

JANUARY 2001

\$4.95

WWW.SCIAM.COM

SCIENTIFIC AMERICAN

BRAVE NEW COSMOS

A SPECIAL REPORT

Can the Universe get any stranger?

Oh, yes.

Wrinkles in Spacetime • Gravity That Repels • Galaxy-Size Particles

COVER STORY

BRAVE NEW COSMOS

37

Observational cosmology is about to become a mature science. Explanations for the universe's unexpectedly odd behaviors may then be around the corner.

Echoes from the Big Bang 38

Robert R. Caldwell and Marc Kamionkowski

A Cosmic Cartographer 44

Charles L. Bennett, Gary F. Hinshaw and Lyman Page

The Quintessential Universe 46

Jeremiah P. Ostriker and Paul J. Steinhardt

Making Sense of Modern Cosmology 54

P. James E. Peebles

Plan B for the Cosmos 58

João Magueijo

SPECIAL REPORT

The Ultimate Optical Networks

The Triumph of the Light 80

Gary Stix, staff writer

Extensions to fiber-optic technologies will supply network capacity that will border on the infinite.



The Rise of Optical Switching 88

David J. Bishop, C. Randy Giles and Saswato R. Das

Eliminating electronic switches will free networks to transmit trillions of bits of data per second.

Routing Packets with Light 96

Daniel J. Blumenthal

The ultimate optical network will depend on novel systems for processing information with lightwaves.

The Cultures of Chimpanzees 60

Andrew Whiten and Christophe Boesch

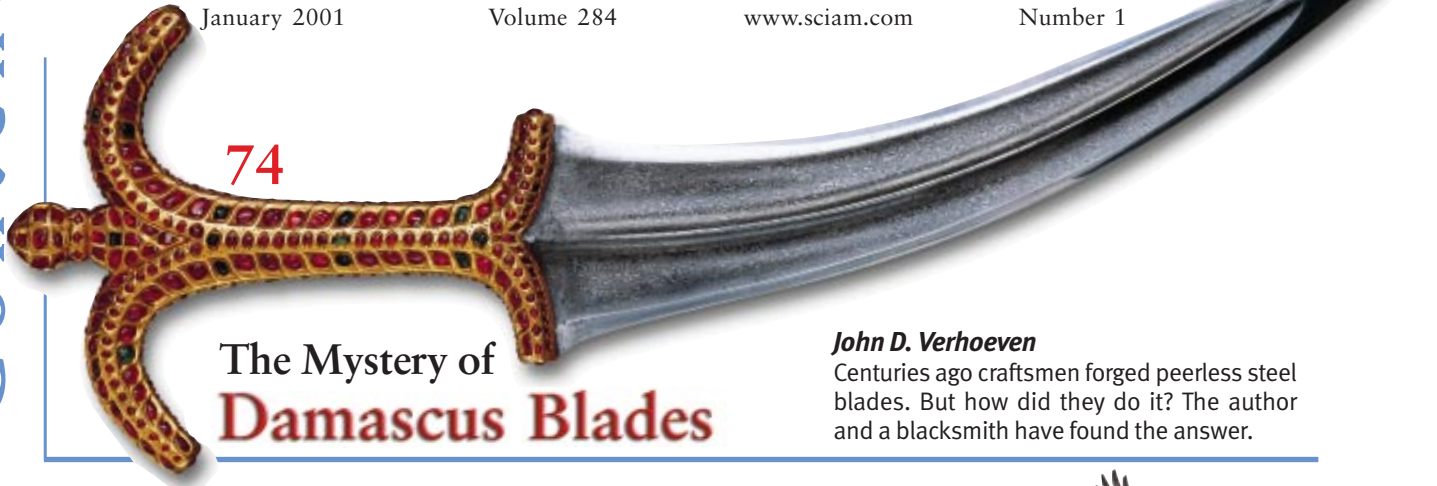


Groups of wild chimpanzees display what can only be described as social customs, a trait that had been considered unique to humans.

The Cellular Chamber of Doom 68

Alfred L. Goldberg, Stephen J. Elledge and J. Wade Harper

Cellular structures called proteasomes recycle old proteins. Some common diseases result when proteins are broken down too zealously—or not at all.



74

The Mystery of Damascus Blades

John D. Verhoeven

Centuries ago craftsmen forged peerless steel blades. But how did they do it? The author and a blacksmith have found the answer.

FROM THE EDITORS 10

LETTERS TO THE EDITORS 12

50, 100 & 150 YEARS AGO 16

PROFILE 29

Thomas R. Cech,
Nobelist with a
\$400-million checkbook.



TECHNOLOGY & BUSINESS 31

Complexity theory helps companies save—and make—millions.

CYBER VIEW 36

2001: Rating HAL against reality.

WORKING KNOWLEDGE 100

The rounded tones of flat-panel speakers.



MATHEMATICAL RECREATIONS 102

by *Ian Stewart*
Becoming a dots-and-boxes champion.

THE AMATEUR SCIENTIST 104

by *Shawn Carlson*
Viewing charged particles.

BOOKS 106

The Sibley Guide to Birds is a new classic in both ornithology and good design. Also, **The Editors Recommend.**



WONDERS by *the Morrison*s 109
Information technology, 2500 B.C.

CONNECTIONS by *James Burke* 110

ANTI GRAVITY by *Steve Mirsky* 112

END POINT 112

NEWS & ANALYSIS 18

How much precaution is too much? 18

Congress ignores genetic prejudice. 19

New planets may be stars. 21

Saving coral reefs. 22

Physics gets granular. 23

Synching the brain's hemispheres. 24

By the Numbers 26
Illegal drug use.

News Briefs 27



About the Cover
Illustration by Slim Films
and Edward Bell.

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 2000 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Periodicals postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 242764. Canadian BN No. 127387652RT; QST No. Q1015332537. Subscription rates: one year \$34.97, Canada \$49, International \$55. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537. Reprints available: write **Reprint Department, Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111; (212) 451-8877; fax: (212) 355-0408** or send e-mail to sacust@sciam.com **Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631.** Printed in U.S.A.

EDITOR JOHN RENNIE

The First Optical Internet

Thanks to fiber optics, the future of communications will be written in lines of light. Yet optical networks are not a completely new development. Although it has largely been forgotten, by the middle of the 19th century Europe was tied together by a high-speed communications network that relied entirely on optical signals.

Sketchy references to the Greeks, Romans and other cultures having used “heliographs” or mirror-polished shields to flash signals date back more than 2,000 years. The first certifiable long-distance network, however, can be traced to the end of the 18th century, when it was born amid the French Revolution. Claude Chappe, a clergyman-turned-physicist, invented a system for conveying information from one tower to another. (Given the dominance that electromagnetic communications later attained, it’s ironic that Chappe built this optical system after frustrating failures to send signals practically by wire.) Chappe’s success quickly inspired Abraham Niclas Edelcrantz, a Swedish nobleman, along a similar course.

These devices introduced *télégraphe* to the lexicons of the world. By 1850 nearly all European countries had at least one optical telegraph line, and a network crisscrossing France connected all its corners. The French system transmitted information through a type of semaphore, whereas the Swedish one employed a grid of swinging panels. Perhaps these sound quaint now, but optical telegraphs worked according to principles at the heart of today’s telecommunications, too: digital codes, data compression, error recovery, and encryption. Even their speeds were respectable. Chappe’s telegraph would probably have had an effective transmission speed of about 20 characters a minute—no threat to a modem but comparable to that of the earliest wired telegraphs of the 1830s.

(For readers who would like to know more about these early optical telegraphs, I recommend “The First Data Networks,” by Gerard J. Holzmann and Björn Pehrson, in our January 1994 issue, or the authors’ site at www.it.kth.se/docs/early_net/ on the World Wide Web.)

A weak link in that 18th-century Internet was the human element. At every tower or node, a fallible human operator had to be alert to incoming signals, to transcribe or repeat them, and to route them along the right line. In modern telecommunications, those functions have been taken over by fantastically quick, reliable electronic switches—but those components are still the weak links. The backbones of the Internet are fiber-optic cables, and photons are faster than electrons. Consequently, optical data networks will never be able to live up to their potential, or meet our future needs, until purely optical switches can replace these electronic bottlenecks. The special report on optical networking beginning on page 80 outlines the best prospects for doing so.



*More than 150 years ago
Europe was blanketed
by an optical
communications system.*

John Rennie
editors@sciam.com

SCIENTIFIC AMERICAN®

Established 1845

EDITOR IN CHIEF: John Rennie**MANAGING EDITOR:** Michelle Press**ASSISTANT MANAGING EDITOR:** Ricki L. Rusting**NEWS EDITOR:** Philip M. Yam**SPECIAL PROJECTS EDITOR:** Gary Stix**SENIOR WRITER:** W. Wayt Gibbs**EDITORS:** Mark Alpert, Graham P. Collins, Carol Ezzell, Steve Mirsky, George Musser, Sasha Nemecek, Sarah Simpson**CONTRIBUTING EDITORS:** Mark Fischetti, Marguerite Holloway, Madhusree Mukerjee, Paul Wallich**ON-LINE EDITOR:** Kristin Leutwyler**ASSOCIATE EDITOR, ON-LINE:** Kate Wong**ART DIRECTOR:** Edward Bell**SENIOR ASSOCIATE ART DIRECTOR:** Jana Brenning**ASSISTANT ART DIRECTORS:** Johnny Johnson,

Heidi Noland, Mark Clemens

PHOTOGRAPHY EDITOR: Bridget Cerety**PRODUCTION EDITOR:** Richard Hunt**COPY DIRECTOR:** Maria-Christina Keller**COPY CHIEF:** Molly K. Frances**COPY AND RESEARCH:** Daniel C. Schlenoff, Rina Bander, Sherri A. Liberman**EDITORIAL ADMINISTRATOR:** Jacob Lasky**SENIOR SECRETARY:** Maya Harty**ASSOCIATE PUBLISHER, PRODUCTION:** William Sherman**MANUFACTURING MANAGER:** Janet Cermak**ADVERTISING PRODUCTION MANAGER:** Carl Cherebin**PREPRESS AND QUALITY MANAGER:** Silvia Di Placido**PRINT PRODUCTION MANAGER:** Georgina Franco**PRODUCTION MANAGER:** Christina Hippeli**ASSISTANT PROJECT MANAGER:** Norma Jones**CUSTOM PUBLISHING MANAGER:** Madelyn Keyes**ASSOCIATE PUBLISHER/VICE PRESIDENT, CIRCULATION:**

Lorraine Leib Terlecki

CIRCULATION MANAGER: Katherine Robold**CIRCULATION PROMOTION MANAGER:** Joanne Guralnick**FULFILLMENT AND DISTRIBUTION MANAGER:** Rosa Davis**ASSOCIATE PUBLISHER, STRATEGIC PLANNING:** Laura Salant**PROMOTION MANAGER:** Diane Schube**RESEARCH MANAGER:** Aida Dadurian**PROMOTION DESIGN MANAGER:** Nancy Mongelli**SUBSCRIPTION INQUIRIES** sacust@sciam.com

U.S. and Canada (800) 333-1199,

Outside North America (515) 247-7631

GENERAL MANAGER: Michael Florek**BUSINESS MANAGER:** Marie Maher**MANAGER, ADVERTISING ACCOUNTING AND****COORDINATION:** Constance Holmes**DIRECTOR, ELECTRONIC PUBLISHING:** Martin O. K. Paul**OPERATIONS MANAGER:** Luanne Cavanaugh**ASSISTANT ON-LINE PRODUCTION MANAGER:** Heather Malloy**DIRECTOR, ANCILLARY PRODUCTS:** Diane McGarvey**PERMISSIONS MANAGER:** Linda Hertz**MANAGER OF CUSTOM PUBLISHING:** Jeremy A. Abbate**CHAIRMAN EMERITUS**

John J. Hanley

CHAIRMAN

Rolf Grisebach

PRESIDENT AND CHIEF EXECUTIVE OFFICER

Gretchen C. Teichgraber

VICE PRESIDENT AND MANAGING DIRECTOR, INTERNATIONAL

Charles McCullagh

VICE PRESIDENT

Frances Newburg

Scientific American, Inc.

415 Madison Avenue

New York, NY 10017-1111

PHONE: (212) 754-0550**FAX:** (212) 755-1976**WEB SITE:** www.sciam.com

A BANANA A DAY ...

Could vaccine-carrying foods ["Edible Vaccines," by William H. R. Langridge] lead to oral tolerance, which would depress immunity? How do you ensure that each child eats exactly enough of the enriched foods to deliver a safe and effective dose of the vaccine, without eating too much? If the modified bananas look and taste like ordinary bananas and they are grown locally to reduce distribution costs, how do you prevent their overconsumption as a normal food crop during famines or control their widespread proliferation as a result of, say, civil disorder? What effects will vaccine-laden bananas have on nonhuman consumers? (The image of a group of monkeys confronting a box labeled "Eat only one banana per person" comes to mind.) Once released into the ecosystem, it will be impossible to issue a recall order.

PAUL PERKOVIC
Montara, Calif.

What about the problem of saturating the environment with low levels of vaccines in foods, thereby promoting resistant strains?

BEN GOODMAN
Menlo Park, Calif.

Langridge replies:

These questions require intensive study in humans, but laboratory results in rodents are encouraging. When the vaccine in the foods consists of pieces from a virus or

bacterium (foreign antigens), as opposed to substances naturally made by rodents (autoantigens), the animals develop an immune response against any infectious agent displaying the foreign antigen. And repeated feedings strengthen the response. Equally fortunate, eating autoantigens shuts down unwanted immune activity against an animal's own tissues. Because human pathogens do not replicate in or attack plants, the presence of a vaccine antigen in a plant is unlikely to promote resistance. Worldwide dissemination of the vaccine plants would be prevented by confining the plants to regions of the world where a particular pathogen is a persistent problem.

RACING HEARTS

The genetic enhancement of skeletal muscle need not be limited to advancing the fortunes of professional athletes ["Muscle, Genes and Athletic Performance," by Jesper L. Andersen, Peter Schjerling and Bengt Saltin]. Researchers in the field of biomechanical cardiac assist (myself included) could benefit mightily from this new technology as we seek to train skeletal muscle for an even greater task: helping the heart to pump blood. Complete conversion of skeletal muscle to high-endurance type I fibers is now routinely achieved via chronic electrical stimulation, but steady-state power output has been limited by relatively slow contractile speeds and reductions in fiber size. This problem could potentially be solved by activating dormant genes within skeletal muscle that code for features normally found only in cardiac muscle.



FOREST/MULLIN

EACH BITE OF BANANA harvested from these trees will contain vaccine.

Such "souped-up" biological engines could be applied directly to the heart or used to drive a mechanical blood pump, providing an effective means of treating end-stage heart disease and improving the lives of millions. Now *there's* something we can all root for.

DENNIS R. TRUMBLE
Cardiothoracic Surgery Research
Allegheny General Hospital
Pittsburgh, Pa.

PLANET DETECTIVE

In "Searching for Shadows of Other Earths," the authors [Laurance R. Doyle, Hans-Jörg Deeg and Timothy M. Brown] state that "photometric transit measurements are potentially far more sensitive to smaller planets than other detection methods are." Actually, the gravitational microlensing technique is even more sensitive to low-mass planets than the transit technique. It can reveal planets with masses as small as a tenth of Earth's. The main difficulty is that the precise stellar alignment needed to see this effect is quite rare, but a wide field-of-view space-based telescope could overcome this problem. Such a mission, the Galactic Exoplanet Survey Telescope (GEST) is currently under consideration by NASA's Discovery Program.

DAVID P. BENNETT
GEST Mission principal investigator
University of Notre Dame

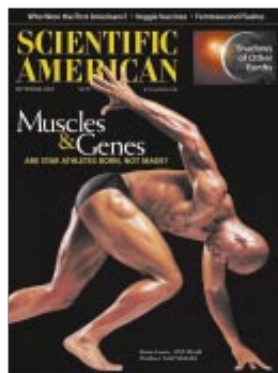
DATA COPYRIGHTS: OUTDATED?

It's true that it is illegal to give away copyrighted materials ["Brace for Impact," Cyber View, by W. Wayt Gibbs]; however, it is not illegal to copy them. Restricting data-manipulation systems because they might be used to break

THE MAIL**COACH-CLASS PASSENGERS OF THE WORLD, UNITE!**

You have nothing to lose but ... your life? Many of us learned last October of a potentially fatal medical condition known as "economy-class syndrome": deep-vein thrombosis, a circulatory problem caused by immobility. In a timely response to Phil Scott's News and Analysis article "Supersized," Mathieu Federspiel of Corvallis, Ore., writes: "It is incredible that Airbus is planning to build a 1,000-seat airplane. I question the feasibility of loading and unloading 1,000 people en masse. Scott describes the airport infrastructure 'box' that the A3XX must be engineered to fit into. I would like to see the 'box' for passenger seats enlarged a bit, to include some comfort and personal space in its specs." Hear, hear. In the meantime, though, don't forget to get out of seat #999 and stretch your legs.

Located above *this* box (in its full upright position): additional reader feedback to the September 2000 issue.



copyright laws is logically equivalent to restricting crowbars because they might be used to break into someone's house.

The entire concept of intellectual property is becoming outdated. It almost made sense at a time when inventors and artists would be discouraged from publishing their works if they didn't have some kind of guarantee of compensation. This guarantee was flimsy then and is nonexistent now. Information can be copied without harming the original.

If I have a fish and I give it to someone, I no longer have the fish. If I know of a way to get fish, and I tell someone about it, I still know how to get fish. Also, if the other person comes up with a way to refine the concept and tells me about it, the information has improved for both of us. This distinction between things and data is seemingly very difficult for people to comprehend. Not everyone who transfers compressed audio is a freeloader. Not all information duplication is theft.

ROBERT DE FOREST
via e-mail

**LIFE, HAZARDOUS;
CELL PHONES, NOT SO MUCH**

Re "Worrying about Wireless" [News and Analysis, by Mark Alpert]: I would like to see a comparison of the harmful effects of sunbathing versus using a cellular phone. Perhaps that would put the "dangers" of cellular phone use into perspective. This unwarranted fear on the part of the public is perhaps caused by the use of the word "radiation" to describe the microwave power from cellular phones. People equate the word with nuclear radiation, which definitely has been proved to cause serious health problems. I guess we need to remember that the act of living is detrimental to our health and that things need to be kept in perspective.

BENJAMIN WHITE
Beaver Dam, Wis.

Letters to the editors should be sent by e-mail to editors@sciam.com or by post to Scientific American, 415 Madison Ave., New York, NY 10017. Letters may be edited for length and clarity. Because of the considerable volume of mail received, we cannot answer all correspondence.

ERRATUM

Taima-Taima is located in Venezuela, not in Brazil ["Who Were the First Americans?"; September 2000].

SCIENTIFIC AMERICAN

Denise Anderman
PUBLISHER
danderman@sciam.com

Gail Delott
ASSOCIATE PUBLISHER
gdelott@sciam.com

NEW YORK ADVERTISING OFFICES
415 Madison Avenue
New York, NY 10017
212-451-8893 fax 212-754-1138

David Tirpack
Sales Development Manager
dtirpack@sciam.com

Wanda R. Knox
wknox@sciam.com

Hunter Millington
hmillington@sciam.com

Darren Palmieri
dpalmieri@sciam.com

Stan Schmidt
sschmidt@sciam.com

DETROIT
Edward A. Bartley
Midwest Manager
248-353-4411 fax 248-353-4360
ebartley@sciam.com

LOS ANGELES
Stephen Dudley
Los Angeles Manager
310-234-2699 fax 310-234-2670
sdudley@sciam.com

SAN FRANCISCO
Debra Silver
San Francisco Manager
415-403-9030 fax 415-403-9033
dsilver@sciam.com

CHICAGO
ROCHA & ZOELLER MEDIA SALES
312-782-8855 fax 312-782-8857
mrocha@aol.com
kzoeller1@aol.com

DALLAS
THE GRIFFITH GROUP
972-931-9001 fax 972-931-9074
lowcpm@onramp.net

CANADA
FENN COMPANY, INC.
905-833-6200 fax 905-833-2116
dfenn@canadads.com

U.K.
Anthony Turner
Simon Taylor
THE POWERS TURNER GROUP
+44 207 592-8323 fax +44 207 592-8324
aturner@publicitas.com

FRANCE AND SWITZERLAND
Patricia Goupy
+33-1-4143-8300
pgoupy@compuserve.com

GERMANY
Rupert Tonn
John Orchard
PUBLICITAS GERMANY GMBH
+49 69 71 91 49 0 fax +49 69 71 91 49 30
rtonn@publicitas.com
jorchard@publicitas.com

MIDDLE EAST AND INDIA
PETER SMITH MEDIA & MARKETING
+44 140 484-1321 fax +44 140 484-1320

JAPAN
PACIFIC BUSINESS, INC.
+813-3661-6138 fax +813-3661-6139

KOREA
BISCOM, INC.
+822 739-7840 fax +822 732-3662

HONG KONG
HUTTON MEDIA LIMITED
+852 2528 9135 fax +852 2528 9281

**OTHER EDITIONS OF
SCIENTIFIC AMERICAN**

Spektrum
DER WISSENSCHAFT

SPEKTRUM DER WISSENSCHAFT
Verlagsgesellschaft mbH
Vangerowstrasse 20
69115 Heidelberg, GERMANY
tel: +49-6221-50460
redaktion@spektrum.com

POUR LA SCIENCE

POUR LA SCIENCE
Éditions Belin
8, rue Fétou
75006 Paris, FRANCE
tel: +33-1-55-42-84-00

LE SCIENZE

LE SCIENZE
Piazza della Repubblica, 8
20121 Milano, ITALY
tel: +39-2-29001753
redazione@lescienze.it

**INVESTIGACION
CIENCIA**

INVESTIGACION Y CIENCIA
Prensa Científica, S.A.
Muntaner, 339 pral. 1.^a
08021 Barcelona, SPAIN
tel: +34-93-4143344
precisa@abaforum.es

المجلة

MAJALLAT AL-OLOOM
Kuwait Foundation for
the Advancement of Sciences
P.O. Box 20856
Safat 13069, KUWAIT
tel: +965-2428186

ŚWIAT NAUKI

ŚWIAT NAUKI
Proszynski i Ska S.A.
ul. Garazowa 7
02-651 Warszawa, POLAND
tel: +48-022-607-76-40
swiatnauki@proszynski.com.pl

日経サイエンス

NIKKEI SCIENCE, INC.
1-9-5 Otemachi, Chiyoda-ku
Tokyo 100-8066, JAPAN
tel: +813-5255-2821

CBIT HAYKH

SVIT NAUKY
Lviv State Medical University
69 Pekarska Street
290010, Lviv, UKRAINE
tel: +380-322-755856
zavadka@meduniv.lviv.ua

**ΕΛΛΗΝΙΚΗ ΕΚΔΟΣΗ
SCIENTIFIC AMERICAN HELLAS SA**

35-37 Sp. Mercouri Street
Gr 116 34 Athens, GREECE
tel: +301-72-94-354
sciam@otenet.gr

科学

KE XUE

Institute of Scientific and
Technical Information of China
P.O. Box 2104
Chongqing, Sichuan
PEOPLE'S REPUBLIC OF CHINA
tel: +86-236-3863170

Vaccines in 1901, The Mosquito's Demise

JANUARY 1951

HUMAN BODY IN SPACE—"How will the human explorer fare in his spaceship? Weightlessness evokes a pleasant picture—to float freely in space under no stress at all seems a comfortable and even profitable arrangement. But it will not be as carefree as it seems. Most probably nature will make us pay for the free ride. There is no experience on the Earth that can tell us what it will be like. It appears that we need not anticipate any serious difficulties in the functions of blood circulation and breathing. It is in the nervous system of man, his sense organs and his mind, that we can expect trouble when the body becomes weightless."

DIANETICS—[Book Review] "*Dianetics: The Modern Science of Mental Health*, by L. Ron Hubbard. Hermitage House (\$4.00). This volume probably contains more promises and less evidence per page than has any publication since the invention of printing. Briefly, its thesis is that man is intrinsically good, has a perfect memory for every event of his life, and is a good deal more intelligent than he appears to be. However, something called the engram prevents these characteristics from being realized in man's behavior. . . . By a process called dianetic revery, which resembles hypnosis and which may apparently be practiced by anyone trained in dianetics, these engrams may be recalled. Once thoroughly recalled, they are 'refiled,' and the patient becomes a 'clear' The system is presented without qualification and without evidence."

JANUARY 1901

SMALLPOX VACCINE PRODUCTION—"Until 1876 arm-to-arm vaccination was usually practiced in New York, the lymph being taken only from a vesicle of a previously vaccinated child a few months old. But human lymph has always been objectionable, in that it is a possible source of infection of a most serious blood disease. In 1876 the city Health Department laid the groundwork for the present vaccine laboratory. A calf has

vaccine (cowpox) virus smeared into superficial linear incisions made on the skin. In a few days, vesicles appear, and it is from these that the virus is obtained. Virus that has been emulsified in glycerine is drawn up into small capillary glass tubes, each tube containing enough virus for one vaccination."

STEAM TURBINE—"Just as the turbine, when installed [for electrical generation] on land, in such places as England and at Elberfeld, Germany, has surpassed the best triple-expansion reciprocating engines in economy of steam; so in marine work the steam turbine is destined to replace the reciprocating engine in all fast vessels, from moderate up to the largest tonnage.—Charles A. Parsons" [Editors' note: Parsons is considered the inventor of the modern steam turbine.]

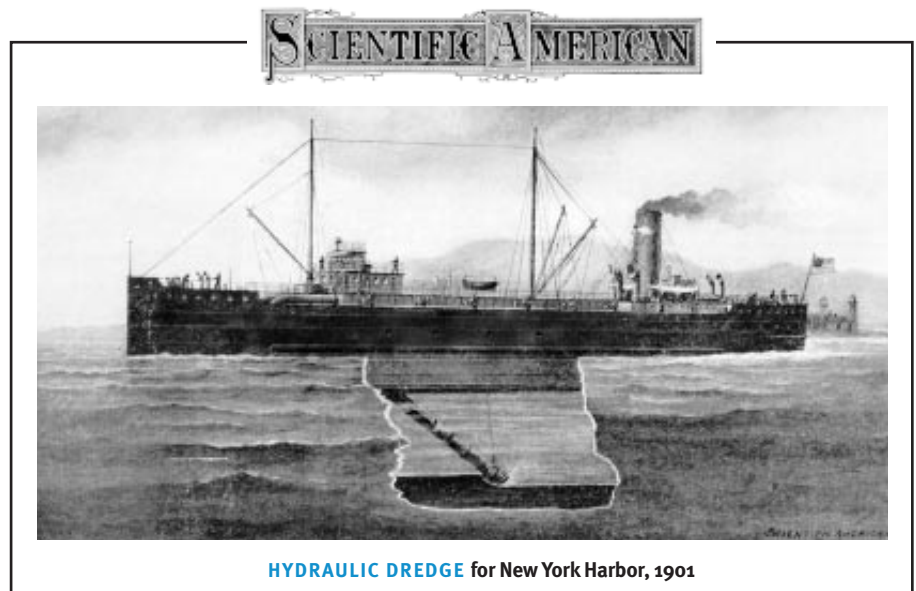
MOSQUITO EXTERMINATION—"It should not be surprising to make this prediction for the next century: Insect screens will be unnecessary. Mosquitoes will be practically exterminated. Boards of health will have destroyed all the mosquito haunts and breeding grounds, drained all stagnant pools, filled in all swamp lands and chemically treated all still-water streams."

INSURING ANARCHY—"King Alexander, of Servia [sic], has tried to have his life insured for \$2,000,000 by several companies, but one company to whom he applied for \$300,000 worth of insurance refused to write a policy on the ground of the great frequency of anarchist crimes."

HYDRAULIC DREDGE—"The rapid increase which has taken place in recent years in the size and draught of ocean steamers has necessitated considerable deepening of the channels both in the approach to New York Harbor and in the harbor itself. We illustrate herewith one of the two hydraulic hopper-type dredges (the most powerful of their kind in the world) that will excavate the estimated 39,020,000 cubic yards of the new Ambrose Channel. Sand and water are drawn up through the pipe by means of a centrifugal dredging pump of 48-inch suction and delivery, and discharged into hoppers within the hull."

JANUARY 1851

MEDICINE IN NAPLES—"The Neapolitans entertain an opinion that bloodletting is indicated in many diseases in which, among us, it would be thought fatal. Bleeding is a distinct profession, and in narrow lanes it is quite common to find painted signs, representing a nude man, tapped at several points—a stream of blood flowing from the arm, the neck, the foot, all at the same moment. In the spring, every body is supposed to require bleeding, just as, in some parts of New England, whole neighborhoods at that season take physic."



HYDRAULIC DREDGE for New York Harbor, 1901

The New Uncertainty Principle

For complex environmental issues, science learns to take a backseat to political precaution

Observe before you project yourself on a parabolic trajectory. The weight of 28.35 grams of prevention is worth 454 grams of cure. Science certainly has much to say on taking precautions. But for the enormously complex and serious problems that now face the world—global warming, loss of biodiversity, toxins in the environment—science doesn't have all the answers, and traditional risk assessment and management may not be up to the job. Indeed, given the scope of such problems, they may *never* be.

Given the uncertainty, some politicians and activists are insisting on caution first, science second. Although there is no consensus definition of what is termed the precautionary principle, one oft-mentioned statement, from the so-called Wingspread conference in Racine, Wis., in 1998 sums it up: "When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically."

In other words, actions taken to protect the environment and human health take precedence. Therefore, some advocates say, governments should immediately ban the planting of genetically modified crops, even though science can't yet say definitively whether they are a danger to the environment or to consumers.

This and other arguments surfaced at a recent conference on the precautionary principle at the Harvard University Kennedy School of Government, which drew more than 200 people from governments, industry, and research institutions of several countries. The participants grappled with the

meaning and consequences of the principle, especially as it relates to biotechnology. "Governments everywhere are confronted with the need to make decisions in the face of ignorance," pointed out Konrad von Moltke, a senior fellow at the International Institute for Sustainable Development, "and this dilemma is growing."

Critics asserted that the principle's definition and goals are vague, leaving its application dependent on the regulators in charge at the moment. All it does, they alleged, is stifle trade and limit innova-

tion. "If someone had evaluated the risk of fire right after it was invented," remarked Julian Morris of the Institute of Economic Affairs in London, "they may well have decided to eat their food raw."

A matter of law in Germany and Sweden, the precautionary principle may soon guide the policy of all of Europe: last February the European Commission outlined when and how it intends to use the precautionary principle. Increasingly, the principle is finding its way into international agreements. It was incorporated for the first time in a fully fledged inter-

national treaty last January—namely, the United Nations Biosafety Protocol regulating trade in genetically modified products. Gradually it has begun to work its way into U.S. policy. In an October speech at the National Academy of Sciences in Washington, D.C., New Jersey governor Christine Todd Whitman averred that "policymakers need to take a precautionary approach to environmental protection.... We must acknowledge that uncertainty is inherent in managing natural resources, recognize it is usually easier to prevent environmental damage than to repair it later, and shift the burden of proof away from those advocating protection toward those proposing an action that may be harmful."

Although the U.S. has taken such an approach for years—the 1958 Delaney Clause overseeing pesticide residues in food, for instance, and requirements for environmental impact statements—the more stringent requirements of the precautionary principle have not generally been welcome. During negotiations of the Biosafety Protocol in Montreal, Senator John Ashcroft of



CITING THE PRECAUTIONARY PRINCIPLE, protesters like these in Oakland, Calif., rally against "Frankenfoods." Genetically modified crops may be able to spread insecticide-laced pollen and kill nontarget species such as the monarch butterfly.

ERIC RISBERG AP Photo

Missouri criticized the incorporation of the principle, writing in a letter to President Bill Clinton that it “would, in effect, endorse the idea of making nonscience-based decisions about U.S. farm exports.”

Is the precautionary principle consistent with science, which after all can never prove a negative? “A lot of scientists get very frustrated with consumer groups, who want absolute confidence that transgenic crops are going to be absolutely safe,” says Allison A. Snow, an ecologist at Ohio State University. “We don’t scrutinize regular crops, and a lot of inventions, that carefully.”

Others don’t see the precautionary principle as antithetical to the rigorous approach of science. “The way I usually think about it is that the precautionary principle actually shines a bright light on science,” states Ted Schettler, science director for the Science and Environmental Health Network (SEHN), a consortium of environmental groups that is a leading proponent of the principle in North America. “We’re talking about enormously complex interactions among a number of systems. Now we’re starting to think that some of these things are probably unknowable and indeterminate,” he says, adding that “the precautionary principle doesn’t tell you what to do, but it does tell you [what] to look at.”

The precautionary principle requires a different kind of science, maintains Carolyn Raffensperger, SEHN’s executive director. “Science has been commodified. What we’ve created in the last 10 or 15 years is a science that has a goal of global economic competitiveness.” As examples, Raffensperger cites a relative lack of National Institutes of Health spending on allergenicity and the environmental consequences of biotechnology, compared with funding for the development of transgenic products and cancer medicines. “Our public dollars go toward developing more drugs to treat cancer rather than doing some of the things necessary to prevent cancer,” she complains.

For science to evolve along the lines envisioned by Raffensperger, researchers will have to develop a broader base of skills to handle the multifaceted data from complicated problems. National Science Foundation director Rita Colwell has been a strong proponent of the type of interdisciplinary work required to illuminate the complex scientific issues of today. The NSF specifically designed the Biocomplexity in the Environment Initiative in 1999 to address interacting sys-

tems such as global warming, human impacts on the environment, and biodiversity. Outlays have grown from an initial \$25.7 million to \$75 million for 2001.

Raffensperger also thinks the precautionary principle will require researchers to raise their social consciousness. “We need a sense of the public good” among scientists, she says. “I’m a lawyer, obligated to do public service. What if scientists shared that same obligation to use their skills for the good, pro bono? We think the precautionary principle invites us to put ethics back into science.”

In fact, Jane Lubchenco called for just such a reorientation in her presidential address at the annual meeting of the American Association for the Advancement of Science in 1997. “Urgent and unprecedented environmental and social changes challenge scientists to define a new social contract,” she said, “a commitment on the part of all scientists to devote their energies and talents to the most pressing problems of the day, in proportion to their importance, in exchange for public funding.” Raffensperger notes that the U.S. has mobilized science in this way in the past with programs on infectious diseases and national defense, such as the Manhattan Project.

What is more, scientists whose work butts up against the precautionary princi-

ple will have “to do a very good job of expressing the uncertainty in their information,” points out William W. Fox, Jr., director of science and technology for the National Marine Fisheries Service. This is difficult for some scientists, Fox notes, particularly in fisheries science, where uncertainty limits can be quite large. “You can’t always collect data exactly like your statistical model dictates, so there’s a bit of experience involved, not something that can be repeated by another scientist. It’s not really science; it’s like an artist doing it—so a large part of your scientific advice comes from art,” he comments.

Those wide limits are the crux of the issue, the point at which proponents of the precautionary principle say decisions should be taken from the realm of science and into politics. “The precautionary principle is no longer an academic debate,” Raffensperger stated at the Harvard conference. “It is in the hands of the people,” as displayed, she argued, by demonstrations against economic globalization, seen most violently in Seattle at the 1999 meeting of the World Trade Organization. “This is [about] how they want to live their lives.”

—David Appell

DAVID APPELL is a freelance science writer based in Gilford, N.H.

GENETICS _ DISCRIMINATION

Pink Slip in Your Genes

Evidence builds that employers hire and fire based on genetic tests; meanwhile protective legislation languishes

In April 1999 Terri Sargent went to her doctor with slight breathing difficulties. A simple genetic test confirmed her worst nightmare: she had alpha-1 deficiency, meaning that she might one day succumb to the same respiratory disease that killed her brother. The test probably saved Sargent’s life—the condition is treatable if detected early—but when her employer learned of her costly condition, she was fired and lost her health insurance.

Sargent’s case could have been a shining success story for genetic science. Instead it exemplifies what many feared

would happen: genetic discrimination. A recent survey of more than 1,500 genetic counselors and physicians conducted by social scientist Dorothy C. Wertz at the University of Massachusetts Medical Center found that 785 patients reported having lost their jobs or insurance because of their genes. “There is more discrimination than I uncovered in my survey,” says Wertz, who presented her findings at the American Public Health Association meeting in Boston in November. Wertz’s results buttress an earlier Georgetown University study in which 13 percent of patients surveyed said they had been denied or let go

from a job because of a genetic condition.

Such worries have already deterred many people from having beneficial predictive tests, says Barbara Fuller, a senior policy adviser at the National Human Genome Research Institute (NHGRI), where geneticists unveiled the human blueprint last June. For example, one third of women contacted for possible inclusion in a recent breast cancer study refused to participate because they feared losing their insurance or jobs if a genetic defect was discovered. A 1998 study by the National Center for Genome Resources found that 63 percent of people would not take genetic tests if employers could access the results and that 85 percent believe employers should be barred from accessing genetic information.

So far genetic testing has not had much effect on health insurance. Richard Coorsh, a spokesperson for the Health Insurance Association of America, notes that health insurers are not interested in genetic tests, for two reasons. First, they already ask for a person's family history—for many conditions, a less accurate form of genetic testing. Second, genetic tests cannot—except for a few rare conditions such as Huntington's disease—predict if someone with a disease gene will definitely get sick.

Public health scientist Mark Hall of Wake Forest University interviewed insurers and used fictitious scenarios to test the market directly. He found that a presymptomatic person with a genetic predisposition to a serious condition faces little or no difficulty in obtaining health insurance. "It's a nonissue in the insurance market," he concludes. Moreover, there is some legislation against it. Four years ago the federal government passed the Health Insurance Portability and Accountability Act (HIPAA) to prevent group insurers from denying coverage based on genetic results. A patchwork of state laws also prohibit insurers from doing so.

Genetic privacy for employees, however, has been another matter. Federal workers

are protected to some degree; last February, President Bill Clinton signed an executive order forbidding the use of genetic testing in the hiring of federal employees. But this guarantee doesn't extend to the private sector. Currently an employer can ask for, and discriminate on the basis of,



DETECTING A MISPRINT in your genes can alert you to potential diseases early enough for you to take preventive measures. But it can also get you fired, as surveys are showing. Legislation protecting private-sector employees has not gone anywhere.

medical information, including genetic test results, between the time an offer is made and when the employee begins work. A 1999 survey by the American Management Association found that 30 percent of large and midsize companies sought some form of genetic information about their employees, and 7 percent used that information in awarding promotions and hiring. As the cost of DNA testing goes down, the number of businesses testing their workers is expected to skyrocket.

Concerned scientists, including Francis S. Collins, director of the NHGRI and the

driving force behind the Human Genome Project, have called on the Senate to pass laws that ban employers from using DNA testing to blacklist job applicants suspected of having "flawed" genes. Despite their efforts, more than 100 federal and state congressional bills addressing the issue have been repeatedly shelved in the past two years. "There is no federal law on the books to protect [private-sector] employees, because members of Congress have their heads in the sand," contends Joanne Husted, a policy director at the National Partnership for Women and Families, a nonprofit group urging support of federal legislation. "Your video rental records are more protected," she claims.

Wertz also believes that more laws are simply Band-Aids on the problem: "We need a public health system to fix this one." And she may be right. In nations such as Canada and the U.K., where a national health service is in place, the thorny issue of genetic discrimination is not much of a concern.

While policymakers play catch-up with genetic science, Seargent and others are hoping that the Equal Employment Opportunity Commission (EEOC) will help. The EEOC considers discrimination based on genetic traits to be illegal under the Americans with Disabilities Act of 1990, which safeguards the disabled from employment-based discrimination. The commission has made Seargent its poster child and is taking her story to court as a test case on genetic discrimination.

Seargent, who now works at home for Alpha Net, a Web-based support group for people with alpha-1 deficiency, doubts she'll be victorious, because all but 4.3 percent of ADA cases are won by the employer. She does not regret, however, having taken the genetic test. "In the end," she says, "my life is more important than a job." Ideally, it would be better not to have to choose. —Diane Martindale

DIANE MARTINDALE is a freelance science writer based in New York City.

Lost Worlds

Evidence for the maverick view that extrasolar planets are really small stars

PASADENA, CALIF.—“It’s not even wrong” was physicist Wolfgang Pauli’s famous putdown for a theory he regarded as implausible and inconsequential. For the past several years, it has been most astronomers’ response to the ideas of David C. Black. The researcher from the Lunar and Planetary Institute in Houston is the most outspoken skeptic of the discovery of planets around other sunlike stars. He thinks the planets are actually misidentified stars, and he has stuck to that position despite the failure of his predictions, the weight

er the parent stars of the purported planets swayed from side to side, the sign of a cosmic do-si-do with partners too small to be seen directly. In many cases, the team concluded, the swaying motion was strong enough that the partners must be fairly heavy—brown dwarfs or other smallish stars, it would seem. At the least, the group has stirred a debate over selection biases in the planet searches and spiced up the broader discussion over what exactly a planet is.

In the 1980s the name of David Black was practically synonymous with extrasolar planets. He was once the head of the National Aeronautics and Space Administration’s search. But his reputation started to slide in 1995 when planet hunting became planet finding. None of the new worlds resembled anything in our solar system. Black took this as a sign that they weren’t planets after all. Their mass distribution and orbital characteristics, he asserted, look rather like those of stars. But most astronomers—including ones who used to share his views, such as William D. Heacox of the University of Hawaii at Hilo—now say Black is clinging to outmoded ideas. If nature created odd planets, even ones with starlike orbits, so be it. Accept it and move on.

To be fair, there was always a loophole in the observations.

The swaying motion of the parent stars has two components, one along the line of sight (the radial velocity) and the other across the sky (the astrometric motion). Today’s instruments can spot the latter only if the partner is fairly massive, like a star, so nearly all planet discoveries rely on the former. But radial velocity alone can merely put a lower limit on the planet masses, and if the orientation is just right, the true mass might be much greater.

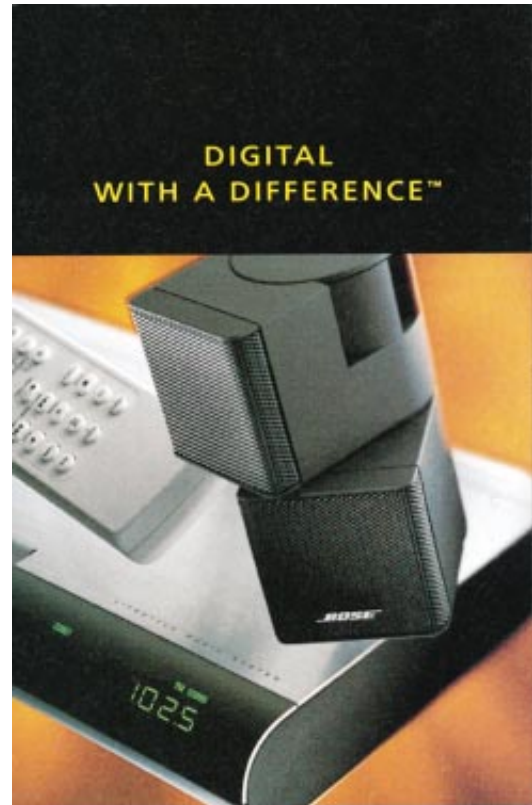
Han, Gatewood and Black have extended previous work that merged radial ve-



POSSIBLE PROTOPLANET, hanging on at the lower left from a star system in Taurus, has several times Jupiter’s mass. Such direct, infrared views are needed to determine whether, in other systems, massive planets are really brown dwarf stars.

of scientific opinion and an almost total lack of observational support. His colleagues whisper that his planet doesn’t go all the way around his star.

Now, for the first time, some evidence for Black’s view has emerged. At the Division for Planetary Sciences conference in Pasadena last October, veteran planet hunter George D. Gatewood of the University of Pittsburgh Allegheny Observatory presented the results of a study he conducted with Black and then graduate student Inwoo Han. They checked wheth-

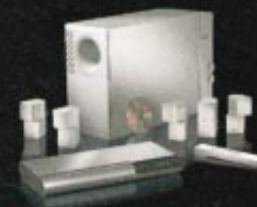


DIGITAL WITH A DIFFERENCE™

LIFESTYLE® HOME THEATER SYSTEMS

give you 5-channel surround sound from any source—5.1 encoded DVDs, any VHS tape, even single-channel TV shows. And only Bose®, the most respected name in sound, has the technology to deliver this performance with small size, elegance and simplicity.

To learn which system is best for your needs, or to find a dealer near you, call 1-800-ASK-BOSE ext. X24



BOSE
Better sound through research®

ask.bose.com/wx24

locities with astrometric data from the Hipparcos satellite. They found that out of 30 stars with companions, 15 showed astrometric motion, which implies that the partners are brown dwarfs or stars. “If that’s right, it sure does make life interesting,” Heacox says.

The response from other planet people has been swift and vigorous. “The claim by David Black is completely incorrect,” says famed planet finder Geoffrey W. Marcy of the University of California at Berkeley. He and others argue that the inferred orientations are incredibly improbable. Four of the partners were said to orbit within one degree of perfect alignment with the line of sight. Yet the chance of any single partner of a given mass having that orientation is about 1 in 5,000. Conversely, for every partner with that orientation, there should be

5,000 or so with less extreme orientations. No such bodies are seen. Marcy is so convinced that he says *Scientific American* “will be doing science a bum steer” simply by mentioning Black’s work.

Two independent groups have weighed in. Tsevi Mazeh and Shay Zucker of Tel Aviv University suggest that the truth lies somewhere in the middle. They confirm that two of the bodies indeed have the heft of a star—but only two. They see no astrometric motions for the other bodies. Hipparcos expert Dimitri Pourbaix of the Free University of Brussels initially got similar results but now suspects that the analyses have fallen prey to subtle computational biases that overestimate the mass and underestimate the error bar. To resolve the dispute, astronomers will need higher-precision astrometry (as at least two teams now intend) and direct

searches for infrared light from the stellar companions (as Mazeh plans this month at the Keck Observatory on Mauna Kea in Hawaii).

Although it looks as if Black is wrong, planet hunters can’t go scot-free just yet. Even two stellar interlopers would be two too many. Brown-dwarf expert Gibor Basri of Berkeley and others say it is quite plausible that searchers have unwittingly skewed their sample. No matter what, the theorists still have their work cut out for them. What could possibly account for the amazing diversity of worlds, from the mannerly ones in our solar system to the errants traipsing through interstellar space? Do they all deserve the label “planet”? Basri quotes from Lewis Carroll: “‘When I make a word do a lot of work like that,’” said Humpty-Dumpty, ‘I always pay it extra.’” —George Musser

CONSERVATION_BIODIVERSITY

Aquatic Homebodies

New evidence that baby fish and shrimp stick close to home may be the key to saving coral reef biodiversity

BALI, INDONESIA—I have descended only about 10 feet below the boat when I notice another diver pointing frantically at my feet. I look down to see a moray eel—giant, toothy mouth with tail—undulating quickly in my direction. A bubbly squeal escapes through my regulator as I squeeze my eyes shut and wait for the demonic creature to bore through my belly.

When I realize that my entrails are *not* scattered like tinsel across the branching corals below, I scurry after Stephen R. Palumbi, the Harvard University marine biologist who is leading this dive at Lembongan Island, just off the west coast of Bali. Eels are just as important to reef biodiversity as are pretty fish and corals, I remind myself—and that is what Palumbi and his colleagues are trying to protect. Saving coral reefs, they have found, may rely on the juvenile desires of its inhabitants.

Long-touted as the heart of marine biodiversity, Indonesian waters are home to more than 93,000 species of animals and plants. But threats such as global warming and overfishing are destroying coral



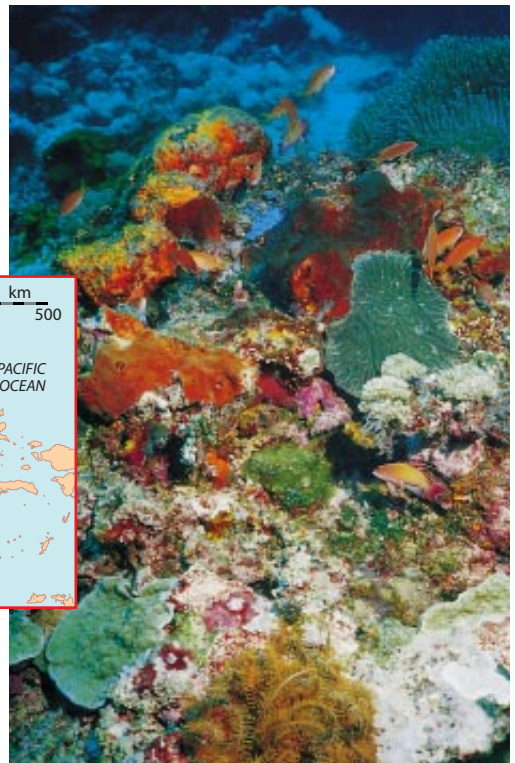
HEALTHY CORAL REEFS in Indonesia might be able to rejuvenate damaged ones if baby animals can get from one marine park (green dots on map) to another.

reefs worldwide. Along the Indonesian archipelago alone, a mere 6.5 percent are still in good condition, according to Indonesia’s vice president Megawati Sukarnoputri. That damage could hurt the nation’s 220 million people, many of whom rely on reef fish as a source of protein and economic livelihood.

To help reefs recover, officials have set up marine sanctuaries where fishing and tourism are prohibited. The key assumption is that animals from healthy parks can

repopulate devastated ones. But studies of a type of mantis shrimp—aggressive, territorial crustaceans that live at the reefs’ edges—suggest that the scheme is flawed.

