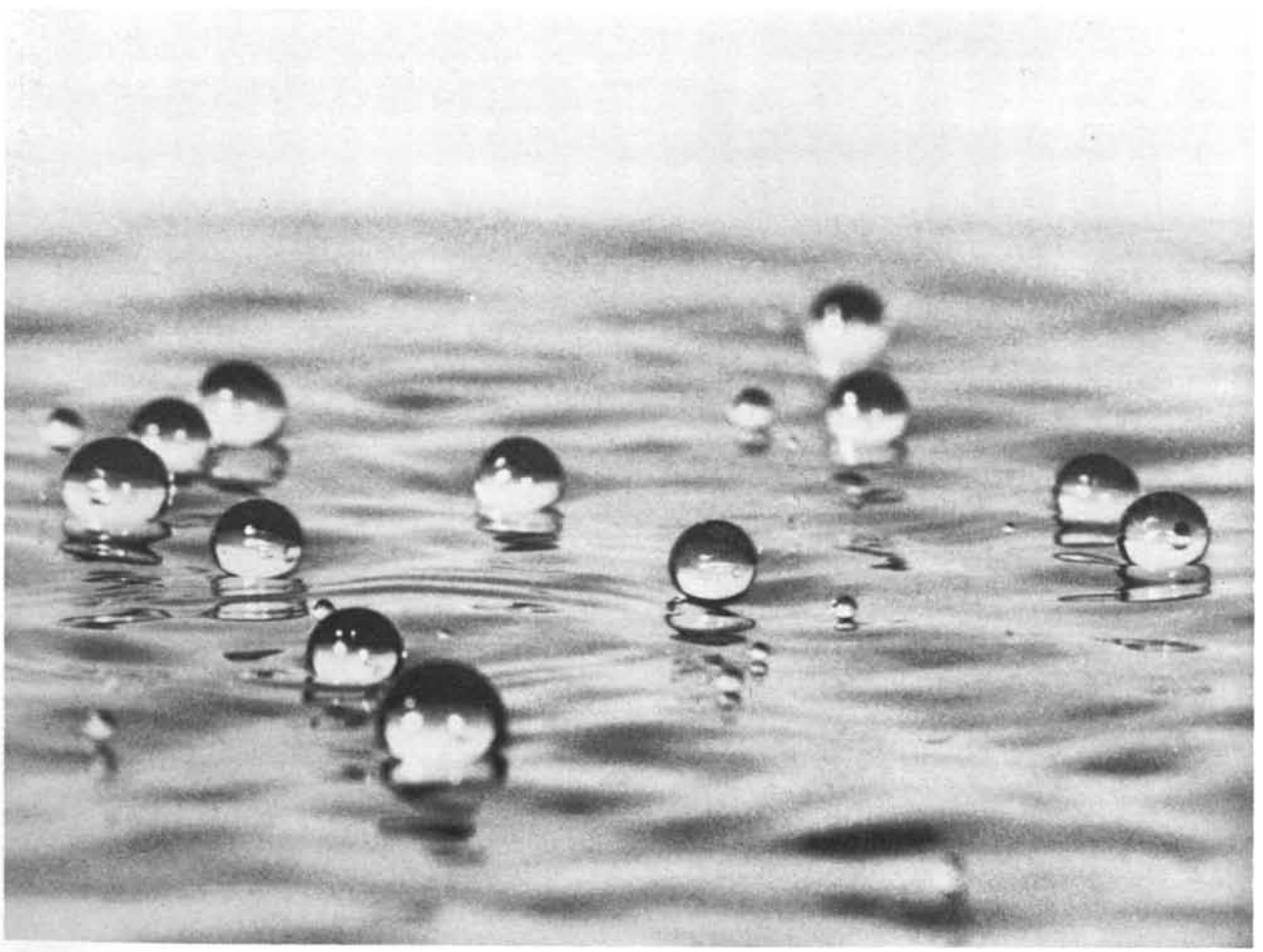
Water droplets that float on water, and Lissajous figures made with a pendulum

Conducted by C. L. Stong

Experiments of seemingly endless variety can be performed with water. Some of them, such as the fountain constructed by Hero of Alexandria in 150 B.C. to demonstrate that water can rise above its own level, are little more than illusions. Others dramatize the remarkable properties of water; an example is Lord Kelvin's electrostatic generator in which drops of falling water develop potentials of thousands of volts. Still others are based on the peculiar behavior of water. An experiment of this kind has been devised by Gerard Schol, a physics teacher of Drachten in the Netherlands. Schol's experiment involves the commonly observed but long neglected phenomenon that drops of water can exist as independent floating objects on the surface of clean water. Schol writes:

"I was rowing a boat about two years ago and happened to hold the oars still for a few seconds. A stream of water that ran from one oar struck the surface at a point from which scores of droplets darted in a radial pattern. They resembled glittering gems and remained on the surface for as long as 15 seconds before vanishing.

"I had seen floating drops on previous occasions on pools and surf and even while pouring a drink. They can easily be mistaken for bubbles. Close examination discloses obvious differences, however, particularly when there is a bit of wind. The drops race across the surface



Globules afloat on the surface of water

in response to even a gentle breeze because they make almost frictionless contact with the supporting water, whereas bubbles are anchored to the water by surface tension.

"Moreover, in contrast to bubbles, the drops are massive and their higher index of refraction causes them to glitter characteristically, particularly when they are viewed in sunlight at about the angle at which rainbows are observed. Finally, the weight of the drops pushes shallow depressions into the supporting surface of the water. A depression can be seen easily by examining a floating drop that stands still or a photograph of a moving drop. All floating drops have the form of oblate spheroids. To distinguish floating drops from ordinary drops I call the floaters water globules.

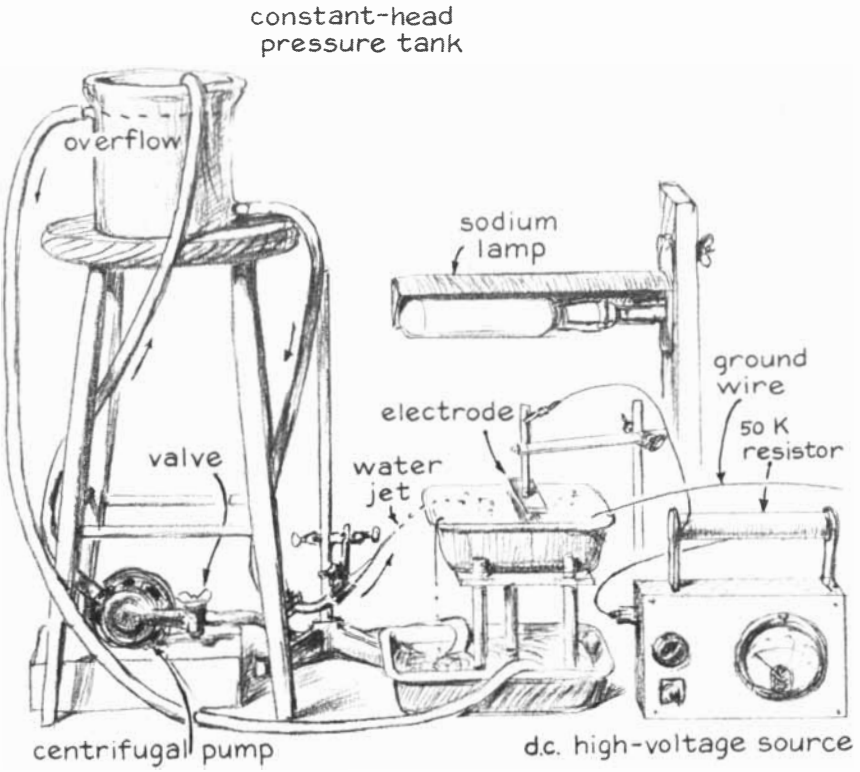
"Water globules can be made readily. Punch a hole about two millimeters in diameter in the bottom of a tin can. Fill the can with water, hold it about a meter above any still body of water and let the contents drain through the hole. A great many drops will spatter radially away from the point where the stream hits the surface. If the surface is clean, water globules will form among the drops that spatter. If the surface is dirty, globules may not form immediately, but a few will appear within seconds and the number will increase as the jet sweeps the surface clean.

"The globules will form in a range of sizes from less than a millimeter in diameter to as much as five millimeters. They are easily seen in light from any direction but particularly when they are lighted from the rear by sunlight. Many globules remain afloat for 15 seconds and then merge with the surface water.