The shrimp study began with Mark V. Erdmann, now with the U.S. Agency for International Development. About four years ago he enlisted fellow graduate student Paul H. Barber, now a postdoctoral fellow working with Palumbi, to confirm his identification of a handful of shrimp by analyzing their genes. In doing so, Bar-



ber stumbled on a startling pattern: the shrimp were indeed all the same species, *Haptosquilla pulchella*, but the individuals' genetic signatures differed markedly depending on where they lived. The team reported in *Nature* last August that a strong pattern of segregation exists among shrimp populations in 11 reefs around Bali and islands to the north.

Such segregation was unexpected, because "if there's any set of coral islands that's likely to be homogenized by rapid currents, it's Indonesia," Palumbi says. "It's like a washing machine." Water drains from the Pacific Ocean into the Indian Ocean through the Makassar Strait, then squeezes through the narrow waterway between Bali and its nearest western neighbor, Lombok. Tiny critters like baby shrimp could be carried hundreds of kilometers in a matter of days.

One explanation is that the babies go far but get beaten out by genetically different shrimp that want to protect their own turf. Or perhaps they are not adapted to subtle differences in the environment. More intriguing—and most likely, the researchers say—is that the shrimp are like salmon. Although they spend their earliest days at sea—as do most other crustaceans, fish and corals—it seems that they can navigate strong ocean currents to return to their birthplaces. By changing their depth at the right time, they can ride one current out from an island and take a different one back. These

larvae "are not the dumb, little floating creatures that people once thought," says Gustav Paulay of the Florida Museum of Natural History in Gainesville, Fla.

Evidence that reef animals stick close to home is turning up in other parts of the world as well. Research reported in 1999 found that fish and invertebrate larvae in the Caribbean and off the coast of Australia travel surprisingly short distances from their origins. This work, like the shrimp study, suggests that the repopulation scenario may work only for marine parks near one another.

These findings may be especially important for managing Indonesia's more than 35 widely scattered parks, whose animal populations were presumably linked by the local ocean currents. "Learning how our protected areas might be related to each other and what the minimum distance requirement is helps us define what will be an effective network for the region," says Ghislaine Llewellyn, marine conservation biologist for the World Wildlife Fund in Indonesia.

Forty minutes into our dive, the sights and sounds of this underwater paradise have overwhelmed my eel concerns. The snapping claws of mantis shrimp call to mind another important implication of my guide's research: if healthy places like this can be made into parks before they are destroyed, the local animals' tendency to stick close to home will keep them thriving.

—Sarah Simpson

PHYSICS GRANULAR MATERIALS

A Gas of Steel Balls

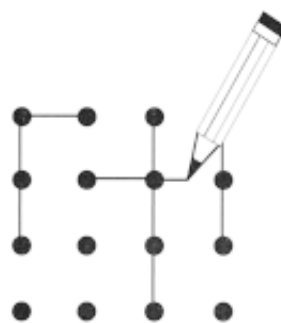
Marbles are more difficult to understand than atoms or molecules

The air that surrounds you and fills your lungs with each breath is accurately described by a detailed, microscopic theory, the kinetic theory of gases. That theory, dating back to the late 1800s, correctly predicts the macroscopic features of an ideal gas, such as its temperature and pressure, based on the motions of all its atoms or molecules. No such comprehensive theory exists for granular gases—collections of larger particles such as dust grains in space. Another baby step on the way to such a theory was taken recently by experimental physicists Florence Rouyer

and Narayanan Menon of the University of Massachusetts at Amherst, who studied the motions of a "gas" of steel ball bearings and determined that a consistent distribution of ball velocities was maintained over a range of conditions.

The study of granular materials has burgeoned over the past two decades or so. The motion of soil in an earthquake or avalanche is granular, as are many industrial processes involving foodstuffs, pharmaceuticals and other chemicals. The rings of Saturn and the interstellar dust and particles that formed the planets are granular gases. Although they move in a

Make a Move...



and Win the Game!

The Dots-and-Boxes Game
Elwyn Berlekamp
144 pp; \$14.95



"Compulsory reading for all mathematical game-players, fun for non-mathematicians, and fascinating even for specialists. The definitive work on Dots-and-Boxes."

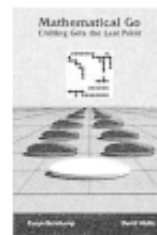
—Ian Stewart, author of *Nature's Numbers*

Improve your game strategy



Hex Strategy
Cameron Browne
384 pp; \$38.50

Mathematical Go
Elwyn Berlekamp,
David Wolfe
256 pp; \$39.00



To purchase these titles, visit your local bookstore or go to: www.akpeters.com

A K Peters, Ltd., 63 South Ave.
Natick, MA 01760-4626
Tel: (508) 655-9933
Fax: (508) 655-5847

Publishers of Science and Technology

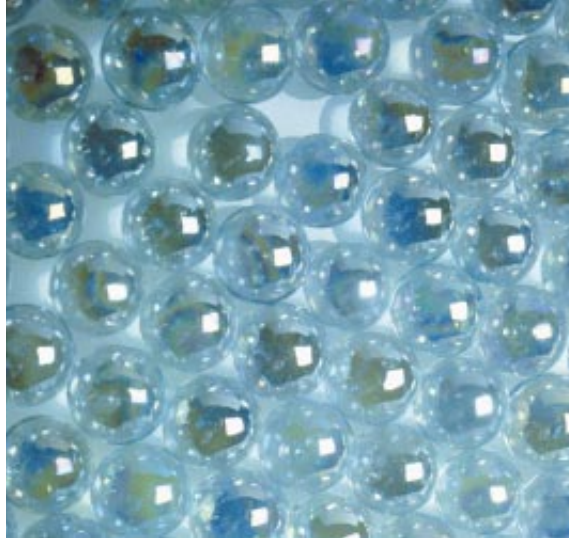
mixture of gas and liquid, powdered catalyst particles used in the multibillion-dollar petrochemical industry also behave in some ways as a granular gas. Yet granular materials remain poorly understood compared with conventional solids, liquids and gases.

The gas studied by Rouyer and Menon consisted of several hundred steel spheres, each 1.6 millimeters in diameter. These balls were enclosed in a clear plastic box, which was continuously shaken up and down a few millimeters, up to a maximum acceleration of about 60 gravities.

The need for shaking illustrates the essential differences between granular and ideal gases. The thermal motion of molecules in a gas at room temperature is great enough that the gas easily overcomes gravity and fills a container. The thermal motion of a steel ball or a dust grain, in contrast, is infinitesimal. The equilibrium state is a pile of balls or dust on the floor of the container. If the shaking is turned off, the balls fall in a heap in less than a second because at each collision some kinetic energy is lost as heat. This energy loss means that a granular gas is in a nonequilibrium state, which is much harder to analyze than an equilibrium state. James Clerk Maxwell deduced the distribution of velocities of molecules in an ideal gas in 1859 without having to measure the movement of individual molecules. For granular gases, such experiments are needed.

Rouyer and Menon obtained their velocity distributions by means of a video camera capturing 2,000 frames a second. Computer software tracked the movement of the balls in a rectangular region away from the walls of the box. To avoid the problem of balls overlapping along the camera's line of sight, they had to study their granular gas in a two-dimensional container. The box was like a double-glazed window, made of two vertical, clear plastic panes separated by slightly more than a ball's diameter.

The Maxwell distribution of velocities in an ideal gas is the familiar bell curve of statistics for which the values nearest the average occur most often. More technically, the curve is known as Gaussian, and its equation has an exponent of 2. Rouyer and Menon's granular gas consistently had a distribution with an exponent of 1.5, a distorted bell curve with fatter tails—that is, more molecules have ex-



DOUGLAS PEEBLES Corbis

MARBLES AT REST are simple enough, but when they bounce around as a “granular gas,” they behave less like a conventional gas than physicists had expected.

treme velocities. Jerry P. Gollub and his co-workers at Haverford College also obtained an exponent of 1.5 in a previous experiment that was oriented horizontally. Menon calls it “surprising and encouraging that the results are similar,” considering the very different geometries of the experiments. The 1.5 value also partially agrees with theoretical calculations made in 1998 by Twan van Noije and Matthieu H. Ernst of the University of Utrecht.

But all is not clear. Georgetown University physicists Jeffrey S. Urbach and Jeffrey S. Olafsen (now at the University of Kansas) previously conducted an experiment similar to Gollub's and obtained

somewhat different results. For some conditions, they also saw an exponent of 1.5. But for low shaking, the exponent dropped to 1, an exponential distribution, and for strong shaking, it rose to 2, the familiar Gaussian of ideal gases. (Gollub's experiment also dropped to 1 at very low shaking.) The Gaussian case occurred in the Georgetown experiment when the balls were starting to bounce through the full three dimensions instead of remaining close to the experiment's vibrating horizontal plate.

A computer simulation of the Georgetown experiment by Eli Ben-Naim of Los Alamos National Laboratory modeled that range of behavior with reasonable accuracy. Olafsen points out that the shaker in the Amherst experiment excites the particles much more strongly than the other two experiments, putting it in “a different region of parameter space.”

What's needed now, he says, are experiments and corresponding simulations that connect the different regions.

Ben-Naim says most theoreticians believe that effects such as clustering of grains and shock waves are important in some circumstances. “You're not going to get a single law that covers all the conditions,” he predicts. —Graham P. Collins

NEUROSCIENCE _ OPTICAL ILLUSIONS

Side Splitting

Jokes, ice water and magnetism can change your view of the world—literally

In John D. Pettigrew's lab, there is less to human experience than meets the eyes. Over the past several years, dozens of test subjects have stared through goggles and pressed keys while the neuroscientist squirted ice water into the volunteers' ear canals, fired strong magnetic pulses into their heads or told jokes that made them giggle. These unusual experiments, which were reported in part last March in *Current Biology* and presented more fully in November at a neuroscience conference in New Orleans, confirmed that people often cannot see what is plainly before their eyes. More important, the studies

suggest that many optical illusions may work not by deceiving our visual system, as long suspected, but rather by making visible a natural contention between the two hemispheres of the human brain. If Pettigrew's theory is correct, then the reason an optical illusion such as the Necker cube outline, which seems to turn inside out periodically, works is that, in some deep biological sense, you are of two minds on the question of what to see.

Reversible figures, such as the Necker cube and drawings of a white vase between black faces, have been curiosities for centuries. And it was in 1838 that Charles Wheatstone first reported an

even more peculiar phenomenon called binocular rivalry. When people look through a stereoscope that presents irreconcilable patterns, such as horizontal stripes before one eye and vertical bars before the other, most don't perceive a blend of the two. Instead they report seeing the left pattern, then the right, alternating every few seconds. "Every couple of seconds something goes 'click' in the brain," Pettigrew says. "But where is the switch?" The answer is still unknown.

For many years, scientists believed that neurons connected to each eye were fighting for dominance. But this theory never explained why reversible illusions work even when one eye is closed. And in monkey studies during the late 1990s, only higher-cognitive areas—parts of the brain that process patterns and not raw sensory data—consistently fired in sync with changes in the animals' perception. That discovery buttressed a new theory: that the brain constructs conflicting representations of the scene and that the representations compete somehow for attention and consciousness.

Pettigrew, a neurobiologist at the Uni-

There are several ways to do this. Ice-cold water dribbled against one eardrum causes vertigo and makes the eyes sway woozily. After the vertigo passes, however, the half of the brain opposite the chilled ear practically hums with activity. Conversely, zapping the parietal lobe on one side of the brain with a highly focused, one-tesla magnetic field temporarily interrupts much of the neural activity in just that hemisphere.

And then there is laughter. No one knows very precisely what a good guffaw does to the brain. But long bouts can cause weakness, lack of coordination, difficulty breathing, and even embarrassing wetness. Those afflicted with cataplexy, a form of narcolepsy, sometimes suffer partial or complete paralysis for several minutes after a good laugh. These seizurelike effects suggested to Pettigrew that mirth might involve neural circuits that connect the two hemispheres.

The results were "astounding," wrote Frank Sengpiel of the Cardiff School of Biosciences in Wales in a recent review. Although every test subject showed a different bias—some seeing bars for longer periods than stripes, others vice versa—most showed a statistically significant change in that bias after ice water stimulated their left hemisphere. Control subjects, who got earfuls of tepid water, showed no such change. Magnetic pulses beamed at the left hemisphere similarly allowed five of seven people tested to interrupt their perceptive cycles, effectively controlling whether they saw bars or stripes.

And among all the 20 volunteers tested, a good belly laugh either obliterated the binocular rivalry phenomenon altogether—so that subjects saw a crosshatch of both bars and stripes—or significantly reduced whatever natural bias the individuals showed toward one of the two forms, for up to half an hour.

The result seems to support, though hardly prove, Pettigrew's theory that when the brain is faced with conflicting or ambiguous scenes, the left hemisphere constructs one interpretation, the right hemisphere forms another, and an "interhemispheric switch" waffles between the two. Laughter, he speculates, either short-circuits the switch or toggles it so fast that we see both interpretations at once. "It rebalances the brain," Pettigrew

REVERSIBLE FIGURE ILLUSIONS, such as the disappearing bust of Voltaire in this Salvador Dali painting, can be short-circuited by a hearty laugh.

versity of Queensland in Brisbane, Australia, came up with a different theory: it is not just clusters of neurons that compete in binocular rivalry, but the left and right hemispheres of the cerebral cortex. To test this ambitious hypothesis, Pettigrew, Steven M. Miller and their colleagues measured how long volunteers dwelled on each possible perception of either a Necker cube or a bars-and-stripes stereoscopic display. Their plan was to fiddle with one hemisphere to see how that affected what the subjects saw.



SLAVE MARKET WITH THE DISAPPEARING BUST OF VOLTAIRE (1940), OIL ON CANVAS, 18 1/4 x 25 7/8 INCHES, COLLECTION OF THE SALVADOR DALI MUSEUM, ST. PETERSBURG, FLA. © 2000 SALVADOR DALI MUSEUM, INC.

Single and science-philic?

You needn't be a rocket scientist to join Science Connection (though we have some as members); all science friendly singles are welcome.

Our North America-wide membership includes science professionals and people with a wide range of occupations (law, teaching, music, etc.) whose interests include science or natural history. (Ages 20s-80s.)



Science Connection

(800) 667-5179
info@sciconnect.com
www.sciconnect.com

Relevant, authoritative and thought- provoking.

Scientific American and the Scientific American Teacher's Kit offer you and your class material that shows the relevance of science and technology to everyday life.

Take your class on a journey that they'll never forget!



For info about special classroom rates and the FREE Teacher's Kit, call 1 (800) 377-9414

SCIENTIFIC AMERICAN

415 Madison Avenue, New York NY 10017

www.sciam.com

says, "and literally creates a new state of mind."

Pettigrew, who has bipolar disorder, found that his own brain took 10 times longer than normal to switch between bars and stripes, an anomaly borne out by stud-

ies on his bipolar patients. A clinical trial is gearing up in Australia to test whether this may offer the first simple physical diagnostic for manic depression. Meanwhile Keith D. White of the University of Florida has discovered that many schizophrenics have

distinctly abnormal binocular rivalry. "It is much too early to say whether this might serve as a diagnostic test," White cautions. "But I wonder whether this isn't the only perceptual difference that we can measure in schizophrenia." —*W. Wayt Gibbs*

SOCIOLOGY_DRUG ABUSE

Coke, Crack, Pot, Speed et al.

In 1999 illegal drug use resulted in 555,000 emergency room visits, of which 30 percent were for cocaine, 16 percent for marijuana or hashish, 15 percent for heroin or morphine, and 2 percent for amphetamines. Alcohol in combination with other drugs accounted for 35 percent. This is not the first time that the U.S. has suffered a widespread health crisis brought on by drug abuse. In the 1880s (legal) drug companies began selling medications containing cocaine, which had only recently been synthesized from the leaves of the coca plant. Furthermore, pure cocaine could be bought legally at retail stores. Soon there were accounts of addiction and sudden death from cardiac arrest and stroke among users, as well as cocaine-related crime. Much of the blame for crime fell on blacks, although credible proof of the allegations never surfaced. Reports of health and crime problems associated with the drug contributed to rising public pressure for reform, which led in time to a ban on retail sales of cocaine under the Harrison Narcotic Act of 1914. This and later legislation contributed to the near elimination of the drug in the 1920s.

Cocaine use revived in the 1970s, long after its deleterious effects had faded from memory. By the mid-1980s history repeated itself as the U.S. rediscovered the dangers of the drug, including its new form, crack. Crack was cheap and could be smoked, a method of delivery that intensified the pleasure and the risk. Media stories about its evils, sometimes exaggerated, were apparently the key element in turning public sentiment strongly in favor of harsh sentences, even for possession. The result was one of the most important federal laws of recent years, the Anti-Drug Abuse Act of 1986. It was enacted hurriedly without benefit of committee hearings, so great was the pressure to do something about the problem. Because crack was seen as uniquely addictive and destructive, the law specified that the penalty for possession of five grams would be the same as that for possession of 500 grams of powder cocaine.

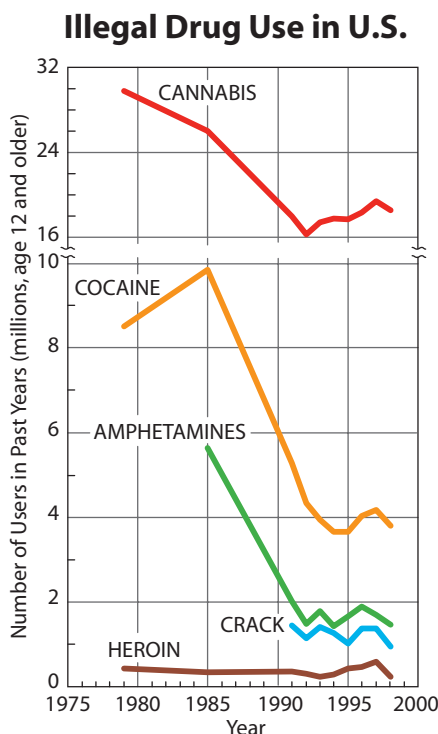
African-Americans were much more likely than whites to use crack, and so, as in the first drug epidemic, they came under greater obloquy. Because of the powder cocaine/crack penalty differential and oth-

er inequities in the justice system, blacks were far more likely to go to prison for drug offenses than whites, even though use of illicit drugs overall was about the same among both races. Blacks account for 13 percent of those who use illegal drugs but 74 percent of those sentenced to prison for possession. In fact, the 1986 federal law and certain state laws led to a substantial rise in the number of people arrested for possession of illegal drugs, at a time when arrests for sale and manufacture had stabilized.

The data in the chart catch the declining phase of the U.S. drug epidemic that started in the 1960s with the growing popularity of marijuana and, later, cocaine. Use of illegal drugs in the U.S. has fallen substantially below the extraordinarily high levels of the mid-1980s and now appears to have steadied, but hidden in the overall figures is a worrisome trend in the number of new users of illegal drugs in the past few years, such as an increase in new cocaine users from 500,000 in 1994 to 900,000 in 1998. In 1999 an estimated 14.8 million Americans were current users of illegal drugs, and of these 3.6 million were drug-dependent.

The decline in overall use occurred for several reasons, including the skittishness of affluent cocaine users, who were made wary by negative media stories. The drop in the number of people in the 18-to-25 age group, in which drug use is greatest, was probably also a factor, and prevention initiatives by the Office of National Drug Control Policy, headed by Gen. Barry McCaffrey, may have had some beneficial effect. The decrease in illegal drug use in the 1980s and early 1990s was part of a broad trend among Americans to use less psychoactive substances of any kind, including alcohol and tobacco.

Even with the decline, the U.S. way of dealing with illegal drugs is widely seen by experts outside the government as unjust, far too punitive and having the potential for involving the country in risky foreign interventions. The system has survived for so many years because the public supports it and has not focused on the defects. Surveys show that most Americans favor the system, despite calls by several national figures for drug legalization, and there is little evidence that support is softening. —*Rodger Doyle (rdoyale2@aol.com)*



SOURCE: National Household Survey on Drug Abuse, Department of Health and Human Services. Latest available data are from 1998.

ECONOMICS

Jobless in the U.S.

The Americans with Disabilities Act (ADA), which is designed to safeguard the disabled from employment-based discrimination, may have backfired. According to economist Richard V. Burkhauser of Cornell University, one group, the nearly 10 percent of working-age people with disabilities, has suffered an unprecedented decline in employment during the past 10 years, while the remainder of healthy Americans have experienced the biggest boost in jobs and financial well-being during that same time. Burkhauser suggests that lawsuits and costly workplace accommodations



Has protection backfired?

DAVID YOUNG-WOLFF/Stone

under ADA rules have made employers less than willing to hire people with disabilities. He also notes, however, that relaxed eligibility standards, which make it easier to receive Social Security benefits, might also be to blame for the drop. Burkhauser's findings will appear in the upcoming book *Ensuring Health and Security for an Aging Workforce*.

—D.M.

DYNAMICS

That Ball Is Gone

Intrigued by the home-run barrage of recent seasons, a University of Rhode Island forensic science team compared today's major league baseballs with older versions. The vintage balls, saved by fans, date back to 1963 and 1970. Investigators announced last October that the new balls' hard rubber cores bounced higher, probably because of a greater concentration of rubber, than the old ones. (The researchers believe the comparison is legitimate because the inner cores of the old balls, protected by the outer layers, did not degrade significantly over time.) Moreover, newer balls incorporate synthetic material in the wool windings, which may make the balls livelier. One researcher, a Red Sox rooter, was quoted as saying that the tests were "probably the most fun I have ever had doing science." The study may be the most fun the Sox fan ever has with baseball as well.

—Steve Mirsky



MIKE NEVELUX/Corbis

DATA POINTS

Have You Got the Right Stuff?

Requirements for space shuttle pilots:

Vision: **no worse than 20/70, correctable to 20/20**

Height: **5'4" to 6'4"**

Education: **bachelor's degree in engineering, math or science**

Jet flight experience: **1,000 hours' minimum**

Blood pressure while sitting: **no higher than 140/90**

Duration of basic training: **1 year**

Odds that a first-timer on the "Vomit Comet," a zero-g-simulating aircraft, will vomit: **1 in 3**

Number of times space shuttle can be sent into space: **100**

Shuttle's orbital speed: **17,322 miles per hour**

Landing speed: **235 mph**

Average shuttle launch cost: **\$450 million**

Frequency of astronauts' underwear changes: **every 2 days**



SOURCES: NASA Marshall Space Flight Center and Johnson Space Center; Scientific American, Vol. 281, No. 5, November 1999

MATT COLLINS

MEDICINE

Cholesterol 1, Aspirin 0

Aspirin can reduce the risk of heart attack by up to 30 percent, but it works in only three quarters of people with heart disease. High cholesterol may be a reason why it fails in the other 25 percent. At the November meeting of the American Heart Association, researchers from the University of Maryland Medical Center reported that daily doses of 325 milligrams of aspirin, a blood thinner, did not reduce the ability of platelets to clump in 60 percent of those with high cholesterol (220 milligrams per deciliter or higher). In contrast, aspirin failed in only 20 percent of those with cholesterol levels of 180 or lower. A cholesterol-controlling agent may be necessary for heart patients who don't respond to aspirin alone.

—P.Y.

BIOLOGIST THOMAS R. CECH

The \$13-Billion Man

Why the head of the Howard Hughes Medical Institute could be the most powerful individual in biomedicine

CHEVY CHASE, MD.—What's it like to lead the largest private supporter of basic biomedical research in the nation? "Very stimulating," replies Thomas R. Cech with a wry smile. "Sometimes I have trouble sleeping at night because it's so intense."

Last January, Cech (pronounced "check") became president of the Howard Hughes Medical Institute (HHMI), which spends more money on fundamental biomedical science than any other organization in the U.S. besides the federal government. In his post, he commands a research enterprise that includes a select group of 350 scientists sprinkled across the country who are generally considered to be the *crème de la crème* in their respective fields. He also oversees the distribution of millions of dollars every year in grants, primarily for science education at levels ranging from elementary school to postdoctoral training. Those two responsibilities, plus his own notable scientific findings, arguably make Cech one of the most preeminent people in biomedicine today.

Cech has assumed the stewardship of HHMI at a critical time for biomedicine. There is more funding available for biomedical research than ever before: the National Institutes of Health's annual budget is at an all-time high of \$18 billion, and that could double over the next five years based on results of proposals pending in Congress.

When added to the \$575 million provided in 2000 by HHMI, U.S. biomedical scientists will have a veritable embarrassment of riches. (The London-based Wellcome Trust, with its endowment of \$17.9

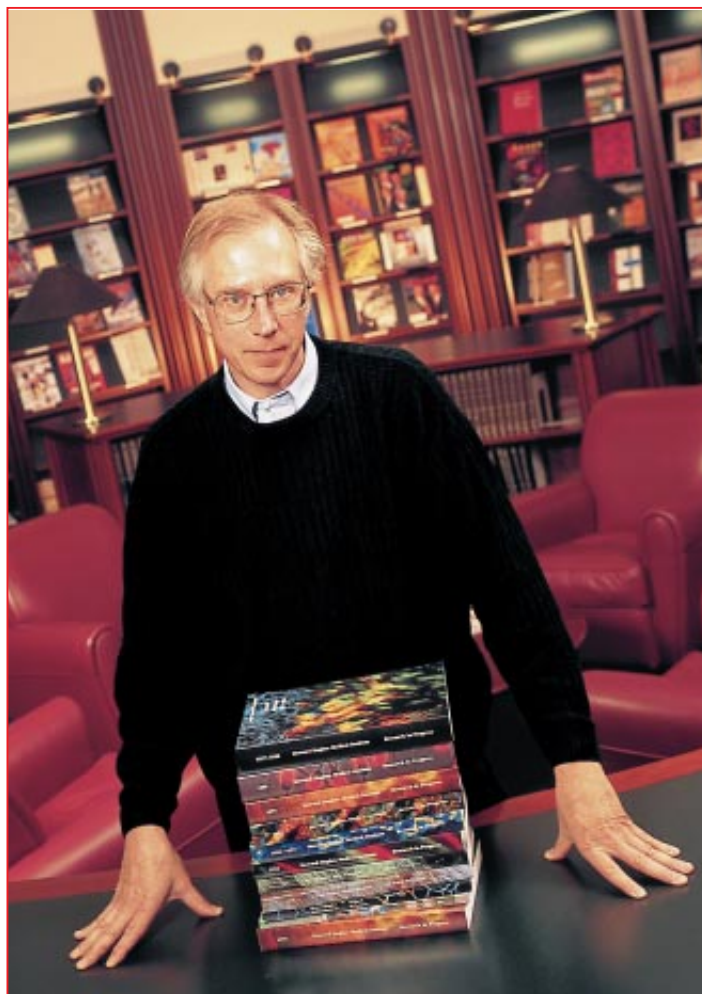
billion, is the largest medical philanthropic organization in the world and spends \$550 million a year on research.)

Cech has also taken over HHMI in an era of rapid change in biomedical science.

There are abundant ethical issues that will need to be addressed surrounding new biotechnologies such as cloning and the derivation of stem cells from human embryos. And the increasing ties between academic scientists and biopharmaceutical companies are raising questions about the propriety of such relationships and how they affect the outcome of science.

HHMI officials like to describe the organization as "an institute without walls." Instead of hiring the best people away from the universities where they work and assembling them in one huge research complex, HHMI employs scientists while allowing them to remain at their host institutions to nurture the next generation of researchers. The institute prides itself on supporting scientists' overall careers, not just particular projects, as most NIH grants do. HHMI emphasizes research in six areas: cell biology, genetics, immunology, neuroscience, computational biology and structural biology, which involves studying the three-dimensional structures of biological molecules. HHMI also has a policy of disclosing business interests in research and has forbidden certain kinds of researcher-company relationships.

As one of the world's richest philanthropies, HHMI—

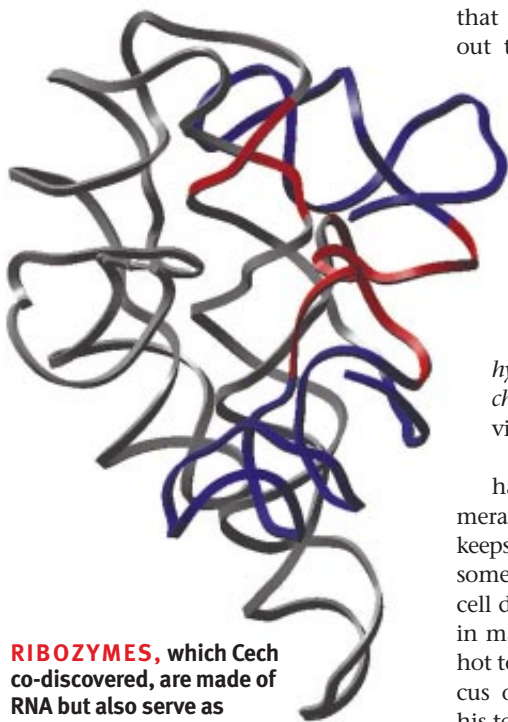


THOMAS R. CECH: FROM BERKELEY TO BIOMEDICAL GURU

- **Shared the 1989 Nobel Prize for Chemistry for discovering ribozymes**
- **Worst job: Worked in a box factory in Iowa as a young man**
- **Recent book read: *The Lexus and the Olive Tree: Understanding Globalization*, by Thomas L. Friedman**
- **Attended the University of California at Berkeley in the 1970s but "never burned anything down"**
- **Starred as "Mr. Wizard" in science education skits at the University of Colorado**
- **Met wife, Carol, over the melting-point apparatus in a chemistry lab at Grinnell College**

which is headquartered in Chevy Chase, Md., just down the road from the NIH—boasts an endowment of a whopping \$13 billion. (Founded by aviator/industrialist Howard Hughes, the organization has been funded since 1984 from the sale of Hughes Aircraft following Hughes's death.) In the past the institute sometimes had a hard time just spending enough of the interest its capital generates to satisfy the Internal Revenue Service.

HHMI's strong finances have enabled it to find top-notch researchers. Cech, for instance, won the Nobel Prize for Chemistry (shared with Sidney Altman of Yale University) in 1989 while he was an



RIBOZYMES, which Cech co-discovered, are made of RNA but also serve as enzymes, cutting and splicing genetic material.

HHMI scientist. Five other Nobelists are currently on the institute's payroll, including Eric R. Kandel of Columbia University, who shared the 2000 Nobel Prize for Physiology or Medicine.

Despite their relatively few numbers, HHMI investigators also have a disproportionate influence on biomedical research. According to a report in the September/October issue of *ScienceWatch*, which tracks research trends, scientists referenced journal articles written by HHMI scientists more frequently than articles by scientists employed by any other institution. HHMI work was cited 76,554 times between 1994 and 1999, more than twice as often as studies done at Harvard University, which at 37,118 ranked sec-

ond in overall citations during that period. The same *ScienceWatch* article reported that nine of the 15 authors with the most "high-impact" papers, as measured by the number of citations, were HHMI investigators.

Cech has written some top-cited articles himself. His papers demonstrating that the genetic material RNA can have enzymatic properties—the finding that earned him the Nobel Prize—are becoming classics. The discovery of the enzymatic RNAs, also known as ribozymes, has spawned inquiries into the origin of life.

Before Cech and Altman discovered ribozymes (during experiments they conducted independently), scientists thought that RNAs only played roles in reading out the information contained in the

DNA of an organism's genes and using those data to make proteins. The dogma also dictated that the proteins were the sole molecules that could serve as enzymes to catalyze biochemical reactions—that is, to break apart and recombine compounds. But Cech and Altman found that RNAs isolated from the ciliated protozoan *Tetrahymena* and from the bacterium *Escherichia coli* could splice themselves in vitro—a clearly enzymatic function.

More recently, Cech's laboratory has branched out to study telomerase, the RNA-containing enzyme that keeps telomeres, the ends of chromosomes, from shrinking a bit each time a cell divides. Telomerase and its function in maintaining telomeres has become a hot topic in research on aging and is a focus of new-drug development. During his tenure as president of HHMI, Cech is maintaining a scaled-down laboratory at the University of Colorado, where he has spent a few days or a week every month.

Cech was a science prodigy from an early age, although his first abiding interest was geology, not biology. He recalls that he began collecting rocks and minerals in the fourth grade and that by the time he was in junior high school in Iowa City, where he grew up, he was knocking on the doors of geology professors at the University of Iowa, pestering them with questions about meteorites and fossils.

After he entered Grinnell College, Cech says, he was drawn to physical chemistry but soon realized that he "didn't have a long enough attention span for the elaborate plumbing and electronics" of the discipline. Instead he turned to molecu-

lar biology and a career that would take him from the Ph.D. program at the University of California at Berkeley to a postdoctoral fellowship at the Massachusetts Institute of Technology to faculty positions at the University of Colorado.

As president of HHMI, Cech says that one of his first priorities concerns bioinformatics (also called computational biology), the use of computers to make sense of biological data. "Bioinformatics is really going to transform biomedical research and health care," he predicts. HHMI has already sponsored new initiatives supporting scientists using bioinformatics to study the structures of biological molecules, to model the behavior of networks of nerve cells and to compare huge chunks of DNA-sequence information arising from the Human Genome Project. "A few years ago biologists used computers only for word processing and computer games," he recalls. "The computer was late coming into biology, but when it hit, did it ever hit."

Cech is also very interested in bioethics. This summer he established a committee to organize a bioethics advisory board to help HHMI investigators negotiate some of the thornier dilemmas of biotechnology. The board, he anticipates, will meet with investigators and develop educational materials. When it comes to cloning, Cech has a specific position. So-called reproductive human cloning—generating a cloned embryo and implanting it into a human womb to develop and be born—is out of bounds for HHMI-supported researchers, he states. But cloning for medical purposes, in which cells from a cloned human fetus would be used to grow replacement tissues for an individual, "would depend on the host institution."

Overall, the 53-year-old Cech cuts quite a different figure from his predecessor at HHMI, Purnell W. Choppin, who retired at the end of 1999 at age 70. Where the courtly Choppin was never seen without a coat and tie, Cech favors open collars, sweaters, and Birkenstock sandals with socks. And where Choppin rarely mingled with his nonscientific employees at HHMI headquarters, Cech hosts a monthly social hour in the institute's enormous flower-trellised atrium. He is also encouraging HHMI investigators to bring a graduate student when they come to the meetings in which HHMI scientists share results. "My style personally," he comments, "is to be open and embracing."

—Carol Ezzell

Complexity's Business Model

Part physics, part poetry—the fledgling un-discipline finds commercial opportunity

About a year ago bottlenecks were plaguing Southwest Airlines's cargo operations, frustrating handlers and delaying flights. Known for unconventional approaches such as open seating, Southwest turned to the **Bios Group**, founded in 1996 by Santa Fe Institute luminary Stuart A. Kauffman to transform academic notions about complexity into practical know-how. Bios simulated Southwest's entire cargo operation to decipher so-called emergent behaviors and lever points—the key elements in complexity science. The goal was to find which local interactions lead to global behaviors and, specifically, what part of a system can be tweaked to control runaway effects.

Bios deftly built an agent-based model, the favored device of complexity researchers. Software agents—essentially autonomous programs—replaced each freight forwarder, ramp personnel, airplane, package and so on. The detailed computerized model revealed that freight handlers were offloading and storing many packages needlessly, ignoring a plane's ultimate destination. To counteract the emergent logjam, Bios devised a “same plane” cargo-routing strategy. Instead of shuffling parcels like hot potatoes onto the most direct flights, handlers began simply leaving them onboard to fly more circuitous routes. The result: Southwest's freight-transfer rate plummeted by roughly 70 percent at its six busiest cargo stations, saving millions in wages and overnight storage rental.

In this age of genomic gigabytes, miracle molecules and e-everything, more and more companies are finding that complexity applications can boost efficiency and profits. It hardly matters that neither a central theory nor an agreed-on definition of complexity exists. Generally speaking, “if you're talking about the real world, you're talking about complex adaptive systems,” explains Santa Fe's John L. Casti. Immune systems, food chains, computer networks and steel production all hint at the variety of both natural and civil systems. Trouble is, the real world seldom reduces to clean



BOOSTING EFFICIENCY in cargo handling and transfer is one application of complexity-based software, which often resembles biological systems.

mathematical equations. So complexologists resort to numerical simulations or models of one type or another, incorporating tools such as genetic algorithms, artificial neural networks and ant systems.

“Thanks to the computational power now available,” researchers can move beyond the reductionist approach and tackle “the inverse problem of putting the pieces back together to look at the complex system,” Kauffman expounds. Backed by **Cap Gemini Ernst & Young**, his 115-member, doctorate-rich Bios Group has advised several firms, including some 40 Fortune 500 companies, modeling everything from supply chains to shop floors to battlefields. Although Bios just released its first software shrink-wrap, called MarketBrain, most of its models are tailored for each client. “Application of complexity to the real world is not a fad,” Kauffman says.

Computer scientist John H. Holland, who holds a joint appointment at the **University of Michigan** and at Santa Fe, sees historical analogies. “Before we had a theory of electromagnetism, we had a lot

of experiments by clever people” like English physicist Michael Faraday, Holland says. “We sprinkled iron on top of magnets and built a repertoire of tools and effects.” While academicians search for an elusive, perhaps nonexistent, overarching theory of complexity, many derivative tools are proving profitable in industry.

Probably no company better illustrates this trend than **i2 Technologies** in Irving, Tex., a leading e-commerce software producer. Customers include **Abbott Laboratories**, **Dell Computer** and **Volvo**, and annual revenues top \$1 billion. Since it acquired **Optimax**, a scheduling-software design start-up, in 1997, i2 has woven complexity-based tools across its product lines. Much of i2's software uses genetic algorithms to optimize production-scheduling models. Hundreds of thousands of details, including customer orders, material and resource availability, manufacturing and distribution capability, and delivery dates are mapped into the system. Then the genetic algorithms introduce “mutations” and “crossovers”

to generate candidate schedules that are evaluated against a fitness function, explains i2 strategic adviser Gilbert P. Syswerda, an Optimax co-founder. "Genetic algorithms have proved important in generating new solutions across a lot of areas," Holland says. "There isn't any counterpart to this type of crossbreeding in traditional optimization analyses."

International Truck and Engine (formerly **Navistar**), for example, recently installed i2 software. By introducing adaptive scheduling changes, the software effectively irons out snags in production that can whipsaw through a supply chain and contribute to dreaded "lot rot." In fact, the software cut costly schedule disruptions by a stunning 90 percent at five International Truck plants, according to Kurt Satter, a systems manager with the transportation Goliath. Genetic-algorithm optimization software can also find pinch points in manufacturing and forecast effects of production-line changes, new product introductions and even advertising campaigns, Syswerda asserts. The thousands of constraints under which businesses operate can be readily encoded as well. Such nonlinear modeling is basically impossible with conventional programming tools, he maintains.

"Many of the tools that come from complexity theory have essentially become mainstream and integrated into product suites, so they are not nearly as visible anymore," explains William F. Fulkerson, an analyst at **Deere & Co.** At his suggestion, Deere's seed division tried Optimax software in its Moline, Ill., plant in the early 1990s, about the time chaos theory hit Wall Street. (A subset of complexity, chaos pertains to phenomena that evolve in predictably unpredictable ways.) Production surged, and Deere now uses the software in several plants. "Five years ago the tool itself was the message," Fulkerson observes. "Now it's the result—how much money can you make" with complexity.

Indeed, a flurry of firms playing complexity have sprouted. And the applications run the gamut. Companies such as **Artificial Life** in Boston are using neural patterning in "smart" bots to model biological processes. Their bots are essentially computer programs that use artificial intelligence to analyze the repetitive content of speech patterns on the Internet so they can interact with humans. The bots,

for example, can automate most of a company's e-mail, cutting costs by one third. The newly released line is ideal for businesses oriented toward customer service, such as the insurance industry, according to Eberhard Schoneburg, Artificial Life's chairman and CEO.

For now, financial applications generate the lion's share of Artificial Life's business, which reached nearly \$9 million in the first nine months of 2000. Its portfolio-management software, used by **Credit Suisse First Bank** and **Advance Bank**, relies on cellular automata to simulate communities of brokers and their reaction to market changes. Each cell can either buy, sell or hold a stock, its action guided by its neighbor's behavior. "When you then add simple rules governing

terns that signal market shifts and now embrace broader tenets of complexity, using filter theory, genetic algorithms, neural nets and other tools.

Complexity will most likely mesh well with the quick, data-intensive world of the Internet. Jeffrey O. Kephart, manager of IBM's agents and emergent phenomena division at its **Thomas J. Watson Research Center**, uses complex computer simulations and intelligent agents to model the development of specialized markets and cyclical price-war behavior. Eventually the Internet may enable real-time feedback of data into models. "Ultimately it's the ability to adapt at the pace of customer order that's going to be a major component of success. Complexity enables that radical view of customer focus," comments Deere & Co.'s Fulkerson.

Some researchers wonder, though, if complexity is being pushed too far. "There's still a great deal of art in the abstraction of the agents and how they interact," says David E. Goldberg, director of the Illinois Genetic Algorithms Laboratory at the **University of Illinois**. "Agent-based modeling is only as good as what the agents know and what they can learn." And currently most of the agents in models rank low on the intelligence curve. Moreover, most models fail to consider how people make decisions, notes Herbert A. Simon of **Carnegie Mellon University**, a Nobel laureate in economics who has also advanced the fields of artificial intelligence and sociobiology. "It will be a long time before the human interfaces

are smooth," he predicts.

Supporters like Casti take this criticism in stride. "Complexity science is a lot closer to physics than it is to poetry," he remarks. "But that doesn't mean there's not a lot of poetry involved." And even though the fledgling field has probably picked the low-hanging fruit, much potential remains. "Probing the boundaries—what complexity can and cannot be successfully applied to—is one of the biggest intellectual tasks the scientific endeavor has faced, and we're still in the middle of it," Goldberg says. "The process may give insight into human innovation and provide an intellectual leverage like never before."

—Julie Wakefield

JULIE WAKEFIELD, based in Washington, D.C., writes frequently about science and technology.

A Complexity Toolbox Sampler

Genetic algorithms take their cue from natural selection, creating "mutations" and "crossovers" of the "fittest" solutions to generate new and better solutions.

Intelligent agents are autonomous programs that can modify their behavior based on their experiences.

Neural networks mimic biological neurons, enabling them to learn and making them ideal for recognizing patterns in speech, images, fingerprints and more.

Cellular automata consist of a checkerboard array of cells, each obeying simple rules, that interact with one another and produce complex behavior.

Ant algorithms use a colony of cooperative agents to explore, find and reinforce optimal solutions by laying down "pheromone" trails.

Fuzzy systems model the way people think, approximating the gray areas between yes and no, on and off, right and wrong.

how to fix a market price of a stock depending on the current bids, a very realistic stock-price development can be simulated," Schoneburg says.

Companies such as **Prediction Co.**, founded in 1991 by Doyme Farmer and Norman Packard, report wild successes in using complexity applications to predict movements in financial markets. "Our results might be comparable to the biggest and best-performing hedge funds," claims CEO Packard, who won't divulge hard numbers because of confidentiality agreements. He also remains tight-lipped about how the company does it, saying that full disclosure would undermine their predictions because other firms would change their behaviors. Packard will say that their tools and models have evolved in sophistication: the duo started with chaos to decipher underlying pat-

2001: A Scorecard

How close are we to building HAL? I'm sorry, Dave, I'm afraid we can't do that

It will always be easier to make organic brains by unskilled labor than to create a machine-based artificial intelligence. That joke about doing things the old-fashioned way, which appears in the book version of *2001: A Space Odyssey*, still has an undeniable ring of truth. The science-fiction masterpiece will probably be remembered best for the finely honed portrait of a machine that could not only reason but also experience the epitome of what it means to be human: neurotic anxiety and self-doubt.

The Heuristically programmed Algorithmic Computer, a.k.a. HAL, may serve as a more fully rounded representation of a true thinking machine than the much vaunted Turing test, in which a machine proves its innate intelligence by fooling a human into thinking that it is speaking to one of its own kind. In this sense, HAL's abilities—from playing chess to formulating natural speech and reading lips—may serve as a better benchmark for measuring machine smarts than a computer that can spout vague, canned maxims that a human may interpret as signs of native intelligence.

Surprisingly, perhaps, computers in some cases have actually surpassed writer Arthur C. Clarke's and film director Stanley Kubrick's vision of computing technology at the turn of the millennium. Today's computers are vastly smaller, more portable and use software interfaces that forgo the type of manual controls found on the spaceship *Discovery 1*. But by and large, computing technology has come nowhere close to HAL. David G. Stork, who edited *Hal's Legacy: 2001's Computer as Dream and Reality*, a collection of essays comparing the state of computing with HAL's capabilities, remarks that for some defining characteristics of intelligence—language, speech recognition and understanding, common sense, emotions, planning, strategy, and lip reading—we are incapable of rendering even a rough facsimile of a HAL. "In all of the human-type problems, we've fallen far, far short," Stork says.

Even computer chess, in which seeming progress has been made, deceives. In 1997 IBM's Deep Blue beat then world champion Garry Kasparov. Deep Blue's

victory, though, was more a triumph of raw processing power than a feat that heralded the onset of the age of the intelligent machine. Quantity had become quality, Kasparov said in describing Deep Blue's ability to analyze 200 million chess positions a second. In fact, Murray F. Campbell, one of Deep Blue's creators, notes in *Hal's Legacy* that although Kasparov, in an experiment, sometimes failed to distinguish between a move by Deep Blue and one of a human grandmaster, Deep Blue's overall chess style did not exhibit human qualities and therefore was not "intelligent." HAL, in contrast, played like a real person. The computer with the unblinking red eye seemed to intuit from the outset that its opponent, *Discovery* crewman Frank Poole, was a patzer, and so it adjusted its strategy accordingly. HAL would counter with a move that was not the best one possible, to draw Poole into a trap, unlike Deep Blue, which assumes that its opponent always makes the strongest move and therefore counters with an optimized parry.

The novel of *2001* explains how the HAL 9000 series developed out of work by Marvin Minsky of the Massachusetts Institute of Technology and another researcher in the 1980s that showed how "neural networks could be generated automatically—self-replicated—in accordance with an arbitrary learning program. Artificial brains could be grown by a process strikingly analogous to the development of the human brain." Ironically, Minsky, one of the pioneers of neural networks who was also an adviser to the filmmakers (and who almost got killed by a falling wrench on the set), says today that this approach should be relegated to a minor role in modeling intelligence, while criticizing the amount of research devoted to it.

"There's only been a tiny bit of work on commonsense reasoning, and I could almost characterize the rest as various

sorts of get-rich-quick schemes, like genetic algorithms [and neural networks] where you're hoping you won't have to figure anything out," Minsky says.

Meanwhile Clarke, ensconced in his Sri Lankan home, has begun to experience an onslaught of press inquiries. "2001 is rearing its ugly head," he says. "I'm absolutely bombed out of my mind with interviews and TV." (George Orwell, who died in 1950, probably would have been glad that he never lived to see January 1, 1984.) On the morning of November 8, Clarke, 83, who suffers from a progressive neurological condition that prevents him from walking, had already received 10 e-mails, most from journalists requesting interviews. At the time, Clarke was preparing to put on scuba gear (something he not done in several years) so that he could be pho-

tographed in a local swimming pool by noted photojournalist Peter Menzel for the German magazine *Stern*. Asked if he regrets putting "2001" in the title of the screenplay, Clarke replies, "I think it was Stanley's idea."

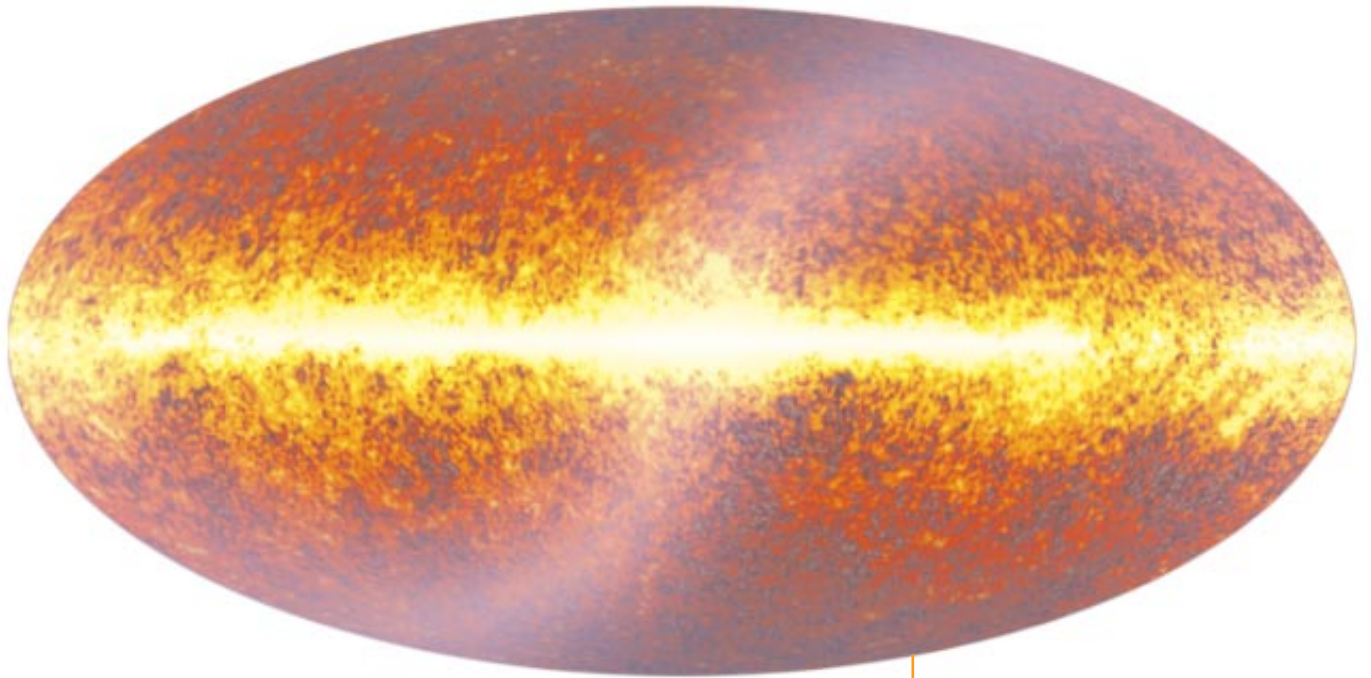
In any case, Clarke remains undeterred by how far off the mark his vision has strayed. Machine intelligence will

become more than science fiction, he believes, if not by the year marked on the cover of this magazine. "I think it's inevitable; it's just part of the evolutionary process," he says. Errors in prediction, Clarke maintains, get counterbalanced over time by outcomes more fantastic than the original insight. "First our expectations of what occurs outrun what's actually happening, and then eventually what actually happens far exceeds our expectations."

Quoting himself (Clarke's third law), Clarke remarks that "any sufficiently advanced technology is indistinguishable from magic; as technology advances it creates magic, and [AI is] going to be one of them." Areas of research that target the ultimate in miniaturization, he adds, may be the key to making good minds. "When nanotechnology is fully developed, they're going to churn [artificial brains] out as fast as they like." Time will tell if that's prediction, like Clarke's speculations about telecommunications satellites, or just a prop for science fiction. —Gary Stix



Brave New Cosmos



In recent years the field of cosmology has gone through a radical upheaval. New discoveries have challenged long-held theories about the evolution of the universe. Through it all, though, scientists have known one thing for certain: that answers to some of their most urgent questions would be coming soon from a new spacecraft, the Microwave Anisotropy Probe, or MAP. With unprecedented precision, the probe would take pictures of the material that filled the early universe, back when stars and galaxies were just a gleam in nature's eye. Encoded in the pictures would be the vital statistics of the universe: its shape, its content, its origins, its destiny.

At last, the day is almost upon us. After some delays, MAP is scheduled for launch this summer. Not since the Hubble Space Telescope have so many hopes rested on a space-based observatory.

Such instruments have turned cosmology from a largely theoretical science into an observational one. "It used to be, 'Let's do cosmology, bring a six-pack,'" says Max Tegmark of the Uni-

versity of Pennsylvania. "Now it's much more quantitative." It was the improvement in observational precision that triggered the revolution in cosmology three years ago, when supernova observers concluded that cosmic expansion is accelerating—an idea once considered laughable, even after a few beers.

The maturing of observational cosmology is the subject of the first two articles in this special section. Robert Caldwell and Marc Kamionkowski, fast-rising stars in the field, discuss how MAP and its successors could finally put the theory of inflation—widely accepted yet poorly corroborated—on a firm footing. Then, three members of MAP's science team—Charles Bennett, Gary Hinshaw and Lyman Page—outline the inner workings of their contraption, which must sift a tiny signal from seas of confounding noise.

The third article describes how the revolution is moving into a new stage. Now that observers have made a strong case for cosmic acceleration, theorists must explain it. The usual hypothesis—Einstein's cosmological constant—is rid-

■ WINDOW ON THE PAST

The Microwave Anisotropy Probe will provide a full-sky map of the cosmic microwave background radiation that was emitted nearly 15 billion years ago.

dled with paradoxes, so renowned astrophysicists Jeremiah Ostriker and Paul Steinhardt have turned to an odd kind of energy known as quintessence. The nice thing about quintessence is that it may reconcile cosmic acceleration to life. The two seem antithetical: acceleration, driven by the relentless force of the cosmological constant, would be the celestial equivalent of nuclear war—a catastrophe from which no living thing could emerge. But quintessence leaves open the possibility of a happier ending.

Finally, James Peebles, the father of modern cosmology, sorts it all out, and João Magueijo, one of the field's most innovative thinkers, mulls alternative theories. If the recent turmoil is anything to go by, we had better keep our options open.

—George Musser and
Mark Alpert, staff writers



■ **SMOOTH UNIVERSE**

In a universe with neither density variations nor gravitational waves, the cosmic microwave background (CMB) would be perfectly uniform.

Scientists may soon glimpse the universe's beginnings by studying the subtle ripples made by gravitational waves

Echoes from the Big Bang

by Robert R. Caldwell and Marc Kamionkowski

Cosmologists are still asking the same questions that the first stargazers posed as they surveyed the heavens. Where did the universe come from? What, if anything, preceded it? How did the universe arrive at its present state, and what will be its future? Although theorists have long speculated on the origin of the cosmos, until recently they had no way to probe the universe's earliest moments to test their hypotheses. In recent years, however, researchers have identified a method for observing the universe as it was in the very first fraction of a second after the big bang. This method involves looking for traces of gravitational waves in the cosmic microwave background (CMB), the cooled radiation that has permeated the universe for nearly 15 billion years.

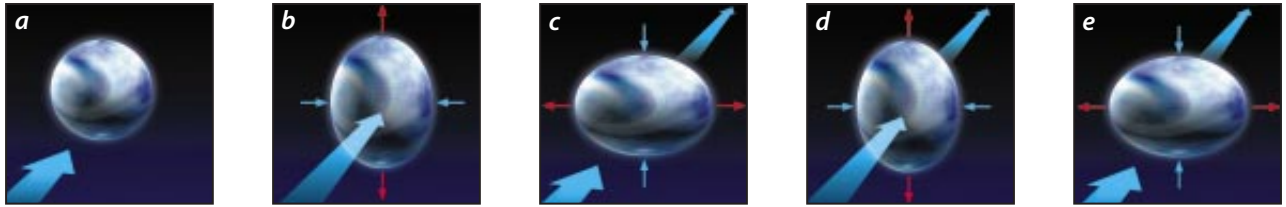
The CMB was emitted about 500,000 years after the big bang, when electrons and protons in the primordial plasma—the hot, dense soup of subatomic particles that filled the early universe—first combined to form hydrogen atoms. Because this radiation provides a snapshot of the universe at that time, it has become the Rosetta stone of cosmology. After the CMB was discovered in 1965, researchers found that its temperature—a measure of the intensity of the black

body radiation—was very close to 2.7 kelvins, no matter which direction they looked in the sky. In other words, the CMB appeared to be isotropic, which indicated that the early universe was remarkably uniform. In the early 1990s, however, a satellite called the Cosmic Background Explorer (COBE) detected minuscule variations—only one part in 100,000—in the radiation's temperature. These variations provide evidence of small lumps and bumps in the primordial plasma. The inhomogeneities in the distribution of mass later evolved into the large-scale structures of the cosmos: the galaxies and galaxy clusters that exist today.

In the late 1990s several ground-based and balloon-borne detectors observed the CMB with much finer angular resolution than COBE did, revealing structures in the primordial plasma that subtend less than one degree across the sky. (For comparison, the moon subtends about half a degree.) The size of the primordial structures indicates that the geometry of the universe is flat [see “Special Report: Revolution in Cosmology,” *SCIENTIFIC AMERICAN*, January 1999]. The observations are also consistent with the theory of inflation, which postulates that an epoch of phenomenally rapid cosmic expansion took place in the first few moments af-

ter the big bang. This year the National Aeronautics and Space Administration plans to launch the Microwave Anisotropy Probe (MAP), which will extend the precise observations of the CMB to the entire sky [see “A Cosmic Cartographer,” on page 44]. The European Space Agency's Planck spacecraft, scheduled for launch in 2007, will conduct an even more detailed mapping. Cosmologists expect that these observations will unearth a treasure trove of information about the early universe.

In particular, researchers are hoping to find direct evidence of the epoch of inflation. The strongest evidence—the “smoking gun”—would be the observation of inflationary gravitational waves. In 1918 Albert Einstein predicted the existence of gravitational waves as a consequence of his theory of general relativity. They are analogues of electromagnetic waves, such as x-rays, radio waves and visible light, which are moving disturbances of an electromagnetic field. Gravitational waves are moving disturbances of a gravitational field. Like light or radio waves, gravitational waves can carry information and energy from the sources that produce them. Moreover, gravitational waves can travel unimpeded through material that absorbs all forms of electromagnetic radiation. Just as x-rays allow doc-



GRAVITATIONAL WAVES

Although gravitational waves have never been directly observed, theory predicts that they can be detected because they stretch and squeeze the space they travel through. On striking a spherical mass (a), a wave first stretches the mass in one direction and squeezes it in a perpendicular direction (b). Then the effects are reversed (c), and the distortions oscillate at the wave's frequency (d and e). The distortions shown here have been greatly exaggerated; gravitational waves are usually too weak to produce measurable effects.



DISTORTED UNIVERSE

The fantastically rapid expansion of the universe immediately after the big bang should have produced gravitational waves. These waves would have stretched and squeezed the primordial plasma, inducing motions in the spherical surface that emitted the CMB radiation. These motions, in turn, would have caused redshifts and blueshifts in the radiation's temperature and polarized the CMB. The figure here shows the effects of a gravitational wave traveling from pole to pole, with a wavelength that is one quarter the radius of the sphere.

tors to peer through substances that visible light cannot penetrate, gravitational waves should allow researchers to view astrophysical phenomena that cannot be seen otherwise. Although gravitational waves have never been directly detected, astronomical observations have confirmed that pairs of extremely dense objects such as neutron stars and black holes generate the waves as they spiral toward each other.

The plasma that filled the universe during its first 500,000 years was opaque to electromagnetic radiation, because any emitted photons were immediately scattered in the soup of subatomic particles. Therefore, astronomers cannot observe any electromagnetic signals dating

from before the CMB. In contrast, gravitational waves could propagate through the plasma. What is more, the theory of inflation predicts that the explosive expansion of the universe 10^{-38} second after the big bang should have produced gravitational waves. If the theory is correct, these waves would have echoed across the early universe and, 500,000 years later, left subtle ripples in the CMB that can be observed today.

Waves from Inflation

To understand how inflation could have produced gravitational waves, let's examine a fascinating consequence of quantum mechanics: empty space is

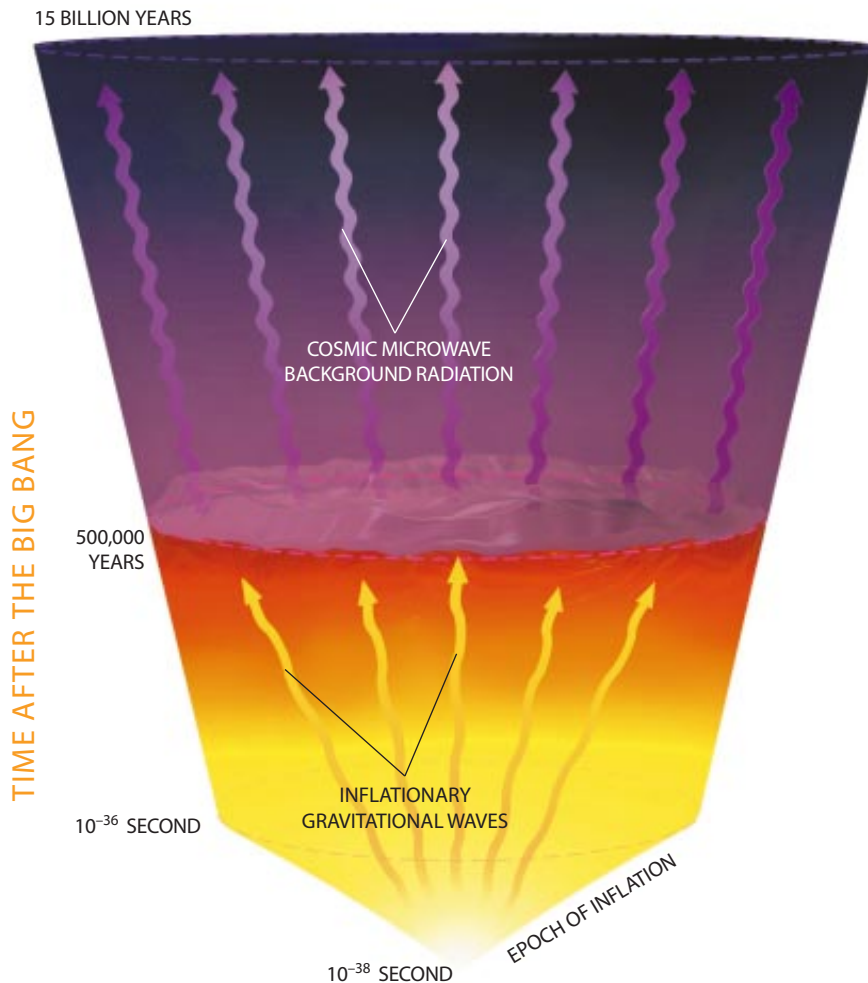
not so empty. Virtual pairs of particles are spontaneously created and destroyed all the time. The Heisenberg uncertainty principle declares that a pair of particles with energy ΔE may pop into existence for a time Δt before they annihilate each other, provided that $\Delta E \Delta t < \hbar/2$ where \hbar is the reduced Planck's constant (1.055×10^{-34} joule-second). You need not worry, though, about virtual apples or bananas popping out of empty space, because the formula applies only to elementary particles and not to complicated arrangements of atoms.

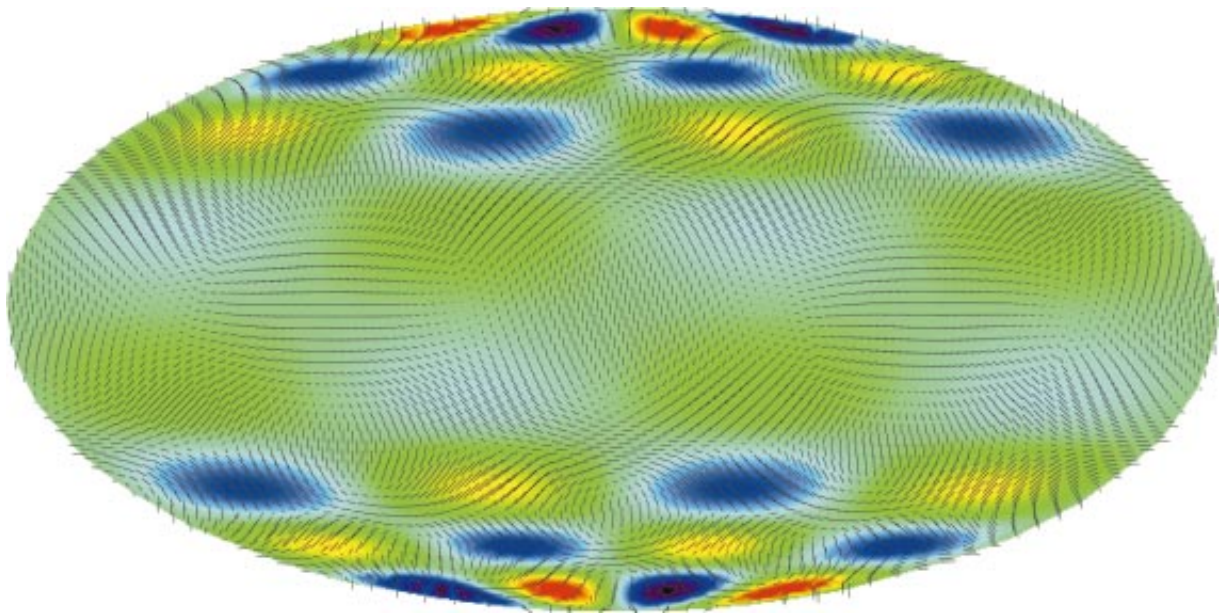
One of the elementary particles affected by this process is the graviton, the quantum particle of gravitational waves (analogous to the photon for electromagnetic waves). Pairs of virtual gravitons are constantly popping in and out of existence. During inflation, however, the virtual gravitons would have been pulled apart much faster than they could have disappeared back into the vacuum. In essence, the virtual particles would have become real particles. Furthermore, the fantastically rapid expansion of the universe would have stretched the graviton wavelengths from microscopic to macroscopic lengths. In this way, inflation would have pumped energy into the production of gravitons, generating a spectrum of gravitational waves that reflected the conditions in the universe in those first moments after the big bang. If inflationary gravitational waves do indeed exist, they would be the oldest relic in the universe, created 500,000 years before the CMB was emitted.

Whereas the microwave radiation in the CMB is largely confined to wavelengths between one and five millimeters (with a peak intensity at two millimeters), the wavelengths of the inflationary gravitational waves would span a much broader range: one centimeter to 10^{23} kilometers, which is the size of the present-day observable universe. The theory of inflation stipulates that the gravitational waves with the longest wavelengths would be the most intense and that their strength would depend on the rate at which the universe expanded during the inflationary epoch. This rate is proportional to the energy scale of inflation, which was determined by the temperature of the universe when inflation began. And because the universe was hotter at earlier times, the strength of the gravitational waves ultimately depends on the time at which inflation started.

■ COSMIC TIMELINE

During the epoch of inflation—the tremendous expansion of the universe that took place in the first moments after the big bang—quantum processes generated a spectrum of gravitational waves. The waves echoed through the primordial plasma, distorting the CMB radiation that was emitted about 500,000 years later. By carefully observing the CMB today, cosmologists may detect the plasma motions induced by the inflationary waves.





RELIC IN THE RADIATION

Inflationary gravitational waves would have left a distinctive imprint on the CMB. The diagram here depicts the simulated temperature variations and polarization patterns that would result from the distortions shown in the bottom illustration on page 39. The red and blue spots represent colder and hotter regions of the CMB, and the small line segments indicate the orientation angle of the polarization in each region of the sky.

Unfortunately, cosmologists cannot pinpoint this time, because they do not know in detail what caused inflation. Some physicists have theorized that inflation started when three of the fundamental interactions—the strong, weak and electromagnetic forces—became dissociated soon after the universe’s creation. According to this theory, the three forces were one and the same at the very beginning but became distinct 10^{-38} second after the big bang, and this event somehow triggered the sudden expansion of the cosmos. If the theory is correct, inflation would have had an energy scale of 10^{15} to 10^{16} GeV. (One GeV is the energy a proton would acquire while being accelerated through a voltage drop of one billion volts. The largest particle accelerators currently reach energies of 10^3 GeV.) On the other hand, if inflation were triggered by another physical phenomenon occurring at a later time, the gravitational waves would be weaker.

Once produced during the first fraction of a second after the big bang, the inflationary gravitational waves would propagate forever, so they should still be running across the universe. But how can cosmologists observe them? First consider how an ordinary stereo

receiver detects a radio signal. The radio waves consist of oscillating electrical and magnetic fields, which cause the electrons in the receiver’s antenna to move back and forth. The motions of these electrons produce an electric current that the receiver records.

Similarly, a gravitational wave induces an oscillatory stretching and squeezing of the space it travels through. These oscillations would cause small motions in a set of freely floating test masses. In the late 1950s physicist Hermann Bondi of King’s College, London, tried to convince skeptics of the physical reality of such waves by describing a hypothetical gravitational-wave detector. The idealized apparatus was a pair of rings hanging freely on a long, rigid bar. An incoming gravitational wave of amplitude h and frequency f would cause the distance L between the two rings to alternately contract and expand by an amount $h \times L$, with a frequency f . The heat from the friction of the rings rubbing against the bar would provide evidence that the gravitational wave carries energy.

Researchers are now building sophisticated gravitational-wave detectors, which will use lasers to track the tiny motions of suspended masses [see *box*

on next page]. The distance between the test masses determines the band of wavelengths that the devices can monitor. The largest of the ground-based detectors, which has a separation of four kilometers between the masses, will be able to measure the oscillations caused by gravitational waves with wavelengths from 30 to 30,000 kilometers; a planned space-based observatory may be able to detect wavelengths about 1,000 times longer. The gravitational waves generated by neutron star mergers and black hole collisions have wavelengths in this range, so they can be detected by the new instruments. But the inflationary gravitational waves in this range are much too weak to produce measurable oscillations in the detectors.

The strongest inflationary gravitational waves are those with the longest wavelengths, comparable to the diameter of the observable universe. To detect these waves, researchers need to observe a set of freely floating test masses separated by similarly large distances. Serendipitously, nature has provided just such an arrangement: the primordial plasma that emitted the CMB radiation. During the 500,000 years between the epoch of inflation and the emission of the CMB, the ultralong-wavelength gravitational waves echoed across the early universe, alternately stretching and squeezing the plasma [see *illustration on opposite page*]. Researchers can observe these oscillatory motions today by looking for slight Doppler shifts in the CMB.

If, at the time when the CMB was emitted, a gravitational wave was

Wave Hunters

New detectors will soon be ready

The gravitational waves produced by quantum processes during the inflationary epoch are by no means the only ones believed to be traveling across the universe. Many astrophysical systems, such as orbiting binary stars, merging neutron stars and colliding black holes, should also emit powerful gravitational waves. According to the theory of general relativity, the waves are generated by any physical system with internal motions that are not spherically symmetric. So a pair of stars orbiting each other will produce the waves, but a single star will not.

The problem with detecting the waves is that their strength fades as they spread outward. Although neutron star mergers and black hole collisions are among the most violent cataclysms in the universe, the gravitational waves produced by these events become exceedingly feeble after traveling hundreds of millions of light-years to Earth. For example, the waves from a black hole collision a billion light-years away would cause the distance between two freely floating test masses to alternately stretch and contract by a fraction of only 10^{-21} —a billionth of a trillionth.

To measure such minuscule oscillations, researchers are preparing the Laser Interferometer Gravitational-Wave Observatory (LIGO), which consists of facilities in Livingston, La., and Hanford, Wash. (photographs at right). At each site, a pair of four-kilometer-long tubes are joined at right angles in a gigantic L shape. Inside the tubes, beams of laser light will bounce back and forth between highly polished mirrors. By adjusting the laser beams so that they interfere with one another, scientists will be able to record minute changes in the distances between the mirrors, measuring oscillations as small as 10^{-17} centimeter (about a billionth the diameter of a hydrogen atom). Results from the Livingston and Hanford facilities will be compared to rule out local effects that mimic gravitational waves, such as seismic activity, acoustic noise and laser instabilities.

Physicists are also building smaller detectors that will be able to work in tandem with LIGO, allowing researchers to triangulate the sources of gravitational waves. Examples of



Livingston, La.

ALEX DESSELLE/Skyview Technologies



Hanford, Wash.

G. WHITE

these observatories are TAMA (near Tokyo), Virgo (near Pisa, Italy) and GEO (near Hannover, Germany). And to monitor gravitational waves with longer wavelengths, NASA and the European Space Agency are planning to launch the Laser Interferometer Space Antenna in 2010. This detector would consist of three identical spacecraft flying in a triangular formation and firing five-million-kilometer-long laser beams at one another. Unfortunately, none of these proposed observatories will be sensitive enough to detect the gravitational waves produced by inflation. Only the cosmic microwave background radiation can reveal their presence. —R.R.C. and M.K.

stretching a region of plasma toward us—that is, toward the part of the universe that would eventually become our galaxy—the radiation from that region will appear bluer to observers because it has shifted to shorter wavelengths (and hence a higher temperature). Conversely, if a gravitational wave was squeezing a region of plasma away from us when the CMB was emitted, the radiation will appear redder because it has shifted to longer wavelengths (and a lower temperature). By surveying the blue and red spots in the CMB—which correspond to hotter and colder radiation temperatures—researchers could conceivably see

the pattern of plasma motions induced by the inflationary gravitational waves. The universe itself becomes a gravitational-wave detector.

The Particulars of Polarization

The task is not so simple, however. As we noted at the beginning of this article, mass inhomogeneities in the early universe also produced temperature variations in the CMB. (For example, the gravitational field of the denser regions of plasma would have redshifted the photons emitted from those regions, producing some of the tempera-

ture differences observed by COBE.) If cosmologists look at the radiation temperature alone, they cannot tell what fraction (if any) of the variations should be attributed to gravitational waves. Even so, scientists at least know that gravitational waves could not have produced any more than the one-in-100,000 temperature differences observed by COBE and the other CMB radiation detectors. This fact puts an interesting constraint on the physical phenomena that gave rise to inflation: the energy scale of inflation must be less than about 10^{16} GeV, and therefore the epoch could not have occurred earlier

than 10^{-38} second after the big bang.

But how can cosmologists go further? How can they get around the uncertainty over the origin of the temperature fluctuations? The answer lies with the polarization of the CMB. When light strikes a surface in such a way that the light scatters at nearly a right angle from the original beam, it becomes linearly polarized—that is, the waves become oriented in a particular direction. This is the effect that polarized sunglasses exploit: because the sunlight that scatters off the ground is typically polarized in a horizontal direction, the filters in the glasses reduce the glare by blocking lightwaves with this orientation. The CMB is polarized as well. Just before the early universe became transparent to radiation, the CMB photons scattered off the electrons in the plasma for the last time. Some of these photons struck the particles at large angles, which polarized the radiation.

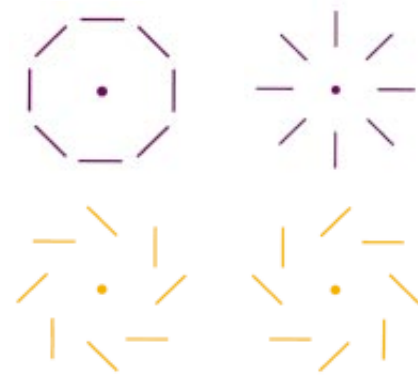
The key to detecting the inflationary gravitational waves is the fact that the plasma motions caused by the waves produced a different pattern of polarization than the mass inhomogeneities did. The idea is relatively simple. The linear polarization of the CMB can be depicted with small line segments that show the orientation angle of the polarization in each region of the sky [see illustration on page 41]. These line segments are sometimes arranged in rings or in radial patterns. The segments can also appear in rotating swirls that are either right- or left-handed—that is, they seem to be turning clockwise or counterclockwise [see illustration at right].

The “handedness” of these latter patterns is the clue to their origin. The mass inhomogeneities in the primordial plasma could not have produced such polarization patterns, because the dense and rarefied regions of plasma had no right- or left-handed orientation. In

contrast, gravitational waves do have a handedness: they propagate with either a right- or left-handed screw motion. The polarization pattern produced by gravitational waves will look like a random superposition of many rotating swirls of various sizes. Researchers describe these patterns as having a curl, whereas the ringlike and radial patterns produced by mass inhomogeneities have no curl.

Not even the most keen-eyed observer can look at a polarization diagram, such as the one shown on page 41, and tell by eye whether it contains any patterns with curls. But an extension of Fourier analysis—a mathematical technique that can break up an image into a series of waveforms—can be used to divide a polarization pattern into its constituent curl and curl-free patterns. Thus, if cosmologists can measure the CMB polarization and determine what fraction came from curl patterns, they can calculate the amplitude of the ultralong-wavelength inflationary gravitational waves. Because the amplitude of the waves was determined by the energy of inflation, researchers will get a direct measurement of that energy scale. This finding, in turn, will help answer the question of whether inflation was triggered by the unification of fundamental forces.

What are the prospects for the detection of these curl patterns? NASA’s MAP spacecraft and several ground-based and balloon-borne experiments are poised to measure the polarization of the CMB for the very first time, but these instruments will probably not be sensitive enough to detect the curl component produced by inflationary gravitational waves. Subsequent experiments may have a better chance, though. If inflation was indeed caused by the unification of forces, its gravitational-wave signal might be strong enough to be detected by the Planck spacecraft, although an



■ POLARIZATION PATTERNS

The polarization of the CMB may hold important clues to the history of the early universe. Density variations in the primordial plasma would cause ringlike and radial patterns of polarization (top). Gravitational waves, in contrast, would produce right- and left-handed swirls (bottom).

even more sensitive next-generation spacecraft might be needed. But if inflation was triggered by other physical phenomena occurring at later times and lower energies, the signal from the gravitational waves would be far too weak to be detected in the foreseeable future.

Because cosmologists are not certain about the origin of inflation, they cannot definitively predict the strength of the polarization signal produced by inflationary gravitational waves. But if there is even a small chance that the signal is detectable, then it is worth pursuing. Its detection would not only provide incontrovertible evidence of inflation but also give us the extraordinary opportunity to look back at the very earliest times, just 10^{-38} second after the big bang. We could then contemplate addressing one of the most compelling questions of the ages: Where did the universe come from? SA

THE AUTHORS

ROBERT R. CALDWELL and **MARC KAMIONKOWSKI** were both physics majors in the class of 1987 at Washington University. Caldwell earned his Ph.D. in physics at the University of Wisconsin–Milwaukee in 1992. One of the chief formulators of the theory of quintessence, Caldwell is now assistant professor of physics and astronomy at Dartmouth College. Kamionkowski earned his doctorate in physics at the University of Chicago in 1991. Now a professor of theoretical physics and astrophysics at the California Institute of Technology, he received the Warner Prize in 1998 for his contributions to theoretical astronomy.

FURTHER INFORMATION

FIRST SPACE-BASED GRAVITATIONAL-WAVE DETECTORS. Robert R. Caldwell, Marc Kamionkowski and Leven Wadley in *Physical Review D*, Vol. 59, Issue 2, pages 27101–27300; January 15, 1999.

Recent observations of the cosmic microwave background are described at these Web sites: pupgg.princeton.edu/~cmb/; www.physics.ucsb.edu/~boomerang/; cfpa.berkeley.edu/group/cmb/

Details of the MAP and Planck missions are available at map.gsfc.nasa.gov/; astro.estec.esa.nl/astrogen/planck/mission_top.html

More information on gravitational-wave detectors is available at www.ligo.caltech.edu; lisa.jpl.nasa.gov

A Cosmic Cartographer


The Microwave Anisotropy Probe will give cosmologists a much sharper picture of the early universe

by Charles L. Bennett,
Gary F. Hinshaw and Lyman Page

This summer the National Aeronautics and Space Administration is planning to launch a Delta 2 rocket carrying an 830-kilogram, four-meter-high spacecraft. Over the next three months the Microwave Anisotropy Probe (MAP) will maneuver into its target orbit around the sun, 1.5 million kilometers beyond Earth's orbit. Then the probe will begin its two-year mission, observing the cosmic microwave background (CMB) radiation in exquisite detail over the entire sky. Because this radiation was emitted nearly 15 billion years ago and has not interacted significantly with anything since then, getting a clear picture of the CMB is equivalent to drawing a map of the early universe. By studying this map, scientists can learn the composition, geometry and history of the cosmos.

As its name suggests, MAP is designed to measure the anisotropy of the CMB—the minuscule variations in the temperature of the radiation coming from different parts of the sky. MAP will be able to record differences of only 20 millionths of a kelvin from the radiation's average temperature of 2.73 kelvins. What is more, the probe can detect hot and cold spots that subtend less than 0.23 degree across the sky, yielding a total of about one million measurements. Thus, MAP's observations of the CMB will be far more detailed than the previous full-sky map, produced in the early 1990s by the Cosmic Background Explorer (COBE), which was limited to a seven-degree angular resolution.

One reason for the improvement is that MAP will employ two microwave telescopes, placed back-to-back, to focus the incoming radiation. The signals from the telescopes will feed



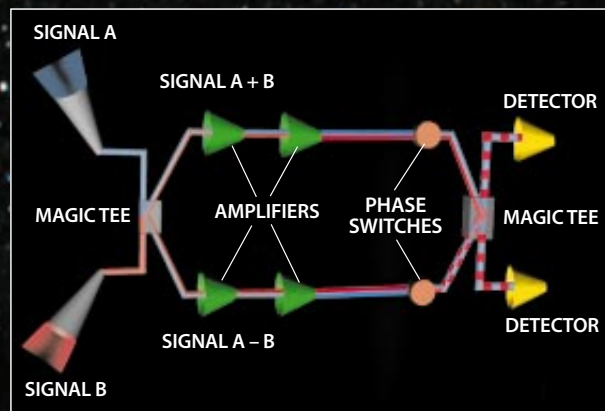
MAP'S BACK-TO-BACK TELESCOPES use primary and secondary reflectors to focus the microwave radiation (*red beams*). The primary reflectors measure 1.6 by 1.4 meters, and the secondary reflectors are one meter wide. Shielding on the back of the solar array (*orange*) blocks radiation from the sun, Earth and moon, preventing stray signals from entering the instruments. The microwaves from each telescope stream into 10 “feed horns” (*beige cones*) designed to sample five frequency bands. The four narrow horns at the center operate at 90 gigahertz, taking in microwaves with a three-millimeter wavelength. The wider horns at the periphery receive microwaves of 22, 30, 40 and 60 gigahertz. At the base of each horn is a device that splits the radiation into two orthogonal polarizations, which then feed into independent differencing assemblies (*inset at bottom of opposite page*).

into 10 “differencing assemblies” that will analyze five frequency bands in the CMB spectrum. But rather than measure the absolute temperature of the radiation, each assembly will record the temperature difference between the signals from the two telescopes. Because the probe will rotate, spinning once every two minutes and precessing once every hour, the differencing assemblies will be able to compare the temperature at each point in the sky with 1,000 other points, producing an interlocking set of data. The strategy is analogous to measuring the relative heights of bumps on a high plateau

MAP'S OBSERVATION POST will be near the L2 Lagrange point, which lies on the sun-Earth line about 1.5 million kilometers beyond our planet. The probe will orbit the sun at the same rate Earth does. This orbit ensures that MAP's telescopes will always have an unobstructed view of deep space.



DIFFERENCING ASSEMBLY combines the radiation from the two telescopes (A and B) in a device called a "magic tee," which yields $A + B$ and $A - B$ outputs. The signals are then amplified and phase-switched. Another magic tee transforms the signals back to their A and B components, and detectors record the difference in their temperatures. Because each amplifier acts on both signals, the process minimizes errors that could arise from changes in the amplifiers. The phase-switching interleaves the signals so that they can be measured precisely.



rather than recording each bump's elevation above sea level.

This method will cancel out errors resulting from slight changes in the temperature of the spacecraft itself. The overall calibration of the data will be done through a continuous measurement of the CMB dipole moment, the change in radiation temperature caused by Earth's motion through the cosmos. The guiding principle of MAP's design is to eliminate any spurious signals that might contaminate its measurements of the CMB. If all goes as planned, the probe will produce a full-sky cosmic map of unprecedented fidelity. SA

MAP SCIENCE TEAM includes Charles L. Bennett (NASA Goddard Space Flight Center), Mark Halpern (University of British Columbia), Gary F. Hinshaw (NASA GSFC), Norman C. Jarosik (Princeton University), Alan J. Kogut (NASA GSFC), Michele Limon (Princeton), Stephan S. Meyer (University of Chicago), Lyman Page (Princeton), David N. Spergel (Princeton), Gregory S. Tucker (Brown University), David T. Wilkinson (Princeton), Edward J. Wollack (NASA GSFC) and Edward L. Wright (University of California, Los Angeles).

The Quintessential

The universe has recently been commandeered by an invisible energy field, which is causing its expansion to accelerate outward

Universe

by Jeremiah P. Ostriker and Paul J. Steinhardt

Is it all over but the shouting? Is the cosmos understood aside from minor details? A few years ago it certainly seemed that way. After a century of vigorous debate, scientists had reached a broad consensus about the basic history of the universe. It all began with gas and radiation of unimaginably high temperature and density. For 15 billion years, it has been expanding and cooling. Galaxies and other complex structures have grown from microscopic seeds—quantum fluctuations—that were stretched to cosmic size by a brief period of “inflation.” We had also learned that only a small fraction of matter is composed of the normal chemical elements of our everyday experience. The majority consists of so-called dark matter, primarily exotic elementary particles that do not interact with light. Plenty of mysteries remained, but at least we had sorted out the big picture.

Or so we thought. It turns out that we have been missing most of the story. Over the past five years, observations have convinced cosmologists that the chemical elements and the dark matter, combined, amount to less than half the content of the universe. The bulk is a ubiquitous “dark energy” with a strange and remarkable feature: its gravity does not attract. It repels. Whereas gravity pulls the chemical elements and dark matter into stars and galaxies, it pushes the dark energy into a nearly uniform haze that permeates space. The universe is a battleground between the two tendencies, and repulsive gravity is winning. It is gradually overwhelming the attractive force of ordinary matter—causing the universe to accelerate to ever larger rates of expansion, perhaps lead-

ing to a new runaway inflationary phase and a totally different future for the universe than most cosmologists envisioned a decade ago.

Until recently, cosmologists have focused simply on proving the existence of dark energy. Having made a convincing case, they are now turning their attention to a deeper problem: Where does the energy come from? The best-known possibility is that the energy is inherent in the fabric of space. Even if a volume of space were utterly empty—without a bit of matter and radiation—it would still contain this energy. Such energy is a venerable notion that dates back to Albert Einstein and his attempt in 1917 to construct a static model of the universe. Like many leading scientists over the centuries, including Isaac Newton, Einstein believed that the universe is unchanging, neither contracting nor expanding. To coax stagnation from his general theory of relativity, he had to introduce vacuum energy or, in his terminology, a cosmological constant. He adjusted the value of the constant so that its gravitational repulsion would exactly counterbalance the gravitational attraction of matter.

Later, when astronomers established that the universe is expanding, Einstein regretted his delicately tuned artifice, calling it his greatest blunder. But perhaps his judgment was too hasty. If the cosmological constant had a slightly larger value than Einstein proposed, its repulsion would exceed the attraction of matter, and cosmic expansion would accelerate.

Many cosmologists, though, are now leaning toward a different idea, known as quintessence. The translation is “fifth element,” an allusion to ancient Greek philosophy, which suggested that the

universe is composed of earth, air, fire and water, plus an ephemeral substance that prevents the moon and planets from falling to the center of the celestial sphere. Three years ago Robert R. Caldwell, Rahul Dave and one of us (Steinhardt), all then at the University of Pennsylvania, reintroduced the term to refer to a dynamical quantum field, not unlike an electrical or magnetic field, that gravitationally repels.

The dynamism is what cosmologists find so appealing about quintessence. The biggest challenge for any theory of dark energy is to explain the inferred amount of the stuff—not so much that it would have interfered with the formation of stars and galaxies in the early universe but just enough that its effect can now be felt. Vacuum energy is completely inert, maintaining the same density for all time. Consequently, to explain the amount of dark energy today, the value of the cosmological constant would have to be fine-tuned at the creation of the universe to have the proper value—which makes it sound rather like a fudge factor. In contrast, quintessence interacts with matter and evolves with time, so it might naturally adjust itself to reach the observed value today.

Two Thirds of Reality

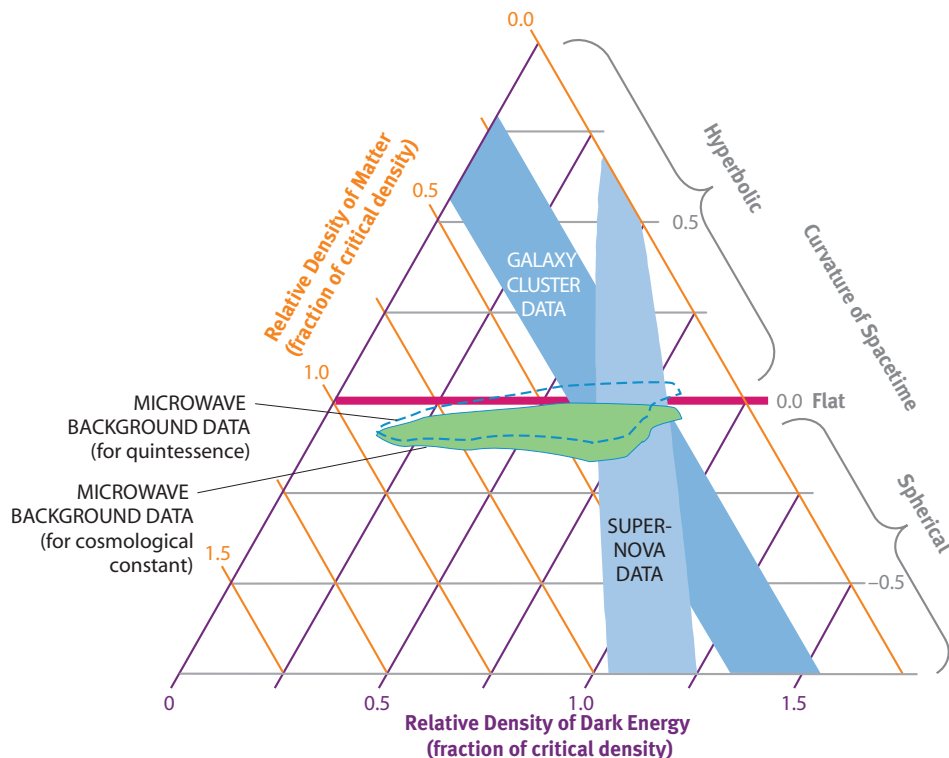
Distinguishing between these two options is critically important for physics. Particle physicists have depended on high-energy accelerators to discover new forms of energy and matter. Now the cosmos has revealed an unanticipated type of energy, too thinly spread and too weakly interacting for accelerators to probe. Whether the energy is inert or dynamical may be cru-



DON DIXON

MEET THE NEW BOSS

On scales where even galaxies are mere smidgens, a bizarre “dark energy” now appears to call the shots.



■ COSMIC TRIANGLE

In this graph of cosmological observations, the axes represent possible values of three key characteristics of the universe. If the universe is flat, as inflationary theory suggests, the different types of observations (colored areas) and the zero-curvature line (red line) should overlap. At present, the microwave background data produce a slightly better overlap if dark energy consists of quintessence (dashed outline) rather than the cosmological constant (green area).

cial to developing a fundamental theory of nature. Particle physicists are discovering that they must keep a close eye on developments in the heavens as well as in the accelerator laboratory.

The case for dark energy has been building brick by brick for nearly a decade. The first brick was a thorough census of all matter in galaxies and galaxy clusters using a variety of optical, x-ray and radio techniques. The unequivocal conclusion was that the total mass in chemical elements and dark matter accounts for only about one third of the quantity that most theorists expected—the so-called critical density.

Many cosmologists took this as a sign that the theorists were wrong. In that case, we would be living in an ever expanding universe where space is curved hyperbolically, like the horn on a trumpet [see “Inflation in a Low-Density Universe,” by Martin A. Bucher and David N. Spergel; *SCIENTIFIC AMERICAN*, January 1999]. But this interpretation has been put to rest by measurements of hot and cold spots in the cosmic microwave background radiation, whose distribution has shown that space is flat and that the total energy density equals the

critical density. Putting the two observations together, simple arithmetic dictates the necessity for an additional energy component to make up the missing two thirds of the energy density.

Whatever it is, the new component must be dark, neither absorbing nor emitting light, or else it would have been noticed long ago. In that way, it resembles dark matter. But the new component—called dark energy—differs from dark matter in one major respect: it must be gravitationally repulsive. Otherwise it would be pulled into galaxies and clusters, where it would affect the motion of visible matter. No such influence is seen. Moreover, gravitational repulsion resolves the “age crisis” that plagued cosmology in the 1990s. If one takes the current measurements of the expansion rate and assumes that the expansion has been decelerating, the age of the universe is less than 12 billion years.

Yet evidence suggests that some stars in our galaxy are 15 billion years old. By causing the expansion rate of the universe to accelerate, repulsion brings the inferred age of the cosmos into agreement with the observed age of celestial bodies [see “Cosmological Antigravity,”

by Lawrence M. Krauss; *SCIENTIFIC AMERICAN*, January 1999].

The potential flaw in the argument used to be that gravitational repulsion should cause the expansion to accelerate, which had not been observed. Then, in 1998, the last brick fell into place. Two independent groups used measurements of distant supernovae to detect a change in the expansion rate. Both groups concluded that the universe is accelerating and at just the pace predicted [see “Surveying Space-time with Supernovae,” by Craig J. Hogan, Robert P. Kirshner and Nicholas B. Suntzeff; *SCIENTIFIC AMERICAN*, January 1999].

All these observations boil down to three numbers: the average density of matter (both ordinary and dark), the average density of dark energy, and the curvature of space. Einstein’s equations dictate that the three numbers add up to the critical density. The different possible combinations of the numbers can be succinctly represented on a triangular plot [see illustration at left]. The three distinct sets of observations—matter census, cosmic microwave background, and supernovae—correspond to strips inside the triangle. Remarkably, the three strips overlap at the same position, which makes a compelling case for dark energy.

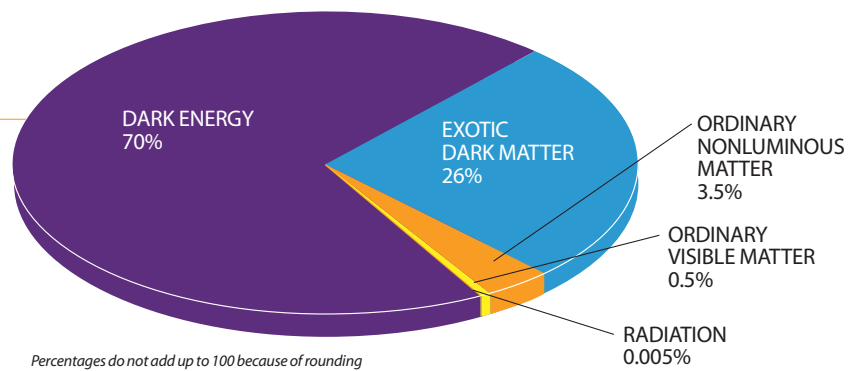
From Implosion to Explosion

Our everyday experience is with ordinary matter, which is gravitationally attractive, so it is difficult to envisage how dark energy could gravitationally repel. The key feature is that its pressure is negative. In Newton’s law of gravity, pressure plays no role; the strength of gravity depends only on mass. In Einstein’s law of gravity, however, the strength of gravity depends not just on mass but also on other forms of energy and on pressure. In this way, pressure has two effects: direct (caused by the action of the pressure on surrounding material) and indirect (caused by the gravitation that the pressure creates).

The sign of the gravitational force is determined by the algebraic combination of the total energy density plus three times the pressure. If the pressure is positive, as it is for radiation, ordinary matter and dark matter, then the combination is positive and gravitation is attractive. If the pressure is sufficiently negative, the combination is negative and gravitation is repulsive. To put it quantitatively, cosmologists consider the ratio of pressure to energy density, known as

RECIPE FOR THE UNIVERSE

The main ingredient of the universe is “dark energy,” which consists of either the cosmological constant or the quantum field known as quintessence. The other ingredients are dark matter composed of exotic elementary particles, ordinary matter (both nonluminous and visible), and a trace amount of radiation.



the equation of state, or w . For an ordinary gas, w is positive and proportional to the temperature. But for certain systems, w can be negative. If it drops below $-1/3$, gravity becomes repulsive.

Vacuum energy meets this condition (provided its density is positive). This is a consequence of the law of conservation of energy, according to which energy cannot be destroyed. Mathematically the law can be rephrased to state that the rate of change of energy density is proportional to $w + 1$. For vacuum energy—whose density, by definition, never changes—this sum must be zero. In other words, w must equal precisely -1 . So the pressure must be negative.

What does it mean to have negative pressure? Most hot gases have positive pressure; the kinetic energy of the atoms and radiation pushes outward on the container. Note that the direct effect of positive pressure—to push—is the opposite of its gravitational effect—to pull. But one can imagine an interaction among atoms that overcomes the kinetic energy and causes the gas to implode. The implosive gas has negative pressure. A balloon of this gas would collapse inward, because the outside pressure (zero or positive) would exceed the inside pressure (negative). Curiously, the direct effect of negative pressure—implosion—can be the opposite of its gravitational effect—repulsion.

Improbable Precision

The gravitational effect is tiny for a balloon. But now imagine filling all of space with the implosive gas. Then there is no bounding surface and no external pressure. The gas still has negative pressure, but it has nothing to push against, so it exerts no direct effect. It has only the gravitational effect—namely, repulsion. The repulsion stretches space, increasing its volume and, in turn, the amount of vacuum energy. The tendency to stretch is therefore self-reinforcing. The universe expands at an accelerating pace. The growing vacuum

energy comes at the expense of the gravitational field.

These concepts may sound strange, and even Einstein found them hard to swallow. He viewed the static universe, the original motivation for vacuum energy, as an unfortunate error that ought to be dismissed. But the cosmological constant, once introduced, would not fade away. Theorists soon realized that quantum fields possess a finite amount of vacuum energy, a manifestation of quantum fluctuations that conjure up pairs of “virtual” particles from scratch. An estimate of the total vacuum energy produced by all known fields predicts a huge amount—120 orders of magnitude more than the energy density in all other matter. That is, though it is hard to picture, the evanescent virtual particles should contribute a positive, constant energy density, which would imply negative pressure. But if this estimate were true, an acceleration of epic proportions would rip apart atoms, stars and galaxies. Clearly, the estimate is wrong. One of the major goals of unified theories of gravity has been to figure out why.

One proposal is that some heretofore undiscovered symmetry in fundamental physics results in a cancellation of large effects, zeroing out the vacuum energy. For example, quantum fluctuations of virtual pairs of particles contribute positive energy for particles with half-integer spin (like quarks and electrons) but negative energy for particles with integer spin (like photons). In standard theories, the cancellation is inexact, leaving behind an unacceptably large energy density. But physicists have been exploring models with so-called supersymmetry, a relation between the two particle types that can lead to a precise cancellation. A serious flaw, though, is that supersymmetry would be valid only at very high energies. Theorists are working on a way of preserving the perfect cancellation even at lower energies.