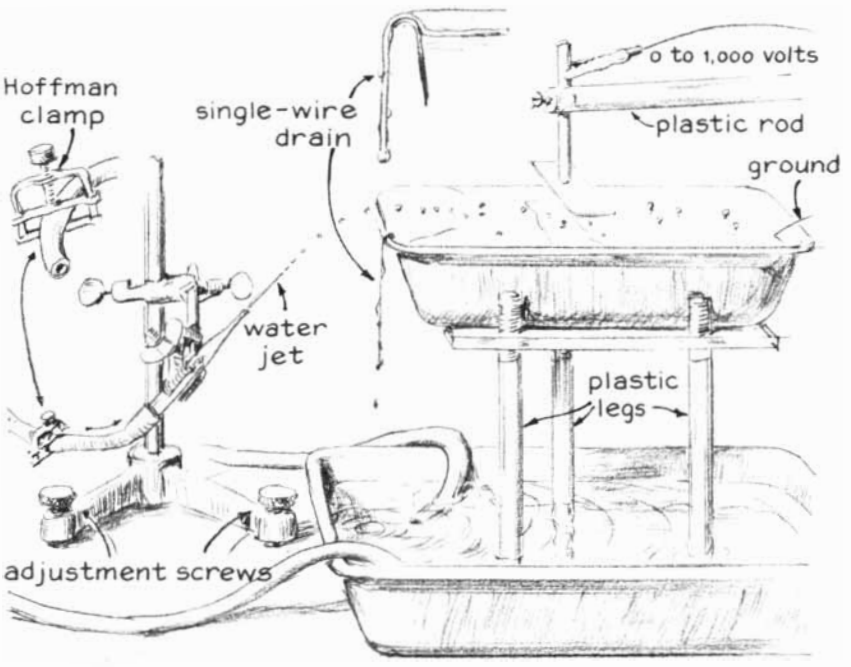
"Some globules disappear sooner, in two or three steps, shrinking to half their diameter every five seconds or so. In my opinion this effect occurs when the drops encounter particles of dirt that protrude from the surface of the water. A particle of dirt pierces the thin layer of molecules constituting the 'skin' of the globule. The interior of the water globule is under considerable pressure, which is maintained by surface tension. Hence when the globule is pierced, water rushes out along the path created by the dirt particle. The escaping jet creates a force of reaction identical with that of a rocket, which lifts the shrinking globule above the surface. The jet necks down after the shrunken globule pulls away from the particle of dirt. The small remaining globule subsequently exists for a time as an independent entity.

"Having done a number of experiments with the tin can, I devised an apparatus for launching a continuous stream of water globules indoors [*see illustration above*]. The launcher consists of a small nozzle that directs a jet of water upward at an angle of approximately 45 degrees. The diameter of the nozzle and the velocity of the jet are adjusted so that the stream breaks into drops just before it reaches the top of its parabolic trajectory.

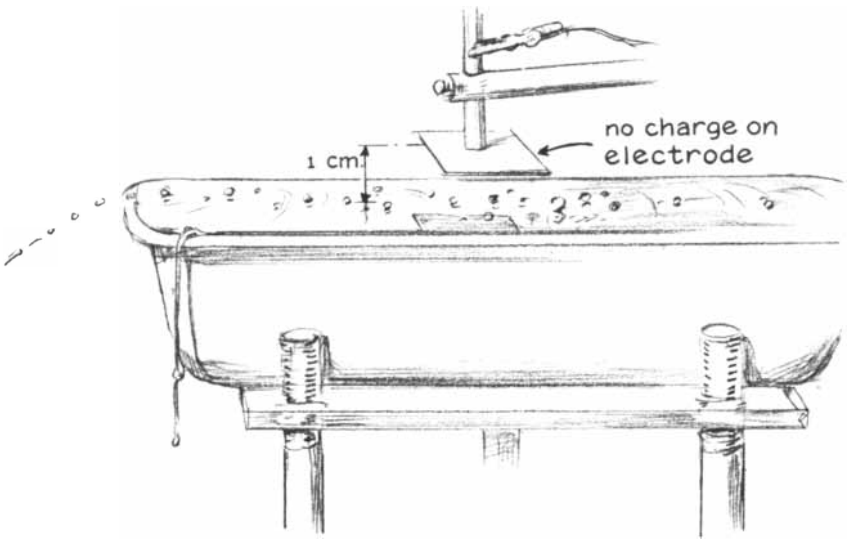
"Immediately after the trajectory bends downward the drops land on the surface of water in a plastic tray of the



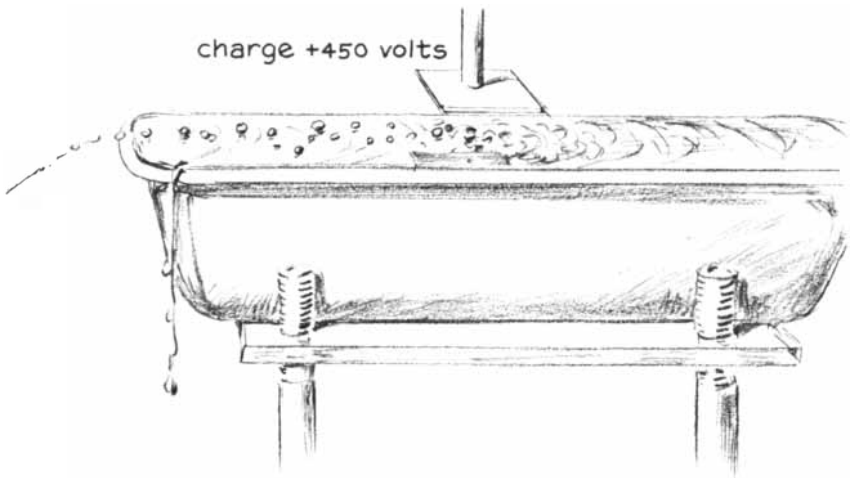
Gerard Schol's apparatus for generating water globules



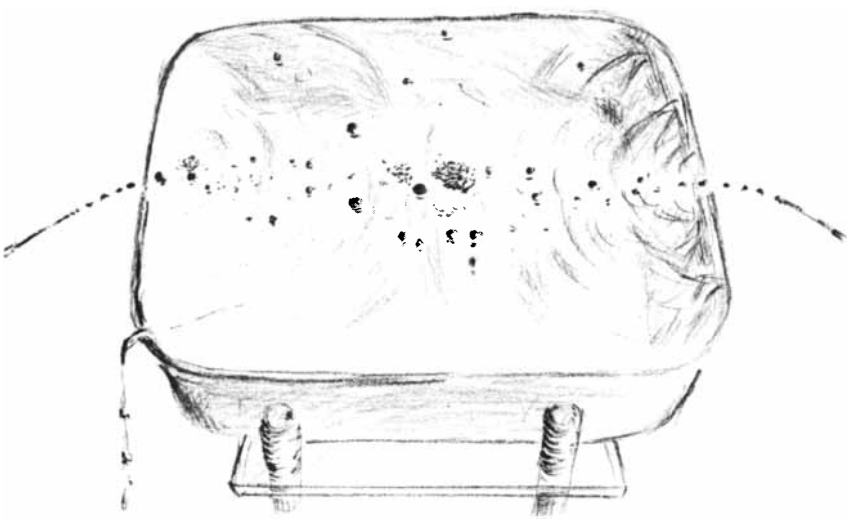
Details of the apparatus



Absence of effect of uncharged electrode on globules



Effect of charged electrode on globules



Globules launched on collision course by opposed nozzles

kind used in photographic darkrooms. The angle at which the drops strike the surface amounts to only a few minutes of arc. The plastic tray is filled to the brim. Indeed, the water rises above the edges as a meniscus that extends around the four sides of the tray. After the apparatus has operated until all dust is carried from the surface the efficiency of globule production approaches 100 percent. Substantially all drops from the jet coast across the surface to the opposite end of the tray. They are easily photographed.

"Overflow from the tray drains into a catch basin from which it is pumped to an elevated reservoir of constant head. The apparatus also contains a source of high voltage. A source of monochromatic light is provided for investigating the nature of the globules.

"The nozzles were made of soft glass tubing seven millimeters in diameter. The middle of a 15-centimeter length of the tubing was rotated in a gas flame until a zone approximately 15 millimeters long became plastic. The tube was then removed from the flame and the ends were promptly drawn apart to form a constriction about three millimeters in outside diameter. One side of the constriction was scratched lightly with the corner of a sharp file at the narrowest point, and then the ends of the tubing were pulled (without bending) until the glass broke squarely at the scratch.

"The angle of the nozzle must be adjustable with respect to the horizontal and vertical planes. I clamp the nozzle to the vertical rod of an apparatus stand that has a base with three feet in the form of two knurled screws at one end and, at the middle of the opposite end, a metal projection of fixed length. The clamp that attaches the nozzle to the vertical rod can be turned in any direction, moved to any height and locked to the rod with a thumbscrew.

"The tray on which the drops are launched rests on an elevated platform of plastic eight millimeters thick, 15 centimeters wide and 20 centimeters long [*see bottom illustration on preceding page*]. The platform has three legs that can be adjusted to alter the height of the tray and to level the tray. The adjustable tray support is made by drilling and threading holes at the two corners of one side of the platform and making a similar hole at the center of the opposite side. Three supporting legs of the plastic rod were threaded to mate with the holes in the platform. (Prethreaded iron rod, which is available from hardware stores in the U.S. as stud-bolt stock, would also be suitable for the legs.)

"The platform stands in a catch basin that is 40 centimeters long, 30 centimeters wide and six centimeters deep. This container can be a baking pan made of sheet metal or plastic. It receives the overflow from the tray. Water is raised from the catch basin to the elevated reservoir by a small centrifugal pump of the kind normally used for circulating water in aquariums.

"Surplus water from the reservoir returns to the catch basin through a hose that connects to an overflow port near the top of the reservoir. The overflow scheme maintains a constant head of pressure in the reservoir. The reservoir is a tin can with a capacity of approximately three liters. The velocity of water in the hose that interconnects the reservoir and the nozzle is controlled by a pinch clamp of the Hoffman type that has an adjustment screw.