Another thought is that the vacuum energy is not exactly nullified after all. Perhaps there is a cancellation mecha-

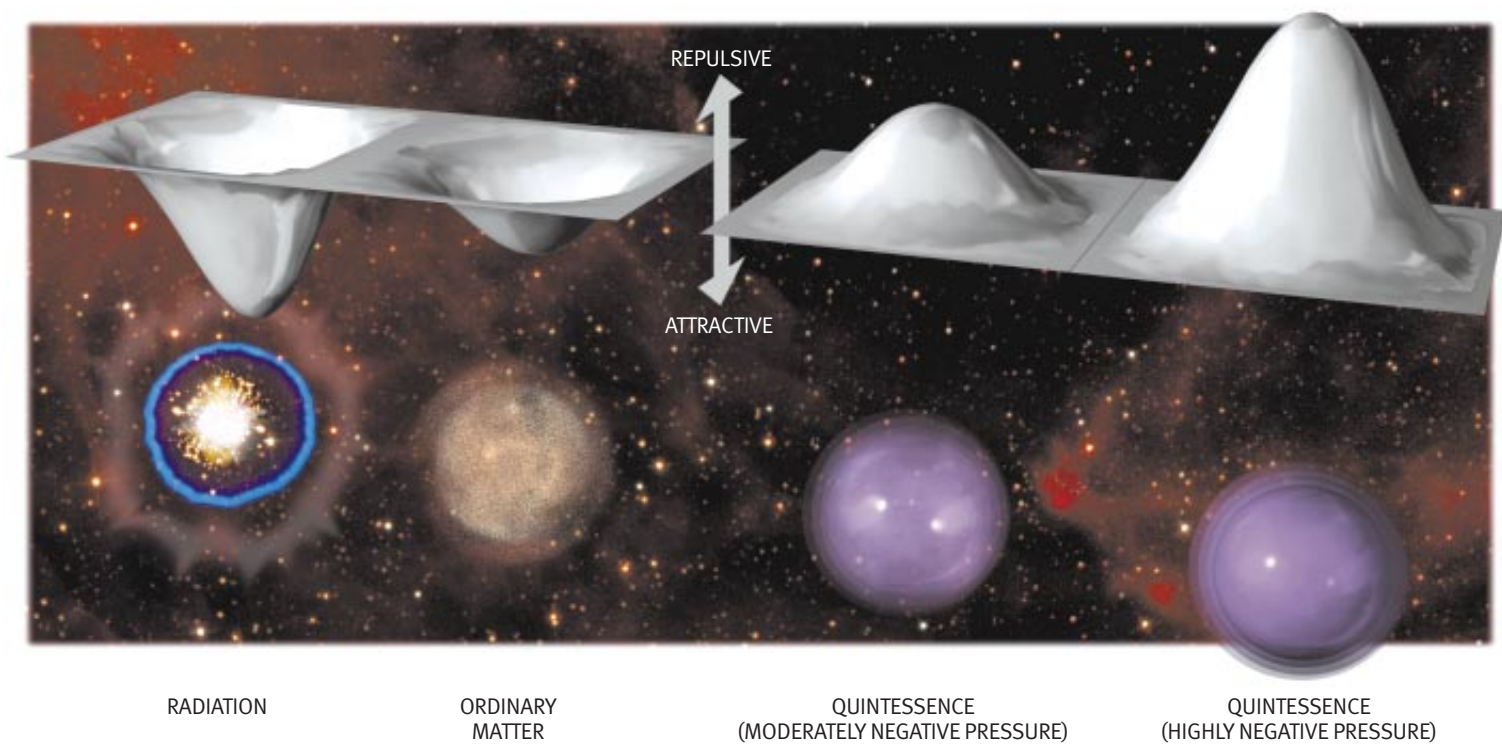
nism that is slightly imperfect. Instead of making the cosmological constant exactly zero, the mechanism only cancels to 120 decimal places. Then the vacuum energy could constitute the missing two thirds of the universe. That seems bizarre, though. What mechanism could possibly work with such precision? Although the dark energy represents a huge amount of mass, it is spread so thinly that its energy is less than four electron volts per cubic millimeter—which, to a particle physicist, is unimaginably low. The weakest known force in nature involves an energy density 10^{50} times greater.

Extrapolating back in time, vacuum energy gets even more paradoxical. Today matter and dark energy have comparable average densities. But billions of years ago, when they came into being, our universe was the size of a grapefruit, so matter was 100 orders of magnitude denser. The cosmological constant, however, would have had the same value as it does now. In other words, for every 10^{100} parts matter, physical processes would have created one part vacuum energy—a degree of exactitude that may be reasonable in a mathematical idealization but that seems ludicrous to expect from the real world. This need for almost supernatural fine-tuning is the principal motivation for considering alternatives to the cosmological constant.

Fieldwork

Fortunately, vacuum energy is not the only way to generate negative pressure. Another means is an energy source that, unlike vacuum energy, varies in space and time—a realm of possibilities that goes under the rubric of quintessence. For quintessence, w has no fixed value, but it must be less than $-1/3$ for gravity to be repulsive.

Quintessence may take many forms. The simplest models propose a quantum field whose energy is varying so slowly that it looks, at first glance, like a constant vacuum energy. The idea is bor-



RADIATION

ORDINARY
MATTER

QUINTESSENCE
(MODERATELY NEGATIVE PRESSURE)

QUINTESSENCE
(HIGHLY NEGATIVE PRESSURE)

■ THE POWER OF POSITIVE (AND NEGATIVE) THINKING

Whether a lump of energy exerts a gravitationally attractive or repulsive force depends on its pressure. If the pressure is zero or positive, as it is for radiation or ordinary matter, gravity is attractive. (The downward dimples represent the potential energy wells.) Radiation has greater pressure, so its gravity is more attractive. For quintessence, the pressure is negative and gravity is repulsive (the dimples become hills).

rowed from inflationary cosmology, in which a cosmic field known as the inflaton drives expansion in the very early universe using the same mechanism [see “The Inflationary Universe,” by Alan H. Guth and Paul J. Steinhardt; *SCIENTIFIC AMERICAN*, May 1984]. The key difference is that quintessence is much weaker than the inflaton. This hypothesis was first explored a decade ago by Christof Wetterich of the University of Heidelberg and by Bharat Ratra, now at Kansas State University, and P. James E. Peebles of Princeton University.

In quantum theory, physical processes can be described in terms either of fields or of particles. But because quintessence has such a low energy density and varies so gradually, a particle of quintessence would be inconceivably lightweight and large—the size of a supercluster of galaxies. So the field description is rather more useful. Conceptually, a field is a continuous distribution of energy that assigns to each point in space a numerical value known as the field strength. The energy embodied by the field has a kinetic component, which depends on the time variation of the field strength, and a potential component, which depends

only on the value of the field strength. As the field changes, the balance of kinetic and potential energy shifts.

In the case of vacuum energy, recall that the negative pressure was the direct result of the conservation of energy, which dictates that any variation in energy density is proportional to the sum of the energy density (a positive number) and the pressure. For vacuum energy, the change is zero, so the pressure must be negative. For quintessence, the change is gradual enough that the pressure must still be negative, though somewhat less so. This condition corresponds to having more potential energy than kinetic energy.

Because its pressure is less negative, quintessence does not accelerate the universe as strongly as vacuum energy does. Ultimately, this will be how observers decide between the two. If anything, quintessence is more consistent with the available data, but for now the distinction is not statistically significant. Another difference is that, unlike vacuum energy, the quintessence field may undergo all kinds of complex evolution. The value of w may be positive, then negative, then positive again. It may have different

values in different places. Although the nonuniformity is thought to be small, it may be detectable by studying the cosmic microwave background radiation.

A further difference is that quintessence can be perturbed. Waves will propagate through it just as sound waves can pass through the air. In the jargon, quintessence is “soft.” Einstein’s cosmological constant is, in contrast, stiff—it cannot be pushed around. This raises an interesting issue. Every known form of energy is soft to some degree. Perhaps stiffness is an idealization that cannot exist in reality, in which case the cosmological constant is an impossibility. Quintessence with w near -1 may be the closest reasonable approximation.

Quintessence on the Brane

Saying that quintessence is a field is just the first step in explaining it. Where would such a strange field come from? Particle physicists have explanations for phenomena from the structure of atoms to the origin of mass, but quintessence is something of an orphan. Modern theories of elementary particles include many kinds of fields that might have the requisite behavior, but not enough is known about their kinetic and potential energy to say which, if any, could produce negative pressure today.

An exotic possibility is that quintessence springs from the physics of extra dimensions. Over the past few decades, theorists have been exploring string the-

ory, which may combine general relativity and quantum mechanics in a unified theory of fundamental forces. An important feature of string models is that they predict 10 dimensions. Four of these are our familiar three spatial dimensions, plus time. The remaining six must be hidden. In some formulations, they are curled up like a ball whose radius is too small to be detected (at least with present instruments). An alternative idea is found in a recent extension of string theory, known as M-theory, which adds an 11th dimension: ordinary matter is confined to two three-dimensional surfaces, known as branes (short for membranes), separated by a microscopic gap along the 11th dimension [see “The Universe’s Unseen Dimensions,” by Nima Arkani-Hamed, Savas Dimopoulos and Georgi Dvali; SCIENTIFIC AMERICAN, August 2000].

We are unable to see the extra dimensions, but if they exist, we should be able to perceive them indirectly. In fact, the presence of curled-up dimensions or nearby branes would act just like a field. The numerical value that the field assigns to each point in space could correspond to the radius or gap distance. If the radius or gap changes slowly as the universe expands, it could exactly mimic the hypothetical quintessence field.

What a Coincidence

Whatever the origin of quintessence, its dynamism could solve the thorny problem of fine-tuning. One way to look at this issue is to ask, Why has cosmic acceleration begun at this particular moment in cosmic history? Created when the universe was 10^{-35} second old, dark energy must have remained in the shadows for nearly 10 billion years—a factor of more than 10^{50} in age. Only then, the data suggest, did it overtake matter and cause the universe to begin accelerating. Is it not a coincidence that, just when thinking beings evolved, the universe suddenly shifted into overdrive? Somehow the fates of matter and of dark energy seem to be intertwined. But how?

If the dark energy is vacuum energy, the coincidence is almost impossible to account for. Some researchers, including Martin Rees of the University of Cambridge and Steven Weinberg of the University of Texas at Austin, have pursued an anthropic explanation. Perhaps our universe is just one among a multitude of universes, in each of which the vacu-

um energy takes on a different value. Universes with vacuum energy much greater than four electron volts per cubic millimeter might be more common, but they expand too rapidly to form stars, planets or life. Universes with much smaller values might be very rare. Our universe would have the optimal value. Only in this “best of all worlds” could there exist intelligent beings capable of contemplating the nature of the universe. But physicists disagree whether the anthropic argument constitutes an acceptable explanation [see “Exploring Our Universe and Others,” by Martin Rees; SCIENTIFIC AMERICAN, December 1999].

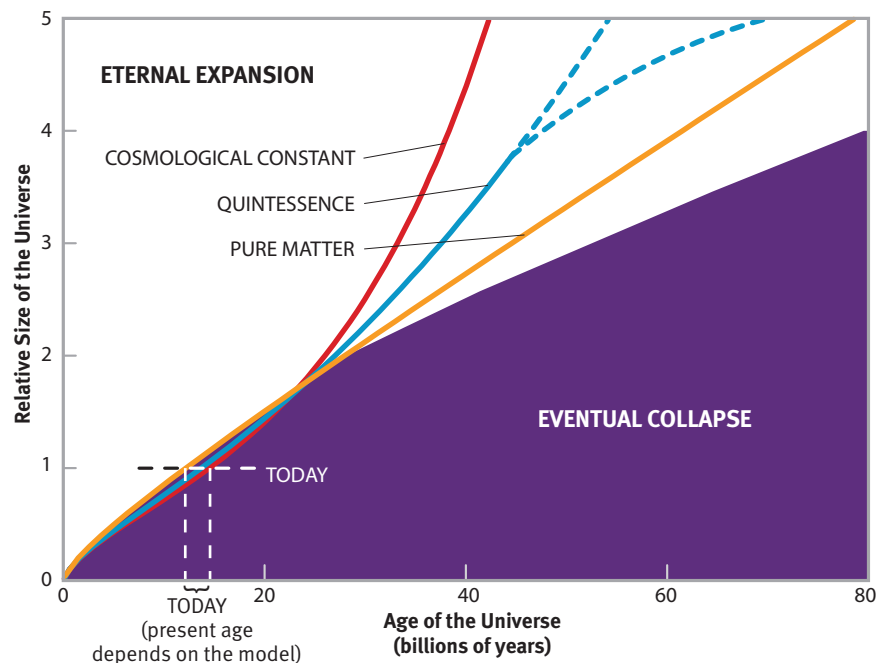
A more satisfying answer, which could involve a form of quintessence known as a tracker field, was studied by Ratra and Peebles and by Steinhardt, Ivaylo Zlatev and Limin Wang of the University of Pennsylvania. The equations that describe tracker fields have classical attractor behavior like that found in some chaotic systems. In such systems, motion converges to the same result for a wide range of initial conditions. A marble put into an empty bathtub, for example, ultimately falls into the drain whatever its starting place.

Similarly, the initial energy density of the tracker field does not have to be tuned to a certain value, because the field rapidly adjusts itself to that value. It locks into a track on which its energy density remains a nearly constant fraction of the density of radiation and matter. In this sense, quintessence imitates matter and radiation, even though its composition is wholly different. The mimicking occurs because the radiation and matter density determine the cosmic expansion rate, which, in turn, controls the rate at which the quintessence density changes. On closer inspection, one finds that the fraction is slowly growing. Only after many millions or billions of years does quintessence catch up.

So why did quintessence catch up when it did? Cosmic acceleration could just as easily have commenced in the distant past or in the far future, depending on the choices of constants in the tracker-field theory. This brings us back to the coincidence. But perhaps some event in the relatively recent past unleashed the acceleration. Steinhardt, along with Christian Armendáriz Picon and Viatcheslav Mukhanov of the Ludwig Maximilians University in Munich, has proposed one such recent event: the

■ GROWING PAINS

The universe expands at different rates depending on which form of energy predominates. Matter causes the growth to decelerate, whereas the cosmological constant causes it to accelerate. Quintessence is in the middle: it forces the expansion to accelerate, but less rapidly. Eventually the acceleration may or may not switch off (dashed lines).



JAMA BRENNING SOURCE: ROBERT R. CALDWELL, Dartmouth College AND PAUL J. STEINHARDT

transition from radiation domination to matter domination.

According to the big bang theory, the energy of the universe used to reside mainly in radiation. As the universe cooled, however, the radiation lost energy faster than ordinary matter did. By the time the universe was a few tens of thousands of years old—a relatively short time ago in logarithmic terms—the energy balance had shifted in favor of matter. This change marked the beginning of the matter-dominated epoch of which we are the beneficiaries. Only then could gravity begin to pull matter together to form galaxies and larger-scale structures. At the same time, the expansion rate of the universe underwent a change.

In a variation on the tracker models, this transformation triggered a series of events that led to cosmic acceleration today. Throughout most of the history of the universe, quintessence tracked the radiation energy, remaining an insignificant component of the cosmos. But when the universe became matter-dominated, the change in the expansion rate jolted quintessence out of its copycat behavior. Instead of tracking the radiation or even the matter, the pressure of quintessence switched to a negative

value. Its density held nearly fixed and ultimately overtook the decreasing matter density. In this picture, the fact that thinking beings and cosmic acceleration came into existence at nearly the same time is not a coincidence. Both the formation of stars and planets necessary to support life and the transformation of quintessence into a negative-pressure component were triggered by the onset of matter domination.

Looking to the Future

In the short term, the focus of cosmologists will be to detect the existence of quintessence. It has observable consequences. Because its value of w differs from that of vacuum energy, it produces a different rate of cosmic acceleration. More precise measurements of supernovae over a longer span of distances may separate the two cases. Astronomers have proposed two new observatories—the orbiting Supernova Acceleration Probe and the Earth-based Large-Aperture Synoptic Survey Telescope—to resolve the issue. Differences in acceleration rate also produce small differences in the angular size of hot and cold spots in the cosmic microwave background radiation, as the Microwave Anisotropy

Probe and Planck spacecraft should be able to detect.

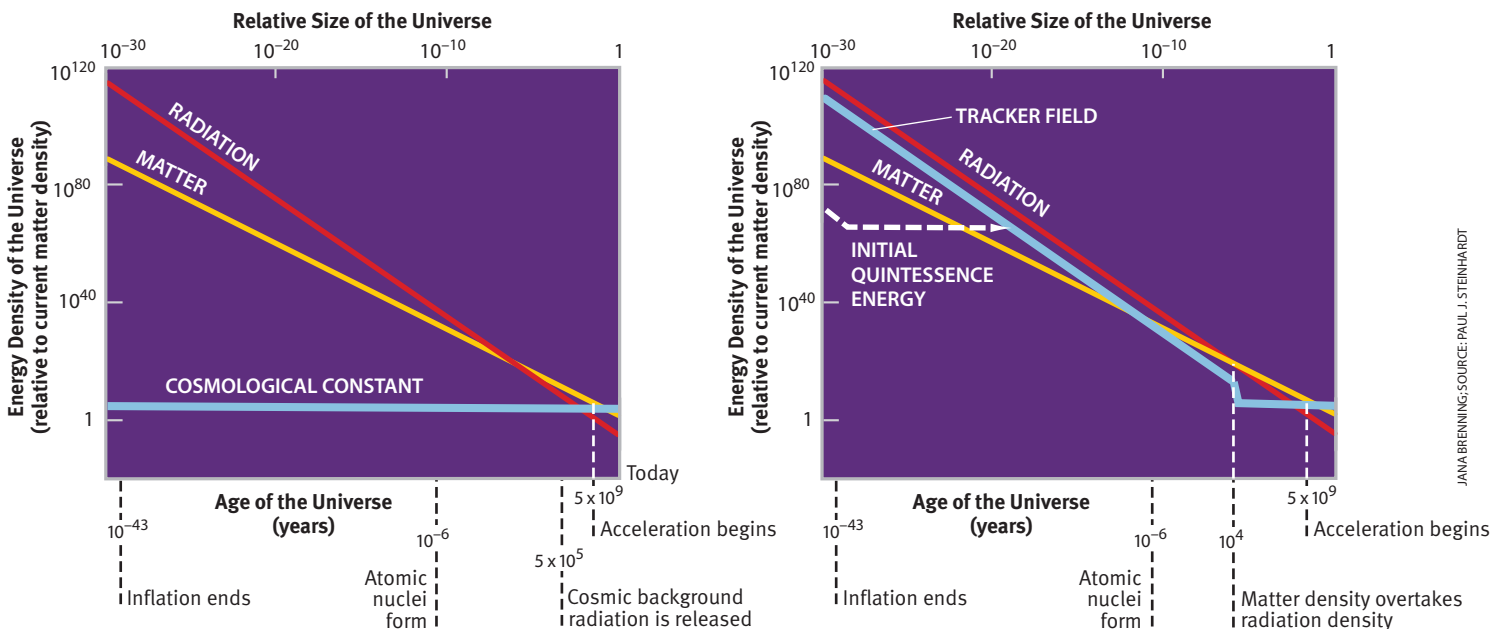
Other tests measure how the number of galaxies varies with increasing redshift to infer how the expansion rate of the universe has changed with time. A ground-based project known as the Deep Extragalactic Evolutionary Probe will look for this effect.

Over the longer term, all of us will be left to ponder the profound implications of these revolutionary discoveries. They lead to a sobering new interpretation of our place in cosmic history. In the beginning (or at least the earliest for which we have any clue), there was inflation, an extended period of accelerated expansion during the first instants after the big bang. Space back then was nearly devoid of matter, and a quintessencelike quantum field with negative pressure held sway. During that period, the universe expanded by a greater factor than it has during the 15 billion years since inflation ended. At the end of inflation, the field decayed to a hot gas of quarks, gluons, electrons, light and dark energy.

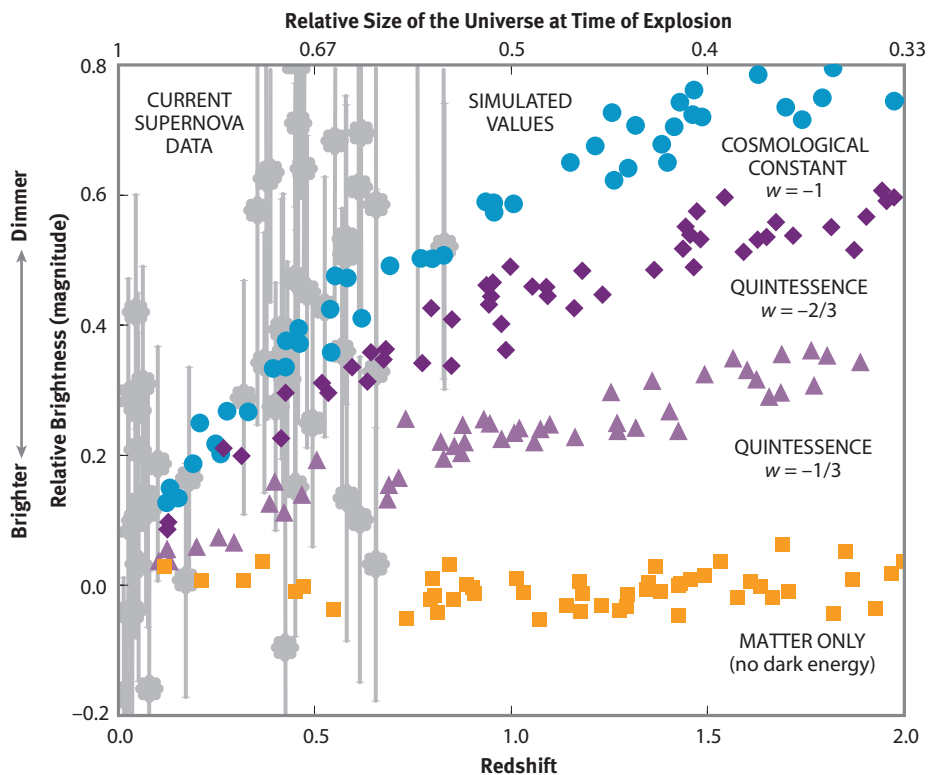
For thousands of years, space was so thick with radiation that atoms, let alone larger structures, could never form. Then matter took control. The next stage—our epoch—has been one of steady cooling, condensation and the evolution of intricate structure of ever increasing size. But this period is coming to an end. Cosmic acceleration is back. The universe as we know it, with shining stars, galaxies and clusters, appears to have been a brief interlude. As acceleration takes hold over the next tens of billions of years, the matter and

KEEPING TRACK

If dark energy consists of the cosmological constant, the energy density must be fine-tuned so that it overtakes the matter density in recent history (*left*). For the type of quintessence known as a tracker field (*right*), any initial density value (*dashed line*) converges to a common track (*blue line*) that runs in lockstep with the radiation density until the matter density overtakes it. This causes the tracker density to freeze and to trigger cosmic acceleration.



JANA BRENNING; SOURCE: PAUL J. STEINHARDT



SEEING WILL BE BELIEVING

Supernova data may be one way to decide between quintessence and the cosmological constant. The latter makes the universe speed up faster, so supernovae at a given redshift would be farther away and hence dimmer. Existing telescopes (*data shown in gray*) cannot tell the two cases apart, but the proposed Supernova Acceleration Probe should be able to. The supernova magnitudes predicted by four models are shown in different colors.

energy in the universe will become more and more diluted and space will stretch too rapidly to enable new structures to form. Living things will find the cosmos increasingly hostile [see “The Fate of Life in the Universe,” by Lawrence M. Krauss and Glenn Starkman;

SCIENTIFIC AMERICAN, November 1999]. If the acceleration is caused by vacuum energy, then the cosmic story is complete: the planets, stars and galaxies we see today are the pinnacle of cosmic evolution.

But if the acceleration is caused by

quintessence, the ending has yet to be written. The universe might accelerate forever, or the quintessence could decay into new forms of matter and radiation, repopulating the universe. Because the dark-energy density is so small, one might suppose that the material derived from its decay would have too little energy to do anything of interest. Under some circumstances, however, quintessence could decay through the nucleation of bubbles. The bubble interior would be a void, but the bubble wall would be the site of vigorous activity. As the wall moved outward, it would sweep up all the energy derived from the decay of quintessence. Occasionally, two bubbles would collide in a fantastic fireworks display. In the process, massive particles such as protons and neutrons might arise—perhaps stars and planets.

To future inhabitants, the universe would look highly inhomogeneous, with life confined to distant islands surrounded by vast voids. Would they ever figure out that their origin was the homogeneous and isotropic universe we see about us today? Would they ever know that the universe had once been alive and then died, only to be given a second chance?

Experiments may soon give us some idea which future is ours. Will it be the dead end of vacuum energy or the untapped potential of quintessence? Ultimately the answer depends on whether quintessence has a place in the basic workings of nature—the realm, perhaps, of string theory. Our place in cosmic history hinges on the interplay between the science of the very big and that of the very small.

THE AUTHORS

JEREMIAH P. OSTRIKER and **PAUL J. STEINHARDT**, both professors at Princeton University, have been collaborating for the past six years. Their prediction of accelerating expansion in 1995 anticipated the groundbreaking supernova results by several years. Ostriker was one of the first to appreciate the prevalence of dark matter and the importance of hot intergalactic gas. In 2000 he won the U.S. National Medal of Science. Steinhardt was one of the originators of the theory of inflation and the concept of quasicrystals. He reintroduced the term “quintessence” after his youngest son Will and daughter Cindy picked it out from several alternatives.

FURTHER INFORMATION

THE OBSERVATIONAL CASE FOR A LOW-DENSITY UNIVERSE WITH A NON-ZERO COSMOLOGICAL CONSTANT. Jeremiah P. Ostriker and Paul J. Steinhardt in *Nature*, Vol. 377, pages 600–602; October 19, 1995. Preprint at xxx.lanl.gov/abs/astro-ph/9505066

COSMOLOGICAL IMPRINT OF AN ENERGY COMPONENT WITH GENERAL EQUATION OF STATE. Robert R. Caldwell, Rahul Dave and Paul J. Steinhardt in *Physical Review Letters*, Vol. 80, No. 8, pages 1582–1585; February 23, 1998; astro-ph/9708069

COSMIC CONCORDANCE AND QUINTESSENCE. Limin Wang, R. R. Caldwell, J. P. Ostriker and Paul J. Steinhardt in *Astrophysical Journal*, Vol. 530, No. 1, Part 1, pages 17–35; February 10, 2000; astro-ph/9901388

DYNAMICAL SOLUTION TO THE PROBLEM OF A SMALL COSMOLOGICAL CONSTANT AND LATE-TIME COSMIC ACCELERATION. C. Armendáriz Picon, V. Mukhanov and Paul J. Steinhardt in *Physical Review Letters*, Vol. 85, No. 21, pages 4438–4441; November 20, 2000; astro-ph/0004314

WHY COSMOLOGISTS BELIEVE THE UNIVERSE IS ACCELERATING. Michael S. Turner in *Type Ia Supernovae: Theory and Cosmology*. Edited by Jens C. Niemeyer and James W. Truran. Cambridge University Press, 2000; astro-ph/9904049



Confused by all those theories? Good

Making Sense of Modern Cosmology

by P. James E. Peebles

This is an exciting time for cosmologists: findings are pouring in, ideas are bubbling up, and research to test those ideas is simmering away. But it is also a confusing time. All the ideas under discussion cannot possibly be right; they are not even consistent with one another. How is one to judge the progress? Here is how I go about it.

For all the talk of overturned theories, cosmologists have firmly established the foundations of our field. Over the past 70 years we have gathered abundant evidence that our universe is expanding and cooling. First, the light from distant galaxies is shifted toward the red, as it should be if space is expanding and galaxies are pulled away from one another. Second, a sea of thermal radiation fills space, as it should if space used to be denser and hotter. Third, the universe contains large amounts of deuterium and helium, as it should if temperatures were once much higher. Fourth, galaxies billions of years ago look distinctly younger, as they should if they are closer to the time when no galaxies existed. Finally, the curvature of spacetime seems to be related to the material content of the universe, as it should be if the universe is expanding according to the predictions of Einstein's gravity theory, the general theory of relativity.

That the universe is expanding and cooling is the essence of the big bang theory. You will notice I have said nothing about an "explosion"—the big bang theory describes how our universe is evolving, not how it began.


I compare the process of establishing such compelling results, in cosmology or any other science, to the assembly of a framework. We seek to reinforce each piece of evidence by adding cross bracing from diverse measurements. Our framework for the expansion of the universe is braced tightly enough to be solid. The big bang theory is no longer seriously questioned; it fits together too well. Even the most radical alternative—the latest incarnation of the steady state theory—does not dispute that the universe is expanding and cooling. You still hear differences of opinion in cosmology,

to be sure, but they concern additions to the solid part.

For example, we do not know what the universe was doing before it was expanding. A leading theory, inflation, is an attractive addition to the framework, but it lacks cross bracing. That is precisely what cosmologists are now seeking [see "Echoes from the Big Bang," on page 38]. If measurements in progress agree with the unique signatures of inflation, then we will count them as a persuasive argument for this theory. But until that time, I would not settle any bets on whether inflation really happened. I am not criticizing the theory; I simply mean that this is brave, pioneering work still to be tested.

More solid is the evidence that most of the mass of the universe consists of dark matter clumped around the outer parts of galaxies. We also have a reasonable case for Einstein's infamous cosmological constant or something similar; it would be the agent of the acceleration that the universe now seems to be undergoing. A decade ago cosmologists generally welcomed dark matter as an elegant way to account for the motions of stars and gas within galaxies. Most researchers, however, had a real distaste for the cosmological constant. Now the majority accept it, or its allied concept, quintessence [see "The Quintessential Universe," on page 46]. Particle physicists have come to welcome the challenge that the cosmological constant poses for quantum theory. This shift in opinion is not a reflection of some inherent weakness; rather it shows the subject in a healthy state of chaos around a slowly growing fixed framework. We are students of nature, and we adjust our concepts as the lessons continue.

The lessons, in this case, include the signs that cosmic expansion is accelerating: the brightness of supernovae near and far; the ages of the oldest stars; the bending of light around distant masses; and the fluctuations of the temperature of the thermal radiation across the sky [see "Special Report: Revolution in Cosmology," *SCIENTIFIC AMERICAN*, January 1999]. The evidence is impressive, but I am still uneasy about details



Our framework for the big bang theory is braced tightly enough to be solid.


REPORT CARD FOR MAJOR THEORIES

Concept	Grade	Comments
The universe evolved from a hotter, denser state	A+	Compelling evidence drawn from many corners of astronomy and physics
The universe expands as the general theory of relativity predicts	A-	Passes the tests so far, but few of the tests have been tight
Dark matter made of exotic particles dominates galaxies	B+	Many lines of indirect evidence, but the particles have yet to be found and alternative theories have yet to be ruled out
Most of the mass of the universe is smoothly distributed; it acts like Einstein's cosmological constant, causing the expansion to accelerate	B-	Encouraging fit from recent measurements, but more must be done to improve the evidence and resolve the theoretical conundrums
The universe grew out of inflation	Inc	Elegant, but lacks direct evidence and requires huge extrapolation of the laws of physics

ROBERT GENDLER

of the case for the cosmological constant, including possible contradictions with the evolution of galaxies and their spatial distribution. The theory of the accelerating universe is a work in progress. I admire the architecture, but I would not want to move in just yet.

How might one judge reports in the media on the progress of cosmology? I feel uneasy about articles based on an interview with just one person. Research is a complex and messy business. Even the most experienced scientist finds it hard to keep everything in perspective. How do I know that this individual has managed it well? An entire community of scientists can head off in the wrong direction, too, but it happens less often. That is why I feel better when I can see that the journalist has consulted a cross section of the community and has found agreement that a certain result is worth considering. The result becomes more interesting when others reproduce it. It starts to become convincing when independent lines of evidence point to the same conclusion. To my mind, the best media reports on science describe not only the latest discoveries and ideas but also the essential, if sometimes tedious, process of testing and installing the cross bracing.

Over time, inflation, quintessence and other concepts now under debate either will be solidly integrated into the central framework or will be abandoned and replaced by something better. In a sense, we are working ourselves out of a job. But the universe is a complicated place, to put it mildly, and it is silly to think we will run out of productive lines of research anytime soon. Confusion is a sign that we are doing something right: it is the fertile commotion of a construction site. 

P. JAMES E. PEEBLES is one of the world's most distinguished cosmologists, a key player in the early analysis of the cosmic microwave background radiation and the bulk composition of the universe. He has received some of the highest awards in astronomy, including the 1982 Heineman Prize, the 1993 Henry Norris Russell Lectureship of the American Astronomical Society and the 1995 Bruce Medal of the Astronomical Society of the Pacific. Peebles is currently an emeritus professor at Princeton University.

FURTHER INFORMATION

THE EVOLUTION OF THE UNIVERSE. P. James E. Peebles, David N. Schramm, Edwin L. Turner and Richard G. Kron in *Scientific American*, Vol. 271, No. 4, pages 52–57; October 1994.

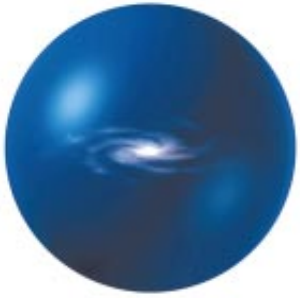
THE INFLATIONARY UNIVERSE: THE QUEST FOR A NEW THEORY OF COSMIC ORIGINS. Alan H. Guth. Perseus Press, 1997.

BEFORE THE BEGINNING: OUR UNIVERSE AND OTHERS. Martin Rees. Perseus Press, 1998.

THE ACCELERATING UNIVERSE: INFINITE EXPANSION, THE COSMOLOGICAL CONSTANT, AND THE BEAUTY OF THE COSMOS. Mario Livio and Allan Sandage. John Wiley & Sons, 2000.

CONCLUDING REMARKS ON NEW COSMOLOGICAL DATA AND THE VALUES OF THE FUNDAMENTAL PARAMETERS. P. James E. Peebles in *IAU Symposium 201: New Cosmological Data and the Values of the Fundamental Parameters*, edited by A. N. Lasenby, A. W. Jones and A. Wilkinson; August 2000. Preprint available at xxx.lanl.gov/abs/astro-ph/0011252 on the World Wide Web.

If the new cosmology fails, what's the backup plan?



Plan B for the Cosmos

by João Magueijo

Although cosmic inflation has acquired an aura of invincibility, alternative theories continue to attract some interest among cosmologists. The steady state theory, which until the 1960s was widely regarded as the main alternative to the big bang, has been kept alive by a small band of proponents. The pre-big bang theory, a reworking of inflation that has been motivated by string theory, also turns some heads. But the most promising and provocative alternative may be the varying-speed-of-light theory (VSL), which my colleagues and I have been developing for several years. If nothing else, these dissenting views add color and variety to cosmology. They also give expression to a nagging doubt: Could the enthusiasm generated by inflation and its offshoots conceal a monstrous error?

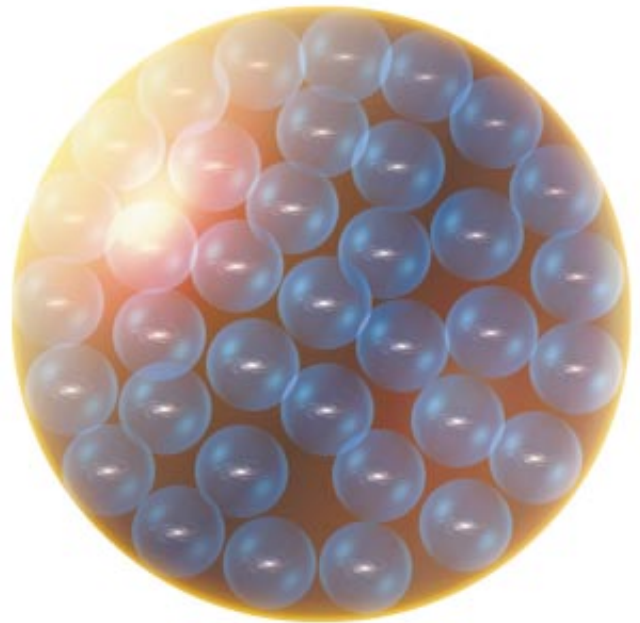
Mainstream cosmological theories such as inflation are based on a crucial assumption: that the speed of light and other fundamental physical parameters have had the same values for all time. (They are, after all, known as constants.) This assumption has forced cosmologists to adopt inflation and all its fantastic implications. And sure enough, experiments show that the presumed constants are not aging dramatically. Yet researchers have probed their values only over the past billion years or so. Postulating their constancy over the entire life of the universe involves a massive extrapolation. Could the presumed constants actually change over time in a big bang universe, as do its temperature and density?

Theorists find that some constants are more agreeable than others to giving up their status. For instance, the gravitational constant, G , and the electron's charge, e , have often been subjected to this theoretical ordeal, causing little scandal or uproar. Indeed, from Paul Dirac's groundbreaking work on varying constants in the 1930s to the latest string theories, dethroning the constancy of G has been exquisitely fashionable. In contrast, the speed of light, c , has remained inviolate. The reason is clear: the constancy of c and its status as a universal speed limit are the foundations of the theory of relativity. And relativity's spell is so strong that the constancy of c is now woven into all the mathematical tools available to the physicist. "Varying c " is not even a swear word; it is simply not present in the vocabulary of physics.

Yet it might behoove cosmologists to expand their vernacular. At the heart of inflation is the so-called horizon problem of big bang cosmology, which stems from a simple fact: at any given time, light—and hence any interaction—can have traveled only a finite distance since the big bang. When the universe was one year old, for example, light

could have traveled just one light-year (roughly). The universe is therefore fragmented into horizons, which demarcate regions that cannot yet see one another.

The shortsightedness of the universe is enormously irritating to cosmologists. It precludes physical explanations—that is, ones based on physical interactions—for puzzles such as why the early universe was so uniform. Within the framework of the standard big bang theory, the uniformity can be explained only by fine-tuning the initial conditions—essentially a recourse to metaphysics.



■ TROUBLE ON THE HORIZON

At the strapping age of one year, the universe was subdivided into isolated pockets, demarcated by "horizons" one light-year in radius (*blue spheres*). Today the horizon is about 15 billion light-years in radius (*red sphere*), so it takes in zillions of these pockets. The odd thing is that despite their initial isolation, all the pockets look pretty much the same. Explaining this mysterious uniformity is the great success of the theory of inflation.

Inflation cunningly gets around this problem. Its key insight is that for a light wave in an expanding universe, the distance from the starting point is greater than the distance traveled. The reason is that expansion keeps stretching the space already covered. By analogy, consider a driver who travels at 60 kilometers an hour for one hour. The driver has covered 60 kilometers, but if the road itself has elongated in the meantime, the distance from the point of departure is greater than 60 kilometers. Inflationary theory postulates that the early universe expanded so fast that the range of light was phenomenally large. Seemingly disjointed regions could thus have communicated with one another and reached a common temperature and density. When the inflationary expansion ended, these regions began to fall out of touch.

It does not take much thought to realize that the same thing could have been achieved if light simply had traveled faster in the early universe than it does today. Fast light could have stitched together a patchwork of otherwise disconnected regions. These regions could then have homogenized themselves. As the speed of light slowed, those regions would have fallen out of contact.

This was the initial insight that led Andreas Albrecht of the University of California at Davis, John Barrow of the University of Cambridge and me to propose the VSL theory. Contrary to popular belief, our motivation was not to annoy the proponents of inflation. (Indeed, Albrecht is one of the fathers of inflationary theory.) We felt that the successes and

shortcomings of inflation would become clearer if an alternative existed, no matter how crude.

Naturally, VSL requires rethinking the foundations and language of physics, and for this reason many different implementations are possible. What we first proposed was a reckless act of extreme violence against relativity, albeit with the redeeming merit of solving many puzzles besides the flatness problem. For example, our theory accounts for the minuscule yet nonzero value of the cosmological constant in today's universe. The reason is that the vacuum-energy density represented by the cosmological constant depends very strongly on c . A suitable drop in c reduces the otherwise domineering vacuum energy to innocuous levels. In standard theories, on the other hand, the vacuum energy cannot be diluted.

But our formulation is just one possibility, and the urge to reconcile VSL to relativity is motivating much ongoing work. The more cautious implementations of VSL pioneered by John Moffat of the University of Toronto and later by Ian T. Drummond of Cambridge are easier for relativity theorists to swallow. It now appears that the constancy of c is not so essential to relativity after all; the theory can be based on other postulates. Some have pointed out that if the universe is a three-dimensional membrane in a higher-dimensional space, as string theory suggests, the apparent speed of light in our world could vary while the truly fundamental c remains constant.

Whether nature chose to inflate or to monkey with c can only be decided by experiment. The VSL theory is currently far less developed than inflation, so it has yet to make firm predictions for the cosmic microwave background radiation. On the other hand, some experiments have indicated that the so-called fine structure constant may not be constant [see "Inconstant Constants," by George Musser; News and Analysis, *SCIENTIFIC AMERICAN*, November 1998]. Varying c would explain those findings.

It remains to be seen whether these observations will withstand further scrutiny; meanwhile VSL remains a major theoretical challenge. It distinguishes itself from inflation by plunging deeper into the roots of physics. For now, VSL is far from being mainstream. It is a foray into the wild. SA

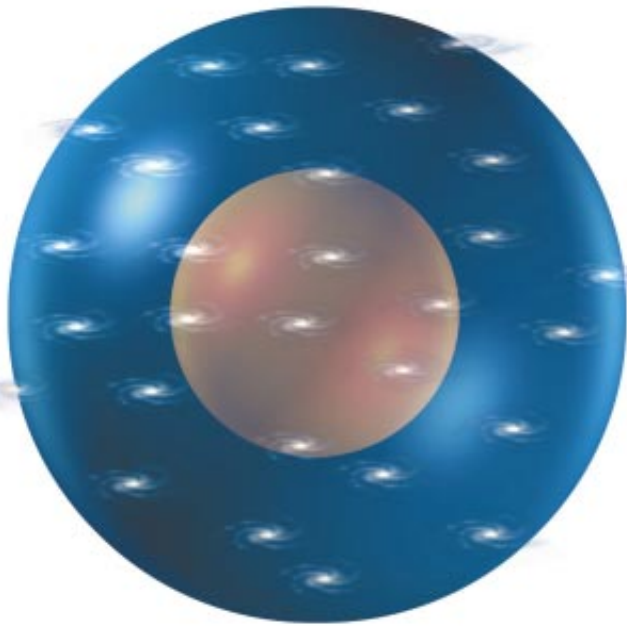
JOÃO MAGUEIJO is a lecturer in theoretical physics at Imperial College, London. His research interests have oscillated between the observational and the "lunatic" (his word) aspects of cosmology, from analysis of the cosmic microwave background radiation to topological defects, quintessence and black holes.

FURTHER INFORMATION

TIME VARYING SPEED OF LIGHT AS A SOLUTION TO COSMOLOGICAL PUZZLES. Andreas Albrecht and João Magueijo in *Physical Review D*, Vol. 59, No. 4, Paper No. 043516; February 15, 1999. Preprint available at xxx.lanl.gov/abs/astro-ph/9811018 on the World Wide Web.

BIG BANG RIDDLES AND THEIR REVELATIONS. João Magueijo and Kim Baskerville in *Philosophical Transactions of the Royal Society A*, Vol. 357, No. 1763, pages 3221–3236; December 15, 1999; xxx.lanl.gov/abs/astro-ph/9905393

COVARIANT AND LOCALLY LORENTZ-INVARIANT VARYING SPEED OF LIGHT THEORIES. João Magueijo in *Physical Review D*, Vol. 62, No. 10, Paper No. 103521; November 15, 2000; xxx.soton.ac.uk/abs/gr-qc/0007036



■ BROADENING THE HORIZON

Inflation is not the only answer to the horizon problem. Instead maybe conditions in the early universe allowed light to travel faster than its present speed—a billion times faster or more. Zippy light made for bigger pockets (*blue sphere*). As light slowed to its present speed, the horizon shrank (*red sphere*). In that way, we are now able to see just a part of one of the initial pockets, so it is no longer a mystery why the universe looks so uniform.





the Cultures of Chimpanzees

Humankind's nearest relative is even closer than we thought: chimpanzees display remarkable behaviors that can only be described as social customs passed on from generation to generation

by Andrew Whiten and Christophe Boesch

GOING FISHING for ants is a handy way to find dinner for some chimpanzees. This chimpanzee, from Mahale National Park in Tanzania, inserts a stick into an ant nest located within a tree; once the ants climb up the stick, the chimpanzee removes the stick and picks the ants off with its lips.

GUNTER ZIESLER/Peter Arnold, Inc.

As researchers quietly approach a clearing in the Tai Forest of Ivory Coast, they hear a complex pattern of soft thuds and cracks. It sounds as though a small band of people are busy in the forest, applying some rudimentary technology to a routine task. On entering the clearing, the scientists observe several individuals working keenly at anvils, skillfully wielding wooden hammers. One or two juveniles have apprenticed themselves to the work and—more clumsily and with less success—are struggling to lift the best hammer they can find. All this activity is directed toward cracking rock-hard but nutritious coula nuts. Intermittently, individuals set aside their tools to gather more handfuls of nuts. An infant sits with her mother, gathering morsels of broken nuts.

In many ways, this group could indeed be a family of foraging people. The hammers and anvils they leave behind, some made of stone, would excite the imagi-

The Culture Club

How an international team of chimpanzee experts conducted the most comprehensive survey of the animals ever attempted

Scientists have been investigating chimpanzee culture for several decades, but too often their studies contained a crucial flaw. Most attempts to document cultural diversity among chimpanzees have relied solely on officially published accounts of the behaviors recorded at each research site. But this approach probably overlooks a good deal of cultural variation for three reasons.

First, scientists typically don't publish an extensive list of all the activities they do *not* see at a particular location. Yet this is exactly what we need to know—which behaviors were and were not observed at each site. Second, many reports describe chimpanzee behaviors without saying how common they are; without this information, we can't determine whether a particular action was a once-in-a-lifetime aberration or a routine event that should be considered part of the animals' culture. Finally, researchers' descriptions of potentially significant chimpanzee behaviors frequently lack sufficient detail, making it difficult for scientists working at other spots to record the presence or absence of the activities.

To remedy these problems, the two of us decided to take a new approach. We asked field researchers at each site for a list of all the behaviors they suspected were local traditions. With this information in hand, we pulled together a comprehensive list of 65 candidates for cultural behaviors.

Then we distributed our list to the team leaders at each site.

In consultation with their colleagues, they classified each behavior in terms of its occurrence or absence in the chimpanzee community studied. The key categories were customary behavior (occurs in most or all of the able-bodied members of at least one age or sex class, such as all adult males), habitual (less common than customary but occurs repeatedly in several individuals), present (seen at the site but not habitual), absent (never seen), and unknown.

Our inquiry concentrated on seven sites with chimpanzees habituated to human onlookers; all told, the study compiled a total of more than 150 years of chimpanzee observation. The behavior patterns we were particularly interested in, of course, were those absent in at least one community, yet habitual or customary in at least one other; this was our criterion for denoting any behavior a cultural variant. (Certain behaviors are absent for specific local reasons, however, and we excluded them from consideration. For example, although chimpanzees at Bossou scoop tasty algae from pools of water with a stick, chimpanzees elsewhere don't do this, simply because algae are not present.)

The extensive survey turned up no fewer than 39 chimpanzee patterns of behavior that should be labeled as cultural variations, including numerous forms of tool use, grooming techniques and courtship gambits, several of which are illustrated throughout this article. This cultural richness is far in excess of anything known for any other species of animal. —A.W. and C.B.

nation of any anthropologist searching for signs of a primitive civilization. Yet these forest residents are not humans but chimpanzees.

The similarities between chimpanzees and humans have been studied for years, but in the past decade researchers have determined that these resemblances run much deeper than anyone first thought. For instance, the nut cracking observed in the Taï Forest is far from a simple chimpanzee behavior; rather it is

a singular adaptation found only in that particular part of Africa and a trait that biologists consider to be an expression of chimpanzee culture. Scientists frequently use the term "culture" to describe elementary animal behaviors—such as the regional dialects of different populations of songbirds—but as it turns out, the rich and varied cultural traditions found among chimpanzees are second in complexity only to human traditions.

During the past two years, an unprecedented scientific collaboration, involving every major research group studying chimpanzees, has documented a multitude of distinct cultural patterns extending across Africa, in actions ranging from the animals' use of tools to their forms of communication and social customs. This emerging picture of chimpanzees not only affects how we think of these amazing creatures but also alters human beings' conception of



MICHAEL NICHOLS/National Geographic Image Collection

Today's Lesson includes a demonstration of how to crack open a coula nut. A mother chimpanzee in the Taï Forest of Ivory Coast uses a stone hammer to cleave a nut while a youngster watches. Not all chimpanzees in this area have developed this behavior: on the eastern bank of the Sassandra-N'Zo River, chimpanzees do not crack nuts even though members of the same species on the other side of the river, just a few miles away, do. All the required raw materials are available on both sides, and the nuts could be cracked using the technique habitual at Taï. The river serves as a literal cultural barrier.

our own uniqueness and hints at very ancient foundations for humankind's extraordinary capacity for culture.

Contemplating Culture

H*omo sapiens* and *Pan troglodytes* have coexisted for hundreds of millennia and share more than 98 percent of their genetic material, yet only 40 years ago we still knew next to nothing about chimpanzee behavior in the wild. That began to change in the 1960s, when Toshisada Nishida of Kyoto University in Japan and Jane Goodall began their studies of wild chimpanzees at two field sites in Tanzania. (Goodall's research station at Gombe—the first of its kind—is more famous, but Nishida's site at Mahale is the second-oldest chimpanzee research site in the world.)

In these initial studies, as the chimpanzees became accustomed to close observation, the remarkable discoveries began. Researchers witnessed a range of unexpected behaviors, including fashioning and using tools, hunting, meat eating, food sharing and lethal fights between members of neighboring communities. In the years that followed, other primatologists set up camp elsewhere, and, despite all the financial, political and logistical problems that can beset African fieldwork, several of these outposts became truly long-term projects. As a result, we live in an unprecedented time, when an intimate and comprehensive scientific record of chimpanzees' lives at last exists not just for one but for several communities spread across Africa.

As early as 1973, Goodall recorded 13 forms of tool use as well as eight social activities that appeared to differ between the Gombe chimpanzees and chimpanzee populations elsewhere. She ventured that some variations had what she termed a cultural origin. But what exactly did Goodall mean by "culture"? According to the *Oxford Encyclopedic English Dictionary*, culture is defined as "the customs . . . and achievements of a particular time or people." The diversity of human cultures extends from technological variations to marriage rituals, from culinary habits to myths and legends. Animals do not have myths and legends, of course. But

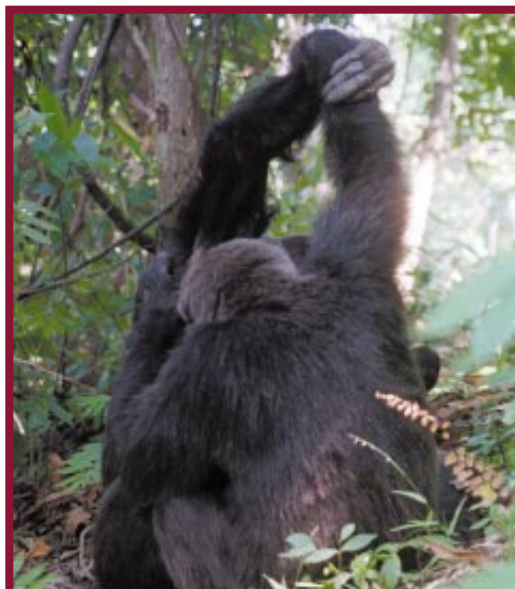
they do have the capacity to pass on behavioral traits from generation to generation, not through their genes but by learning. For biologists, this is the fundamental criterion for a cultural trait: it must be something that can be learned by observing the established skills of others and thus passed on to future generations [see box on page 66].

By the 1990s the discovery of new behavioral differences among chimpanzees made it feasible to begin assembling comprehensive charts of cultural variations for these animals. William C. McGrew, in his 1992 book *Chimpanzee Material Cultures*, was able to list 19 different kinds of tool use in distinct communities. One of us (Boesch), along with colleague Michael Tomasello of

any other animal studied to date. Of course, chimpanzees also remain distinct from humans, for whom cultural variations are simply beyond count. (We must point out, however, that scientists are only beginning to uncover the behavioral complexity that exists among chimpanzees—and so the number 39 no doubt represents a minimum of cultural traits.)

Multicultural Chimpanzees

When describing human customs, anthropologists and sociologists often refer to "American culture" or "Chinese culture"; these terms encompass a wide spectrum of activities—language, forms of dress, eating habits,



High Five during grooming is commonplace among chimpanzees at Tai Forest, Mahale and Kibale. Here two male chimpanzees at Mahale groom each other while clasp hands. Recent research by William C. McGrew and Linda F. Marchant, both at Miami University, suggests that the two adjacent communities at Mahale display subtle differences in how they clasp hands, with one community avoiding palm-to-palm contact, the style common among their neighbors. In 40 years of observations at Gombe, hand clasping has never been seen; chimpanzees sometimes grasp a branch overhead, but they do not hold their grooming partner's hand.

the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, identified 25 distinct activities as potential cultural traits in wild chimpanzee populations.

The most recent catalogue of cultural variations results from a unique collaboration of nine chimpanzee experts (including the two of us) who pooled extensive field observations that, taken together, amounted to a total of 151 years of chimp watching [see box on opposite page]. The list cites 39 patterns of chimpanzee behavior that we believe to have a cultural origin, including such activities as using sticks to "fish" for ants, making dry seats from leaves, and a range of social grooming habits. At present, these 39 variants put chimpanzees in a class of their own, with far more elaborate customs than

marriage rituals and so on. Among animals, however, culture has typically been established for a single behavior, such as song dialects among birds. Ornithologists haven't identified variation in courtship patterns or feeding practices, for example, to go alongside the differences in dialect.

Chimpanzees, though, do more than display singular cultural traits: each community exhibits an entire set of behaviors that differentiates it from other groups [see illustrations on pages 64 and 65]. As a result, we can talk about "Gombe culture" or "Tai culture." Indeed, once we observe how a chimpanzee behaves, we can identify where the animal lives. For instance, an individual that cracks nuts, leaf-clips during drumming displays, fishes for ants with

Continued on page 66

A Guide to the Cultures of Chimpanzees

In an effort to catalogue cultural variations among chimpanzees, we asked researchers working at six sites across central Africa to classify chimpanzee behaviors in terms of occurrence or absence in seven communities. (There are two communities at Mahale.) The key categories were customary behavior, which occurs in most or all members of one age or sex class; habitual, which is less

common but which still occurs repeatedly; present; absent; and unknown. Certain behaviors are absent for ecological reasons (eco): for example, chimpanzees do not use hammers to open coula nuts at Budongo, because the nuts are not available. The survey turned up 39 chimpanzee rituals that are labeled as cultural variations; 18 are illustrated below. —A.W. and C.B.

Hammering nuts

To crack open nutritious coula nuts, chimpanzees use stones as rudimentary hammers and anvils.



Pounding with pestle

With the stalks of palm trees acting as makeshift pestles, chimpanzees can pound and deepen holes in trees.



Fishing for termites

Chimpanzees insert thin, flexible strips of bark into termite mounds to extract the insects, which they then eat.



Wiping ants off stick manually

Once the ants have swarmed almost half-way up sticks dipped into the insects' nests, chimpanzees pull the sticks through their fists and sweep the ants into their mouths.



Eating ants directly off stick

After a few ants climb onto sticks inserted into the nests, chimpanzees bring the sticks directly to their mouths and eat the ants.



Removing bone marrow

With the help of small sticks, chimpanzees eat the marrow found inside the long bones of monkeys they have killed and eaten.



Sitting on leaves

A few large leaves apparently serve as protection when chimpanzees sit on wet ground.



Fanning flies

To keep flies away, chimpanzees utilize leafy twigs as a kind of fan.



Tickling self

A large stone or stick can be used to probe especially ticklish areas on a chimpanzee's own body.



BOSSOU	TAÏ FOREST	GOMBE	MAHALE M-GROUP	MAHALE K-GROUP	KIBALE	BUDONGO
customary	customary	absent	absent	absent	absent (eco?)	absent (eco)
customary	absent	absent	absent (eco?)	absent (eco?)	absent (eco?)	absent (eco?)
absent	absent (eco)	customary	absent	customary	absent (eco)	absent (eco?)
present	absent	customary	absent	absent	absent	absent
customary	customary	present	absent	absent	absent	absent
absent	customary	absent	absent	absent	absent	absent
present	habitual	absent	absent	absent	present	absent
absent	habitual	present	absent	absent	absent	habitual
absent	absent	habitual	absent	absent	absent	absent



BOSSOU	TAÏ FOREST	GOMBE	MAHALE M-GROUP	MAHALE K-GROUP	KIBALE	BUDONGO
customary	customary	customary	customary	absent	present	present
absent	present	present	absent	absent	customary	absent
customary	customary	absent	customary	customary	habitual	customary
absent	absent	habitual	unknown	unknown	absent	absent
absent	absent	present	unknown	unknown	absent	customary
absent	customary	present	absent	absent	absent	absent
absent	habitual	absent	customary	customary	customary	absent
present	customary	habitual	customary	customary	absent	absent
absent	habitual	customary	customary	customary	customary	habitual



Throwing
Chimpanzees can throw objects such as stones and sticks with clear—though often inaccurate—aim.



Inspecting wounds
When injured, chimpanzees touch wounds with leaves, then examine the leaves. In some instances, chimpanzees chew the leaves first.



Clipping leaves
To attract the attention of playmates or fertile females, male chimpanzees noisily tear leaf blades into pieces without eating them.



Squashing parasites on leaves
While grooming another chimpanzee, an individual removes a parasite from its partner, places it on a leaf and then squashes it.



Inspecting parasites
Parasites removed during grooming are placed on a leaf in the chimpanzee's palm; the animal inspects the insect, then eats or discards it.



Squashing parasites with fingers
Chimpanzees remove parasites from their grooming partners and place the tiny insects on their forearms. They then hit the bugs repeatedly before eating them.



Clasping arms overhead
Two chimpanzees clasp hands above their heads while grooming each other with the opposite hand.



Knocking knuckles
To attract attention during courtship, chimpanzees rap their knuckles on trees or other hard surfaces.



Rain dancing
At the start of heavy rain, adult males perform charging displays accompanied by dragging branches, slapping the ground, beating buttress roots, and pant hooting.

Continued from page 63

one hand using short sticks, and knuckle-knocks to attract females clearly comes from the Tai Forest. A chimp that leaf-grooms and hand-clasps during grooming can come from the Kibale Forest or the Mahale Mountains, but if you notice that he also ant-fishes, there is no doubt anymore—he comes from Mahale.

In addition, chimpanzee cultures go beyond the mere presence or absence of

a particular behavior. For example, all chimpanzees dispatch parasites found during grooming a companion. But at Tai they will mash the parasites against their forearms with a finger, at Gombe they squash them onto leaves, and at Budongo they put them on a leaf to inspect before eating or discarding them. Each community has developed a unique approach for accomplishing the same goal. Alternatively, behaviors may look similar yet be used in different contexts:

at Mahale, males “clip” leaves noisily with their teeth as a courtship gesture, whereas at Tai, chimpanzees incorporate leaf-clipping into drumming displays.

The implications of this new picture of chimpanzee culture are many. The information offers insight into our distinctiveness as a species. When we first published this work in the journal *Nature*, we found some people quite disturbed to realize that the characteristic that had appeared to separate us so

Do Apes Ape?

Recent studies show that chimpanzees and other apes can learn by imitation

The notion that the great apes—chimpanzees, gorillas, orangutans and gibbons—can imitate one another might seem unsurprising to anyone who has watched these animals playing at the zoo. But in scientific circles, the question of whether apes, well, *ape*, has become controversial.

Consider a young chimpanzee watching his mother crack open a coula nut, as has been observed in the Tai Forest of West Africa. In most cases, the youth will eventually take up the practice himself. Was this because he imitated his mother? Skeptics think perhaps not. They argue that the mother’s attention to the nuts encouraged the youngster to focus on them as well. Once his attention had been drawn to the food, the young chimpanzee learned how to open the nut by trial and error, not by imitating his mother.

Such a distinction has important implications for any discussion of chimpanzee cultures. Some scientists define a cultural trait as one that is passed down not by genetic inheritance but instead when the younger generation copies adult behavior. If cracking open a coula is something that chimpanzees can simply figure out how to do on their own once they hold a hammer stone, then it can’t be considered part of their culture. Furthermore, if these animals learn exclusively by trial and error, then chimpanzees must, in a sense, reinvent the wheel each time they tackle a new skill. No cumulative culture can ever develop.

The clearest way to establish how chimpanzees learn is through laboratory experiments. One of us (Whiten), in collaboration with Deborah M. Custance of Goldsmiths College, University of London, constructed artificial fruits to serve as analogues of those the animals must deal with in the wild (right). In a typical experiment, one group of chimpanzees watched a complex technique for opening one of the fruits, while a second group observed a very different method; we then recorded the extent to which the chimpanzees had been influenced by the method they observed. We also conducted similar experiments with

three-year-old human children as subjects. Our results demonstrate that six-year-old chimpanzees show imitative behavior that is markedly like that seen in the children, although the fidelity of their copying tends to be poorer.

In a different kind of experiment, one of us (Boesch), along with some co-workers, gave chimpanzees in the Zurich Zoo in Switzerland hammers and nuts similar to those available in the wild. We then monitored the repertoire of behaviors displayed by the captive chimpanzees. As it turned out, the chimpanzees in the zoo exhibited a greater range of activities than the more limited and focused set of actions we had seen in the wild. We interpreted this to mean that a wild chimpanzee’s cultural environment channeled the behavior of youngsters, steering them in the direction of the most useful skills. In the zoo, without benefit of existing traditions, the chimpanzees experimented with a host of less useful actions.

Interestingly, some of the results from the experiments involving the artificial fruits converge with this idea. In one study, chimpanzees copied an entire sequence of actions they had witnessed, but did so only after several viewings and after trying some alternatives. In other words, they tended to imitate what they had observed others doing at the expense of their own trial-and-error discoveries.

In our view, these findings taken together suggest that apes do ape and that this ability forms one strand in cultural transmission. Indeed, it is difficult to imagine how chimpanzees could develop certain geographic variations in activities such as ant-dipping and parasite-handling without copying established traditions. They must be imitating other members of their group.

We should note, however, that—just as is the case with humans—certain cultural traits are no doubt passed on by a combination of imitation and simpler kinds of social learning, such as having one’s attention drawn to useful tools. Either way, learning from elders is crucial to growing up as a competent wild chimpanzee.

—A.W. and C.B.



SARAH MARSHALL AND ANDREW WHITEN, Ngamba, UMWEX Uganda

PRACTICE MAKES PERFECT as a juvenile chimpanzee experiments with an artificial fruit it has been given to “peel” after watching others do so. Such studies help scientists determine how chimpanzees learn by imitating others.

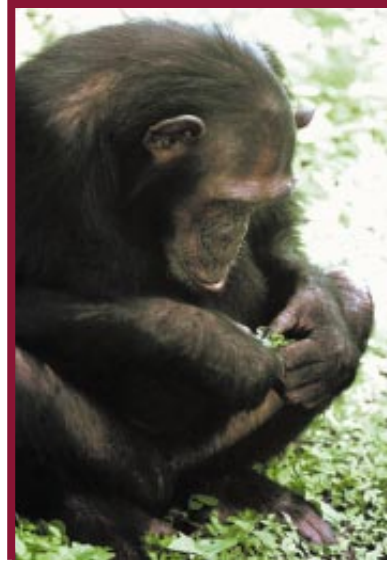
starkly from the animal world—our capacity for cultural development—is not such an absolute difference after all.

But this seems a rather misdirected response. The differences between human customs and traditions, enriched and mediated by language as they are, are vast in contrast with what we see in the chimpanzee. The story of chimpanzee cultures sharpens our understanding of our uniqueness, rather than threatening it in any way that need worry us.

Human achievements have made enormous cumulative progress over the generations, a phenomenon Boesch and Tomasello have dubbed the “ratchet effect.” The idea of a hammer—once simply a crude stone cobbler—has been modified and improved on countless times until now we have electronically controlled robot hammers in our factories. Chimpanzees may show the beginnings of the ratchet effect—some that use stone anvils, for example, have gone a step further, as at Bossou, where they wedge a stone beneath their anvil when it needs leveling on bumpy ground—but such behavior has not become customary and is rudimentary indeed beside human advancements.

The cultural capacity we share with chimpanzees also suggests an ancient ancestry for the mentality that must underlie it. Our cultural nature did not emerge out of the blue but evolved from simpler beginnings. Social learning similar to that of chimpanzees would appear capable of sustaining the earliest stone-tool cultures of human ancestors living two million years ago.

Whether chimpanzees are the sole



DAVID BYGOTT/Kiboyu Perinets

Grooming one another is a source of fascination to chimpanzees in all known populations, but exactly how they deal with nuisances such as ticks and lice differs. Those in East Africa, such as the Gombe chimpanzee shown here, will sometimes turn from grooming a companion's coat to “grooming” leaves. When Gombe chimpanzees find parasites, they may put them on top of a stack of leaves and then carefully, using their thumbnails, squash the insects before eating them. At Budongo they will instead place the insects on leaves and inspect them before eating or discarding them. In the Tai Forest, however, chimpanzees do not use leaves; they place parasites on their forearms and hit them repeatedly with their forefingers until the insects have been smashed. All East African communities incorporate leaves in their grooming habits, suggesting a common eastern origin to the practice.

species on the planet that shares humankind's capacity for culture is too early to judge: nobody has undertaken the comprehensive research necessary to test the idea. Early evidence hints that other creatures should be included in these discussions, however. Carel P. van Schaik and his colleagues at Duke University have found orangutans in Sumatra that habitually use at least two different kinds of tools. Orangutans monitored for years elsewhere have never been seen to do this.

And Hal Whitehead of Dalhousie University and his colleagues have begun to document the ways in which populations of whales that sing in different dialects also hunt in different ways. We hope that our comprehensive approach to documenting chimpanzee cultures may provide a template for the study of these other promising species.

What of the implications for chimpanzees themselves? We must highlight the tragic loss of chimpanzees, whose

populations are being decimated just when we are at last coming to appreciate these astonishing animals more completely. Populations have plummeted in the past century and continue to fall as a result of illegal trapping, logging and, most recently, the bushmeat trade. The latter is particularly alarming: logging has driven roadways into the forest that are now used to ship wild-animal meat—including chimpanzee meat—to consumers as far afield as Europe. Such destruction threatens not only the animals themselves but also a host of fascinatingly different ape cultures.

Perhaps the cultural richness of the ape may yet help in its salvation, however. Some conservation efforts have already altered the attitudes of some local people. A few organizations have begun to show videotapes illustrating the cognitive prowess of chimpanzees. One Zairian viewer was heard to exclaim, “Ah, this ape is so like me, I can no longer eat him.”

SA

The Authors

ANDREW WHITEN and CHRISTOPHE BOESCH have collaborated since 1998 on the cross-cultural study of chimpanzees. Whiten, a fellow of the British Academy, is professor of evolutionary and developmental psychology at the University of St. Andrews in Scotland. Boesch is co-director of the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, and a professor at the University of Leipzig. The chimpanzee field-study directors participating in the research described here are Jane Goodall, Jane Goodall Institute, Washington, D.C.; William C. McGrew, Miami University; Toshisada Nishida, Kyoto University, Japan; Vernon Reynolds, University of Oxford; Yukimaru Sugiyama, Tokaigakuen University, Japan; Caroline E. G. Tutin, University of Stirling, Scotland; and Richard W. Wrangham, Harvard University.

Further Information

CHIMPANZEE MATERIAL CULTURE. William C. McGrew. Cambridge University Press, 1992.
CULTURES IN CHIMPANZEES. A. Whiten, J. Goodall, W. C. McGrew, T. Nishida, V. Reynolds, Y. Sugiyama, C.E.G. Tutin, R. W. Wrangham and C. Boesch in *Nature*, Vol. 399, pages 682–685; 1999.
CHIMPANZEES OF THE TAI FOREST: BEHAVIORAL ECOLOGY AND EVOLUTION. Christophe Boesch and Hedwige Boesch-Aschermann. Oxford University Press, 2000.
PRIMATE CULTURE AND SOCIAL LEARNING. Andrew Whiten in *Cognitive Science*. Special issue on primate cognition, Vol. 24, pages 477–508; 2000.
Chimpanzee Cultures Web site: <http://chimp.st-and.ac.uk/cultures/>
Wild Chimpanzee Foundation Web site: <http://www.wildchimps.org>

The Cellular Chamber of

Structures called proteasomes inside cells continuously destroy proteins. Several common diseases result when the process works too zealously—or not at all

by Alfred L. Goldberg, Stephen J. Elledge and J. Wade Harper

Every minute of every day a scene straight out of an Indiana Jones movie plays out in all our cells. One second a hapless protein is tooling along just trying to do its job. The next instant it is branded for destruction and gets sucked into a dark tunnel, where it is quickly cut to pieces. Unlike Indiana Jones, for the protein there is no escape. Inside the chamber of doom, the protein is stretched out like a medieval prisoner on the rack and fed through a series of enzymatic knives that deliver the Death of a Thousand Cuts. A few seconds later the remnants emerge from the tunnel, only to be pounced on and chewed up further by simpler enzymes.

One might think that this intracellular drama is insignificant (except, perhaps, to the unfortunate protein). But scientists in many laboratories, such as our own, are now finding that these molecular abattoirs, called proteasomes (pronounced “pro-tee-ah-somes”), are crucial players in pathways that regulate an entire repertoire of cellular processes. A typical cell in the body has roughly 30,000 proteasomes. When they malfunction—whether overeagerly gobbling important proteins or failing to destroy those that are damaged or improperly formed—diseases can ensue. Some viruses, such as the human immunodeficiency virus (HIV), have even developed the means to manipulate protein degradation by proteasomes for their own ends. Indeed, several of the next-generation drugs to treat cancer and other dire diseases are expected to consist of chemical compounds that act on proteasomes and the pathways that feed proteins into proteasomes. Several biopharmaceutical companies are now studying compounds that inhibit the proteasome pathway; two such potential drugs are already in clinical trials in humans.

Turnover Is Fair Play

Proteins are the very fabric of which cells are made. Some proteins also act as enzymes, the molecular workhorses that drive the chemical reactions of life. The types of proteins a cell produces depend on which of its genes are active at any

given time. Genes encode how the 20 basic protein subunits, called amino acids, are assembled into chains of various combinations. The chains fold into compact coils and loops to become different kinds of proteins, each with a specific function determined by its shape and chemistry.

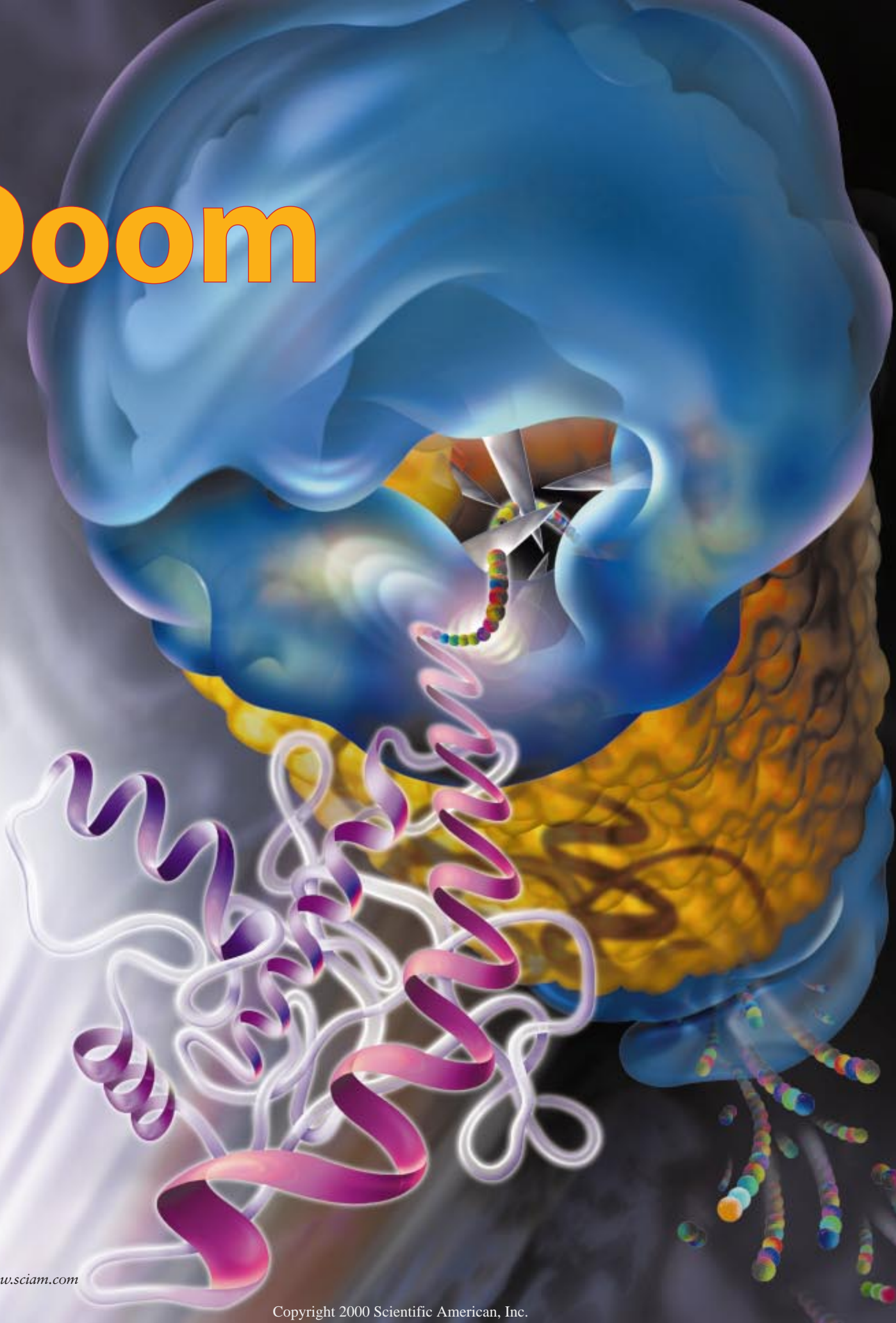
What happens when proteins are no longer needed or fail to fold correctly? For years, scientists presumed that the lion's share of protein degradation occurs in lysosomes, bags of digestive enzymes present in most cells of the body. But in the early 1970s one of us (Goldberg) showed that cells lacking lysosomes, such as bacteria and immature red blood cells, can nonetheless destroy abnormal proteins rapidly. What is more, the process requires energy, whereas other degradative processes do not.

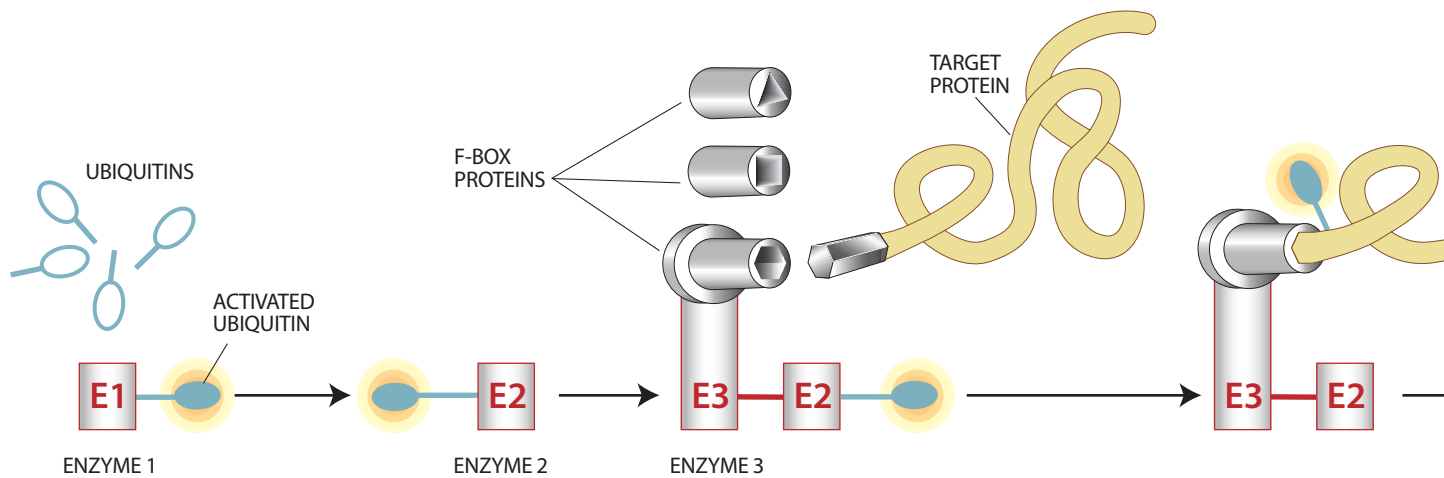
He and his colleagues were able to get the energy-requiring degradation process to work in test tubes, which enabled several research groups in the late 1970s and throughout the 1980s to discover the enzymes responsible. Eventually, in 1988, two groups—one led by Goldberg and the other by Martin C. Rechsteiner of the University of Utah—found that the proteins are broken down by large, multienzyme complexes that Goldberg's group named proteasomes.

Proteasomes were so named because they contain many proteases, enzymes that cut proteins into chunks. But proteasomes are 100 times larger and more complex than other proteases. Once a protein is laid on the doormat of a proteasome, it is taken inside the particle and ultimately disassembled like a Tinker Toy into amino acids that can be reassembled later

PROTEASOME draws a protein (*ribbonlike structure at left in lower half*) into its maw for destruction by six specific enzymes, shown here as knives. An average body cell has thousands of proteasomes, which chop proteins the cell wishes to remove into bits of various sizes. The bits are then broken down by other enzymes into the basic building blocks of proteins—amino acids—which are eventually recycled to make new proteins.

Doom





into other proteins. Most proteins are replaced every few days, even in cells that themselves divide rarely, such as those in the liver or nervous system. And different proteins are degraded at widely differing rates: some have half-lives as short as 20 minutes, whereas others in the same cell may last for days or weeks. These rates of breakdown can change drastically according to changing conditions in our bodies.

At first glance, such continuous destruction of cell constituents appears very wasteful, but it serves a number of essential functions. Degrading a crucial enzyme or regulatory protein, for example, is a common mechanism that cells use to slow or stop a biochemical reaction. On the other hand, many cellular processes are activated by the degradation of a critical inhibitory protein, just as water flows out of a bathtub when you remove the stopper. This rapid elimination of regulatory proteins is particularly important in timing the transitions between the stages of the cycle that drives cell division [see box on page 72].

Protein degradation also plays special roles in the overall regulation of body metabolism. In times of need, such as malnourishment or disease, the proteasome pathway becomes more active in our muscles, providing amino acids that can be converted into glucose and burned for energy. This excessive protein breakdown accounts for the muscle wasting and weakness seen in starving individuals and those with advanced cancer, AIDS and untreated diabetes.

Our immune system, in its constant search to eliminate virus-infected or cancerous cells, also depends on proteasomes to generate the flags that distinguish such dangerous cells. In this process, the immune system functions like a suspicious landlady checking whether

her tenants are doing something undesirable by monitoring what they throw out in their daily trash. Although cell proteins are usually degraded all the way to amino acids, a few fragments composed of eight to 10 amino acids are released by proteasomes, captured, and ultimately displayed on the cell's surface, where the immune system can monitor whether they are normal or abnormal [see illustration on page 73]. Indeed, in disease states and in certain tissues such as the spleen and lymph nodes, specialized types of proteasomes termed immunoproteasomes are produced that enhance the efficiency of this surveillance mechanism.

Protein breakdown by proteasomes also serves as a kind of cellular quality-control system that prevents the accumulation of aberrant—and potentially toxic—proteins. Bacterial and mammalian cells selectively destroy proteins with highly abnormal conformations that can arise from mutation, errors in synthesis or damage.

The degradation of abnormal proteins is important in a number of human genetic diseases. In various hereditary anemias, a mutant gene leads to the production of abnormal hemoglobin molecules, which do not fold properly and are rapidly destroyed by proteasomes soon after synthesis. Similarly, cystic fibrosis is caused by a mutation in the gene encoding a porelike protein that moves chloride across a cell's outer membrane. Because these mutant chloride transporters are slightly misshapen, proteasomes degrade them before they can reach the cell membrane. The sticky mucus that builds up in the lungs and other organs of people with cystic fibrosis results from the lack of normal chloride transporters.

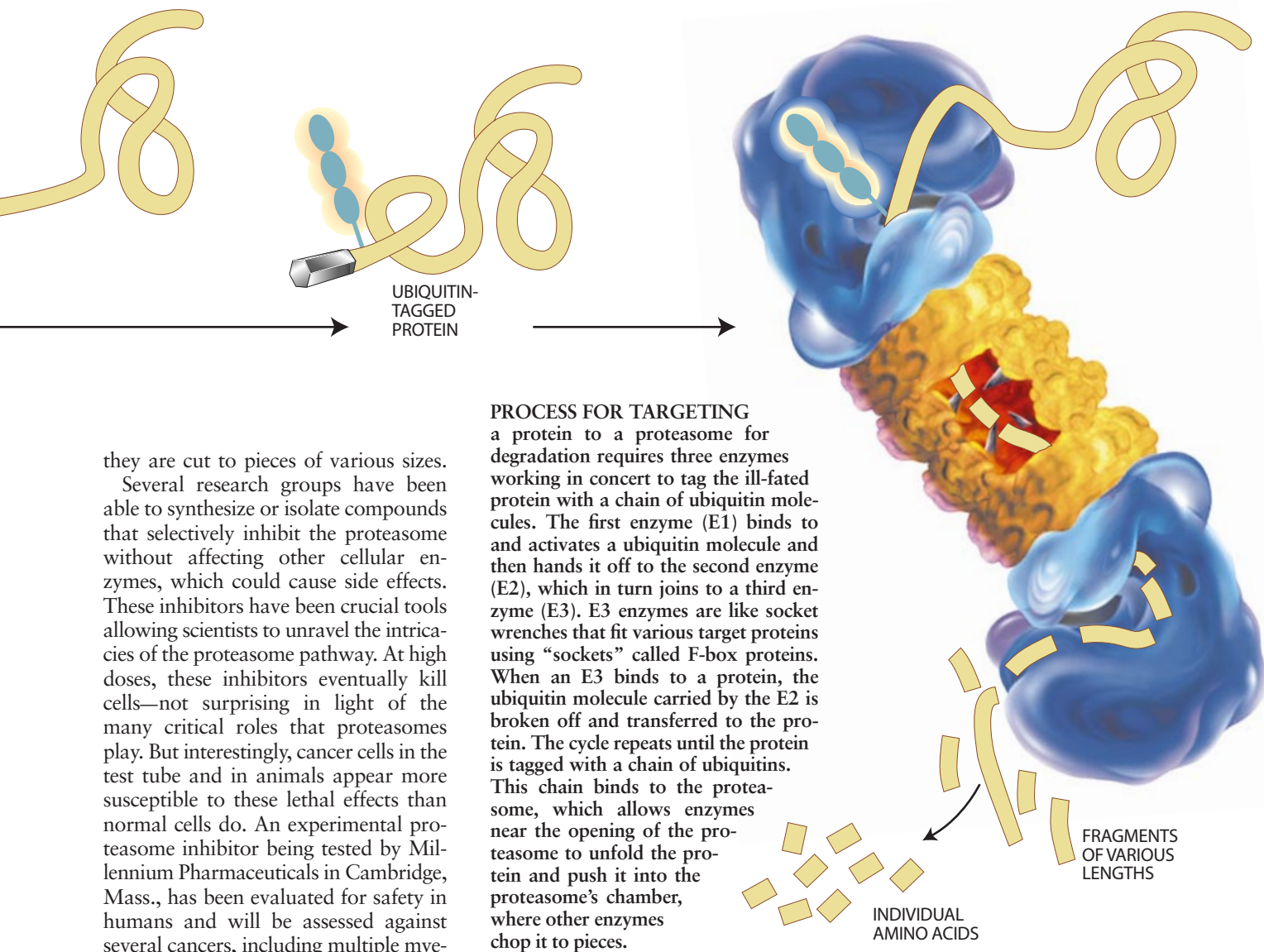
Still other diseases could result in part from the failure of abnormal proteins to

be degraded by proteasomes. Scientists are finding, for example, that clumps of misfolded proteins accumulate in association with proteasomes in certain nerve cells, or neurons, in the brains of people with neurodegenerative disorders such as Parkinson's, Huntington's and Alzheimer's diseases. Why the neurons of individuals stricken with these maladies fail to degrade the abnormal proteins is a burgeoning field of research.

In the Belly of the Beast

From a protein's humble perspective, proteasomes are enormous structures. Whereas the average protein is 40,000 to 80,000 daltons (or 40,000 to 80,000 times the molecular weight of a hydrogen atom), most proteasomes from higher organisms weigh in at a whopping two million daltons. In the mid-1990s scientists led by Wolfgang Baumeister and Robert Huber of the Max Planck Institute for Biochemistry in Martinsried, Germany, used x-ray diffraction and electron microscopy to determine the molecular architecture of proteasomes. Each one consists of a tunnellike core particle with one or two smaller, regulatory particles positioned at either or both ends like caps. The core particle is formed by four stacked rings—each composed of seven subunits—surrounding a central channel that constitutes a proteasome's digestive tract. The outer two rings appear to act as gates to keep stray proteins from accidentally bumping into the degradation chamber.

Similarly, the regulatory “cap” particles are thought to act as highly selective gatekeepers to the core particle. These regulatory particles recognize and bind to proteins targeted for destruction, then use energy to unfold the proteins and inject them into the core particle, where



PROCESS FOR TARGETING
 a protein to a proteasome for degradation requires three enzymes working in concert to tag the ill-fated protein with a chain of ubiquitin molecules. The first enzyme (E1) binds to and activates a ubiquitin molecule and then hands it off to the second enzyme (E2), which in turn joins to a third enzyme (E3). E3 enzymes are like socket wrenches that fit various target proteins using “sockets” called F-box proteins. When an E3 binds to a protein, the ubiquitin molecule carried by the E2 is broken off and transferred to the protein. The cycle repeats until the protein is tagged with a chain of ubiquitins. This chain binds to the proteasome, which allows enzymes near the opening of the proteasome to unfold the protein and push it into the proteasome’s chamber, where other enzymes chop it to pieces.

they are cut to pieces of various sizes. Several research groups have been able to synthesize or isolate compounds that selectively inhibit the proteasome without affecting other cellular enzymes, which could cause side effects. These inhibitors have been crucial tools allowing scientists to unravel the intricacies of the proteasome pathway. At high doses, these inhibitors eventually kill cells—not surprising in light of the many critical roles that proteasomes play. But interestingly, cancer cells in the test tube and in animals appear more susceptible to these lethal effects than normal cells do. An experimental proteasome inhibitor being tested by Millennium Pharmaceuticals in Cambridge, Mass., has been evaluated for safety in humans and will be assessed against several cancers, including multiple myeloma, in trials set to begin this winter. Another of Millennium’s proteasome inhibitors is in early safety trials in humans as a possible treatment for stroke and myocardial infarction.

The Kiss of Death

The proteasome does not just randomly pick out proteins to destroy. Instead a cell points out which proteins are doomed. Scientists have discovered that the vast majority of such proteins are first tagged with another protein called ubiquitin, for its ubiquity among many different organisms. With only 76 amino acids, ubiquitin is a relatively tiny protein that can be attached to larger proteins in long chains. These poly-ubiquitin tails act like postal codes that speed doomed proteins to proteasomes.

What controls the timing of a protein’s demise is not its actual breakdown by the proteasome, but the process of adding the ubiquitin chains, called ubiquitination, which requires energy. The

basic outline for how ubiquitin is attached to a protein has come from Avram Hershko and Aaron Ciechanover of the Technion-Israel Institute of Technology in Haifa, working with Irwin A. Rose of the Fox Chase Cancer Center in Philadelphia.

The ubiquitination process has several steps and involves three enzymes, dubbed E1, E2 and E3 [see illustration above and on opposite page]. The E1 enzyme activates ubiquitin and connects it to E2. The third enzyme, E3, then facilitates the transfer of the activated ubiquitin from the E2 to the protein. The process repeats until a long chain of ubiquitins dangles off the protein. That chain is then recognized by a proteasome, which draws the protein in.

The mystery of how a protein is chosen for ubiquitination revolves around the E3 proteins. Recently researchers, including two of us (Elledge and Harper), have discovered that there are hun-

dreds of distinct E3 proteins that recognize information in the amino acid sequences of other proteins that make them targets for ubiquitination. In response to altered physiological conditions, such as infection or a lack of nutrients, cells can modify proteins by adding phosphate groups. Such phosphorylation can alter the activity of a protein or its ability to bind to E3s. Proteins that fail to fold or that become damaged are also recognized by E3s, which come along and clean up the proteins by marking them for pickup by the proteasome—a little like putting them out on the curb on garbage day. Many key cellular processes rely on protein stability, and finding out how stability is controlled therefore holds the key to many of biology’s secrets.

By controlling the stability of crucial proteins, the E3 proteins regulate many cellular processes, such as limb development, the immune response, cell divi-

sion and cell-to-cell communication. Even circadian rhythms and flowering in plants are dictated by E3 enzymes. What is more, several E3s have been identified as tumor suppressors or oncogenes, tying ubiquitination to the onset of cancer.

A case in point is the Von Hippel Lin-

dau (VHL) tumor suppressor, an E3 that is often mutated in kidney tumors. VHL's job is to retard cell growth by limiting the development of blood vessels in tissues; when it is mutated, newly formed tumors are able to generate a rich blood supply and grow rapidly. Scientists have now found that an inherited

form of Parkinson's disease results from a mutation in the gene for a type of E3 enzyme that can cause proteins to build up in certain brain cells and kill them.

Viruses, which are famous for diverting cellular processes, have evolved the means to hijack the process of ubiquitination and protein degradation for their

Why Cell Division Depends on Protein Death

One of the best examples of why a cell's ability to break down proteins is important for its life and growth comes from studying cell division in *Saccharomyces cerevisiae*, the common baker's yeast. Before a yeast cell—or even a human cell—divides, it must first copy its DNA. And to begin DNA synthesis, a cell needs to activate a particular class of proteins called S-phase Cdk, which are composed of two proteins, a cyclin and a Cdk subunit.

S-phase Cdk's are normally inactive because they are bound to inhibitory proteins (called CKIs) that were made during the previous cell division. To activate the S-phase Cdk's, a cell must get rid of the inhibitory proteins by sending them to a proteasome to be degraded (*below*).

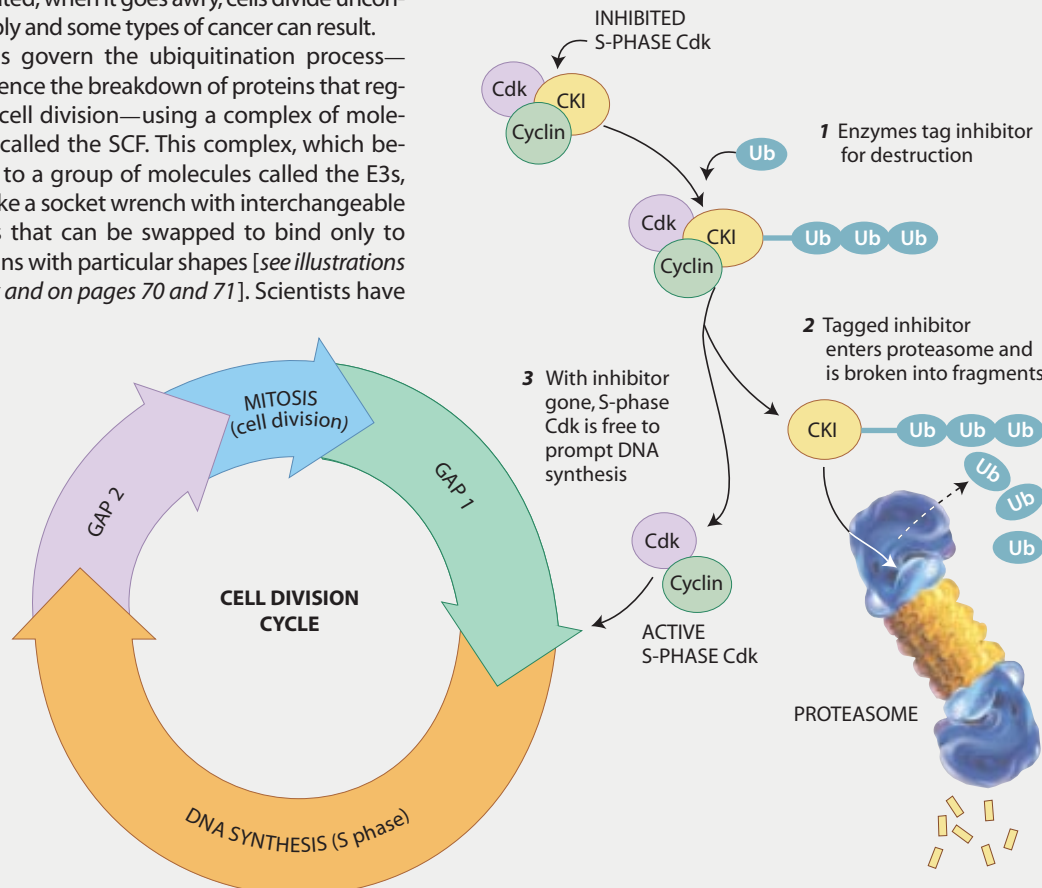
Targeting the inhibitory proteins for destruction by a proteasome requires tagging the proteins with a death signal called ubiquitin (Ub). This tagging process is normally tightly regulated; when it goes awry, cells divide uncontrollably and some types of cancer can result.

Cells govern the ubiquitination process—and hence the breakdown of proteins that regulate cell division—using a complex of molecules called the SCF. This complex, which belongs to a group of molecules called the E3s, acts like a socket wrench with interchangeable heads that can be swapped to bind only to proteins with particular shapes [*see illustrations below and on pages 70 and 71*]. Scientists have

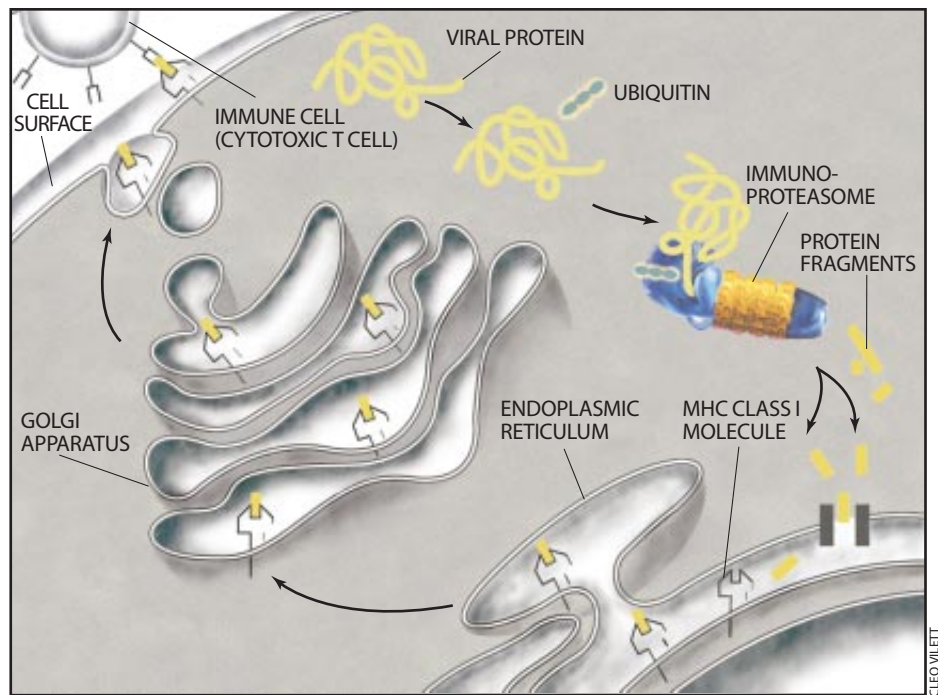
identified more than 217 such socket heads, called F-box proteins, in the nematode worm *Caenorhabditis elegans*; several dozen have been found in human cells so far, and the count is rising.

The SCF complex uses its specific set of socket heads to recognize proteins that should be broken down by the proteasome. Indeed, cells choose which proteins to degrade by adding a phosphate group to them so that they bind to the F-box proteins of the SCF. The SCF also serves as a go-between to bring such ill-fated proteins together with the enzymes that add the ubiquitin death tag.

The variety of SCF complexes gives a cell exquisite control over which types of proteins—and how much of each one—it has on hand at any given time. Proteins regulated by SCF complexes include those that promote or inhibit the cell division cycle and those that turn on genes. —S.J.E. and J.W.H.



IMMUNE SYSTEM relies on specialized proteasomes called immunoproteasomes to help it distinguish healthy cells from cancerous ones or those that have been infected by viruses. In the example shown at the right, a viral protein is tagged with ubiquitin for destruction by the immunoproteasome. Bits of the viral protein that are between eight and 10 amino acids in length then enter the endoplasmic reticulum, where they are loaded onto newly formed, forklike molecules called the major histocompatibility complex (MHC) class I. As the MHC class I molecules are transported through the Golgi complex and float to the cell surface, they take along the viral protein bits. Immune cells called cytotoxic T cells recognize the MHC class I molecules on the cell surface as foreign and kill the infected cell.



own nefarious purposes. Human papillomaviruses (HPVs), which can cause genital warts or cervical or anal cancer, are examples. The transformation to cancerous growth is usually blocked by the defense protein p53, one of the body's tumor suppressor proteins. HPVs use a trick to circumvent this cellular defense system: they make a protein that binds simultaneously to both p53 and an E3 enzyme. This binding leads to the ubiquitination of p53, which destines p53 to be sliced and diced to obliteration by the enzymatic Ginsu knives of the proteasome. The defenseless cells are then more likely to become cancers.

HIV uses a similar ploy to destroy the cell-surface protein CD4, which is nec-

essary for the virus to infect cells but which interferes with the production of more viruses later on. CD4 acts as a docking site for HIV to enter the T cells of the immune system; it binds to the gp160 protein that protrudes from the surface of the virus. But when HIV starts attempting to replicate in the newly infected cells, CD4 can present a problem: it adheres to freshly made gp160 proteins, keeping them from assembling with other viral proteins into new viruses. To circumvent this obstacle, HIV has evolved a protein called Vpu that puts CD4 on the fast track to oblivion. Vpu binds to both CD4 and a complex containing an E3 enzyme, causing CD4 to become ubiquitinated and then dropped down the chute

of the proteasome to be destroyed.