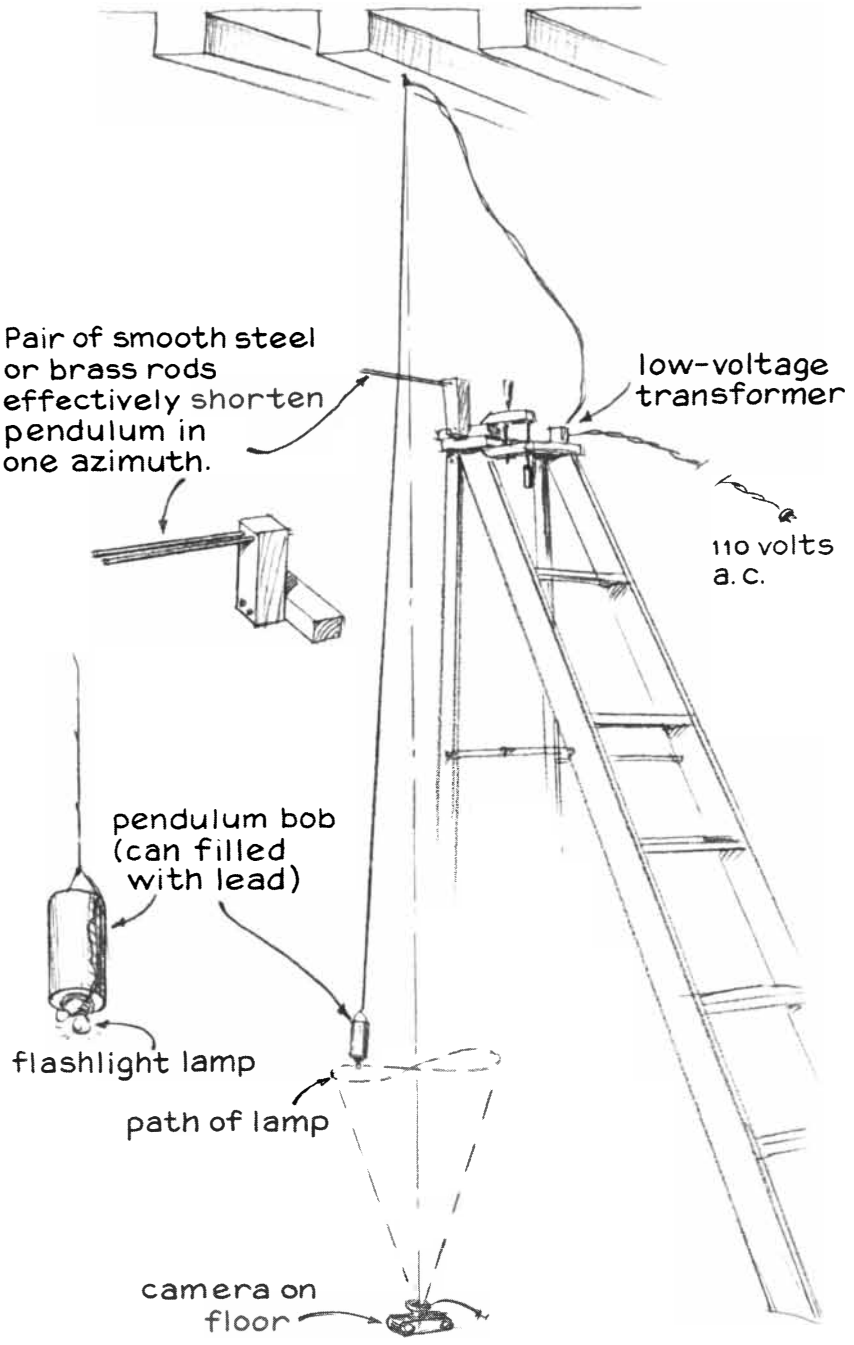
"The production of water globules is critically dependent on three adjustments. The tray must be level and filled to the point at which the surface of the water becomes convex, that is, until a meniscus forms completely around the brim of the tray. The nozzle must project upward a jet that breaks into drops before the parabolic trajectory reaches its maximum height. The path of the impinging drops must meet the surface of the water in the tray at an angle of not more than 15 minutes of arc. Before carrying out an experiment I rinse the apparatus with several changes of tap water. After being refilled the system will run unattended for a number of hours.

"My experiments were designed to probe two questions. What mechanism accounts for the existence of water globules? What are the properties of the globules and how can they be influenced?

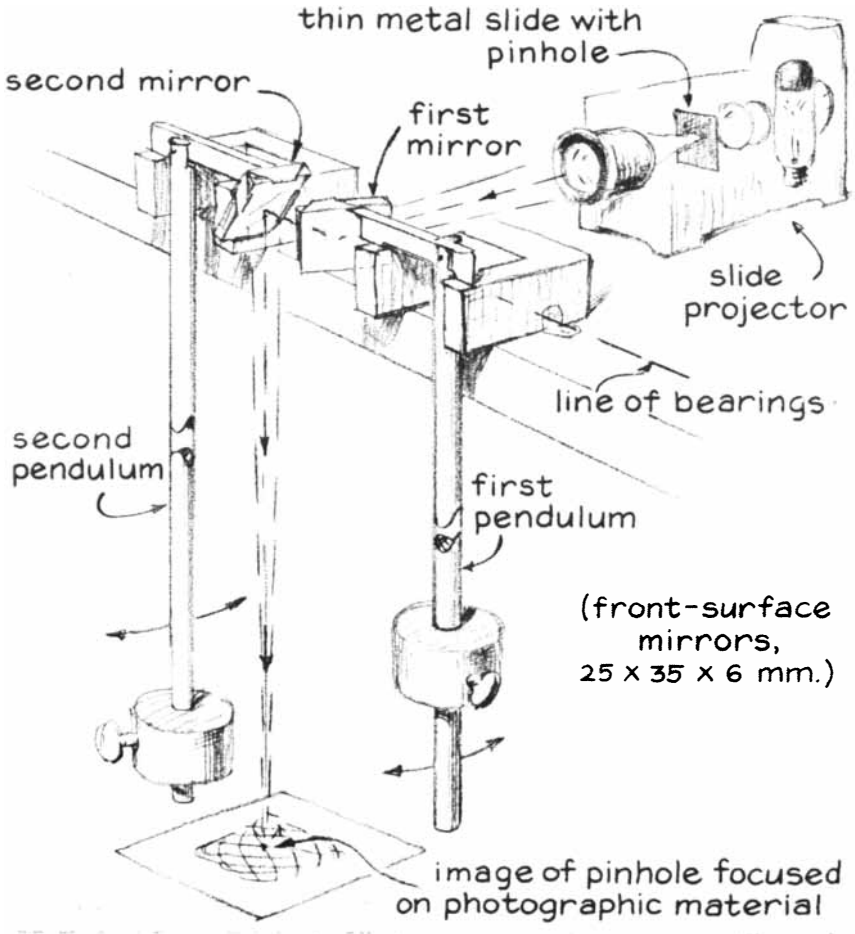
"Quite early in my investigation I discovered more or less by accident that the globules are sensitive to an electric field. While the apparatus was running I happened to pick up a piece of plastic tubing (polyvinyl chloride) about 40 centimeters long and to rub it with a piece of wool. The friction dislodged electrons from the wool and deposited them on the plastic. The rod emitted a faint crackling sound, which indicated that negative discharges in excess of 10,000 volts were occurring. I was standing next to the tray across which the globules were coasting. The globules vanished whenever I brought the rod near the tray. What would be the effect of a comparable positive charge? A glass rod rubbed with silk acquires positive charge. I tried it. The globules vanished as they did under a negative charge.

"These experiments convinced me that water globules owe their existence at least in part to electrical forces. It is well known that although the water molecule is electrically neutral, it forms an electric dipole. The two hydrogen atoms are bonded to the single oxygen atom at an angle of approximately 105 degrees. The asymmetry causes a negative charge to appear on the oxygen side of the molecule and a positive charge to appear on the hydrogen side. The molecule is free to rotate into alignment with an external electric field.

"Moreover, the interaction between neighboring molecules, which are kept in a state of perpetual motion by thermal energy, is such that molecules at the surface of the liquid spend most of their time with the oxygen atoms turned away from the body of the liquid. Hence the surface of water, either in a container or in the form of a drop, is negatively polarized. It is my opinion that when a drop



Philippe Lebrun's apparatus for generating Lissajous figures



Roger Hayward's apparatus for Lissajous figures

from a nozzle approaches the surface of water in the tray at a small angle, and hence at sufficiently low vertical velocity, a force of repulsion exists between the two negatively polarized surfaces and is strong enough to support the weight of the globule. The distance between the polarized surfaces is very small and so maximizes the repulsive force. (An experiment I shall describe below discloses that the separation between the globule and the surface is less than the wavelength of yellow light.) An external electric field of sufficient strength should alter the orientation of the water molecules, causing the molecules to rotate so that the hydrogen atoms point toward the surface, which should then become positively polarized. Such reversals of polarization of the molecules should cause the globules to merge with water in the tray.

"To test this assumption I applied an electric field of adjustable and known potential between an electrode and the surface of the water in the tray. The electrode was an aluminum strip 10 centimeters long, 1.5 centimeters wide and approximately one millimeter thick. The strip was supported at a right angle to the stream of globules one centimeter above the center of the tray by an insulator, which was a rod made of polyvinyl chloride. The electrode was connected through a resistor of 50,000 ohms to a direct-current power supply that was adjustable from 0 to 1,000 volts. The resistor's role was to limit the current to a safe value if the electrode accidentally made contact with a grounded conductor. A voltmeter measured the potential on the electrode. The circuit was completed by hanging a wire connected to the grounded side of the power supply in the water at one corner of the tray.