New discoveries about the importance of E3s in disease are rapidly emerging, and these enzymes are likely to be targets for drug development in the future. Because each E3 is responsible for the destruction of a small number of proteins, specific inhibitors of E3s should be highly specific drugs with few side effects. The recent identification of large families of E3 enzymes have opened up whole new avenues for drug discovery. These are exciting developments that promise to enrich the understanding of diverse regulatory phenomena and human biology. The more we learn about proteasomes and the ubiquitination selection machinery, the more we appreciate how much of life is linked to protein death. SA

The Authors

ALFRED L. GOLDBERG, STEPHEN J. ELLEDGE and J. WADE HARPER have built their careers on studying protein degradation and its role in disease. Goldberg is a professor of cell biology at Harvard Medical School, where he received his Ph.D. in 1968 and where he has spent most of his academic life. He has consulted widely for industry; among his honors are the 1998 Novartis-Drew Award for Biomedical Research. Elledge is the Robert A. Welch Professor of Biochemistry at Baylor College of Medicine and is an investigator for the Howard Hughes Medical Institute. He received his Ph.D. from the Massachusetts Institute of Technology in 1983 and was awarded the Michael E. DeBakey Award for Research Excellence in 1994. Harper is a professor in Baylor's department of biochemistry and molecular biology and its department of molecular physiology and biophysics. He received his Ph.D. from the Georgia Institute of Technology in 1984 and was granted the Vallee Visiting Professorship at the University of Oxford in 2000.

Further Information

HOW THE CYCLIN BECAME A CYCLIN: REGULATED PROTEOLYSIS IN THE CELL CYCLE. Deanna M. Koepf, J. Wade Harper and Stephen J. Elledge in *Cell*, Vol. 97, No. 4, pages 431-434; May 14, 1999.

INTRICACIES OF THE PROTEASOME. Stu Borman in *Chemical and Engineering News*, Vol. 78, No. 12, pages 43-47; March 20, 2000.

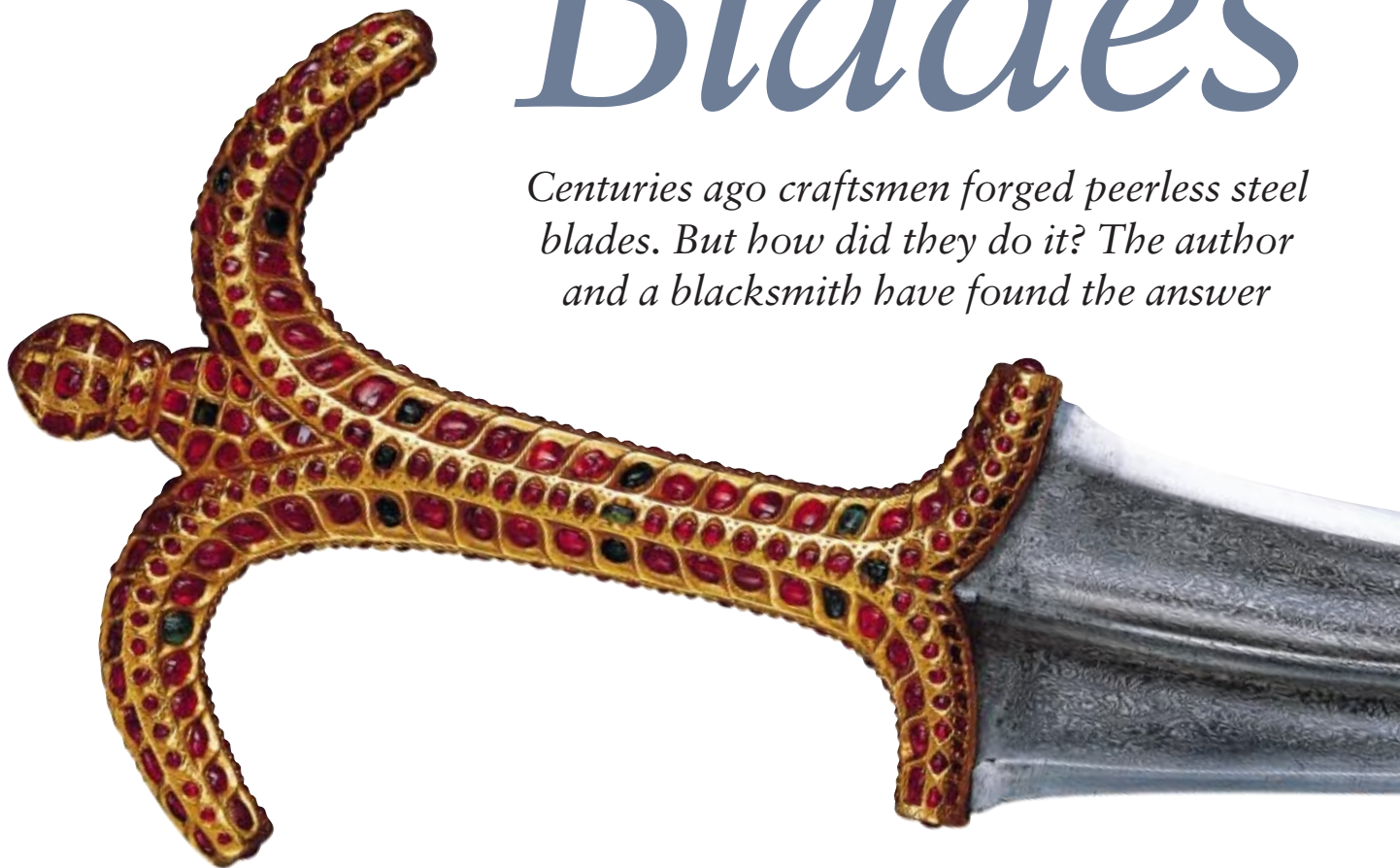
PROBING THE PROTEASOME PATHWAY. Alfred L. Goldberg in *Nature Biotechnology*, Vol. 18, No. 5, pages 494-496; May 18, 2000.

For more information on the history of studies of the proteasome pathway, visit the Albert and Mary Lasker Foundation Web site at www.laskerfoundation.org/library/2000/citation1.html

The Mystery of *Damascus* Blades

by John D. Verhoeven

Centuries ago craftsmen forged peerless steel blades. But how did they do it? The author and a blacksmith have found the answer



From the Bronze Age up to the 19th century, warriors relied on the sword as a weapon. Armies possessing better versions enjoyed a distinct tactical advantage. And those with Damascus swords—which Westerners first encountered during the Crusades against the Muslim nations—had what some consider to be the best sword of all.

Those blades, originally thought to have been fashioned in Damascus (which is now in Syria), featured two qualities not found in European varieties. A wavy pattern known today as damask, or damascene, decorated their surface [see illustration above]. And, more important, the edge could be incredibly

sharp. Legend tells how Damascus swords could slice through a silk handkerchief floating in the air, a feat no European weapon could emulate.

Despite the fame and utility of these blades, Westerners have never been able to figure out how the steel—also used for daggers, axes and spearheads—was made. The most accomplished European metallurgists and bladesmiths could not replicate it, even after bringing specimens home and analyzing them in detail. The art of production has been lost even in the land of origin; experts generally agree that the last high-quality Damascus swords were crafted no later than the early 1800s. Recently, however, an ingenious blacksmith and I have, we believe, unlocked the secret.

We are not the first to have claimed a solution, but we are the first to have proved our case by making faithful replicas of the revered weapons. To validate any theory of how Damascus swords and daggers were made, replicas ought to be fashioned from the same starting materials as the originals. The finished weapons should also bear the same damask pattern and have the same chemistry and microscopic structure.

What Is Real Damascus Steel?

Genuine Damascus blades are known to have been made in that city—and later elsewhere in the Muslim Middle East and Orient—from small ingots made of steel (a mix of iron and carbon) shipped from India; those starting mate-

rials have been called wootz ingots or wootz cakes since around 1800. They were shaped like hockey pucks, about four inches in diameter and a bit less than two inches in height. Early English observers in India established that the wootz Damascus swords were made by forging these ingots directly into a blade shape by many repeated heating and hammering operations. The steel contains around 1.5 percent carbon by weight, plus low levels of other impurities such as silicon, manganese, phosphorus and sulfur.

The attractive surface pattern found on Damascus swords can be created in other ways, however. Modern artist-blacksmiths can “forge weld” together alternate sheets of high- and low-carbon steel into an intricate composite. Such forge welding, or “pattern welding,” has a tradition in the West dating back to ancient Rome, and similar techniques

to produce satisfactory blades that have the exterior appearance and internal structure of the ancient originals.

Efforts to compare the chemistry and microscopic features of modern wootz blades with their older counterparts were long hampered by a curious obstacle. Museum-quality Damascus weapons are valuable art objects and are rarely sacrificed to science for examination of their internal structure. In 1924, though, European collector Henri Moser donated four swords to metallurgist B. Zschokke, who sectioned them for chemical and microstructural analysis. The remaining pieces went to the Berne Museum in Switzerland, which recently donated some of them to me for study.

When I examined the prized specimens, I found that they contained bands of iron carbide particles, Fe_3C , known as cementite. These particles are gener-

soon realized, though, that I would need to work with someone skilled in the art of forging edged weapons. Master bladesmith Alfred H. Pendray had been working independently on the Damascus puzzle. He had been making small ingots in a gas-fired furnace and forging them into blade shapes, and he had often obtained microstructures that were intriguingly close to those of the finer-quality antique blades.

We began collaborating in 1988. Pendray as a youth learned the skills of a farrier from his father and has a deep and patient understanding of the art of forging steel. But to reproduce a technique, we would need to back up our theories with accurate scientific data and rigorous attention to the details of our experiments. In 1993 one of my students at Iowa State University and I



DAGGER with a Damascus steel blade, from Mughal India, was made in about 1585. The fine-quality blade is thickened near the point to pierce armor; the gold hilt is set with emeralds and rubies.

can be found in Indonesia and Japan. The internal structure resulting from these techniques is totally different, though, from that of the wootz blades. To avoid confusion between the two types of manufacture, I refer to the forge-welded blades as “welded” Damascus and reserve the term “wootz” Damascus for the weapons of interest in this article.

As early as 1824, Jean Robert Bréant in France and, slightly later, Pavel Anosoff in Russia announced success at uncovering the secret arts of the Muslim bladesmiths; both claimed to have replicated the originals. In this century other solutions have been advanced, the most recent by Jeffrey Wadsworth and Oleg D. Sherby [see “Damascus Steels,” *SCIENTIFIC AMERICAN*, February 1985]. But in no case have modern artisans been able to use the proposed methods

ally around six to nine microns in diameter, well rounded and tightly clustered into bands spaced 30 to 70 microns apart, which are lined up parallel to the blade surface, like the grain inside a plank of wood. When the blade is etched with acid, the carbides appear as white lines in a dark steel matrix. Just as the wavy growth rings in a tree produce the characteristic swirling patterns on cut wood, undulations in the carbide bands account for the intricate damascene patterns on the blade surfaces. The carbide particles are extremely hard, and it is thought that the combination of these bands of hard steel within a softer matrix of springier steel gives Damascus weapons a hard cutting edge combined with a tough flexibility.

I first attempted to match the microstructures of wootz Damascus steel in the confines of a university laboratory. I

went to Pendray’s blacksmith shop near Gainesville, Fla., where we set up computer-monitored thermocouple and infrared pyrometer equipment to record the temperatures of the melting and forging processes we were trying.

At first we tried to produce blades using the method put forward by Wadsworth and Sherby, but we failed to produce either the internal microstructure or the surface damascene patterns. Then, over a period of several years, we developed a technique that Pendray can routinely use to make reconstructed wootz Damascus steel blades. He can also replicate the pattern known as Mohammed’s ladder [see *illustration on page 79*], found on some of the finest of the old Muslim examples. In this pattern the undulations line up in a ladderlike formation along the length of the blade; it was thought to be symbolic of the way the faithful ascended to heaven.

Our technique is similar to the general method described by the earlier researchers—but with crucial differences. We produce a small steel ingot of a precise composition in a closed crucible and then forge it into a blade shape. Our success—and what enables us to go further than our predecessors—depends critically on the mix of iron, carbon and other elements (such as vanadium and molybdenum, which we refer to as impurity elements) in the steel, how hot and for how long the crucible is fired, and the temperature and skill used in the repeated forging operations.

A Tale of Steel

If you have steel containing about 1.5 percent carbon, add to it one of several impurity elements (at surprisingly low levels, around 0.03 percent), and then put it through five or six cycles of heating to a precise temperature range and cooling to room temperature, you can get groups of clustered carbide particles to form. It is these carbide particles that produce the characteristic surface patterns during forging. Experiments on antique and modern blades show that band formation results from segregation at a microscopic level of some impurity elements as the liquefied ingot cools and solidifies.

Here's how microsegregation happens within the steel. As the hot ingot cools down and freezes, a solid front of crystallized iron extends into the liquid, adopting the shape of pine-tree-like projections called dendrites [see illustration on opposite page]. In the 1.5 percent carbon steel, the type of iron that solidifies from the liquid steel is called austenite. In the regions between these dendrites (called the interdendritic regions), liquid metal becomes briefly trapped. Solid iron can accommodate fewer atoms of carbon and other elements than liquid iron can, so as the metal solidifies into crystalline iron dendrites, carbon and impurity atoms tend to segregate into the remaining liquid. Hence, the concentration of those atoms can become very high in the last interdendritic regions to freeze.

As the iron solidifies and the dendrites grow, the regions between them are left with a lattice of impurity atoms frozen into place like a string of pearls. Later, when the ingot goes through multiple heating and cooling cycles, it is these impurity atoms that encourage the growth of the strings of hard cementite parti-

cles that are the lighter bands in the steel. We can show that this lattice is related to the light and dark steel bands in the wootz steel. The distance between dendrite branches is around half a millimeter, and as the ingot is hammered out and its diameter is reduced, this distance is also reduced. The final spacing between dendrites corresponds closely to the distance between bands in Damascus steel.

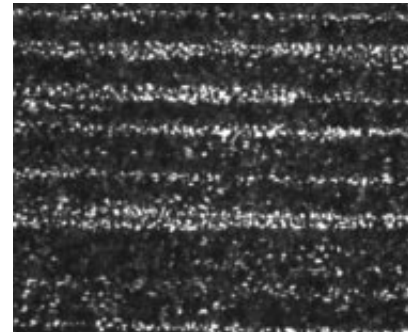
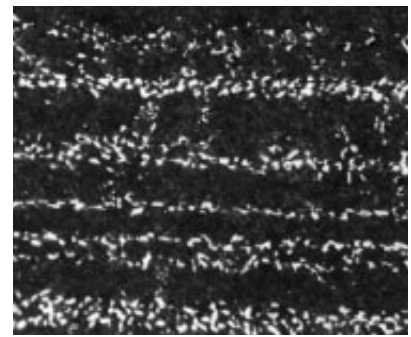
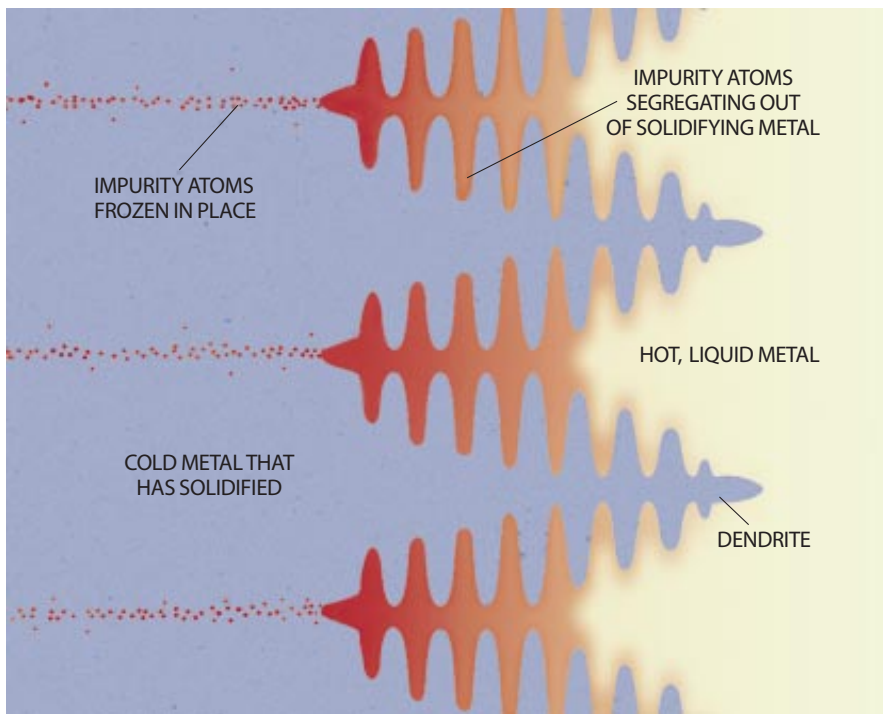
During forging, it is important to get just the right temperature in the steel to obtain a mix of austenite and cementite particles. When the ingot's temperature falls below a critical point, iron carbide particles (the same cementite particles I saw in the Moser blades) start forming. The lowest temperature above which all the cooling steel remains austenite is called the A temperature. In steels with more than 0.77 percent carbon, the A temperature is termed the A_{cm} temperature. Below the A_{cm} , cementite particles begin appearing, randomly spaced within the austenitic steel.

The Trick of Banding

A major mystery of wootz Damascus blades has been how simple forging of small steel ingots into the shape of a blade can cause carbides to line up into distinctive bands. We systematically examined cross sections of the forged ingots as we changed them from hockey-puck shapes to blades. To bring about that change, we heated an ingot to a temperature at which the steel would form a mixture of cementite particles and austenite and then hammered it. While the ingot was being forged, it would cool down from about 50 degrees Celsius below the A_{cm} to about 250 degrees C below the A_{cm} . During this cooling, the proportion of cementite particles increased. We would then put the ingot through another cycle of heating and hammering between the same two temperatures. Based on experience, we found we needed around 50 of these forging cycles to produce a blade close to the size of the originals—45 millimeters wide and five millimeters thick.

This is how we think banding occurs:

DAMASCUS STEEL SWORD from the 17th century shows a classic damascene pattern of swirling light and dark bands. The inscription tells us that this excellent blade was made in 1691 or 1692 by Assad Allah, the most renowned Persian swordsmith of his time.



COOLING INGOT of Damascus steel, on a microscopic level, has a front of freezing metal extending into the molten steel, crystallizing, at first, into pine-tree-like formations called dendrites. Atoms of impurity elements (red) such as vanadium rapidly segregate out of the solid iron into the regions between the dendrites, where they freeze in place lined up like beads on a necklace. In subsequent cycles of heating and cooling, these impurity

atoms are the basis for the growth of particles of hard iron carbide (cementite), which are the light-colored bands in the Damascus blade. The top micrograph shows light and dark bands in a section through an original Damascus sword. The lower micrograph shows a section through the author's modern reconstruction. The similarity between the two structures indicates that the modern technique is an accurate replication of the original process.

During the initial 20 or so cycles, the hard carbide particles form more or less randomly, but with each additional cycle they tend to become more strongly aligned along the latticework of points formed in the interdendritic regions. The reason for the improvement is that each time the steel is heated, some of its carbide particles dissolve. But the atoms of the impurity elements slow the rate of dissolution, causing larger particles of carbide to remain. Each cycle of heating and cooling causes these particles to grow only slightly, which is why it takes so many cycles to form the distinct bands. Because the impurity elements are lined up in the regions between the dendrites, the carbide particles become concentrated there as well.

The Right Elements

Although we long suspected that impurity elements played a key role in the formation of bands, we were not sure which ones were most important. We determined quickly that silicon, sulfur and phosphorus, well known to be present in ancient wootz steels, did not appear to be major players. But that

information did not solve the problem.

We had a lucky breakthrough when we started to use Sorel metal as one ingredient for the ingots. This metal is a high-purity iron-carbon alloy containing 3.9 to 4.7 percent carbon, produced from a large ilmenite ore deposit at Lac Tio on the St. Lawrence River in Quebec. The ore deposit contains traces of vanadium; hence, the Sorel metal comes with 0.003 to 0.014 percent vanadium impurity. Initially we disregarded this impurity because we couldn't believe such a low concentration was significant. But we eventually (after two years of hitting a brick wall) tumbled to the fact that even low levels could be important.

Adding vanadium in such tiny amounts as 0.003 percent to high-purity iron-carbon alloys yielded good banding. Molybdenum also produces the desired effect, and, to a lesser extent, so do chromium, niobium and manganese. Elements that do not promote carbide formation and banding include copper and nickel. Electron-probe microanalysis has confirmed that the effective elements, when present at only 0.02 percent or less in the ingots, become microsegregated into the interdendritic regions and

become much more concentrated there.

To test our conclusion that banding comes from microsegregation of impurity elements leading to microsegregation of cementite particles, we conducted experiments designed to show that if we got rid of the microsegregation of impurity atoms, we could get rid of the bands. We took small pieces of nicely banded antique and modern blades and heated these to around 50 degrees C above the A_{cm} temperature. At this temperature, all the iron carbide particles dissolved away into the austenite. We then quenched the blades in water. The rapid cooling produced the martensite phase of steel—very hard and strong, with no carbide particles. Because the carbide particles had vanished, so had the bands that came from them.

To re-create the cementite particles, we put the blades through several cycles of being heated to 50 degrees C below the A_{cm} temperature and then slowly air-cooled, which gave the particles time to regrow and become segregated. After the first cycle, the carbide particles reappeared but were randomly distributed. But after an additional cycle or two, these particles began to align into

HOW TO MAKE A DAMASCUS BLADE

Master bladesmith Alfred H. Pendray demonstrates the technique in his smithy near Gainesville, Fla.

1 Assemble the ingredients to load into the crucible, including high-purity iron, Sorel iron, charcoal, glass chips and green leaves. The quantity of carbon and impurity elements that end up in the ingot is controlled by the proportions of iron, Sorel iron and charcoal added to the mix.

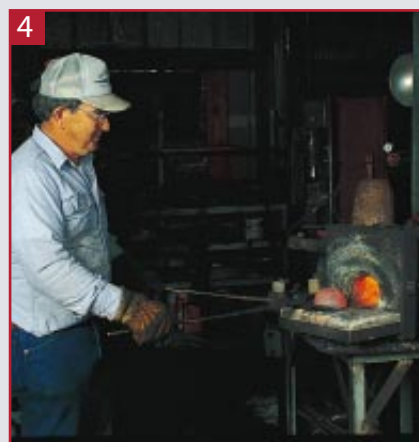
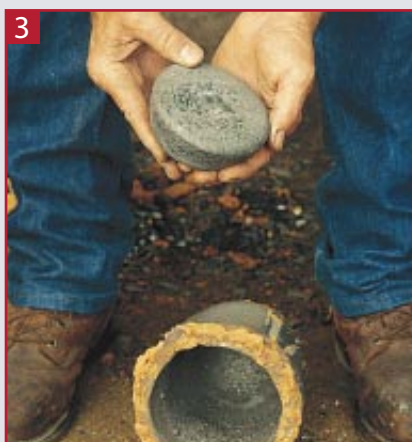
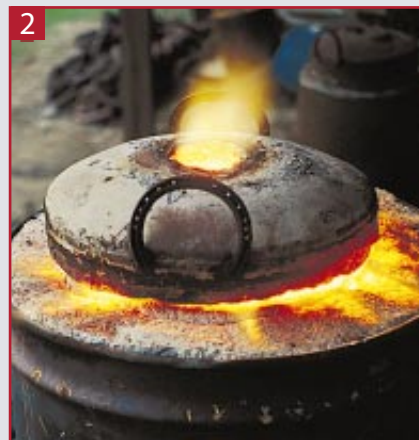
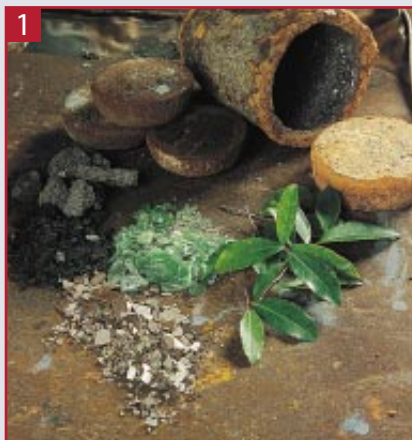
2 Heat the crucible. During this process, the glass melts, forming a slag that protects the ingot from oxidizing. The leaves generate hydrogen, which is known to accelerate carburization of iron. The carbon content of the iron is raised to 1.5 percent, a good proportion for forming the hard iron carbide particles whose accretion into bands gives Damascus blades their characteristic wavy surface pattern. The leaves and glass can be left out, but ingots made without them are more prone to cracking during hammering.

3 When the crucible has cooled, remove the ingot, which bears a resemblance to the wootz cakes used by the ancients.

4 Heat the ingot to a precise temperature. Pendray is using a gas-fired furnace with the propane-to-air ratio adjusted to minimize the formation of oxide scale during forging. Typically, a surface oxide layer of about half a millimeter in thickness forms, and the final grinding operation must be sufficient to remove it.

5 Forge the ingot (deform it slightly with hammer blows while it is still hot). When the ingot gets too cold to deform without cracking, heat it up and forge again. Four separate stages of the ingot are shown here; each stage is the result of several cycles of heating and forging. A total of about 50 cycles may be needed to bang out the blade shape from the ingot—a highly labor-intensive process. Pendray uses a modern air hammer. A handheld hammer works, too, but it takes longer.

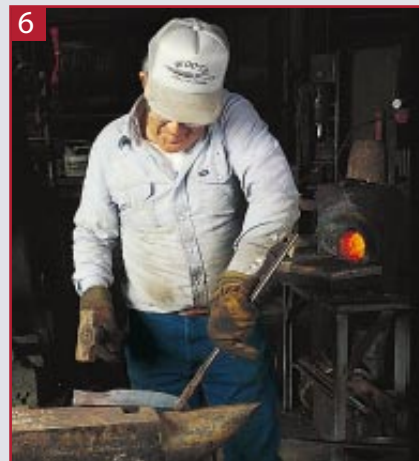
6 Cut the blade to final shape and hand-forge to add finer details.



7 Remove the excess steel and the decarburized surface metal. Pendray is using an electric belt grinder for this step.

8 Cut grooves and drill holes into the surface of the blade to create Mohammed's ladder and rose patterns, if desired. Forge the blade flat again and polish the surface to give the blade its near final form.

9 Etch blade surface with an acid to bring out the pattern; the softer steel darkens, and the harder steel appears as brighter lines.





FINISHED BLADE
shows the Mohammed's
ladder and rose patterns.

weak bands, and after six to eight cycles the bands became quite distinct.

In one test, we cranked up the heat well beyond the A_{cm} —to 1,200 degrees C, just below the melting point of the steel—and held it there for 18 hours. Subsequent thermal cycling of the steel did not bring back the bands of cementite particles. Calculations show that this high-temperature treatment completely removes the microsegregation of impurity atoms by diffusion.

Pendray and I also tried carefully controlled experiments in which we left out the impurity elements altogether. Even after many cycles of heating and slow cooling, these ingots did not produce clusters of carbide particles or bands. When we added the impurity elements to the same ingot and put it through the heating and cooling cycles, the bands appeared.

Our re-creation of the Damascus blade helps us to answer another question: How did the ancient smiths generate the Mohammed's ladder pattern? Our work supports one theory proposed in the past—that the ladder rungs were produced by cutting grooves across the blades. The ladder pattern visible in the bottom photograph above was made by incising small trenches into the blade after it had been forged to near its final thickness [see illustration 8 above], then subsequently forging it to fill in the trenches. Such forging reduces the spacing between light and dark bands on the final surface, especially along the edges of the trenches. The round configuration between the rungs, known as the rose pattern, is also known from older scimitars. It comes from shallow holes drilled in the blade at the same time the grooves are cut.

Why was the art of making these weapons lost sometime around two centuries ago? Perhaps not all iron ores from India contained the necessary carbide-forming elements. The four ancient Moser blades that we studied all contained vanadium impurities, which is probably why the bands formed in these steels. If changes in world trade resulted in the arrival of ingots from India that no longer contained the required impurity elements, bladesmiths and their sons would no longer be able to make the beautiful patterns in their blades and would not necessarily know why. If this state of affairs persisted, after a generation or two the secret of the legendary Damascus sword would have been lost. It is only now, thanks to a partnership between science and art, that the veil has been lifted from this mystery.

The Author

JOHN D. VERHOEVEN is an emeritus Distinguished Professor of Materials Science and Engineering at Iowa State University. He has been interested in the mystery of wootz Damascus swords since he was a graduate student at the University of Michigan. In 1982 he began research experiments on re-creating Damascus steel. The work, which was primarily a hobby, grew into a serious effort as he collaborated with blacksmith Alfred H. Pendray over many years.

Further Information

HISTORY OF METALLOGRAPHY: THE DEVELOPMENT OF IDEAS ON THE STRUCTURE OF METALS BEFORE 1890. Cyril S. Smith. MIT Press, 1988.
ON DAMASCUS STEEL. Leo S. Figiel. Atlantis Arts Press, 1991.
ARCHAEOLOGY: THE KEY ROLE OF IMPURITIES IN ANCIENT DAMASCUS STEEL BLADES. J. D. Verhoeven, A. H. Pendray and W. E. Dauksch in *JOM: A Publication of the Minerals, Metals and Materials Society*, Vol. 50, No. 9, pages 58–64; September 1998. Available at www.tms.org/pubs/journals/JOM/9809/Verhoeven-9809.html on the World Wide Web.

*Extensions to fiber optics
will supply network capacity
that borders on the infinite*

by Gary Stix, *staff writer*

THE TRIUMPH OF THE LIGHT



BERND AUERS

W

as it Britney Spears or Fatboy Slim? The network administrators at Kent State University had not a clue. All they did know last February was that “Rockefeller Skank” and thousands of other downloading hits had gotten intermingled with e-mails from the provost and research data on genetic engineering of *E. coli* bacteria. The university network slowed to a crawl, triggering a decision to block access to Napster, the music file-sharing utility.

As demand for network capacity soars, the Napster craze may mark the opening of only the first of many floodgates. Venture capitalists, in fact, have wagered billions of dollars on technologies that may help telecommunications companies counter the prospect that a video Napster capable of downloading anything from *Birth of a Nation* to *Rocky IV* might bring down the entire Internet.

PowerPoint slides at industry conferences emphasize why the deluge is yet to come. Video Napster is just one hypothesis. A trillion bits a second—the average traffic on the Internet’s backbones, its heaviest links—may fulfill less than a thousandth of future requirements. Online virtual reality could overwhelm the backbones with up to 10 petabits a second, 10,000 times more than today’s traffic. (A petabit is a quadrillion bits, a one with 15 trailing zeros.) Computers that share one another’s computing power across the network—what is called metacomputing—might require 200 petabits.

If these scenarios materialize—and, to be sure, people have been tapping their feet for virtual reality for more than a decade—the only transmission medium that could come close to meeting the seemingly infinite demand is optical fiber, the light pipes trumpeted in commercial interludes about the “pin drop” clarity of a phone connection. Fiber links can channel hundreds of thousands of times the bandwidth of microwave transmitters or satellites, the nearest competitors for long-distance communications. As one wag pointed out, the only other technology that comes close to matching this delivery capacity is a panel truck full of videos.

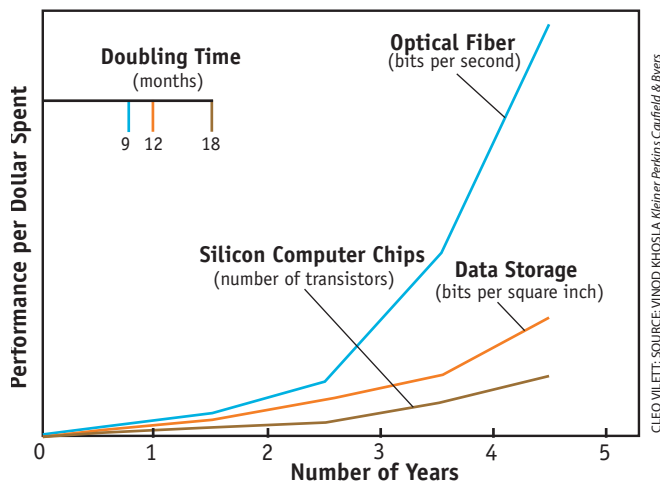
The race to augment the fiber content of the world’s networks has started. Every day installers lay enough new cable to circle the earth three times. If improvements in fiber optics continue, the carrying capacity of a single fiber may reach hundreds of trillions of bits a second just a decade or so from now—and some technoidal utopians foresee the eventual arrival of the vaunted petabit mark. To overcome that barrier, however, will require both fundamental breakthroughs and the deployment of technologies that are still more physics experiments than they are equipment ready to be slotted into the racks on nationwide phone and data networks.

More immediately, new photonic technologies, which literally use mirrors instead of electrons for rerouting signals, will make a whole class of electronic switching systems obsolete. Even now the transmission speeds of the most advanced networks—at 10 billion bits a second—threaten to choke the processing units and memory of microchips in existing switches. As the network becomes faster than the processor, the cost of using electronics with optical transmissions skyrockets. The gigabit torrent contained in a wavelength of light in the fiber must be broken up into slower-flowing data streams that can be converted to electrons for processing—and then reaggregated into a fast-flowing river of bits. The equipment for going from photon to electron and back to photon not only slows traffic on the superhighway but makes equipment costs soar.

While network designers contemplate the prospect of machine overload, hundreds of companies, big and small, now grapple with creating networks that can exploit fiber’s full bandwidth by transmitting, combining, amplifying and switching wavelengths without ever converting the signal to electrons. Photonics is at a stage that electronics experienced 30 years ago—with the development and integration of component parts into larger systems and subsystems. A rising tide of venture capital has emerged to support these endeavors. In the first nine months of 2000, venture funding for optical networking totaled \$3.4 billion, com-

WAVELENGTH carrying 40 billion bits per second flows through this yellow fiber, provided by start-up Enkido, founded by Nayel Shafei.

paring with the \$1.5 billion spent on silicon networking in the same period. The technology is still in its infancy, but the industry is growing rapidly.



FIBER LEADS in performance improvements. The number of bits a second (a measure of fiber performance) doubles every nine months for every dollar spent on the technology. In contrast, the doubling time for the number of transistors on a computer chip occurs every 18 months—a trend known as Moore’s law. Over a five-year period, optical technology far outpaces silicon chips and data storage.

pared with \$1.5 billion for all of 1999, although this pace may have slowed in recent months. The success of a stock like component supplier JDS Uniphase stems in part from the perception that its edge in integrated photonics could make it the next Intel.

Investment in optical communications already yields payoffs, if fiber optics is matched against conventional electronics. The cost of transmitting a bit of information optically halves every nine months, as against 18 months to achieve the same cost reduction for an integrated circuit (the latter metric is famous as Moore's law). "Because of dramatic advances in the capacity and ubiquity of fiber-optic systems and subsystems, bandwidth will become too cheap to meter," predicts A. Arun Netravali, president of Lucent Technologies's Bell Laboratories in a recent issue of *Bell Labs Technical Journal*.

Identical forecasts about a free resource eventually came to haunt the nuclear power industry. And the future of broadband networking, in which a full-length feature film would be transmitted as readily as an e-mail message, is still not a sure bet. A decade ago telecommunications providers and media companies started preparing for the digital convergence of entertainment and networking. Five hundred channels. Video on demand. We're still waiting. Meanwhile the Internet, once viewed as a quaint techno sideshow for the gov-

ernment and schoolkids, has transmuted into the network that ate the world. E-mails and Web sites have triumphed over Mel Gibson and Cary Grant.

And Then There Was Light

Prospects of limitless bandwidth—the basis for speculations about networked virtual reality and high-definition videos—are of relatively recent vintage. AT&T and GTE deployed the first optical fibers in the commercial communications network in 1977, during the heyday of the minicomputer and the infancy of the personal computer. A fiber consists of a glass core and a surrounding layer called the cladding. The core and cladding have carefully chosen indices of refraction (a measure of the material's ability to bend light by certain amounts) to ensure that the photons propagating in the core are always reflected at the interface of the cladding. The only way the light can enter and escape is through the ends of the fiber. To understand the physics behind how a fiber works, imagine looking into a still pool of water. If you look straight down, you see the bottom. At viewing angles close to the water, all that is perceived is reflected light. A transmitter—either a light-emitting diode or a laser—sends electronic data that have been converted to photons over the fiber at a wavelength of between 1,200 and 1,600 nanometers.

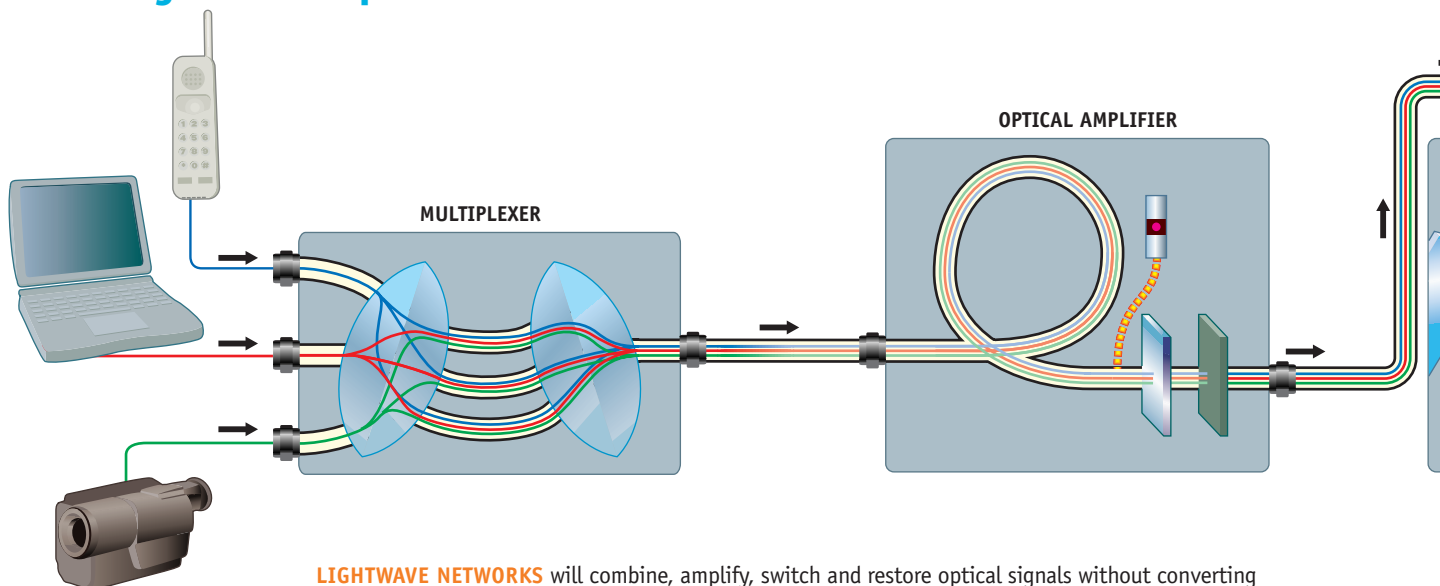
Today some fibers are pure enough

that a light signal can travel for about 80 kilometers without the need for amplification. But at some point the signal still needs to be boosted. The next significant step on the road to the all-optical network came in the early 1990s, a time when the technology made astounding advances. It was then that electronics for amplifying signals were replaced by stretches of fiber infused with ions of the rare-earth element erbium. When these erbium-doped fibers were zapped by a pump laser, the excited ions could revive a fading signal. The amplifiers became much more than plumbing fixtures for light pipes. They restore a signal without any optical-to-electronic conversion and can do so for very high speed signals sending tens of gigabits a second. Perhaps most important, however, they can boost the power of many wavelengths simultaneously.

This ability to channel multiple wavelengths enabled the development of a technology that has helped drive the frenzy of activity for optical-networking companies in the financial markets. Once you can boost the strength of multiple wavelengths, the next thing you want to do is jam as many wavelengths as possible down a fiber, with a wavelength carrying as much data as possible. The technology that does this has a name—dense wavelength division multiplexing (DWDM)—that is a paragon of technospeak.

DWDM set off a bandwidth explo-

Technologies for All-Optical Networks



LIGHTWAVE NETWORKS will combine, amplify, switch and restore optical signals without converting them to an electronic transmission for processing. A dense wavelength division multiplexer (DWDM) will take different wavelengths of light and place them on a single fiber connection. An optical ampli-

LAURIE GRACE

sion. With the multiplexing technology, the capacity of the fiber expands by the number of wavelengths, each of which can carry more data than could be handled previously by a single fiber. Nowadays it is possible to send 160 frequencies simultaneously, supplying a total bandwidth of 400 gigabits a second over a fiber. Every major telecommunications carrier has deployed DWDM, expanding the capacity of the fiber that is in the ground and spending what could be less than half of what it would cost to lay new cable, while the equipment gets installed in a fraction of the time it takes to dig a hole.

In the laboratory, meanwhile, experiments point toward using much of the capacity of fiber—dozens of individual wavelengths, each modulated at 40 gigabits or more a second, for effective transmission rate of a few terabits a second. (One company, Enkido, has already deployed commercial links containing 40-gigabit-a-second wavelengths.) The engorgement of fiber capacity will not stop anytime soon and could reach as high as 300 or 400 terabits a second—and, with new technical advances, perhaps exceed the petabit barrier.

The telecommunications network, however, does not consist of links that tie together point A and point B—switches are needed to route the digital flow to its ultimate destination. The enormous bit conduits that now populate laboratory testbeds will flounder if the light streams

are routed using conventional electronic switches. Doing so would require a multiterabit signal to be converted into dozens or hundreds of lower-speed electronic signals. Finally, switched signals would have to be reconverted to photons and reagggregated into light channels that are then sent out through a designated output fiber.

The cost and complexity of electronic switching have prompted a mad scramble to find a means of redirecting either individual wavelengths or the entire light signal in a fiber from one pathway to another without the optoelectronic conversion. Research teams, often inhabiting tiny start-ups, fiddle with microscopic mirrors, liquid crystals and fast lasers to try to devise all-optical switches [see “The Rise of Optical Switching,” on page 88].

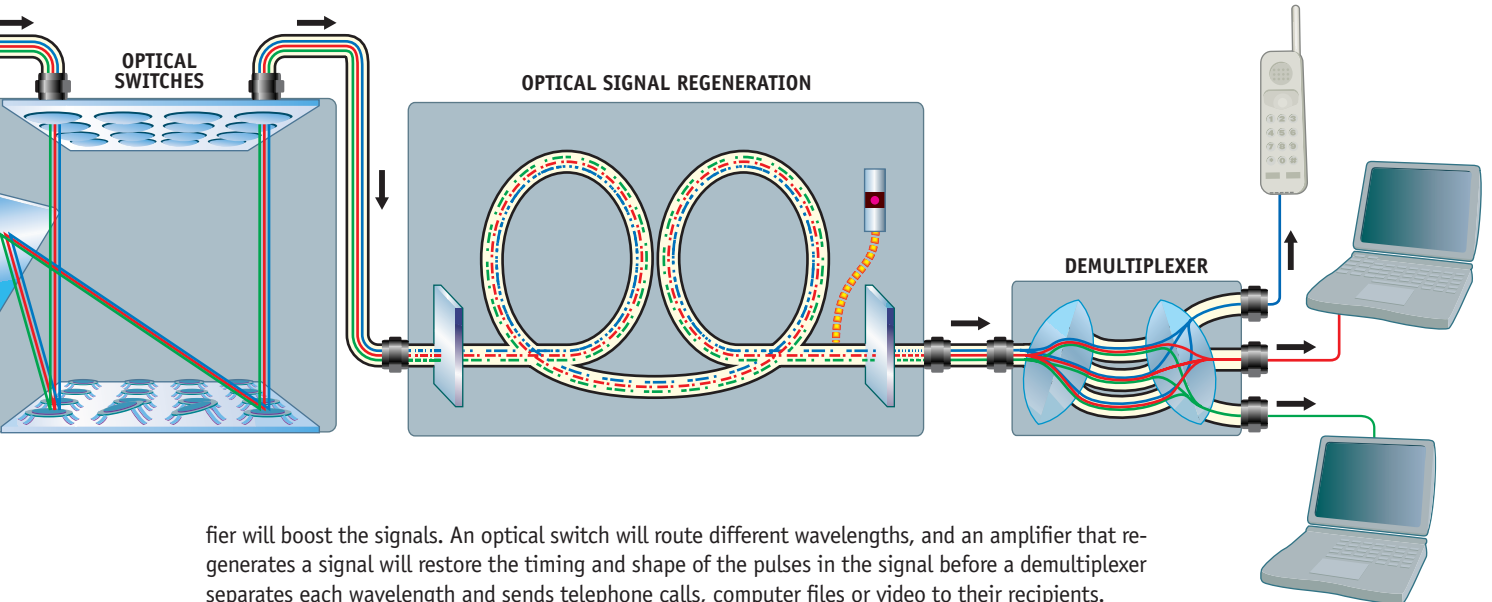
All-optical switching, however, will differ in fundamental ways from existing networks that switch individual chunks of data bits, such as IP (Internet Protocol) packets. It is an easy task for the electronics in routers or large-scale telephone switches to read on a packet the address that denotes its destination. Photonic processors, which are at about the same stage of development that electronics was in the 1960s, have demonstrated the ability to read a packet only in laboratory experiments.

Optical switches heading to the marketplace hark back to earlier generations of electronic equipment. They will switch

a circuit—a wavelength or an entire fiber—from one pathway to another, leaving the data-carrying packets in a signal untouched. An electronic signal will set the switch in the right position so that it directs an incoming fiber—or wavelengths within that fiber—to a given output fiber. But none of the wavelengths will be converted to electrons for processing.

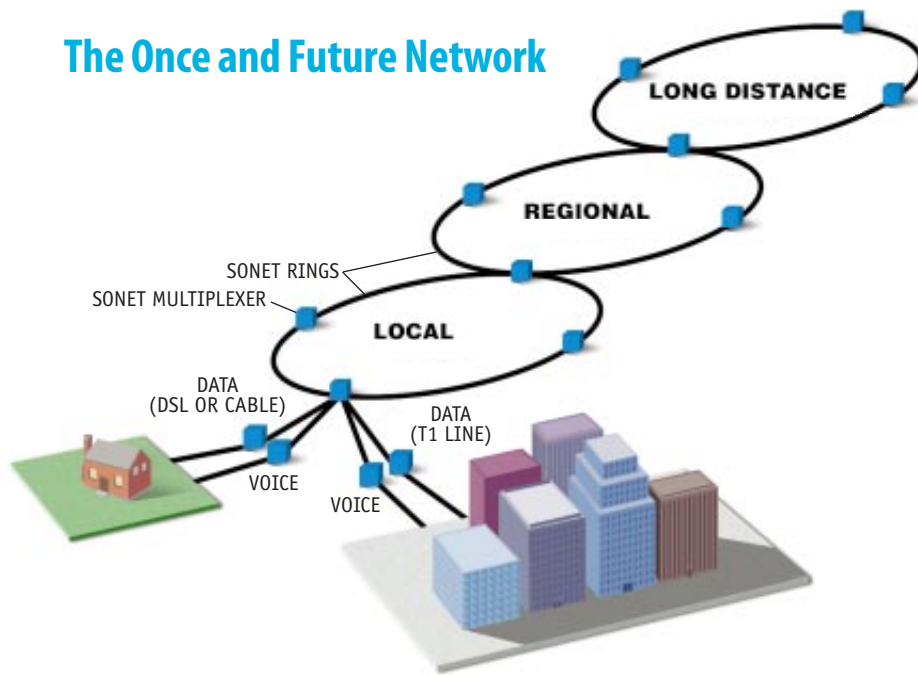
Optical circuit switching may be only an interim step, however. As networks get faster, communications companies may demand what could become the crowning touch for all-optical networking, the switching of individual packets using optical processors [see “Routing Packets with Light,” on page 96].

With the advent of optical packet switching, individual packets will still need to get read and routed at the edges of optical networks—on local phone networks near the points where they are sent or received. For the moment, that task will still fall to electronic routers from companies such as Cisco Systems. Even so, the evolution of optical networking will promote changes in the way networks are designed. Optical switching may eventually make obsolete existing lightwave technologies based on the ubiquitous SONET (Synchronous Optical Network) communications standard, which relies on electronics for conversion and processing of individual packets. And this may proceed in tandem with the gradual withering away of Asynchronous Transfer Mode

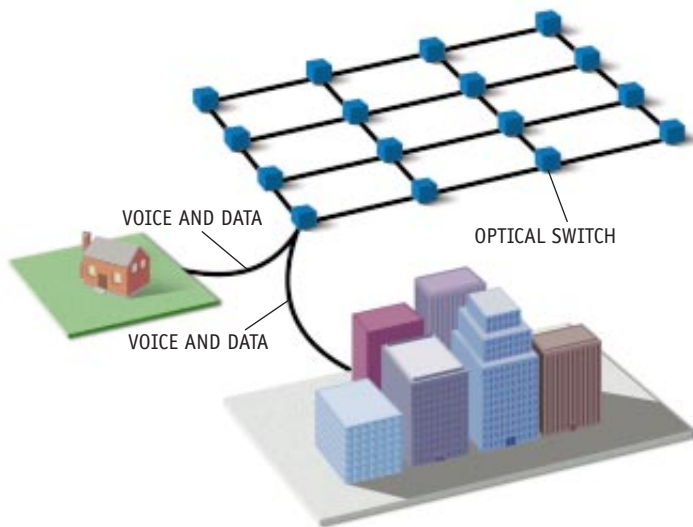


fier will boost the signals. An optical switch will route different wavelengths, and an amplifier that regenerates a signal will restore the timing and shape of the pulses in the signal before a demultiplexer separates each wavelength and sends telephone calls, computer files or video to their recipients.