"All globules directly below the electrode disappear at a characteristic voltage, which varies with the composition of the water. In distilled water the globules merge at a potential of about 450 volts, which is equivalent to an electric-field strength on the order of 45 volts per millimeter. An exact value cannot be given because the behavior of the globules is influenced by several variables, including their mass.

"In distilled water a gradient of 45 volts per millimeter causes 99 percent of the globules below the electrode to disappear. The remaining globules are the largest ones. All globules merge with the tray water when the potential is increased to 550 volts, which is equivalent to a field strength of 55 volts per millimeter.

"The addition of table salt to the water greatly increases the potential required to stop the globules. The same effect is observed when liquid detergent is added to the water. The addition of liquid detergent also causes larger globules to form, perhaps because the extraordinarily long molecules of the detergent form a highly polarized surface layer on the water. It seems reasonable to suppose that the resulting electrical force at the surface is stronger than that of distilled water. If this is the case, a proportionately stronger electric field would be required to make the globules disappear.

"Interesting phenomena can be observed by putting nozzles at each end of the tray to launch drops on a collision course, particularly if detergent is added to the water. Globules that collide elastically bounce apart, often vibrating so violently that they break into clusters of small globules suggestive of what happens when atomic particles collide. Other colliding globules merge to form large globules. Multiple encounters frequently result in globules that grow in diameter to more than three centimeters. The weight of the globules depresses the surface of the supporting water in the immediate vicinity. Relatively large flattened globules almost sink below the surface.

"Part of the light that enters the space between the bottom of a globule and the surface of the supporting water in the tray follows the curvature of the drop by multiple reflection and emerges on the other side. I suspect that this effect may account in part for the glittering ring that appears to surround globules that are lighted from the rear. To estimate the thickness of the film of air that separates the globules from the tray water I examined the globules from the top with monochromatic light from a sodium lamp. Newton's rings appear in each globule. Such rings form a pattern of concentric dark and light bands or fringes that arises from the optical interference of light waves reflected through the globule from the supporting surface and from the lower surface of the globule. The central fringe of the pattern appears as a black disk, which indicates

that the space between the two surfaces is less than a wavelength of yellow light and not more than a fraction of a micron.

"I have found no published explanations of water globules, although the fact that drops of water can float temporarily on larger bodies of water is mentioned in a few references. I hope that my investigations will encourage others to tinker with the globules. Experiments performed with alcohol and nonpolarized liquids of low viscosity should be interesting. I also hope that those who undertake these or any related projects will let me have a report of their results. My address is: Gerard Schol, Sint Jansberg 63, Drachten, The Netherlands."

A pair of sine waves that vary with respect to axes at right angles can combine to generate a family of curves known as Lissajous figures. The figures can take the form of a straight line, a pattern of complex elliptical forms or a circle, depending on the relative amplitude, period and phase of the sine waves. The patterns are named for the 19th-century French physicist Jules Antoine Lissajous, who investigated them both experimentally and mathematically.

To generate the figures Lissajous cemented a small mirror to one tine each of two tuning forks mounted at a right angle. A beam of light that was reflected sequentially by the mirrors came to a focus as a small spot on a distant screen. The mirror attached to one vibrating fork caused the spot to oscillate in the vertical plane. The mirror of the second fork simultaneously contributed to the beam a horizontal component of vibration. Persistence of vision caused observers to perceive the path of the rapidly moving spot as a geometric figure.

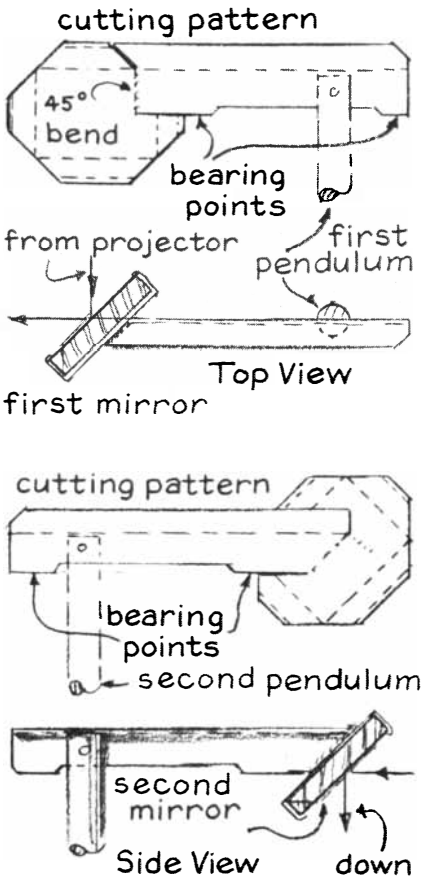
Lissajous could alter the frequency, amplitude and phase of the tuning forks by adding weights to the tines and altering the force with which he set the forks into vibration. He could thus generate all possible forms of the figures. As he reported, waves of the same frequency, which are either in phase or exactly 180 degrees out of phase, generate straight lines. Sine waves of equal amplitude that are 90 degrees out of phase generate a circle. Waves of unequal frequency generate complex patterns, but distinct and easily recognized patterns appear when the frequencies differ in the ratio of whole numbers. For example, the frequency ratio of 1 : 2 generates a pattern in the form of the numeral 8, which has two loops. A ratio of 1 : 3 generates a comparable figure of three loops, and so on. The figures are therefore useful for measuring the phase and frequency relations of sine waves. Such measurements of alternating electric currents are made routinely by observing Lissajous figures with cathode ray oscilloscopes.

A much simpler and more inexpensive apparatus for investigating the figures is submitted by Philippe Lebrun of the School of Mines in Paris (École des Mines de Paris, 60 Boulevard St. Michel, 75 Paris 6e, France). "My apparatus," he writes, "consists of a pendulum made by suspending a tin can filled with lead balls by a slender nylon string about two meters long [*see illustration on page 107*]. To the bottom of the can I attached a miniature incandescent lamp. The lamp is energized through a pair of thin flexible wires twisted around the string. These leads connect to the output of a small transformer that provides appropriate voltage for the lamp.

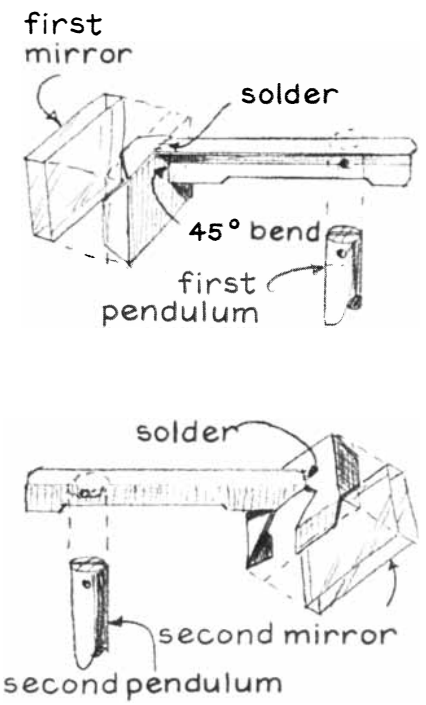
"The path of the swinging pendulum can be recorded photographically in a dark room by putting an appropriately focused camera on its back 50 centimeters directly under the pendulum bob. A pattern of ellipses appears on the film when the pendulum is free to swing in all directions, as in the case of a Foucault pendulum. Normally I work with color film. Various figures generated by altering the amplitude of the pendulum can be recorded in the form of multiple time exposures by using lamp bulbs of distinctive color for each exposure of the series.

"Interesting Lissajous figures can be generated by increasing the frequency at which the pendulum beats in one azimuth. This is accomplished by passing the string between a closely spaced pair of smooth horizontal rods rigidly mounted about 50 centimeters below the point where the string is suspended. The ratio between the two periods at which the pendulum beats can be altered by raising or lowering the horizontal rods."

Lebrun's apparatus works nicely and is easy to set up. On the other hand, it is relatively bulky. The pendulum must be suspended in a room with a ceiling height of at least 2½ meters to provide adequate range for the camera. Roger Hayward, who illustrates these columns, submits a design for an apparatus that is more compact, although it is more difficult to construct. It employs two pendulums [*see illustration on opposite page*]. They cause two mirrors to vibrate at right angles, as in the case of Lissajous's apparatus. The details of the construction are depicted by the accompanying illustrations.

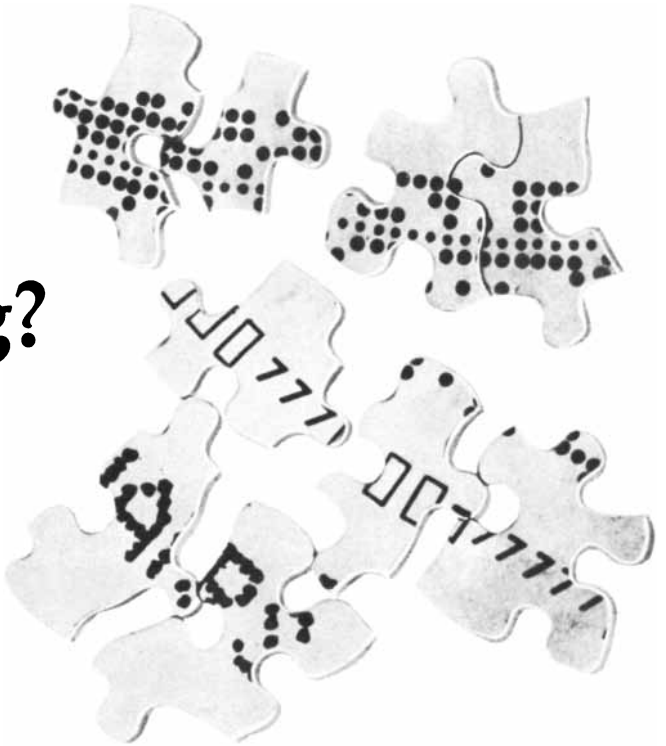


Details of Hayward's apparatus



Mounting of mirrors

How much new computer information are you missing?