The Once and Future Network



TODAY'S ADVANCED NETWORKS maintain mostly separate electronic connections for voice and data and achieve reliability using rings based on the Synchronous Optical Network (SONET) communications standard: if one link is cut, traffic flows down the other half of the ring. The SONET multiplexer aggregates traffic onto the ring.



TOMORROW'S NETWORKS will channel all traffic over the same fiber connection and will provide redundancy using the Internet's mesh of interlocking pathways: when a line breaks, traffic can flow down several alternating pathways. Optical switching will become the foundation for building these integrated networks.

(ATM), another phone company standard for packaging information.

In this new world, any type of traffic, whether voice, video or data, may travel as IP packets. A development heralded in telecommunications for at least 20 years—the full integration of voice, video and data services—will be complete. “It’s going to be a data network, and everything else, whether it’s voice

or video, will be applications traveling over that data network,” says Robert W. Lucky, a longtime observer of the telecommunications scene and director of research for the technology development firm Telcordia.

When you ring home on Mother’s Day, the call may get transmitted as IP packets that move on a Gigabit Ethernet, a made-for-the-superhighway ver-

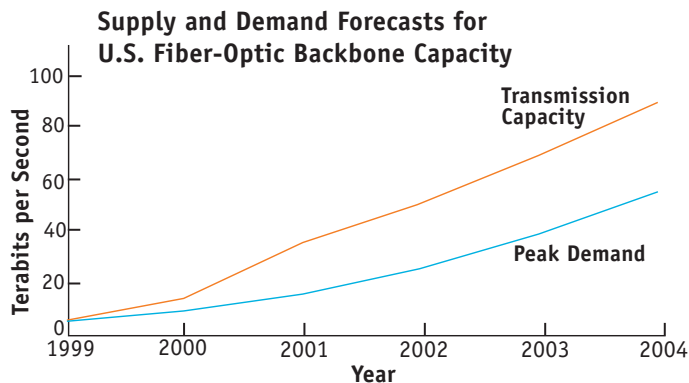
sion of the ubiquitous local-area network (LAN). Gigabit Ethernet would in turn ride on wavelength-multiplexed fiber. Critics of this approach question whether such a network would provide ATM and SONET’s quality of service and their ability to reroute connections automatically when a fiber link is cut.

Life would be simpler, though. The phone network would become just one big LAN. You could simply slot an Ethernet card into a computer, telephone or television, a far cheaper and less time-consuming solution than installing new SONET hardware connections. Some companies are even now preparing for the day when IP reigns. Level 3 Communications, a carrier based in Denver, has laid an international fiber network stretching more than 20,000 miles in both the U.S. and overseas. Although the network still relies on SONET, CEO James Q. Crowe foresees a day when these costly legacies of the voice network will wither into nothingness. “It will be IP over Ethernet over optics,” Crowe says.

Home Light Pipes

Even if network engineers can pare down the stack of protocols that weighs heavy on today’s network, they must still contend with the need to address the “last mile” problem, getting fiber from the curbside utility box into the TV room and home office. Some builders now lay out new housing projects with fiber, presaging the day when households routinely get their own wavelength connection. But cost still hangs over any discussion of fiber to the home. Until recently, advanced optical-networking equipment, such as DWDM, was too expensive to consider for deployment on regional phone networks. Extending the equipment into a wall panel of a split level—at perhaps \$1,500 a line—still costs more than all but a few are willing to pay. Most people have yet to take delivery of their first megabit connection. So it remains unclear when the time will come when the average household will need the gigabits to project themselves holographically into a neighbor’s house rather than just picking up the phone.

Dousing “Help me, Obi-Wan Kenobi” fantasies, engineers are confronting an array of nettlesome technical problems before a seamless all-optical network can become commonplace. Take one example: even with lightwave switching in



FUTURE BANDWIDTH REQUIREMENTS	
Applications	Backbone Bandwidth (terabits per second)
Online virtual reality	1,000 to 10,000
3-D holography/telepresence	30,000 to 70,000
Metacomputing	50,000 to 200,000
Web agents	50,000 to 200,000

1 terabit = 1 trillion bits

DEMAND GAP for optical-fiber backbones—the most heavily used links—emerges in a study by consultant Adventis that shows that supply will overmatch demand. Yet new applications such as virtual

reality and metacomputing could require huge increments in optical bandwidth above the few terabits per second currently needed to satisfy demand on U.S. communications backbones.

place, one critical part of the network requires conversion to electronics. About every 160 kilometers, a wavelength has to be converted back to an electronic signal to restore the shape and timing of individual pulses within the vast train of bits that occupy each lightwave.

Equipment suppliers also struggle mightily with electronics envy. Component suppliers such as JDS Uniphase labor on methods to build modules that combine lasers, fiber and gratings (which separate wavelengths). Building photonic integrated circuits remains difficult. Photons have no charge, as the negatively charged particles called electrons do. So there is no such thing as a charge-storage device, a photonic capacitor, that will store indefinitely the photons that represent zeros and ones. Moreover, it is difficult to build photonic circuitry as small as electronic integrated circuits, because the wavelength of infrared light used in fiber-optic lasers is about 1.5 microns, which places limits on how small you can make a component. Electronic circuits reached that dimension more than a decade ago.

The good news is that companies both small and big are now trying to solve problems such as signal restoration, and a pot of venture money exists to fund them. The field, which has taken on the same aura that genomics now holds and dot-coms once did, has become an exemplar of a new, hyperventilating model of research. Tiny development houses proceed until they can furnish some proof that they can make good on their promises, and then they are bought out by a Nortel, Cisco or Lucent.

“It’s a crazy world,” says Alastair M. Glass, director of photonics at Lucent. “Anyone can go out with the dumbest

ideas and get funding for them, and maybe they’ll be bought for big bucks. And they’ve never made a product.” Glass adds: “This has never happened in the past. Part of it is because companies need people, so they’re buying the people. But other times they’re buying the technology because they don’t have it in the house, and sometimes they don’t know what they’re buying.” From idea to development happens fast: a 1998 paper in *Science* about a “perfect mirror,” a dielectric (insulating) material that reflects light at any angle with little loss of energy, inspired the founding of a company that wishes to create a hollow fiber whose circumference is lined with the reflector. The fibers may increase capacity 1,000-fold, one company official claims.

Will Anybody Come?

What can be done with all this bandwidth? Lucent estimates that if the growth of networks continues at its current pace, the world will have enough digital capacity by 2010 to give every man, woman and child, whether in San Jose or Sri Lanka, a 100-megabit-a-second connection. That’s enough for dozens of video connections or several high-definition television programs. But does each !Kung tribesman in the Kalahari Desert really need to download multiple copies of *The Gods Must Be Crazy*?

Despite estimates of Internet traffic doubling every few months, some industry watchers are not so sure about infinite demand for infinite bandwidth. Adventis, a Boston-based consultancy, foresees only 15 to 20 percent of home Internet users obtaining broadband ac-

cess—either cable modems or digital subscriber lines—by 2004. Moreover, storing frequently accessed Web pages on a server will reduce the burden on the network. In the U.S., according to the firm’s estimate, nearly 40 percent of existing fiber capacity will go unused in 2004, whereas in Europe almost 65 percent will stay dormant. The notion of a capacity glut is by no means a consensus view, however.

In the end, terabit or petabit networking will probably emerge only once some as yet unforeseen use for the bandwidth reveals itself. Like the World Wide Web, originally a project to help particle physicists more easily share information, it may arrive on a tangent, not from a big media company’s focused attempt to repackage networked virtual reality. Vinod Khosla, a venture capitalist with Kleiner Perkins Caufield & Byers, talks of the promise of projects that pool together computers that may be either side by side or distributed across the globe. Metacomputing can download Britney Spears and Fatboy Slim, or it can comb through radio telescope data in search of extraterrestrial life. Khosla sees immense benefit in using this model of networked computing for business, tying together machines to work on, say, the computational fluid dynamics of a 1,000-passenger jumbo jet.

So efforts to pick through the radio emissions from billions and billions of galaxies may yield useful clues about what on earth to do with a network pulsing a quadrillion bits a second. 54

FURTHER INFORMATION

See www.lightreading.com for a wealth of coverage on new technologies and on companies involved in optical networking.

Replacing electronic switches with purely optical ones will become the technological linchpin for networks that transmit trillions of bits each second

In the 19th century, pressures to extend the burgeoning rail network meant that trains would transport passengers even before a link between two cities had reached completion: the laying of tracks often preceded the building of bridges over major rivers. Thus, a section of track would end on one bank, and a new section would start from the other side. Rail passengers would transfer onto a ferry, cross the river by boat, then board different trains on the other bank to continue their journey. This tedious process limited the speed at which people and goods could cover the distance between two points. When bridges finally spanned the rivers, travel times diminished by hours. This trend continues even into the present: dramatic time savings accrued with the opening of the tunnel under the English Channel.

A similar story occurs today in the world of optical networking. Data often move over optical fibers that connect one location to another. Yet the switches that redirect traffic down different pathways of the network are the equivalent of 19th-century ferries. They use electrical signals instead of light, and so voice, video or data transmission must exit the optical highway onto a low-speed interchange when transferring from one route to another. Light from an optical fiber has to be converted into an electric current to pass through the switch and then be reconverted into light to continue its journey along another length of fiber, a process that adds both delay and cost.

Increasingly, this optical-electronic-optical conversion has been exacerbated by another problem—the so-called electronic bottleneck that arises because of different rates of technological progress in electronics and photonics (as optical-networking technology is known). While the speed of electronics continuously improves, the performance of photonics gets better at an even more rapid pace. As a result of this growing disparity, photonics will swiftly outpace the performance of electronics, and the fastest processors will be ever less able to keep up with the flood of bits reaching an electronic switch.

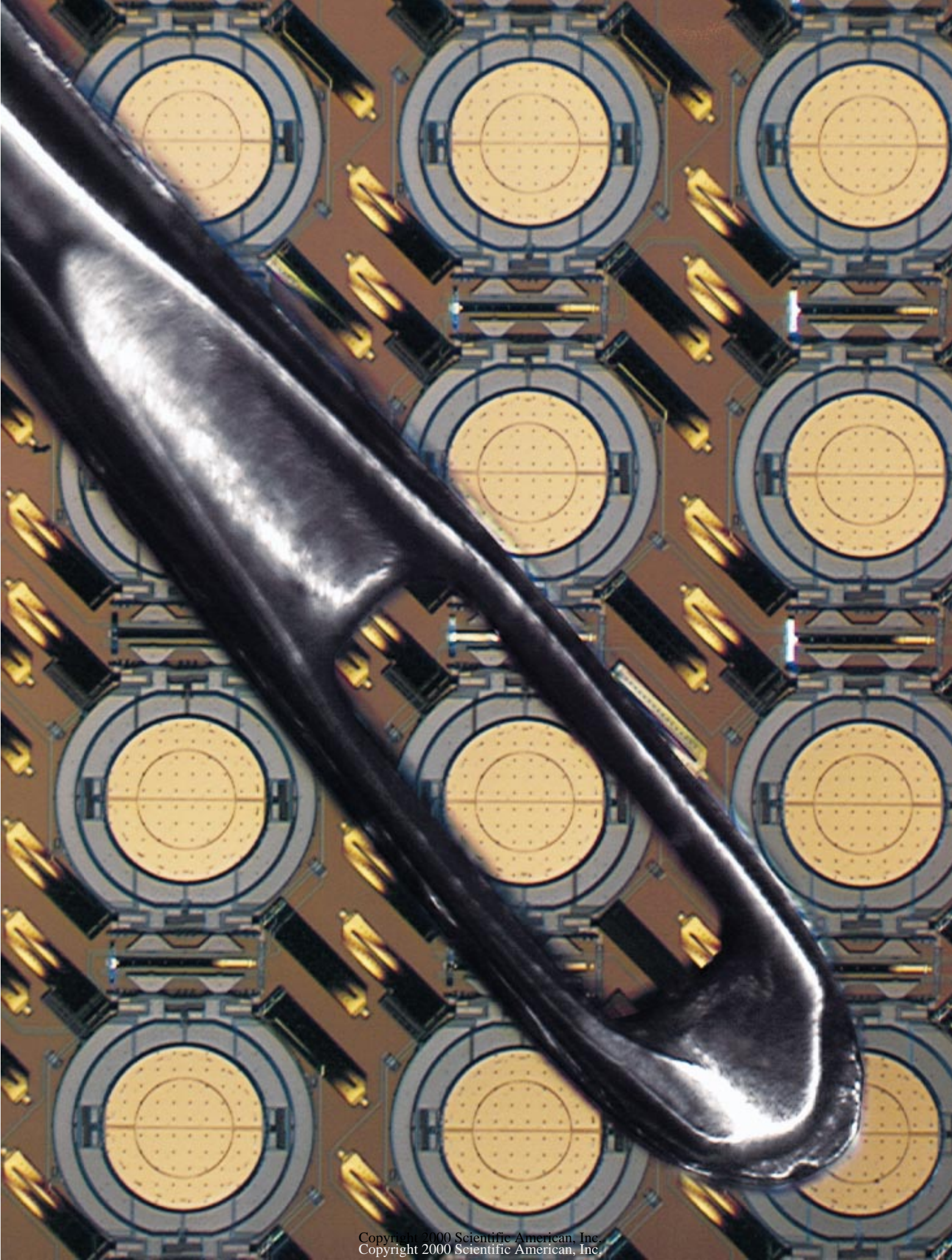
Consequently, in the next generation of telecommunications networks, it will not be enough to use fiber to transmit information from one point to another over long distances. Switching will have to be done optically to avoid the optoelectronic conversion that creates ferrylike choke points in the network. In response to this chal-

MICROMIRRORS, small enough to be seen through the eye of a needle, reflect light-waves from one optical fiber to another.

THE RISE OF

by David J. Bishop, C. Randy Giles and Saswato R. Das

OPTICAL SWITCHING



lenge, photonic switches—built with microscopic mirrors, bubbles or other novel technologies—are now coming of age.

A switch defines the very essence of the word “network.” When Alexander Graham Bell invented the telephone, he had a direct hot line to his assistant, Thomas A. Watson. As telephones became popular, electrical signals needed to be routed down varying paths. Initially, telephone operators connected one circuit with another manually. With the advent of electromechanical relays, automatic switches came into existence. Today’s phone switches and the data switches called routers are essentially specialized electronic computers that track transmissions and send them on their way, whether they are voice, video or the packets of data that make up a digital message or file.

Optical switching, in contrast, portends a fundamental change in the design of telecommunications networks. The sheer volume of traffic means that the switches will be unable to direct individual phone calls or e-mail messages racing through a fiber at billions or trillions of bits every second. Instead they will channel the tens or hundreds of wavelengths in every fiber, each wavelength packaging together thousands of calls or billions of data bits, and send them to one or more of the hundreds of output fibers. Alternatively, the wavelength itself may not be transferred. Rather the information in that wavelength may get imprinted on a different wavelength in an output fiber. Only when phone calls or packets of Internet data arrive near their destination will

they be switched individually by electronic processors—and even those switches may one day eschew the use of electrons as new photonic technologies develop [see “Routing Packets with Light,” on page 96].

Terabits per Second

In the meantime, the technical challenge of switching whole fibers or individual wavelengths within those fibers may prove daunting. The cross-connects, as the optical wavelength switches are called, must route incoming wavelengths of light to any available output fiber. This does not seem very difficult at first—after all, telephone switches do it very efficiently whenever a telephone subscriber wants to call another number in town. But optical fibers have a lot more traffic: an optical switch that has 100 input optical channels, carrying 10 gigabits each, will need to handle a terabit (a trillion bits) per second.

Until very recently, optical switches could accommodate nowhere near this capacity. Back in 1992 Bell Laboratories, the research and development arm of Lucent Technologies, demonstrated a cross-connect made of lithium niobate that could take 16 incoming fibers and direct them to 16 outgoing fibers—and more recently, it tested a 48-input/48-output switch. Such switches had too few input-to-output channels for all but the most limited of emerging optical-networking applications.

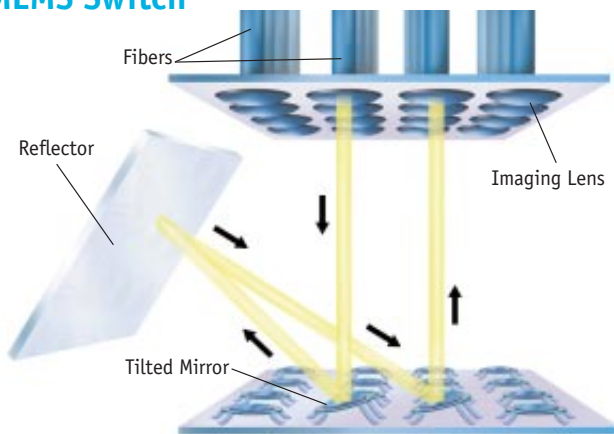
Today many companies have taken a dramatically different approach to optical cross-connects through the use of

MEMS (MicroElectroMechanical Systems) technology. MEMS is a new process for device fabrication, which builds “micromachines” that are finding increasing acceptance in many industries, ranging from telecommunications to automotive, aerospace, consumer electronics and others. They are, in essence, a mechanical integrated circuit. MEMS technology uses photolithographic and etching processes similar to those employed in making large-scale integrated circuits—devices that are deposited and patterned on the surface of a silicon wafer. In MEMS, oxide layers are etched away to sculpt the device’s structural elements. Instead of creating transistors, though, lithographic processes build devices a few tens or hundreds of microns in dimension that move when directed by an electrical signal.

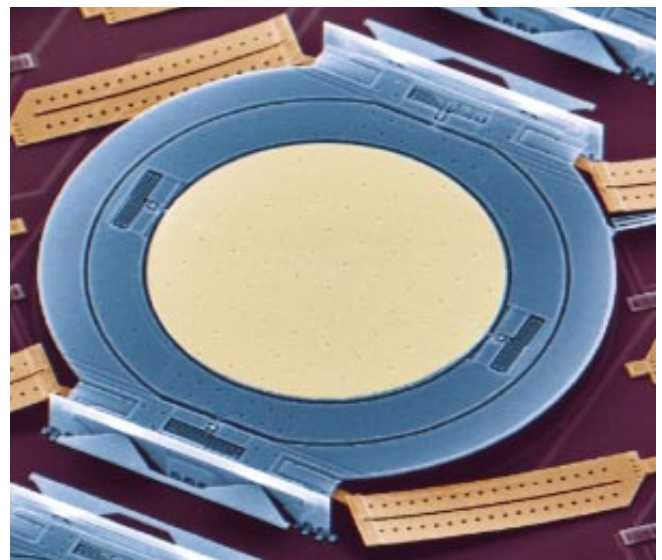
Lucent and companies such as Optical Micro Machines, Calient Networks and Xros (now a part of Nortel Networks) have selected MEMS for building optical cross-connects because it yields small, inexpensive devices that can be incorporated with very large scale integrated circuits. Most important, MEMS can yield micromachines that are robust, long-lived and scalable to large numbers of devices on a wafer. The technology is also exceptionally well matched to optics applications—because it easily accommodates the need to expand or reconfigure the number of pathways through the switch.

To direct a wavelength along a pathway in the network, the MEMS switch uses tiny mirrors positioned so that each is illuminated by one or more of the

MEMS Switch

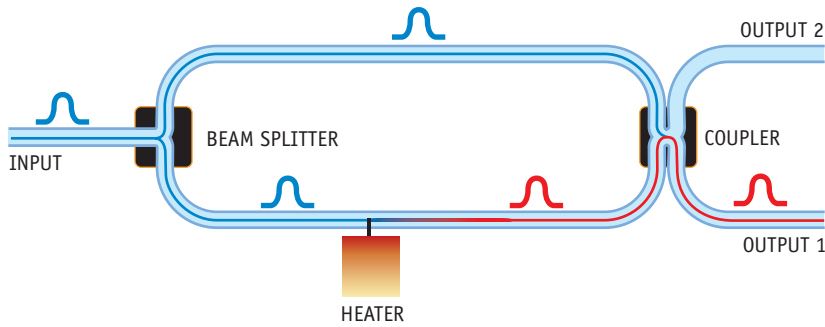


TILTED MIRROR in a MicroElectroMechanical System (MEMS) switch (*photograph shows close-up*) bounces a lightwave from an incoming fiber onto a reflector, off another mirror and into an outgoing fiber.



LAURIE GRACE (left); LUCENT TECHNOLOGIES (right)

Thermo-optic Switch



HEAT is applied in a thermo-optic switch to one of two waveguides into which the light passes after going through a beam splitter. The temperature rise lengthens the pathway slightly, thereby changing the phase of the light. When the two beams recombine, the light exits the first output pathway. Left unheated, the light leaves the other output path (*not shown*).

multiple wavelengths carrying a stream of information within a single fiber. In one type of MEMS switch, the mirrors tilt from top to bottom or side to side to enable any of the multiple wavelengths within 256 incoming fibers to be passed to any of 256 outgoing fibers.

To understand how the switch works, imagine that you are sitting in a room with many windows. If sunlight streams in through one window and there is a movable mirror inside the room, manipulating the mirror allows you to reflect that sunlight through any of the other windows. In the case of a MEMS switch, a stream of photons in a wavelength coming in through an input port hits a series of MEMS mirrors that send it out through one of many output ports, depending on which route it is supposed to take.

More concretely, software in the switch's processor makes a decision about where an arriving stream of photons should go. It sends a signal to an electrode on the chip's surface that generates an electrical field that tilts the mirrors. The incoming lightwave gets filtered into separate wavelengths, each of which hit one of 256 tilted input mirrors. The wavelengths bounce off the input mirrors and get reflected off another mirror onto output mirrors that then direct the wavelength into another fiber [see illustration on opposite page]. The entire process lasts a few milliseconds, fast enough for the most demanding switching applications.

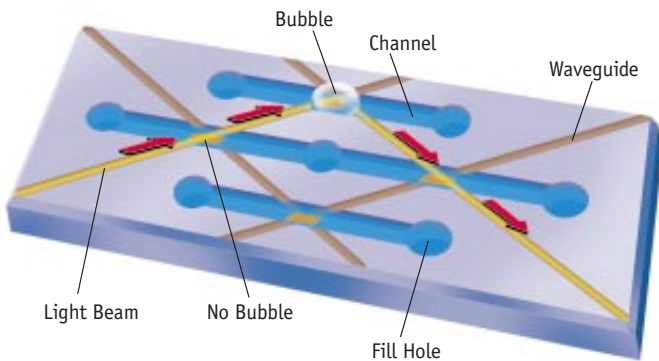
The size of the individual switching elements makes the MEMS approach extraordinarily attractive. Each mirror in one MEMS switch is half a millimeter

in diameter, about the size of the head of a pin. Mirrors rest one millimeter apart, and all 256 mirrors are fabricated on a 2.5-centimeter-square piece of silicon. The entire switch is about the size of a grapefruit. The set of mirrors that make up the switch is about 32 times denser than the equivalent components in an electronic switch. And with no need for signal processing or making the optoelectronic conversion, such switches will provide up to a 100-fold reduction in power consumption over electronic switches.

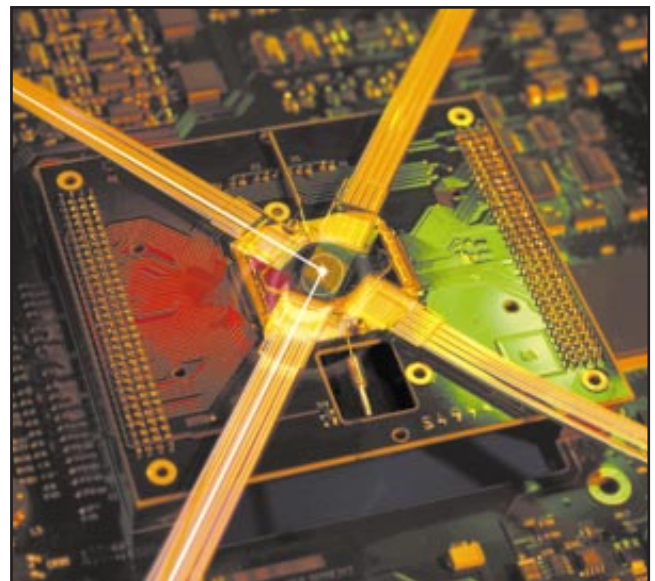
Standard silicon-circuit manufacturing processes make the technology cost-effective. And silicon mirrors afford greater stability than if the mirrors were fabricated from metal. One novel step in making the mirrors uses self-assembly, a process that takes its name from the way amino acids in protein molecules fold themselves into three-dimensional shapes. In the final stages of manufacture, tiny springs on the silicon surface release the mirrors and a frame around each one lifts them and locks them in place, positioning them high enough above the surface to allow for a range of movement.

The design of the mirror array uses one mirror for input and one for output. This approach entails rigorous mechanical demands, because the mirrors must be tilted to different angles. But the use of silicon-fabrication technology results in stiffer mirrors that are less prone to drifting out of alignment. And superior software-control algorithms let the individual element be manipulated precisely. This design will promote the building of much larger switches. The design of ear-

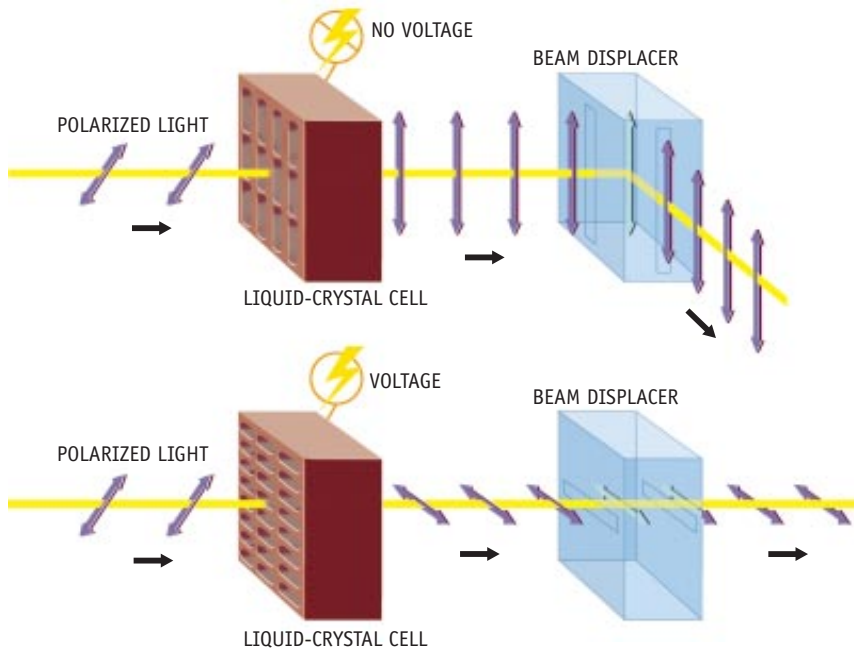
Bubble Switch



BLOWING a microscopic bubble into the junction of a switch using an inkjet-printing head causes the light to make a right turn (*photograph*). The absence of a bubble in the junction of this Agilent switch lets the beam proceed straight. Liquid from fill holes enters channels that intersect with junction points, where the inkjet head (*not shown*) inflates a bubble.



Liquid-Crystal Switch



CELL of liquid crystals in a switch realigns obliquely polarized light into either a vertical or a horizontal orientation, depending on whether a voltage is applied. A prismlike beam displacer lets vertically polarized light pass through to the right; horizontally polarized light moves straight through the crystal. The type of liquid crystal used in a switch is shown below.



lier lithium niobate switches required far more switching elements—a number equal to the square of the number of inputs or outputs, which would have made it too cumbersome to design a large switch. This is in contrast to the simpler architecture used with MEMS switches.

The ability to make switches as large

or as small as needed is of paramount importance to telecommunications carriers, who must be able to accommodate the accelerating growth in demand. The first introduction of a large MEMS switch, the Lucent LambdaRouter, occurred in July 2000. It offered more than 10 terabits per second of total switching

capacity, 10 times the traffic over the most heavily used segments of the Internet. Each of the 256 input-to-output channels can support speeds of 320 gigabits per second—128 times faster than current electronic switches. Eventually such switches might support the petabit (quadrillion-bit) systems that are not very far over the horizon.

Beyond MEMS

More researchers now pursue MEMS than any other optical-switching technology. Over 10 MEMS start-ups populate Silicon Valley alone. But it is by no means the only approach. Another area of investigation focuses on photonic waveguides. Like MEMS, switches built from waveguides use many simple components that control the trajectory of lightwaves, allowing them to be sent down alternative pathways in the network.

Waveguide circuits, which can also be built with standard integrated-circuit processing methods, resemble optical fibers. Waveguides consist of two types of glass, a core and a cladding with different indexes of refraction. An index of refraction is a measure of the material's ability to bend light by a given amount. The appropriate choice of indexes for the two materials ensures that all the light is reflected within the fiber.

A type of waveguide switch currently under development by JDS Uniphase, Nanovation, Lucent and a number of others employs the thermo-optic effect: a temperature change that alters the phase of a lightwave (the position of its oscillation in time) and thus the route down which it travels. It consists of an optical light pipe that gets split into two separate pathways. A change in temperature on one of the divergent pipes, caused by heating an electrical resistor, makes the length of that optical pathway grow slightly longer. The lengthening, in turn, changes the phase of light streaming through. When the two branching signals reconverge, a phase change can cause the light to be switched between one of two output ports [see top illustration on preceding page].

Because they can be built on a common material substrate like silicon, waveguides tend to be small and inexpensive, and they can be manufactured in large batches. The substrates, called wafers, can serve as platforms to attach lasers and detectors that would enable transmission or receipt of optical pulses

that represent individual bits. Integration of various components could lead to photonic integrated circuits, a miniaturized version of the components that populate physics laboratories, one reason the waveguide technology is sometimes called a silicon optical bench.

One fascinating type of switching element combines photonic waveguides with the inkjet technology used in printers. Although it seemed to be an unlikely candidate for an optical switch, this creation of Agilent, a spin-off company of Hewlett-Packard, has begun to make believers out of skeptics.

Forever Blowing Bubbles

The switch consists of a silica waveguide with arrays of intersecting light pipes that form a mesh. A small hole sits at a point where these light pipes intersect. It contains an index-matching fluid (one whose index of refraction is the same as silica). An ink-jet printing head underneath can blow a bubble into the hole, causing light to bend and move into another waveguide. But if no bubble is present, the light proceeds straight. That this switch works at all is

a testament to the extraordinary sophistication of the fluid technology behind inkjet printers [see *bottom illustration on page 91*].

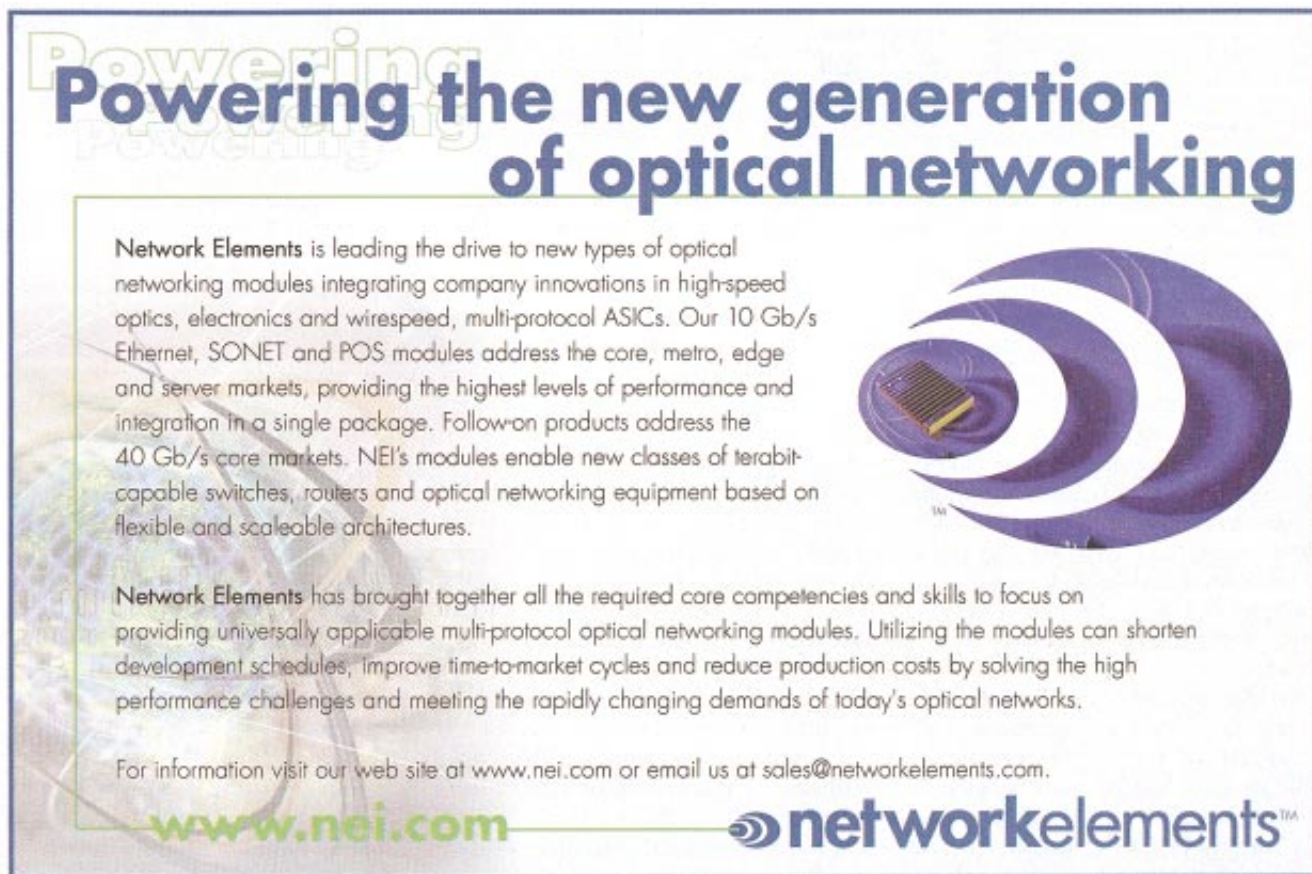
The bubble switch combines small size, reasonable switching speed and good optical performance. As with many other designs, however, engineers may encounter difficulties in building large switches. A switch with a given number of inputs and outputs may need a large number of print heads to insert and remove fluid from the holes: one with 10 inputs and 10 outputs would require 100 print heads. For small numbers of switching connections, this design will prove feasible; for the multithousand port switches envisioned by network architects, however, it seems impractical.

Yet another type of switch uses the electro-optic properties of the liquid crystals commonly found in digital watches and computer displays. Liquid crystals consist of molecules with a pronounced one-dimensional shape, like molecular "hot dogs." These molecules interact with externally applied electrical fields to change their orientation, a characteristic that makes them suitable for optical switching.

When a large electrical field is applied to the slender crystals, it orients them in a particular direction. The change in orientation causes lightwaves passing through them to alter their polarization (vibrations in a given orientation). Other components in the switch only let light of a particular polarization pass into a given output fiber.

The liquid crystal resides in a cell between two glass plates coated with a transparent conducting oxide that serve as electrodes. A voltage applied between the electrodes creates an electrical field that changes the orientation of the liquid-crystal molecules and then the polarization of the light passing through the cell. The light then passes through a displacer, a composite crystal that directs light, depending on its polarization, to a given output port [see *illustration on opposite page*].

Historically, liquid-crystal components have suffered from low switching speeds and poor optical performance. The slightest change, for instance, in the angle of polarization of the light would affect performance. Recent work has minimized these effects, and companies such as Corning and Chorum Technolo-




Powering the new generation of optical networking

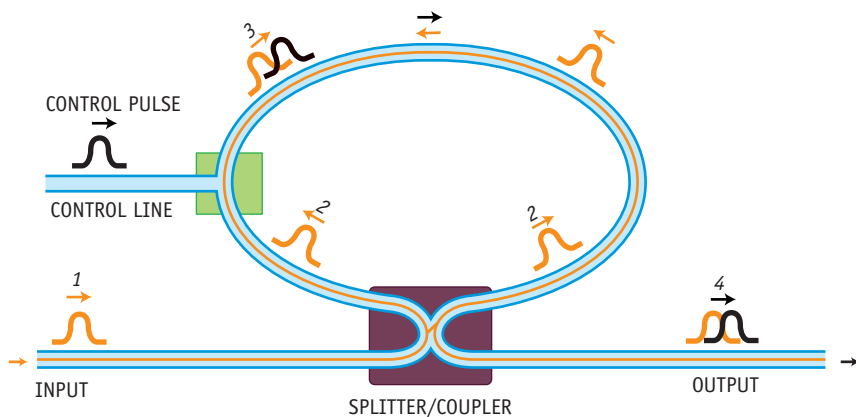
Network Elements is leading the drive to new types of optical networking modules integrating company innovations in high-speed optics, electronics and wirespeed, multi-protocol ASICs. Our 10 Gb/s Ethernet, SONET and POS modules address the core, metro, edge and server markets, providing the highest levels of performance and integration in a single package. Follow-on products address the 40 Gb/s core markets. NEI's modules enable new classes of terabit-capable switches, routers and optical networking equipment based on flexible and scalable architectures.

Network Elements has brought together all the required core competencies and skills to focus on providing universally applicable multi-protocol optical networking modules. Utilizing the modules can shorten development schedules, improve time-to-market cycles and reduce production costs by solving the high performance challenges and meeting the rapidly changing demands of today's optical networks.

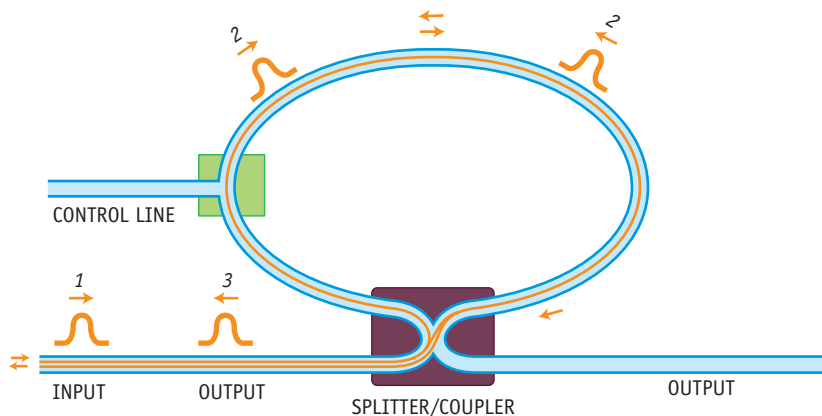
For information visit our web site at www.nei.com or email us at sales@networkelements.com.

www.nei.com 

Nonlinear Optical Switch



LIGHT PULSE entering a nonlinear optical loop mirror (1) gets split into two separate pulses that circulate through in opposite directions (2). When a pulse enters the loop from the control line, it interacts with the pulse moving clockwise, changing its phase (3). The countercirculating pulses recombine in the splitter/coupler. The change in phase of the reconstituted pulse causes a pulse to exit through the right pathway.



NO CONTROL PULSE changes the switch's output. A pulse enters the loop (1), gets split, and circulates in opposite directions (2), causing the two pulses to recombine in the splitter and exit the way they entered (3).

gies have intensive development programs. As with several other design approaches, though, building large switches may prove difficult, because the number of switching elements scales in proportion to the square of the number of either the inputs or outputs, which makes engineering a big switch difficult. Nevertheless, this technology may prove suitable for building a reconfigurable multiplexer, a switchlike device that allows wavelengths to be loaded on and off a network.

Another type of optical switch takes advantage of the way the refractive index of glass changes as the intensity of light varies. Most of the optical phenomena we are familiar with in everyday life are linear. If you shine more light on a mirror, the surface reflects more of the incident light and the im-

aged room appears brighter. A nonlinear optical effect, however, changes the material properties through which the light travels. Think of the mirror becoming transparent when you shine light on it.

Glass optical fibers experience nonlinear effects, some of which can be used to design very fast switching elements, capable of changing their state on a femtosecond (quadrillionth of a second) time scale. Consider a nonlinear optical loop mirror, a type of interferometer in which two light beams interact.

In the mirror, a fiber splitter divides an incoming beam. In one instance, each segment travels through the loop in opposite directions, recombines after completing the circle and exits on the same fiber on which it entered the loop. In other cases, though, after the two beams split, an additional beam is sent

down one side of the loop but not the other. The intensity of light produced by the interaction of the coincident beams changes the index of refraction in the fiber, which in turn changes the phase of the light. The recombined signal, with its altered phase, exits out a separate output fiber [see illustration at left].

In general, nonlinear optical switching requires the use of very short optical pulses that contain sufficient power to elicit nonlinear effects from the glass in the fiber. An optical amplifier incorporated into the switch, however, can reduce the threshold at which these nonlinear effects occur. Although nonlinear switches have yet to reach commercial development, the technology shows promise for the future.

With the current vigorous interest in developing new materials and processes for switching lightwaves, the rule of the electron in telecommunications is reaching its twilight years. We can continue to expect rapid progress in optical switching with the development of new optical materials and systems; some researchers have even begun to explore holograms or acoustic materials as switching elements. The driving force for these diverse efforts targets the complete elimination of the electronic bottleneck in order to provide large systems for high-capacity optical networks that will lead to a telecommunications service provider's dream: a virtually unlimited wealth of bandwidth. SA

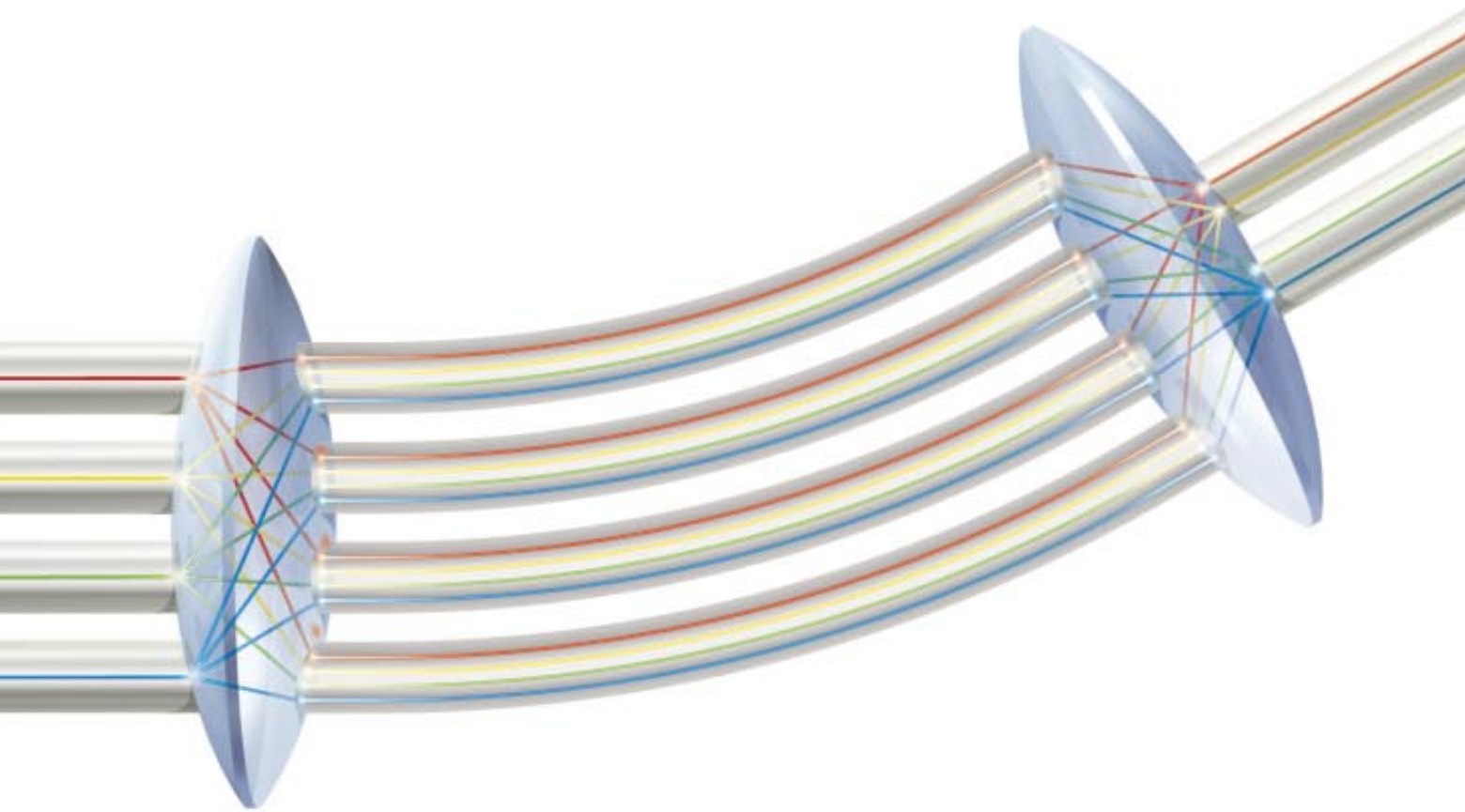
DAVID J. BISHOP and C. RANDY GILES helped to create Lucent Technologies's MEMS optical switch. Bishop is director of the micromechanics research department at Lucent's Bell Laboratories. Giles is technical manager of the photonic subsystems research group in the micromechanics research department at Bell Labs. SASWATO R. DAS is a science and technology writer and spokesperson for Bell Labs. This article would not have been written without the help of many people, including Alastair Glass, Richart Slusher and Alice White.

FURTHER INFORMATION

OPTICAL FIBER TELECOMMUNICATIONS IIIB. Edited by Ivan P. Kaminow and Thomas L. Koch. Academic Press, 1997.

UNDERSTANDING FIBER OPTICS. Jeff Hecht. Prentice Hall, 1998.

BELL LABS TECHNICAL JOURNAL. Various articles on the future of optical communications. Vol. 5, No. 1; January–March 2000. Available at www.lucent.com/minds/tech_journal/ on the World Wide Web.



OPTICAL MULTIPLEXER, a component of an optical packet router, sends four incoming wavelengths to two output ports.

Today's network architects have already begun to build the Optical Internet using technologies that can switch some or all of the light signals in one incoming optical fiber to one or several outgoing fibers. The new-generation Optical Internet will serve as a high-speed mail delivery vehicle that will bring units of data called packets to a point nearby a recipient. There the time-consuming process of sorting out which packet goes where—the Internet equivalent of the local post office—will fall to electronic routers from companies such as Cisco Systems and Juniper Networks. Over time, even this task of switching individual packets may be taken over by routers that use photons, not electrons, for processing.

The IP packet is the Internet's basic unit of currency. In today's networks, every e-mail message gets chopped into thousands of packets, which are switched over different pathways and reassembled at a final destination using the Internet Protocol (IP). The routers, together with other networking equipment, convert data from lightwaves to electronic signals in order to read the packets and send them along to their final destination: a mail server, where they are reassembled into a coherent message before a synthesized voice proclaims, "You've got mail!" The key is that the routers all along the way can easily read the address of each IP packet.


The lightwave network that does the heavy hauling may serve as only one stage in the evolution toward a truly all-optical network, however. When networks routinely shuttle

terabits (trillions of bits) a second, the conversion step into electronics may prove too costly both in time and money. In coming years, a router will routinely have to break down a data stream carrying 40 gigabits (billions of bits) a second over a single wavelength into 16 parallel electronic data streams, each transmitting 2.5 gigabits a second within the router. Moving a massive number of packets every second through multiple layers of electronics in a router can lead to congestion and degrade network performance.

Light to Light

In response to these problems, network engineers are contemplating a solution that will use lightwaves to process lightwaves and optical switches that can redirect packets at blindingly fast speeds. The technology for photonic routing switches is still in its infancy. Creation of a photonic packet-switched network will depend on overcoming multiple technological hurdles equivalent to those that had to be addressed by electronics engineers from the mid-1950s to the mid-1970s, going from individual capacitors and resistors soldered onto a circuit board to monolithic integrated circuits.

Nevertheless, the quest has begun. The Defense Advanced Research Projects Agency (DARPA) has funded programs in optical packet switching at the University of California at Santa Barbara, Telcordia Technologies in Morristown, N.J., Princeton University and Stanford University. Alcatel in France, the Technical University of Denmark and



The ultimate all-optical network will require dramatic advances in technologies that use one lightwave to imprint information on another

by Daniel J. Blumenthal

ROUTING PACKETS WITH LIGHT

the University of Strathclyde in Scotland have also launched research efforts.

A potential communications technology for photonic packet switching in future networks is known as All-Optical Label Swapping (AOLS). In AOLS, individual IP packets or groups of packets get tagged with an optical label. The IP packet is like an envelope with an address on the front (called the header) and contents inside (called the payload). AOLS attaches a label to the packet by placing it, say, in front of the header. The use of labels resembles packing into the same mailbag all letters going through the same set of major cities on the way to their final destination. Postal workers read only the label at each intermediate routing point, not each individual envelope.

In a communications network, an optical router will look only at the smaller label and determine which output fiber or wavelength to send the packet to, instead of reading the header inside the packet. Current photonic technologies do not make this task easy. Optical components that perform the function of the integrated-circuit elements that read and process a label exist mostly as laboratory demonstrations that have not made their way into the commercial mainstream.