Are you up on all the recent breakthroughs in computer science? The right equipment, the correct decision, new programming techniques, the well-thought-out recommendation to management—everything that can advance your career and your professional standing. It all falls into place with the help of the Library of Computer and Information Sciences, the one book club offering you only the most advanced, most significant, most informative books in your field, all at substantial discounts. Our liberal introductory offer makes it hard not to get acquainted.

Take any 3 books
(values to $54.95)

all for only $3.95

with trial membership in the

Library of Computer and Information Sciences

(You simply agree to accept three more books all at substantial member discounts over the next 12 months.)

Library of Computer and Information Sciences 7-37T
Riverside, New Jersey 08075

Please accept my application for membership and send me the three volumes indicated, billing me only $3.95 for all three. I agree to purchase at least three additional Selections or Alternates during the first year I am a member, at special members' prices. Savings range up to 30% and occasionally even more. My membership is cancelable any time after I buy these three books. A shipping and handling charge is added to all shipments.

Indicate by number the 3 books you want.

Some expensive books count as 2 or 3 selections.

Name _____

Address _____

City _____

State _____ Zip _____

Send no money. Members are billed when books arrive.

(retail prices shown)

62741. MINICOMPUTERS IN DATA PROCESSING AND SIMULATION. *Branko Soucek. Counts as 2 of your 3 choices.* **$19.95**

39960. COMPUTER HARDWARE THEORY. *W. J. Poppelbaum.* **$16.95**

66020. OPTICS. *Miles V. Klein.* Traditional topics in geometrical and physical optics *plus* a complete introduction to lasers, holography, and coherence. The best book in its field. **$16.95**

44030. ELECTRONIC COMPUTER MEMORY TECHNOLOGY. *Edited by Wallace B. Riley.* Complete analysis of the "state of the art." **$14.50**

81690. SWITCHING AND FINITE AUTOMATA THEORY. *Zvi Kohavi.* For the first time, the *entire* subject of switching and finite automata is brilliantly covered in a single, volume. **$16.50**

56610. INTRODUCTION TO SCIENTIFIC COMPUTING. *Gabriel A. Pall.* IBM expert Pall provides the professional know-how for using computers to maximum advantage in scientific work. *Counts as 2 of your 3 choices.* **$19.95**

35050. AUTOMATIC DATA PROCESSING: System/360 Edition. *Frederick P. Brooks, Jr. and Kenneth E. Iverson.* "At last, a book written for computing people, by computing people." **$16.50**

70790. PROGRAMMING SYSTEMS AND LANGUAGES. *Edited by Saul Rosen.* A comprehensive, detailed review of computer software that brings together the most authoritative articles from computer journals and proceedings. **$14.95**

77670. SEMANTIC INFORMATION PROCESSING. *Edited by Marvin Minsky.* Exciting presentation of experiments aimed at making intelligent machines. **$15.00**

44110. ELECTRONIC INTEGRATED SYSTEMS DESIGN. *Hans R. Camenzind.* Covers fundamentals of the integrated circuit (IC) plus information on resistors, transistors, diodes, how to design circuits to use them. **$14.95**

65910. OPERATING SYSTEMS. *Harry Katzan, Jr.* The first all-inclusive coverage of operating systems technology, from structure and utilization to functions and operations. **$14.95**

63340. MUSIC BY COMPUTERS. *Edited by Heinz von Foerster and James W. Beauchamp.* An intriguing book—accompanied by eight sides of 33 rpm high-fidelity records—that demonstrates a novel and inventive use of the computer. **$16.95**

84080. THICK FILM HYBRID MICROCIRCUIT TECHNOLOGY. *Donald W. Hamer and James V. Biggers.* Actually gives the reader a "feel" for thick film materials and processes. *Counts as 2 of your 3 choices.* **$19.95**

70770. PROGRAMMER'S GUIDE TO THE IBM SYSTEM/360. *John H. Bradley.* "Highly recommended."—*Computing Reviews.* **$13.50**

39791. COMPILER CONSTRUCTION FOR DIGITAL COMPUTERS. *David Gries.* Treats not only theory but also the practical aspects of compiler writing. **$14.95**

52210. HANDBOOK OF DATA PROCESSING FOR LIBRARIES. *Robert M. Hayes and Joseph Becker.* Almost 900 pages packed with detailed "how to" information. *Counts as 2 of your 3 choices.* **$22.50**

32410. AMERICAN INSTITUTE OF PHYSICS HANDBOOK (3rd Edition). *Coordinating Editor, Dwight E. Gray.* The standard in its field. *Counts as 3 of your 3 choices.* **$49.50**

38160. CIRCUIT DESIGN OF DIGITAL COMPUTERS. *Joseph K. Hawkins.* Principles, applications and limitations of circuit design. **$21.50**

81920. SYSTEMS ANALYSIS FOR DATA TRANSMISSION. *James Martin.* The dean of writers on computer telecommunications and teleprocessing, shows how organization networks can slice data transmission costs in half. *Counts as 2 of your 3 choices.* **$22.50**

HOW THE CLUB OPERATES

• The Book Club **News,** describing the coming Main and Alternate Selections, will be sent to you 15 times a year at three to four week intervals. • If you wish to purchase the Main Selection, do nothing and it will be shipped to you automatically. • If you prefer one of the Alternates, or no book at all, simply indicate your decision on the reply form always enclosed with the **News** and mail it so we receive it by the date specified. • The **News** is mailed in time to allow you at least 10 days to decide if you want the coming Main Selection. If, because of late mail delivery of the **News,** you should ever receive a Main Selection without having had the 10-day consideration period, that Selection may be returned at Club expense. • After completing your trial membership, you will be entitled to take advantage of our bonus plan offering savings up to 70% and more. (Offer good in Continental U.S. and Canada. Prices slightly higher in Canada.)

BOOKS

A guide to modern optical technology and a survey of the near-science of hypnosis

by Philip Morrison

OPTICAL PRODUCTION TECHNOLOGY, by D. F. Horne. Crane, Russak & Company, Inc. ($42.50). Around this magazine for decades there rallied the legion of the amateur mirrormakers, the patient grinders of large precision surfaces in glass that often—not always—provided admirable homemade telescopes. Here is a massive (and costly) volume that surveys the optical working art of today "for a new generation of opticians." Its author, a man of wide experience in the industry and clear pedagogical skill, comes out of the Hilger firm of precision optics; his book is a conscious successor to Frank Twyman's smaller volume on making prisms and lenses published by that London company on the eve of World War II. He writes for the student, the technician and the manager, not for the design engineer. His is the world of the man who must choose and set up the machines, make tools for them and operate them; who watches over the slurry and tests the surfaces; not of the man who nowadays programs the path of rays through thick lenses or the small coterie who design the intricate electronic-mechanical machines the industry now depends on.

The new world of optics is plain in these pages. The glossary is genuinely international: in the fourth column we see the Japanese expression for, say, "Swedish pitch" or "orangewood sticks." Optical surfaces are far from spherical these days; even our spectacle lenses are toroidal, and Schmidt correctors and other aspheric surfaces are sold by the tens of thousands in quality zoom lenses for film and television. Aspheric surfaces made by hand can cost 10 times as much as spherical ones, so adapted is the machine to rotation. Since 1960 machines for generating aspheric lens curves have been available that produce blanks accurate enough to go direct to polishing. Two or three are described here in some detail, and they are subtle enough. The British version is out of the older precision machine-tool tradition: elegant spindles, magnetically driven to avoid distortion, make circular or straight-line cuts with a diamond point, while a punched-tape control superimposes the asphericity. The workshop temperature is regulated by thermostat, of course, and the tolerances achieved are closer than one micron.

The American design is less direct. Its basis is a system of two commercial precision lead screws, one for each of the two directions in the plane. They move carriages sliding on special very stiff ways. The cutting tool is a diamond bur spun by an air turbine at 33,000 revolutions per minute. A four-inch lens can be ready for polishing in half an hour or less. The heart of the accuracy is a servo-corrected "electrostatic nut," an insulating epoxy nut with brass spirals adhering to the thread helixes, arranged in a servo circuit that detects and corrects the nut position by sensing capacity differences. This scheme achieves microinch accuracy in the more critical direction. Polishing can be done on a device in which air pressure holds a nylon diaphragm covered with a thin layer of pitch against a block of aspheric lenses, all mounted in an elegant rocking fixture.

All of this has an air very different from the hand operations of the old optical industry. Development has followed the enlargement of the market, now that high-quality cameras are everywhere, spectacles abound and "the microcircuit industry has required a large number of microscopes." (Nothing is said about reconnaissance satellites, which in one decade must have burned up on atmospheric reentry more yard-wide mirrors than all the astronomers since Galileo have seen.) The big market has given the optical shop materials and machines never before imagined. Diamonds are now the optician's best friend. Diamonds finer than sand, carefully selected for shape, are sintered with bronze or ferrous powders into grinding wheels. Or the diamond powder is held by electroplated nickel bond to a steel tool, to make a drill or a hollow circle-cutter. Synthetic diamond is commonplace. Carborundum and alumina remain the chief abrasives for rough and smooth work; the polisher's rouge is now often replaced by cerium oxide.

Glass comes in a wide range of compositions: thorium, zirconium, lanthanum and tungsten are not foreign to specialty glasses. Optical glass is no longer made potful by clay potful; it is continuously melted in platinum tanks, and "the opinion . . . that bubbles in an optical glass are a sign of quality can be considered, with few exceptions, as being out of date." Glass ceramics, special glasses that have crystallized very uniformly around controlled nuclei, provide mirrors with temperature-expansion coefficients 100 times smaller than the coefficient of fused silica.

Lenses, mirrors and fiber optics of plastics, usually polymethacrylate and polystyrene, are often made by glass-working techniques but are much faster to work; the remarkable flat grooved plastic Fresnel lenses are precision-molded, and even aspheric lenses can be cast in optically worked glass molds ready to use. The text does not ignore the exotic crystals in the world of electromodulated devices: yttrium aluminum garnet, gallium arsenide, potassium tantalate niobate and the like. Growing, cutting, lapping and grinding these substances now require that the optician provide controlled atmospheres and fans for taking away the noxious vapors.

Testing is always the domain of the highest skills. The schemes for making interference-fringe "contour maps" of the surface are still useful; the Twyman-Green interferometer, a development of World War I, is now packaged in a modern folded form, compact on a heavy granite base. The diamond stylus, electronically amplified, can feel and record surface heights along a line at a magnification of a million. The autocollimator, a projecting telescope with an eyepiece scale that measures the deflection of its own light beam in a mirror attached to the work, remains a key device; now it may record photoelectrically on a roll of paper. With such an autocollimator at-

tached to an American special-precision indexing machine, the assembly can measure angles to a tenth of an arc-second over the entire circle—one part in 10 or 12 million. This device, figured and justly praised here (but without much description of its construction), is the culmination of the art of circle division, which goes back to Hero by way of Jesse Ramsden and the 16th-century artisans of Nürnberg. That art once made possible a Kepler. Tycho Brahe would have rejoiced in circles calibrated by a Moore Master 1440 Precision Index and its Hilger & Watts TA 5 autocollimator. (Horne is preparing a "companion volume on *Dividing and Ruling*"—not, in this case, a political-science text.) For all of this, for making a small quantity of optical components—perhaps only one—"there is no alternative but to use skilled operators and, substantially, hand methods." The chapter covering this art is cited nearly unchanged from the older Twyman volume. Indeed, the book contains dozens of such long citations of experts from Jena or Rochester or Ealing, using them for its most highly specialized and up-to-date accounts.

Theory occupies one chapter. Even with modern solid-state physics and the techniques of electron diffraction and imaging we do not know the nature of polishing. Surely some material must be removed, since the tracks of abrasive particles can be seen. The tracks are often hidden under surface material, however, so that we must admit to some plastic flow as well, perhaps even accompanied by local melting. Finally, surface hydrolysis may be involved, and water removal may be part of polishing. A brief account of the modern mathematical theory of imaging by Fourier-transform techniques is given elsewhere, in the context of automatic systems for measuring performance. The book is a very Baedeker into the wonders of the optical shop: informed, detailed—and a little overwhelming.

HYPNOSIS: RESEARCH DEVELOPMENTS AND PERSPECTIVES, edited by Erika Fromm and Ronald E. Shor. Aldine-Atherton ($27.50). The editors of this set of 20 original review papers by American psychiatrists and psychologists remark justly that "the study of hypnosis has not yet reached an advanced stage of scientific development." The contributors and their contributions are as diverse in intention as they are in attitude; they present summaries of theoretical issues, history, surveys of broad areas, lines of individual research and specific topics of research, at differing length and depth. The tone varies too. Some papers are filled with electroencephalograms and correlation indices; others draw heavily on the subtleties of ego and id and regressive adaptation. The learning of the historical chapters, the comprehensive bibliography and the catholicity of the whole make the volume useful for the general reader who wants a glimpse of this strange domain of a "profoundly compelling imaginal fantasy," in which the Scylla and Charybdis of the researcher are "the disciplined skepticism of the scientist and the confident persuasiveness of the hypnotist."

The phenomena are, of course, authentically observed. There is everyday use of hypnosis for analgesia in medical and dental practice and there is a school of psychotherapy that depends for access on hypnosis. Few hold today that the effects arise from any magnets or miracles; they are seated within the mind, a miracle enough. One key issue, which is engaged by the longest (and for a nonpsychiatrist reader the most interesting) papers, is how far subjects in the hypnotic state can transcend "normal volitional capacities" that "highly motivated nonhypnotized subjects in the usual waking state are unwilling or unable to mimic."

Stage hypnosis is not unexpectedly a poor basis for judgment. Dr. Theodore Barber, a Massachusetts psychologist, received training in that art. He explains how the air of fun and the expectancy of the audience help in the selection and the performance of people who are suggestible in a fully normal state. Stage whispers give the subjects instructions: "Do exactly as I secretly tell you. O.K.? Swell!" The human-plank behavior—bridging two chairs with the rigid body while a man sits on the chest—is easy for controls. Six unselected male subjects told to do that in the laboratory all succeeded immediately; they were all surprised, and they all "disagreed vehemently... that they were in a trance." Asked to take part in a test of imagination, the great bulk of subjects could forget, lock hands, hallucinate thirst and do similar things as well as hypnotized people do. Most remarkable are the simulation experiments. Here subjects are simply asked to "work with Dr. *X* and to convince him that you are an excellent hypnotic subject and become deeply hypnotized." Dr. *X*, an experienced clinician, is asked to distinguish the simulators from the fully entranced. He finds it "extremely difficult without special procedures." A piece of drama provided one such procedure. Experimenter and subject are alone in a room, observable through a one-way screen. This time, the subject is told, all instructions, even the starting and ending of hypnosis, are to be administered by tape recording. Suddenly the music stops, the lights go out, the experimenter, silent until now, mutters "The damn fuse" and dashes from the room. Five out of six hypnotized subjects did not rouse when the hypnotist left, although sooner or later they did pull themselves out of the state, somewhat confused. But five out of six simulators too continued to appear hypnotized—over the full half-hour until the experimenter returned. They had caught on, they explained later, to the fact that the power failure was contrived! The conscientious simulators were more faithful than the entranced. The experiment became meaningful, however, once the staging improved, with the hall lights going out along with the room lights, the one-way screen replaced by a hidden means of observation and so on. Now the simulators themselves were fooled. Four out of six of them stopped simulating the moment the experimenter left the room. The hypnotized subjects, on the other hand, behaved just as before.

The skeptic cannot relax with such an easy success. The subjective impact of hypnosis is by no means slight, and some Stanford psychology-student volunteers performed quite unusually under hypnosis. They could—and the matching controls could not, even though they tried hard—change the carefully measured skin temperatures of their two hands, achieving differences of up to four degrees Celsius within two minutes. Now that Neal Miller has found out how to teach such controls by operant conditioning, hypnosis finds a context, but it is clear that an unusual, if not a unique, impact of cognitive processes on body function may be involved. There are other papers of a "softer" kind, relating hypnosis to creativity, to dreams, to memory. The volume is a good lookout from which to view the gains and the tensions within psychology today.

One of the 1,200 papers cited has a moving title: "Does the Heart Learn?" The investigation, however, was literal; the answer was an ambiguous no. The absence of any mention of the analogous situation in animals, the response of immobility, is one shortcoming of the work.

LIFE: THE UNFINISHED EXPERIMENT, by S. E. Luria. Charles Scribner's Sons ($7.95). What quantum mechanics was to science between the wars, molecular biology has been for the past 30 years. Its brilliant success has underwritten re-

flective accounts by a number of its pioneers, men who indeed had something to be arrogant about and who mostly gave in to that temptation without much reluctance. They have provided us with good books—this one hedonistic, that one philosophical and neo-Cartesian, another discursive, even crotchety. This deliberately direct and compact book is another biologist's view of life, but with a difference.

Professor Luria is an authentic pioneer of molecular biology. Even before the first wagon train set out he ventured as a mountain man among bacterial viruses, a man cunning in choice of experiment, almost prophetic. Genetics became molecular with the materials he studied, and he remains a leader in the field. Yet his book is neither self-centered nor formidable; rather it is humane, sensitive and genuinely clear. It is literary in the best sense; the flowing text is personal, informed, honestly meant for the general reader. Not one formula—hardly even a number—occurs in its pages; three structural diagrams of the nucleic acids at work and a table of the amino acid code alone interrupt the lucid prose. Readers who have not come to understand the subtle process of information flow in the living cell will find it explained here, stripped of all nonessentials and surrounded with a helpful context of ideas, both from the standpoint of their history and from that of their implications.

Evolution is the theme, of course. Natural selection "acts blindly but effectively" to adapt each organism with uncanny precision to perform well in its environment. The theory can explain, but "like all historical theories" it cannot predict, since it cannot foresee all future conditions or the occasional role of pure chance. Like human history, life is an incomplete record, "a small sample" of the possibilities of the past. It is the mechanism of heredity, the sorting out of potential genes by the single test of survival to multiplication, that produces the web of actual living things. What genes are, how the genes multiply and how they directly inform protein sequences, the working catalytic determiners of the course of biochemical reactions—all of this is explained with grace and a minimum of biochemical burden. "The harmony of the genes . . . is not immutable. It is rather a flowing chorale, superbly adapted to the present, yet evolving to remain in tune with the uncertain future."

The text also deals with the chief issues of life beyond the gene. It appears that cells may themselves be living communities: the special organelles that produce the energy currency of the cell may once have been free-living bacteria. The orchestration of the genes that alone makes possible differentiated animals or plants is not yet understood; there may be signal systems coded into our more abundant DNA—or the DNA of a flea—that are not known to single-celled life. Some membranes may determine specific hereditary form without genetic coding. Contacts between cell surfaces, free-flowing chemical messengers and the electrochemical signals of the nervous system remain the theater of research. "Memory molecules," however, do not appeal to this molecular biologist: "It is absurd to believe that each achievement" could be written in nucleic acids.

The origins of life are given somewhat cursory attention. Then we come to man, and pure biology trails off into culture. Our primate genes fixed the basis of language and hence of thought, but that new force itself remolded selection. It is interesting that the electric fishes, which use electric fields for perceiving what is near them in the water, show a "spectacular enlargement" of the brain portions invoked in electrical signals. Mankind's left brain, the speech region, must also have evolved; here lies the structural basis for the deep commonality of all human grammars.

Luria's science is not fatalism. Motive and purpose, values and will are as fully expressions of brain as of experience, reflecting of course cultural as well as individual choice. Our world is inescapably dual: "the mind . . . an instrument created by biological evolution and perfected in each individual through learning, experience, and social intercourse." That essential dualism places on human beings the burden of self-awareness, of the transience of life—in sum, the existential absurdity. "Yet most men do not despair." Perhaps our evolution has programmed man "with subtle wisdom" to tap the "innermost sources of optimism—art, and joy, and hope, confidence in the powers of the mind, concern for his fellow men, and pride in the pursuit of the unique human adventure." That adventure may turn out as well to be one of a class, no more unique than any other species is, a way of living pursued near many another sun over all past ages.

THE LIVING ARTS OF NIGERIA, edited by William Fagg. Illustrated by Michael Foreman. Photographs by Harri Peccinotti. The Macmillan Company ($12.95). OLD WAYS OF WORKING WOOD, by Alex W. Bealer. Barre Publishers ($12.50). The displacement of human skills by the machine—more precisely, by the complex productive organization and by the designers—leaves us holding the plastic bag, idle and dissatisfied. These two books, entirely different in approach and topic, share in the celebration of craftsmanship, complex human skills that join utility neatly to art.

The Nigeria book is a beauty, self-consciously so, with some stunning full-page color photographs and six or seven sequences of lively, colorful renderings of craftsmen and their tools. It is a tour de force, a work of commando ethnography. To Nigeria, a land that remains a living exhibit of old crafts plied each day by masters, a team of four talented people went for only two weeks. They were anthropological innocents but were themselves artist-craftsmen, including a first-class fashion photographer and a well-known illustrator. Their charge was to report on the state of the traditional crafts in the country. Theirs is a quick-caught sampling, impressionistic and aimed at the general reader, but carried out with such brilliance as to evoke the vivid glow of the country. "To this nest of eagles, then, I am appointed the straight man," writes the editor, who is a distinguished London authority, Keeper of Ethnography at the British Museum. He supplies a helpful few pages of introduction and an explanatory page of text for each craft. We see bead makers, brassworkers, dyers, leatherworkers, potters, weavers and wood-carvers. Each craft is compactly sampled: a few photographs of the men and women at their worksites, a connected set of renderings that follow the process in more or less detail and then a display of some of the finest pieces of the work, often in magnified detail, the handsome pictures crowding the marginless page. Here is the spot on the brass leopard when it was a circle of yellow wax; here is a clay pot being made into a sphere by the potter sitting at the hollow he dug (the wheel is not traditional in Africa), and the roc's nest of sun-dried pots waiting for the firing; here is one weaver at her broadloom and here another at his narrow ribbon. The crafts are seen north and south, east and west; all the land is sampled. A wood mask decorated with the red seeds of *Abrus* and one bearing a Model-T "with prominent headlamps and klaxon" are two memorable images. The bead makers' workshop is "visually indistinguishable from that painted . . . on the Frobenius Expedition in 1910–11," but they no longer make glass from sand there at Bida. Instead they recycle Guinness, Star, Fanta, Seven-Up, Coca-Cola, Pond's and Mentholatum bottles

INDEX OF ADVERTISERS

AUGUST 1973

into gemlike necklaces, the meltdown fanned by the bowl bellows of ancient Egypt that makes "a thump, thump, thump sound like incredible drumming."

The Bealer book is altogether different. It is the fruit of lifelong study and interest. The author's informants were woodcraftsmen of the depression South, "a remnant of the 'makedo' South of poverty-stricken Reconstruction times." Jim Whitley of Georgia could fell an oak, split it, rive shingles and fix the leaky roof of his log cabin (in which he had been born in 1873) and after his 70th birthday, working from a book, could prepare the small compass and inlay planes he needed and then produce two or three professional fiddles. Bryan Owle of North Carolina was a real woodman. "No tree in the woods passes his notice"; he recalls that a big sassafras, for example, was prized for making oxbows. Its wood is strong, close-grained and light, the best for training young oxen.

To this catalogue of remembered craft the author adds the experience of a lifetime of watching and trying. ("There are no power tools in my shop.") He treats the woodworking crafts of wide distribution, not the stylish details of the cabinetmaker's and carver's crafts, which are perhaps the only living woodworking trades that remain dominated by handcraft. From this book you can learn how trees were felled with ax and crosscut saw, how honest Abe split rails with hardwood glut and maul, what the old workbench was like (a real tool), how men have sawed. The first finishing tools were the adz and the broadax, with which hewers of wood long before mills or even planes could produce "a finely smoothed board with only the tool and their hands, *sans* bench, *sans* vise, *sans* holdfast and clamp. They hewed to the line and let the chips fall where they might." Lathes, augers and drills, planes, chisels of every kind, drawknives and spokeshaves—all are treated. The level is detailed enough so that with effort a person with some skills of hand might learn to use the tools after much practice. Bealer is a true amateur, a devotee of handcraft. It is to be noted that his book, a sort of song to wood, is equally a tribute to edges of steel. His is the craft of the modernizing world, the age of iron. He treats sharpening with the fullness and gravity it deserves. (He has in fact already celebrated the blacksmith in a companion volume.) The book has more than 200 clear and attractive line drawings by the talented author, which indeed supplement the text.

The power chain saw fells most trees

today. It is dangerous to life and limb, it diminishes the pride of the tree-feller in his strength, it is hideously, even damagingly, noisy, it implies an economic nexus that ties the woodman to a larger world, not always to his advantage. It commands its own craft skills, no doubt. Chiefly it is fast. Is there a more humane way to use that speed in our crowded world, hungry for cheap boards and paper? Bealer's book is an eloquent brief for the plaintiff.

THE ECOLOGY OF STRAY DOGS: A STUDY OF FREE-RANGING URBAN ANIMALS, by Alan M. Beck. York Press ($9.50). The plains of the Serengeti are one romantic domain of the special science that studies closely the numbers, behavior, social organization and fate of animals, our modern version of natural history. This fascinating small monograph is the work of a man who—armed with camera, tape recorder and thermometer and driving not a Land-Rover but a used sedan—studied free-ranging dogs among the bricks of Baltimore. Using the best population-sample techniques, he photographed every dog he saw within half a block of his car. The pictures gave him individual identification of the dogs; the frequency of "recapture" by photography (never once did he mark or physically capture a dog) helps him to estimate the number. His range was mainly south Baltimore, an area of row houses, alleyways and narrow streets. The dog population of the city is about 100,000, the human population nine times that. (A survey-based rule of thumb suggests a normal dog-to-people ratio near 1:7.) Free-ranging dogs, the subject of this study, are about a third to a half of the total. These are chiefly dogs that are released by their owners for the day or the night; others escape from their owners, over fences and under them. Free-breeding, truly feral dogs also exist, but it is hard to estimate how numerous they are.

Dogs are active and visible mainly around dawn and dusk; they avoid the midday heat. (Englishmen and mad dogs are rare in this study.) One wholly feral pack was watched; these dogs too ran and fed on garbage at the two ends of the day (it was a summer month). They use the streets for socializing; they feed in the alleys. The pack Beck watched ranged over only one or two dozen city blocks, an area that is small compared with what is reported for rural dogs; Baltimore must be a favorable habitat. Dogs generally spend most of the time near the center of their customary range, which is bounded by no obvious human marks such as streets or fields. Their food is mostly garbage, but some comes from handouts left for them by people. Water is taken from gutters and puddles filled by rain or by the wastes and leaks of human use. The dogs search through trash containers, tip over cans and carry bags away to areas where they can open and rummage through them safely. (Rats may then scavenge on the dogs' leavings.) Poorer districts have more garbage and more free-ranging dogs. On days of garbage collection the dog population in a given area is reduced, but not by much.

Street dogs seem pretty well fed. Healthy strays taken in and pampered gain little weight; the thin street dogs are generally suffering from disease. Since dogs are fully acceptable to humans they can take shelter almost anywhere they choose, under parked cars, on car roofs in hot weather, in a landfill, in and around vacant buildings and so on. Rats, on the other hand, must find shelter away from human view. Stray dogs can easily "disperse rats and cats, and may be the apex consumers of garbage in the urban ecosystem. (Only once have I observed a human being going through garbage, and he probably was not looking for food.)" Stray dogs are genetically overwhelmingly heterogeneous; their social structure is loose, based on small transient groups that can make the best use of resources, including the capricious and complex behavior of people.

Death comes to all, dogs and men. Free-ranging dogs live about 2.3 years, pets more than twice as long. Disease, automobiles, gassing by the pound and killing by individuals "acting without public sanction" are the chief causes of death. ("Death from starvation, cold, or exposure appears to be rare.") William Blake foresaw this fact in times of civil peace: "A dog starv'd at his master's gate / Predicts the ruin of the State."

The report is all too brief. It contains serious recommendations for public health and for dog control in general. Dog bites are common and often place small children at real risk. (Strays are not the main source of bites. In any case, people fear rat bites more than dog bites, although the latter are reported 200 times more often!) The couple of thousand tons of feces spread across Baltimore each year by dogs produces genuine problems, whose solution is impeded by the sharp polarization the issue generally induces among human residents. People in our cities must come to work together "to make cities into places where dog and man can live together in health and peace."

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

DEFENSE AGAINST BOMBER ATTACK

THE MILITARY BALANCE, 1972–73. The International Institute for Strategic Studies, London, 1972.

HEARINGS ON FISCAL YEAR 1973 AUTHORIZATION FOR MILITARY PROCUREMENT, RESEARCH AND DEVELOPMENT. Committee on Armed Services, U.S. Senate, February–April, 1972.

AIR FORCE HOPES FOR IMPROVED INTERCEPTOR THROUGH MODIFICATION. In *Aviation Week & Space Technology,* Vol. 96, No. 26, pages 42–43; June 26, 1972.

U.S. TO MODERNIZE AIR DEFENSE. Cecil Brownlow in *Aviation Week & Space Technology,* Vol. 97, No. 22, page 12; November 27, 1972.

THE ISOLATION OF GENES

RIBOSOMAL CISTRONS AND THE NUCLEOLAR ORGANIZER. H. Wallace and M. L. Birnstiel in *Biochimica et Biophysica Acta,* Vol. 114, No. 2, pages 296–310; February 21, 1966.

SPECIFIC GENE AMPLIFICATION IN OOCYTES. Donald D. Brown and Igor B. Dawid in *Science,* Vol. 160, No. 3825, pages 272–280; April 19, 1968.

PURIFICATION AND SOME CHARACTERISTICS OF 5S DNA FROM *XENOPUS LAEVIS*. Donald D. Brown, Pieter C. Wensink and Eddie Jordan in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 68, No. 12, pages 3175–3179; December, 1971.

A COMPARISON OF THE RIBOSOMAL DNA'S OF *XENOPUS LAEVIS* AND *XENOPUS MULLERI:* THE EVOLUTION OF TANDEM GENES. Donald D. Brown, Pieter C. Wensink and Eddie Jordan in *Journal of Molecular Biology,* Vol. 63, No. 1, pages 57–73; January 14, 1972.

EXPERIMENTS WITH NEUTRINO BEAMS

NEUTRINO PHYSICS. Leon M. Lederman in *High Energy Physics: Vol. II,* edited by E. H. S. Burhop. Academic Press, 1967.

PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON HIGH ENERGY PHYSICS: Vol. IV. Edited by D. H. Perkins. National Accelerator Laboratory, 1972.

FUNDAMENTALS OF ELEMENTARY PARTICLE PHYSICS. Michael Longo. McGraw-Hill Publishing Co., 1973.

METAL-OXIDE-SEMICONDUCTOR TECHNOLOGY

THE SILICON INSULATED-GATE FIELD-EFFECT TRANSISTOR. S. R. Hofstein and F. P. Heiman in *Proceedings of the IEEE,* Vol. 51, No. 9, pages 1190–1202; September, 1963.

C/MOS UNITES WITH SILICON GATE TO YIELD MICROPOWER TECHNOLOGY. R. R. Burgess and R. G. Daniels in *Electronics,* Vol. 44, No. 18, pages 38–43; August 30, 1971.

MOS INTEGRATED CIRCUITS. Engineering Staff of American Micro-systems, Inc. Van Nostrand Reinhold Company, 1972.

CMOS SHRINKS SIZE AND PRICE OF DATA-ACQUISITION SYSTEM. *Electronic Design,* Vol. 20, No. 19, page 104; September 14, 1972.

THE EVOLUTION OF THE ANDES

SEISMOLOGY AND THE NEW GLOBAL TECTONICS. Bryan Isacks, Jack Oliver and Lynn R. Sykes in *Journal of Geophysical Research,* Vol. 73, No. 18, pages 5855–5899; September 15, 1968.

MOUNTAIN BELTS AND THE NEW GLOBAL TECTONICS. John F. Dewey and John M. Bird in *Journal of Geophysical Research,* Vol. 75, No. 14, pages 2625–2647; May 10, 1970.

RELATIONS OF ANDESITES, GRANITES, AND DERIVATIVE SANDSTONES TO ARC-TRENCH TECTONICS. William R. Dickinson in *Reviews of Geophysics and Space Physics,* Vol. 8, No. 4, pages 813–860; November, 1970.

PLATE TECTONIC MODEL FOR THE EVOLUTION OF THE CENTRAL ANDES. David E. James in *Geological Society of America Bulletin,* Vol. 82, pages 3325–3346; December, 1971.

GEOSYNCLINES, MOUNTAINS AND CONTINENT-BUILDING. Robert S. Dietz in *Scientific American,* Vol. 226, No. 3, pages 30–38; March, 1972.

DUET-SINGING BIRDS

DUETTING AND ANTIPHONAL SONG IN BIRDS: ITS EXTENT AND SIGNIFICANCE. W. H. Thorpe in collaboration with J. Hall-Craggs, B. Hooker, T. Hooker and R. Hutchison in *Behaviour: Monograph Supplement XVIII.* E. J. Brill, The Netherlands, 1972.

NON-VERBAL COMMUNICATION. Edited by R. A. Hinde. Cambridge University Press, 1972.

THE STIRLING ENGINE

PROSPECTS OF THE STIRLING ENGINE FOR VEHICULAR PROPULSION. R. J. Meijer in *Philips Technical Review,* Vol. 31, No. 5/6, pages 169–185; 1970.

ANNUAL REPORT—1972. Medical Devices Applications Program, National Heart and Lung Institute. U.S. Department of Health, Education, and Welfare, 1972.

THE STIRLING ENGINE. Compiled by Jafar Zarinchang. Intermediate Technology Development Group, London, 1972.

STIRLING CYCLE MACHINES. Graham Walker. Oxford University Press, 1973.

HEREDITARY FAT-METABOLISM DISEASES

IDENTIFICATION OF HETEROZYGOUS CARRIERS OF LIPID STORAGE DISEASES. Roscoe O. Brady, William G. Johnson and B. William Uhlendorf in *American Journal of Medicine,* Vol. 51, No. 4, pages 423–431; October, 1971.

BIOCHEMICAL AND METABOLIC BASIS OF FAMILIAL SPHINGOLIPIDOSES. Roscoe O. Brady in *Seminars in Hematology,* Vol. 9, No. 3, pages 273–284; July, 1972.

DUNCAN'S DISEASES OF METABOLISM. Edited by P. K. Bondy and L. E. Rosenberg. W. B. Saunders Company, 1973.

HEREDITARY DISEASES—CAUSES, CURES, AND PROBLEMS. Roscoe O. Brady in *Angewandte Chemie—International Edition in English,* Vol. 12, No. 1, pages 1–11; January, 1973.

REPLACEMENT THERAPY FOR INHERITED ENZYME DEFICIENCY: USE OF PURIFIED CERAMIDETRIHEXOSIDASE IN FABRY'S DISEASE. R. O. Brady, J. F. Tallman, W. G. Johnson, A. E. Gal, W. R. Leahy, J. M. Quirk and A. S. Dekaban in *The New England Journal of Medicine,* Vol. 289, No. 1, pages 9–14; July 5, 1973.

THE AMATEUR SCIENTIST

HARMONIC CURVES. William F. Rigge. The Creighton University, 1926.

WATER. Arthur M. Buswell and Worth H. Rodebush in *Scientific American,* Vol. 194, No. 4, pages 76–89; April, 1956.

The Silver Martini.
For people who want a silver lining without the cloud.

Smirnoff Silver. Ninety point four proof. Smirnoff leaves you breathless®

Who has the lead in lead frames?

We think it's the company that can offer volume production of high-quality, open-tooled lead frames—in practically all popular sizes. With maximum economy and consistently high yield.

The production facility that can provide selective and tip gold plating. On precision-stamped frames with the most intricate configurations—in a broad range of materials, finishes, thicknesses and tempers.

The organization that can market a comprehensive family of open-tooled frames, including 8-, 14-, 16- and 18-lead DIP's and TO 66 and TO 92 power and signal transistor types.

Who has the lead in lead frames? We think it's AMP.

16-lead DIP frames.

18-lead DIP frames.

Power transistor/ lead frames.

TO 66 lead frames.

TO 92 lead frames.

For information on Lead Frames write to:
AMP Industrial Division, Harrisburg, Pa. 17105.

AMP
INCORPORATED

AMP is a trademark of AMP Incorporated