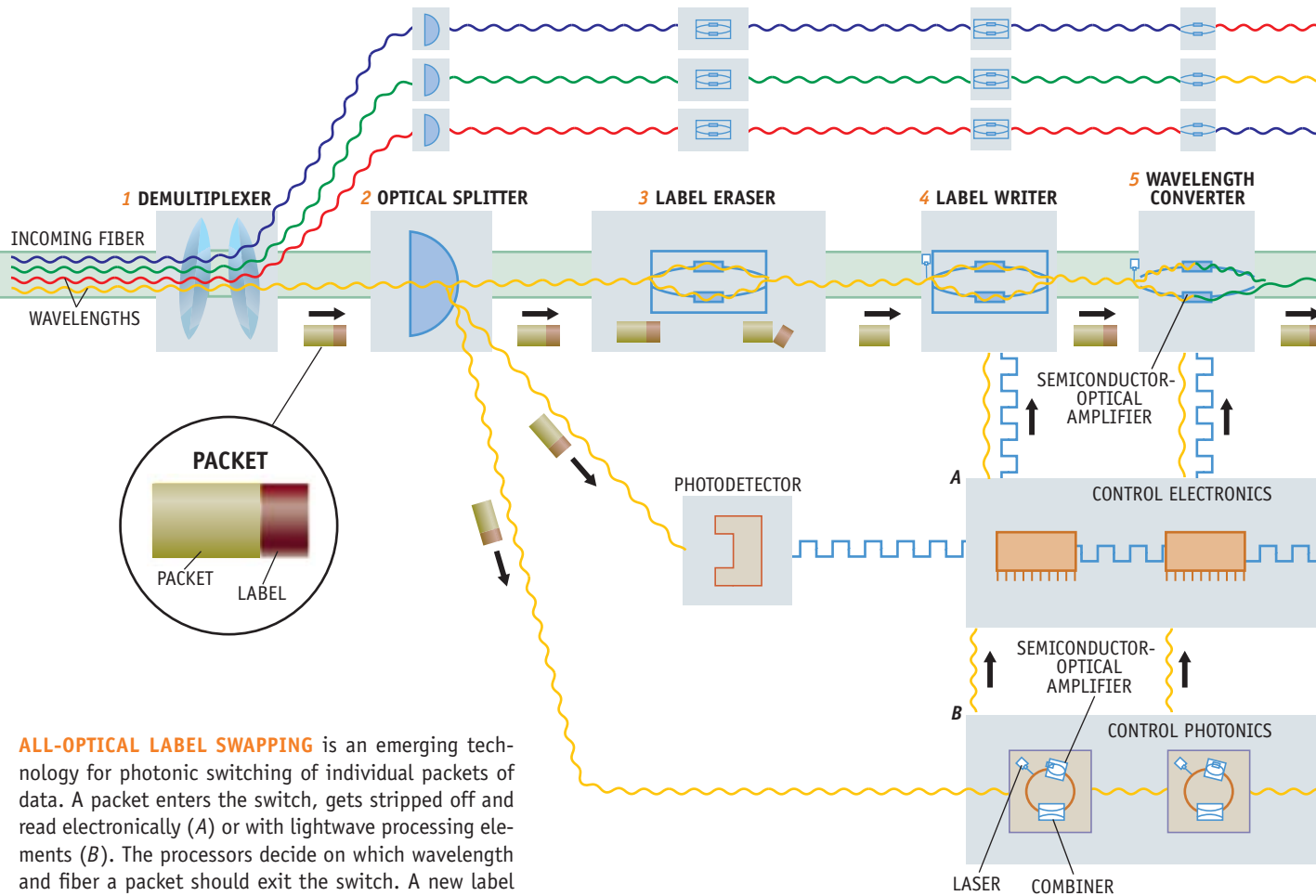
Take something as simple as the buffers used for temporary storage of the bits in a packet. In designing an electronic router, an engineer would hold the packet in a buffer that is similar to the dynamic random-access memory (DRAM) used in computers. Photons cannot be stored easily, however. So designers must then figure out ways to buffer the packet by com-

plex optical circuits that do not hold the photons still but rather corral the pulses of light into a holding area similar to an automobile traffic circle. Other approaches use techniques that match how long it takes the pulses in a packet to move through a switch. They do so by sending the pulses through a predetermined length of fiber. The time it takes to traverse this “detour” fiber equals the time needed by the switch to strip off the packet label and read the routing information it contains.

Optical buffers on laboratory benches are currently hundreds of times larger than the submicron-size dimensions of their electronic buffer counterparts. It is not known if the optical version of the dense conglomeration of transistors found in the Pentium chip will ever come to be a reality.

Besides buffers, logic circuits that can process labels based on light controlling light have also been shown in the lab but are a long way from being engineered into a full-scale router. The first examples of photonic routers will therefore still rely partly on electronics. A packet that enters a first-generation AOLS router will have a small amount of its optical energy diverted down a separate pathway, where it gets converted by a photodetector into an electronic copy of the packet and label. The label goes to circuitry that reads its contents and computes the next pathway for the packet along the network. The electronic processor generates a new label, attaches it to the original packet, and sends a control signal to the switch to specify a particular wavelength or fiber along which the packet should travel [see box on next page]. Despite the need to make an optoelectronic conver-

Optical Router



ALL-OPTICAL LABEL SWAPPING is an emerging technology for photonic switching of individual packets of data. A packet enters the switch, gets stripped off and read electronically (A) or with lightwave processing elements (B). The processors decide on which wavelength and fiber a packet should exit the switch. A new label gets placed on the packet, which then gets switched to a new wavelength and an output fiber.

sion for the tiny label, perhaps 20 bits in size, the optical switches can forgo electronic processing of the much larger packet, which may range from 40 to literally thousands of bytes. The big packet moves through the switch at the higher-speed optical bit rate. Making the conversion for the entire packet might be prohibitive in cost as well as consuming power and space for electronic equipment.

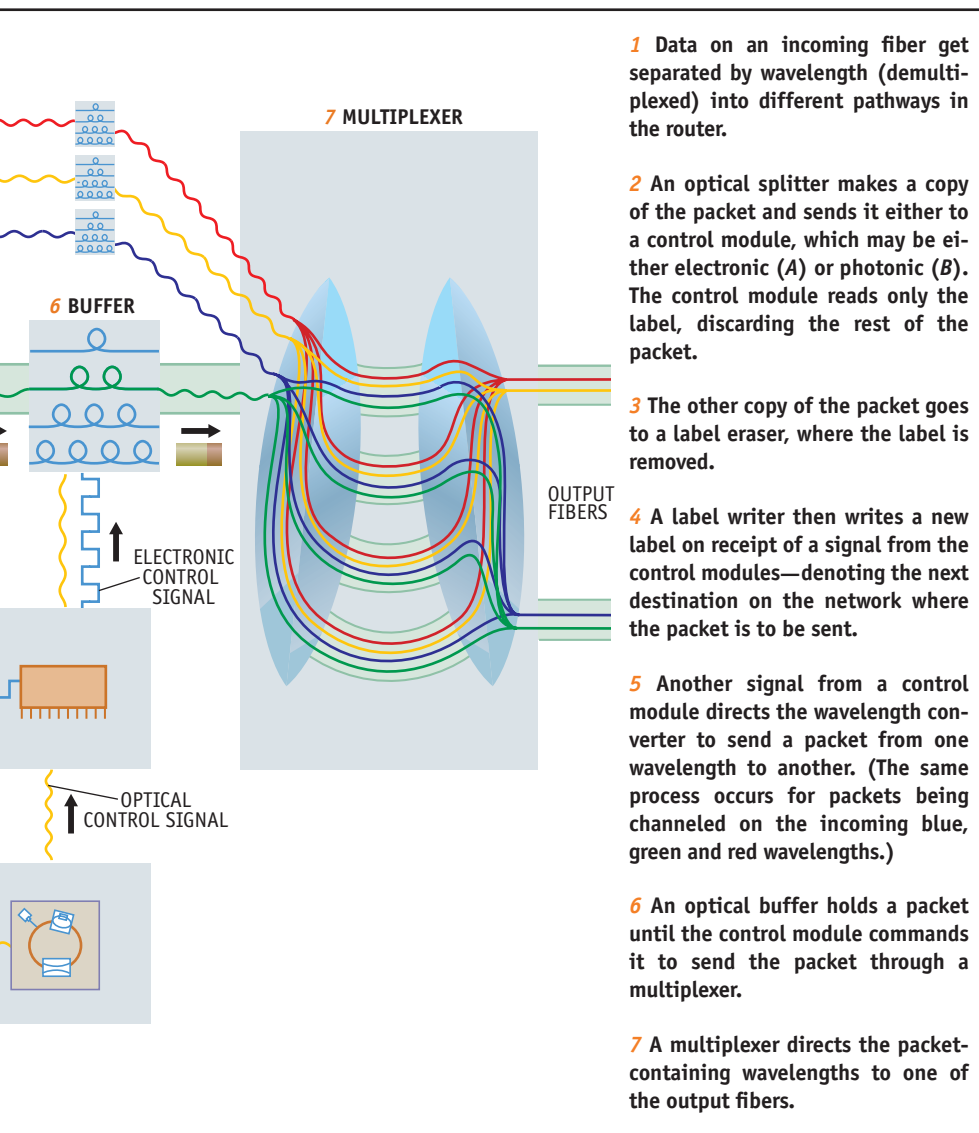
Once the label has been processed, the packet finally arrives at the physical elements that switch it from the input to the output fiber. Optical switching elements pose another major engineering challenge, because the switching components must be able to send a packet to a new fiber or wavelength in a nanosecond or less. Several technologies have emerged that are capable of switching packets at the desired speeds. One example is the semiconductor-optical amplifier, a device that uses stimulated emission of light as the basis for switching, the same process that drives a laser.

To switch a signal on an incoming fiber to any of many outgoing ones, the amplifier forms an optical bridge between the desired input and output fibers. When a data-carrying optical signal reaches the bridge, an electric current in a control signal injects electrons and “holes” (areas where electrons are absent) into the amplifier. Light entering the amplifier causes the electrons and holes to combine with one another, giving off more photons that are exact copies of the optical signal trying to cross the bridge. Once the signal reaches a certain level of power, it moves from one side of the bridge to the other. When the control current is shut off, light at the amplifier input is absorbed and does not make it through to the output fiber. Although the control signal

that opens the gateway is electronic, the stream of photons carrying packets races through the switch without getting converted to an electronic signal.

The technology remains in early development. Investigators have devised fast switches with up to eight incoming and outgoing fibers each, although for marketable products they will need to create switches with hundreds or thousands of input and output ports.

Semiconductor-optical amplifiers may also perform a vital role in realizing a technology that can transfer a stream of bits directly from one lightwave to another. This wavelength converter can switch packets to multiple output fibers by combining the technology with a device called a wavelength multiplexer that, much like a prism, can separate light into its individual colors. The wavelength converter will determine how to route packets to avoid switching



- 1** Data on an incoming fiber get separated by wavelength (demultiplexed) into different pathways in the router.
- 2** An optical splitter makes a copy of the packet and sends it either to a control module, which may be either electronic (A) or photonic (B). The control module reads only the label, discarding the rest of the packet.
- 3** The other copy of the packet goes to a label eraser, where the label is removed.
- 4** A label writer then writes a new label on receipt of a signal from the control modules—denoting the next destination on the network where the packet is to be sent.
- 5** Another signal from a control module directs the wavelength converter to send a packet from one wavelength to another. (The same process occurs for packets being channeled on the incoming blue, green and red wavelengths.)
- 6** An optical buffer holds a packet until the control module commands it to send the packet through a multiplexer.
- 7** A multiplexer directs the packet-containing wavelengths to one of the output fibers.

conflicts. If two packets in different fibers both need to be transferred to the red wavelength in a given output fiber, the wavelength converter will route one of the packets to a green wavelength to ensure that they do not conflict.

Interacting Wavelengths

Experimental converters have employed photonic integrated-circuit (PIC) technology, the optical equivalent of integrating electronic components on microchips. In a converter built with PICs, light from a laser moves along waveguides, similar to optical fibers. In some types of optical processing, a wavelength carrying a stream of bits interacts with another wavelength, imprinting information on it. The light-to-light transfer occurs when the stream of optical bits causes the waveguide to change the phase of the recipient light-

wave, which then exits into an output fiber. A change in phase might represent a digital one and no change a zero.

To make converters feasible, researchers have devoted massive efforts to the development of tunable lasers that can be set to different wavelengths onto which bits from incoming packets can be modulated. Currently laboratory lasers allow an incoming data stream to be switched to any of 80 wavelengths, and that number will grow to hundreds in the future.

Ultimately even the decisions on how to route a packet may by necessity become all-optical. As packet speeds on a given wavelength exceed 160 gigabits a second, a typical packet will race through the switch in a nanosecond, whereas it would take 100 nanoseconds to process a label using electronics. In essence, the address label will have become larger than the envelope (packet) onto which it

is affixed. Increasing the speed at which the label moves through the processor—to beyond 100 gigabits a second—might tax the limits of the electronic control circuitry (although new exotic electronic technologies that could operate at speeds above 100 gigabits per second are being tested).

Anticipating this problem, a few laboratories have fashioned early prototypes of ultrahigh-speed optical logic gates that can be built from the same technology used to build light-controlled switches. These devices—developed at Princeton, the M.I.T. Lincoln Laboratory and British Telecom Labs, among others—utilize different configurations of semiconductor-optical amplifiers or other optical materials to implement simple optical Boolean logic gates (that is, AND, XOR and NOT) that can process light signals moving at speeds in excess of 250 gigabits a second.

These gates have made extremely simple packet-routing decisions in the laboratory but cannot yet be scaled up to the number needed to make complex routing decisions required for a commercial switch. Still, device integration and new optical switching technologies may one day make photonic control possible, marking the true advent of the all-optical router. If such issues can be resolved, a packet might travel from New York City to Los Angeles through IP routers and never pass through electronics. SA

DANIEL J. BLUMENTHAL is co-founder of Calient Networks, headquartered in San Jose, Calif., which makes intelligent optical switches based on microelectromechanical devices. He is on a leave of absence from the University of California, Santa Barbara, where he is associate professor in the department of electrical and computer engineering and associate director for the Center on Multidisciplinary Optical Switching Technology (MOST), which is funded by the Defense Advanced Research Projects Agency.

FURTHER INFORMATION

PHOTONIC PACKET SWITCHING TECHNOLOGIES, TECHNIQUES AND SYSTEMS. Special issue of *Journal of Lightwave Technology* (IEEE/OSA). Vol. 16, No. 12; December 1998.

OPTICAL NETWORKING SOLUTIONS FOR NEXT-GENERATION INTERNET NETWORKS. Marco Listanti and Roberto Sabella in *IEEE Communications Interactive*; September 2000. Available by subscription at www.comsoc.org/~ci/ on the World Wide Web.

The Well-Rounded Flat Speaker

Flat-panel speakers utilize roughly the same technology as their more familiar conical cousins, with one fundamental difference: flat-panel speakers thrive on chaos and interacting sound waves.

Flat speakers can amplify sound with panels that may be only millimeters thick. The best-known versions of flat-panel speakers rely on technology introduced four years ago by the British company NXT. In a typical speaker, an “exciter,” a magnet-and-coil device about the size of a dollar, translates electrical impulses into tiny vibrations. From the point where the magnet assembly is attached, ripples of sound 10 to 15 microns in amplitude radiate across the panel, hit the edge and travel back, until the entire surface of the panel vibrates. Because the waves are so small and similar, they do not interfere noticeably with one another, according to NXT’s Adrian Horne.

The resulting sound waves broadcast uniformly in 360 degrees, with the panel at the center. The sound seems undirected, in contrast to that coming from cone speakers, which tend to “beam” sound toward one ideal location. “The center of a cone is driven farther and faster than the outside is,” explains Olin D. Williford, a sound engineer for Benwin Sound, a subsidiary of Kwong Quest in City of Industry, Calif. (one of the 205 companies to which NXT licenses its technology). “Pushing the air in different waves from the center to the outside gives higher frequencies at the center than at the edge, so the dispersion of sound narrows.” Flat-panel speakers distribute all frequencies equally, from both surfaces. One suggested use of the panels’ so-called bipolarity is for public-address systems, with both sides broadcasting sound.

The speakers do not yet reproduce deep tones as well as more traditional equipment does. Lower frequencies are so few on the flat-panel surfaces that they cancel out more easily, requiring woofers to support bass tones. According to Horne, researchers at NXT and its licensees are continually searching for new materials that will extend the sound range to deeper tones while keeping the speakers small. Other companies, such as SoundLab in New Zealand, are pursuing alternative methods of creating flat-panel speakers.

Flat-panel speakers can be found in guitars and music keyboards as well. Some reviewers have predicted that because the panels use less energy and cost less than other speakers, they will appear in everything from cars and handheld telephones to combined flat-screen/speaker laptops. The dream, though, is to create a flat-panel speaker that reproduces the sound of today’s best—and most expensive—speakers.

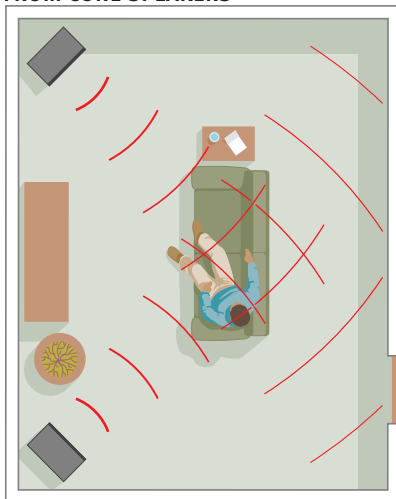
—Naomi Lubick, staff writer

FLAT-PANEL SPEAKERS vibrate in random waves that are set off by electromagnetic impulses. One “exciter” is enough for a small speaker, although larger speakers need several. In the smallest speakers—for example, those used in telephones—electrically responsive crystals act as the exciters. Larger speakers use a magnet-and-coil assembly attached to a carefully chosen point to maximize the spread of tiny sound waves across the panel.

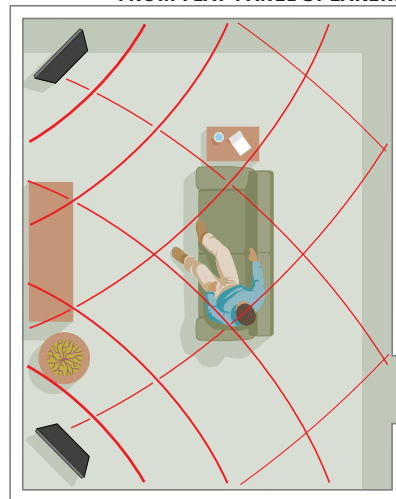


SOUND DISTRIBUTION is more uniform from flat-panel speakers than from cone speakers. Sound waves from cone diaphragms travel in directed beams because the centers and outside edges of the cones vibrate dissimilarly. Flat panels emit sound more evenly because their entire surfaces distribute frequencies equally.

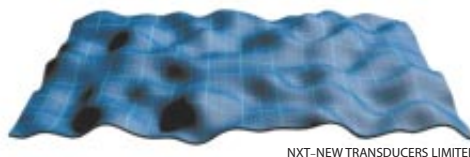
FROM CONE SPEAKERS



FROM FLAT-PANEL SPEAKERS



MODEL of sound waves with 10- to 15-micron amplitudes traveling across a panel.



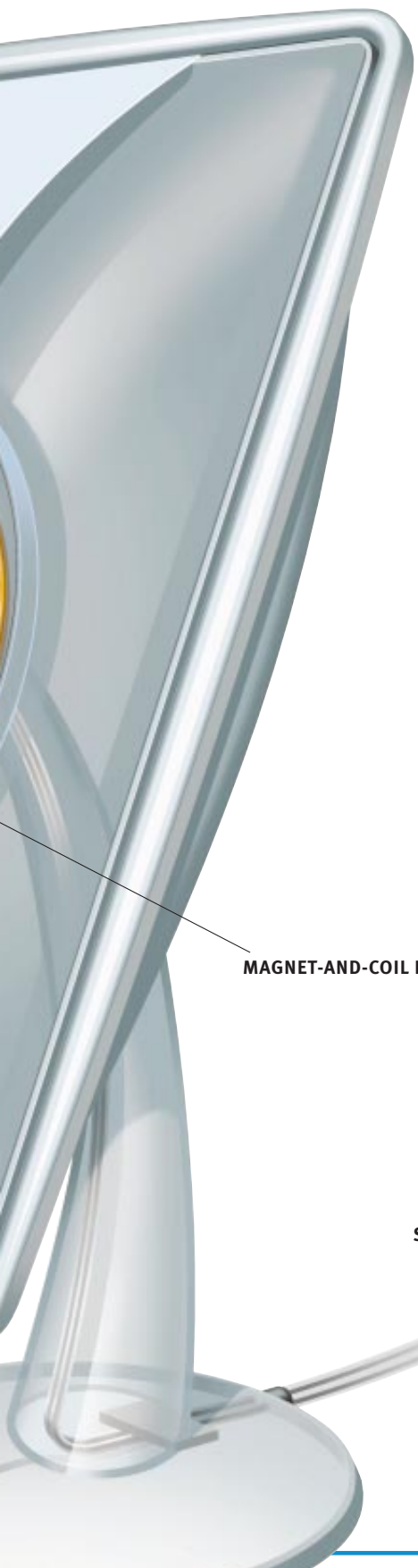
NXT-NEW TRANSDUCERS LIMITED

DID YOU KNOW ...

- The technology for NXT's distributed-mode operating system came from the British Defense Evaluation and Research Agency, whose engineers were attempting to damp sound in military aircraft. Just as the speakers can augment vibrations, they could also theoretically stimulate vibrations in a pattern that would interfere with and cancel out sound waves from unwanted rhythmic background noises.
- Kwong Quest has more than 700 materials logged by density in an acoustical-property database. The materials range from plastics to hard foams to a polycarbonate honeycomb substance. "Even epoxy glass sheets make very good sound," Williford says, although their performance drops as the sheets get thicker. A sheet of glass six-by-six feet and a quarter-inch thick would produce good sound, he notes, but he currently prefers an aerospace foam material.
- Sophisticated electrostatic speakers have flat, vibrating diaphragms poised between two perforated plates that create attractive and repulsive electrical charges, vibrating the entire diaphragm at once. These speakers, which generally stand as tall as a person, are behemoths beside flat-panel speakers but tend to produce higher-quality sound.

MAGNET-AND-COIL EXCITER

SIGNAL FROM AMPLIFIER



Dots-and-Boxes for Experts

Ian Stewart reveals the secret subtleties of a children's game

I never cease to be amazed by the mathematical complexities inherent in what seem to be the simplest of games. Consider, for example, the children's pastime of dots-and-boxes. Generations of kids have played this game in grade school, but I doubt whether one person in a million has played it as well as possible. Mathematician Elwyn Berlekamp of the University of California at Berkeley has outlined the game's many subtleties in his new book, *The Dots-and-Boxes Game* (A. K. Peters, 2000).

First let's review the rules. The game begins with a rectangular grid of dots. Players take turns drawing lines between dots that are adjacent either horizontally or vertically (but not diagonally). When someone draws the fourth side of a box—a square connecting four adjacent dots—the player writes his or her initial in that box and plays again (and continues to play as long as he or she keeps drawing completed boxes). At the end of the game, the player who has initialed the most boxes wins.

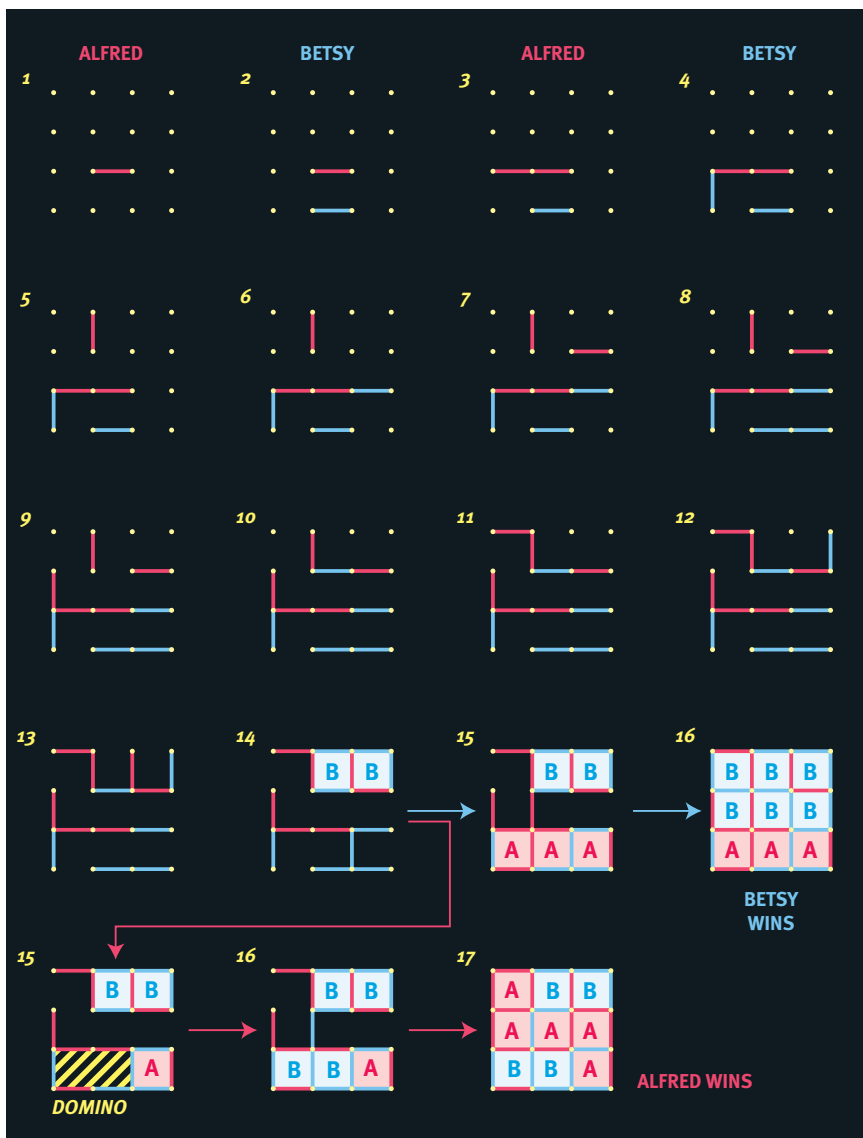
We'll call the first player Alfred and the second Betsy. The illustration at the right shows a sample game on a four-by-four grid in which the players use the most basic strategy, which I call Level Zero play. Alfred and Betsy avoid giving away boxes for as long as they can by making moves that do not create the third side of a potential box. As a result, the grid becomes divided into a series of "chains"—snakelike regions bounded by lines. As soon as a player claims a box in a chain, he or she can continue grabbing boxes until the entire region is taken.

At some point in the game, the whole grid becomes divided into such chains—a state I call gridlock. After gridlock is reached, the next player to move usually draws a line in the shortest chain available, thereby giving his or her opponent the smallest number of boxes. (This move is known as "opening the chain.") The opposing player takes those boxes, then gives away the next smallest chain. In the sample game shown at the right, Betsy creates gridlock in the 12th move.

The grid is divided into three chains of two, three and four boxes each. In the 13th move, Alfred gives away the chain of two boxes to Betsy. Betsy in turn cedes the chain of three boxes to Alfred, who is then obliged to present Betsy with the chain of four boxes. Betsy wins because her six boxes exceed Alfred's three.

In Level Zero play, two factors determine which player will win: whether the

number of chains in the grid is odd or even when gridlock is reached, and what the order of play is immediately after gridlock. Suppose the number of chains at gridlock is even. In this case, the player who opens the first chain will win, because at each move this player will give away a chain that is smaller than the one he or she will receive in his or her next move. But if the number of chains at gridlock



TYPICAL GAME of dots-and-boxes reaches a critical juncture in the 15th move. If Alfred uses basic strategy (blue arrows), he loses. If he uses more advanced strategy (red arrows), he wins.

ALL ILLUSTRATIONS BY BRYAN CHRISTIE



SIX-BY-SIX GRID has four chains and two dominoes.

It is a sacrifice: by giving up two of the boxes in the three-box chain, Alfred puts Betsy in a fatal position. If she draws a line across the middle of the domino in the 16th move, she gains the two boxes. But she has to play again, and whatever move she makes will open the remaining chain of four boxes. Then Alfred will swipe the lot and win by five boxes to Betsy's four. The outcome for Betsy is even worse if she declines to play the domino. Any other move will open the four-box chain, and Alfred will simply take those boxes and the two in the domino as well, winning the game by seven to two.

double-dealing move, taking only two of the boxes in the four-box chain and creating a domino with the other two. This tactic forces Betsy to claim the domino and open the chain of five boxes. Alfred then makes another double-dealing move, taking three of the boxes and leaving another domino for Betsy. As long as the chains contain five or more boxes, Alfred always comes out ahead.

In this case, Alfred is controlling the game because he can keep forcing Betsy to open chains. Thus, a good plan for winning dots-and-boxes is to gain control and retain it by always declining the last two boxes of every chain. (Except when there is only one chain left, of course.) This is called Level Three strategy. But how does one gain control? That task requires Level Four strategy, which can be stated as follows:

- Alfred tries to ensure that the sum of the number of dots in the grid and the number of long chains (those containing three or more boxes) at gridlock is even.
- Betsy tries to ensure that this sum is odd.

lock is odd, the player who opens the first chain will lose, because the opposing player will make the last move of the game. That's why Betsy beats Alfred in the sample game: the number of chains is odd, and Alfred has to open the first one.

Furthermore, the order of play after gridlock depends on whether the number of moves made before then is odd or even. If it takes an even number of moves to reach gridlock, Alfred will open the first chain and Betsy will make the first territorial gain. But if gridlock occurs on an odd move, Betsy will open the first chain and Alfred will make the first gain. So if Alfred wants to win, he must ensure that the number of moves made before gridlock and the number of chains at gridlock are both even or both odd. Conversely, if Betsy wants to win, she must ensure that one of these numbers is odd and the other even. A careful consideration of the grid a few moves away from gridlock can often help achieve these goals. I call this strategy Level One play.

But what if Level One strategy fails? Let's suppose that despite Alfred's best efforts to draw the grid lines to his advantage, he finds himself in the same position shown after the 12th move of the sample game. It turns out that he can still win by following Level Two strategy. In the 13th move, he opens the chain of two boxes. In the next move, Betsy claims the territory and opens the chain of three boxes. But in the 15th move, Alfred declines to accept all three boxes in that chain. Instead he accepts just one box and then draws a line at the bottom of the grid that leaves an enclosed rectangle, which I call a domino.

This is known as a double-dealing move.

Betsy's cause is clearly lost the moment Alfred makes his double-dealing move, because there is only one chain left in the grid. But what if several chains remain? Can Betsy claw back some territory by making double-dealing moves herself?

The answer is "Not always." The illustration above shows a six-by-six grid in which there are two dominoes and four chains. If it's Betsy's turn to move, she may as well grab the dominoes; if she doesn't, Alfred can claim them in his next move without worsening his position. Betsy then opens the shortest chain. Because the number of chains is even, she believes she can win the game using basic Level Zero strategy. But Alfred makes a

You may think this is getting pretty deep, but so far we've only reached page 7 out of 86 pages of strategy in Berlekamp's book. Dots-and-boxes is such a sophisticated game that the complete winning strategy remains unknown. In fact, Berlekamp describes it as "the mathematically richest popular child's game in the world, by a substantial margin." SA

READER_FEEDBACK

Responses to the column on paradoxes ["Paradox Lost," June 2000] continue to flood in. R. B. Burckel of Kansas State University sent a letter concerning the Richard Paradox, which focuses on the tricky statement "The smallest number that cannot be defined by a phrase in the English language containing fewer than 20 words." It seems that such a number must exist: to find it, you simply make a list of all possible phrases with 19 or fewer words that define a unique number and then determine the smallest number that's omitted. But whatever this number may be, the above statement defines it in an English phrase containing only 19 words!

Burckel notes a problem with the paradox that its creator, French logician Jules Antoine Richard, pointed out in 1906: the list of phrases cannot be well defined. For example, consider these two expressions (I have modified Burckel's suggestions and take full responsibility for the result):

"The number named in the next expression, if a number is named there, and zero if not."

"One plus the number named in the preceding expression."

Each phrase on its own seems to define a unique number, so they belong on the list. But the two together are contradictory if the second follows the first. Because the list of phrases cannot be well defined, Richard's phrase does not specify a unique number, and the paradox melts away. —I.S.

A Canteen Cloud Chamber

Shawn Carlson describes a way to view the path of charged particles

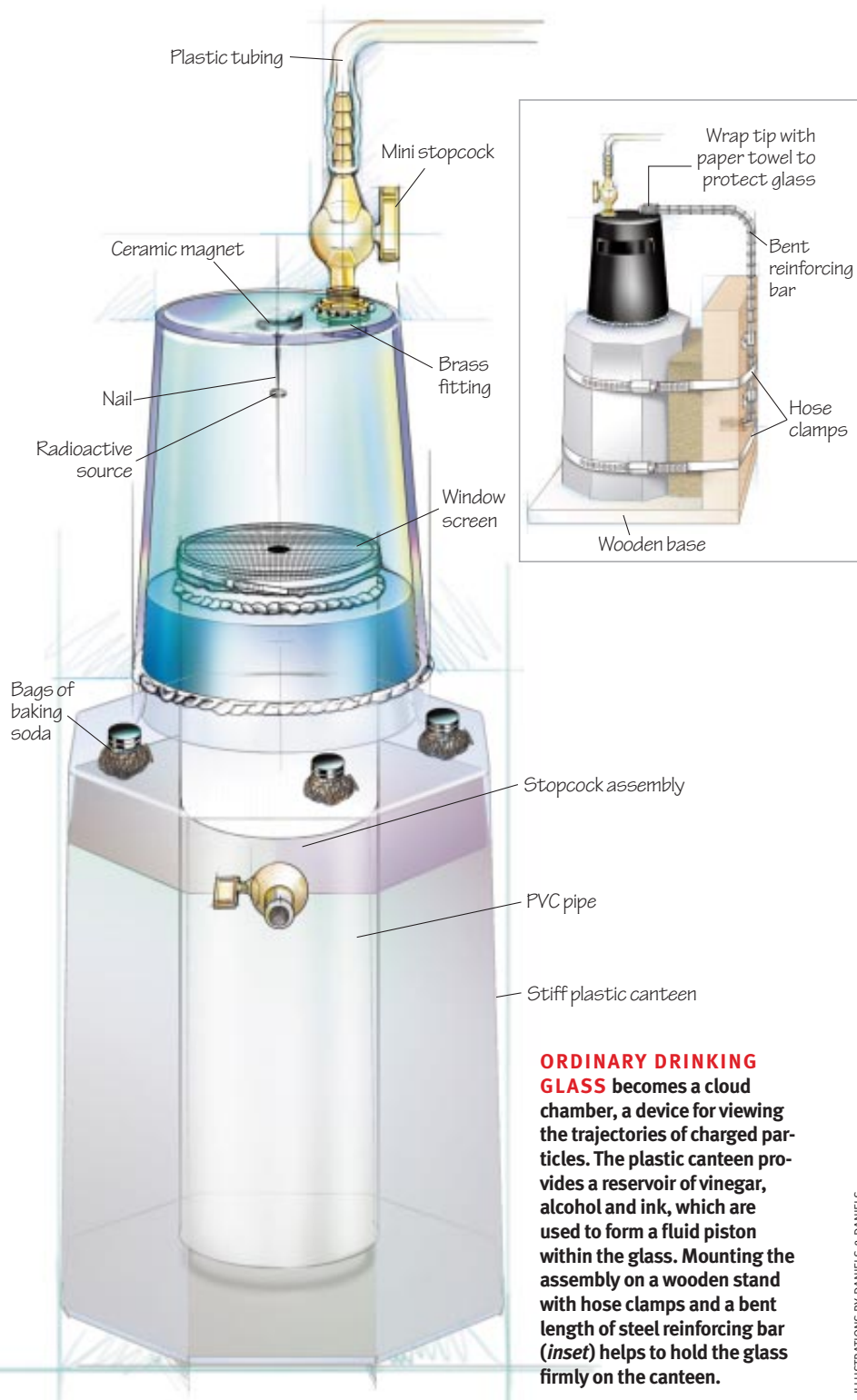
When air cools, the water vapor it holds eventually condenses to form a cloud. But as scientists have long known, air can be chilled well below its normal dew point without such condensation taking place. The trick is to remove the dust particles on which the water droplets normally form. The cooled air can then become “supersaturated.”

So what? Well, in 1896 a University of Cambridge physicist named C.T.R. Wilson discovered that certain subatomic particles leave visible trails when they pass through supersaturated vapors. Why? The particles convert some neutral atoms in the air into charged ions, which, like dust specks, induce droplets to form. Wilson thus was able to fashion the first “cloud chamber” to reveal the trajectories of these ionizing particles.

While in high school, I spent many frustrating hours trying—and failing—to build cloud chambers from instructions I had read in this department. Then, as a college sophomore, my interest was re-kindled when I noticed that a cloud had formed in the neck of a freshly popped bottle of champagne. Within two hours I had converted that bottle into my first working cloud chamber. My design has since evolved, but it remains quite simple and inexpensive to build. The current version costs less than \$30 to put together.

The “generator” (a canteen) is filled with a mixture of vinegar, alcohol and ink. It’s pressurized by adding baking soda. The carbon dioxide given off forces the colored liquid out and into an attached drinking glass, where the fluid acts like a piston to squeeze the gas within. The compression heats the air and causes it to become saturated with vapor from the liquid. Opening a valve allows the fluid piston to drop, which lowers the pressure and temperature of the air, which in turn supersaturates it.

I use a one-liter plastic canteen with flat sides and a wide mouth. The cap, being slightly tapered, fits snugly inside a tall drinking glass. If you find that the cap rests so deeply inside your glass that



ORDINARY DRINKING GLASS becomes a cloud chamber, a device for viewing the trajectories of charged particles. The plastic canteen provides a reservoir of vinegar, alcohol and ink, which are used to form a fluid piston within the glass. Mounting the assembly on a wooden stand with hose clamps and a bent length of steel reinforcing bar (*inset*) helps to hold the glass firmly on the canteen.



WINDOWS for viewing particle tracks and for adjusting the liquid level are formed by placing tape on the glass before spraying it with black paint.

the canteen cannot be attached, ask the folks at a local glass shop to cut off some of the rim. You might also ask them to bore an off-center hole in the base of the glass for a stopcock—or do it yourself. Surprisingly, it's not hard to drill glass. Just cut several notches in the end of a piece of brass tubing. Then put it into the chuck of an electric drill and turn the notched end against the glass while bathing the surface with a slurry of number 120 Carborundum powder and water. Apply a gentle but steady pressure. It's best to use a drill press, but the job can be done with a handheld electric drill. Wear suitable eye protection (as always, when working with power tools) and gloves, in the event the glass should shatter.

Though unlikely, it's conceivable that your glass could break when it is pressurized. So you should also wear your safety glasses when experimenting. And you can add a further level of protection by coating the glass with plastic. Ace Glass in Vineland, N.J. (800-223-4524 or 856-692-3333), sells a special plastic coating (catalogue no. 13100-10) designed to hold the glass shards together in case of a catastrophe. Half a liter costs about \$30.

When your protective coating has fully dried, pass the threaded brass fitting through the hole, seal it carefully with silicone aquarium cement, secure it with a washer and nut, and add the stopcock. Also, find a supply of small ceramic magnets (Radio Shack catalogue no. 64-1883 contains five such magnets) and glue one inside the glass at the top center using silicone cement.

Next obtain a short length of PVC pipe with an outer diameter that is just slightly smaller than the mouth of the can-

teen. Cut a hole in the cap to accommodate the pipe, which should reach down to about two centimeters above the bottom. Glue the top end into the hole in the cap. Then stretch some plastic window screening across the top of the pipe. The mesh reduces turbulence in the fluid, thereby reducing turbulence in the air inside the chamber. Secure the screening with a nylon cable tie. Then punch a small hole in the center of the screen. (You'll need this opening for access to the chamber.) Finally, glue the cap into the glass using plenty of silicone cement and attach the lower stopcock just as you did the upper one.

Now wrap two strips of adhesive tape most of the way around the drinking glass, positioned a few centimeters from the base. Run a third strip vertically along one side. Then spray-paint the glass a flat black to improve contrast and make the particle tracks easier to see. You can now create viewing ports by removing the tape. To monitor the changing fluid level, make a photocopy of a ruler and glue your paper scale beside the vertical window.

Adding baking soda to the liquid while everything is sealed up is simpler than you might think. Just epoxy a ceramic magnet into a small bag fashioned from the toe of a nylon stocking. I suggest you begin by adding 2.5 milliliters (about half a teaspoon) of baking soda to the bag, but you'll probably need to adjust that amount after some trial and error. Place the baking soda on a small piece of paper towel and insert this makeshift holder into the bag. Affix the bag just below the neck of the bottle using a stack of two magnets on the outside; removing the outer magnets releases the baking soda into the solution. I put as many as five bags inside at once so that I can repeat the experiment without opening the canteen.

Mix two liters of liquid by combining equal parts of distilled vinegar and the most concentrated isopropyl alcohol you can find. Squirt in some ink, which (like the black spray paint) will make the tracks easier to see, and add two milliliters of salt. Fill the generator bottle with this solution to within about one centimeter of the nylon bags. Open the top stopcock and then screw the glass cloud chamber on tightly. You may need to cover the threads with Teflon tape or petroleum jelly to get a pressure-tight seal. Insert the whole assembly in its holder. Now open both stopcocks, carefully suck on the tube until liquid fills the cloud chamber about halfway, then close both stopcocks.

Position a bright light to one side of your viewing port. Drop in a bag of baking soda and monitor the fluid level in the cloud chamber, bleeding the pressure with the lower stopcock when the compression ratio rises above about 1.33. (If this ratio is below about 1.25, tracks won't materialize, and if it is above 1.38, a dense cloud forms and obscures everything.) Wait a minute or so, then rapidly open that same stopcock completely.

Particle tracks can appear only during the next brief instant, so the odds of seeing a vagabond cosmic ray are not at all good. A potent source of either alpha particles (helium nuclei) or beta particles (electrons) provides a much more satisfying show. Alpha particles produce short tracks, whereas beta particles leave long ones. You can obtain both an alpha and a beta source suitable for this project from the Society for Amateur Scientists.

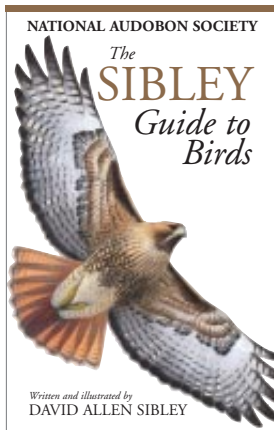
Epoxy your radioactive source to the tip of a nail that has a large, flat head. Lift this assembly through the hole in the window screening and into the cloud chamber using a drinking straw and stick the head of the nail to the magnet at the top of the glass. This arrangement will hold the source securely in place until you want to swap it for another one.

Though easy to build, my cloud chamber has its limitations. The optical quality of most drinking glasses is poor, which can make the tracks hard to see. And reloading the device with packets of baking soda can be tedious. So despite the failure of my teenage attempts, you might want to consider other designs of this kind offered previously in the *Amateur Scientist* (check April and December of 1956). Prospective builders should also consider the so-called diffusion cloud chambers (described in the *Amateur Scientist* columns of September 1952 and June 1959), which require little more than dry ice and alcohol—isopropyl alcohol that is, not champagne. SA

For this project, the Society for Amateur Scientists is making available a set containing one alpha and one beta source. The cost is \$35. To order, call 401-823-7800. You can write the society at 5600 Post Road, Suite 114-341, East Greenwich, RI 02818. To take part in a discussion about this project, surf over to www.sas.org and click on the "Forum" button. For information about a new CD-ROM containing past columns published in this department (over 800 in all), link to www.tinkersguild.com or call 888-875-4255.

Not Only Fine Feathers . . .

... but intelligent organization and design create a new classic



National Audubon Society
The Sibley Guide to Birds
 Written and illustrated
 by David Allen Sibley
 A Chanticleer Press Edition
 Alfred A. Knopf,
 New York, 2000 (\$35)

For birders who cut their teeth on Roger Tory Peterson's *Field Guide to the Birds*, that book seemed definitive, like the King James Bible or Bogart and Bergman in

Casablanca. But it's a new millennium, and David Allen Sibley and the National Audubon Society have produced an impressive new *Guide to Birds*.

How does it differ from earlier guides? When Sibley himself was asked, he replied: "My book relies much more on illustrations. . . . I believe the average field guide user spends the vast majority of time looking at the pictures, and when I was developing this layout I based it on the premise that most of the text in current field guides is redundant. . . . I wanted a book that would condense a huge amount of information into a portable size, and at the same time make the information 'patterned,' logical, and accessible to any reader."

He delivers. Full-color paintings—6,600 of them—show us 810 North American species in an array of shapes, stages, colors, markings and poses (at rest, in flight, perched, swimming and so on). Raptors are shown from below. All significant plumages are depicted: the Laughing Gull, for example, is shown in six different stages. Voice descriptions (songs, flight calls, juvenile begging cries, threats, displays) appear on every page. Full-color range maps show complete distributions, migration routes, and summer, winter and breeding locations. Measurements are there, too: wingspan, length and weight. To facilitate comparison, information and illustrations are arranged in the same way

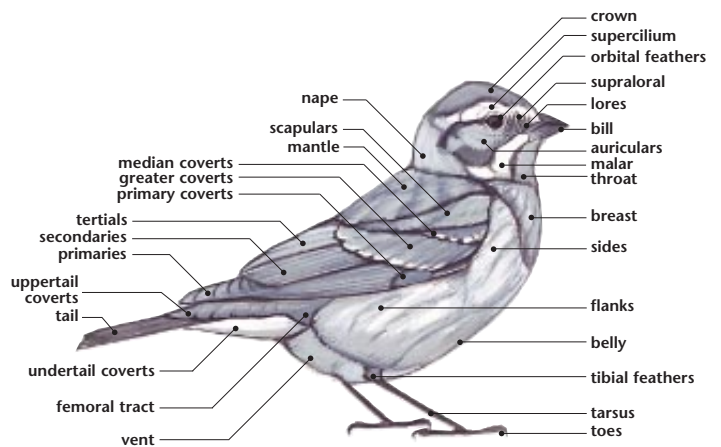
for each species, and birds are shown in similar poses. Happily, the text accompanies the drawings as captions, so you don't have to flip back and forth. Pointers guide your eye to the relevant feature.

The book's introductory material is a primer on how to look at and identify birds, beginning with the parts of a passerine, or songbird [see illustration below]. The introduction also includes Sibley's "rules," the first of which is: "Look at the bird. Don't fumble with a book, because by the time you find the right picture the bird will most likely be gone. Start by looking at the bird's bill and facial markings. Watch what the bird does, watch it fly away, and only then try to find it in your book."

Even the endpapers overflow, in an organized way, with useful tools: metric conversions, rulers, a map of the area the guide covers. A sturdy, flexible cover, sewn-in binding, and heavy, nonreflective paper add to the pleasure of using this book. True, it's a little hefty for the field, but this is a quibble. Better to have all this information than to be able to tuck the book in your pocket. Besides, it fits easily in a backpack. —The Editors

PARTS OF A PASSERINE

This figure shows the basic parts of a passerine, or songbird.



CAROLYNNE BAILEY

DAVID SIBLEY_BORN TO BIRD

THE SON OF YALE ORNITHOLOGIST Fred Sibley, David Sibley taught himself to draw at age six by tracing Arthur Singer's illustrations in *Birds of the World*. After two semesters at Cornell, he dropped out to work at the Cape May Bird Observatory.

Several years later he left Cape May and crisscrossed the U.S. in his pickup truck to study birds, storing his sketching equipment in the cab and sleeping in the back. By his late 20s he was an acknowledged expert, leading tours for WINGS.

Now 39, he is married to Joan Walsh, an ornithologist he met at Manomet Bird Observatory, and the father of two young children.

Sibley starts his drawings in pencil, working from photographs and field sketches. He then puts the illustrations on an opaque projector and adjusts for size so they are exactly proportional to the others on that page of the guide. After correcting for size and shape, he traces the projection and paints the final version in gouache, working in transparent layers until he reaches the desired color and texture.

More of Sibley's art is on view at his Web site, www.sibleyart.com

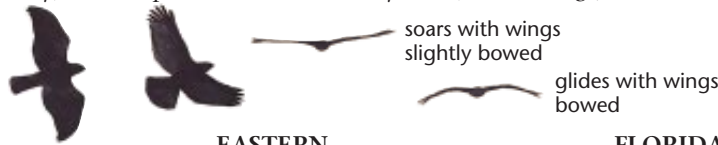
Red-shouldered Hawk

Buteo lineatus

L 17" WS 40" WT 1.4 lb (630 g) ♀ > ♂

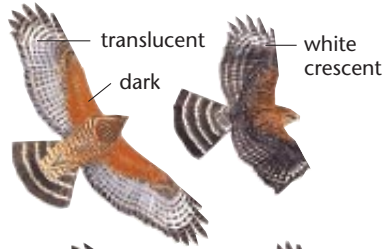
Rather compact, stocky, and accipiter-like with relatively short, broad wings; all show translucent pale crescent across wingtips.

wingbeats choppy, quick (especially California)

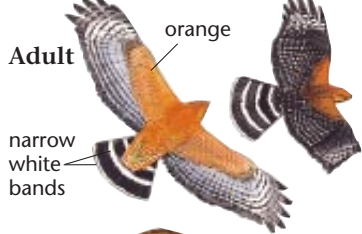


CALIFORNIA

Juvenile



Adult



Juvenile (1st year)



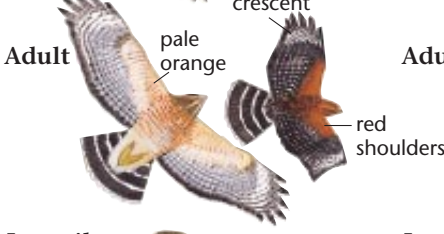
Adult

EASTERN

Juvenile



Adult



Juvenile (1st year)



Adult

FLORIDA

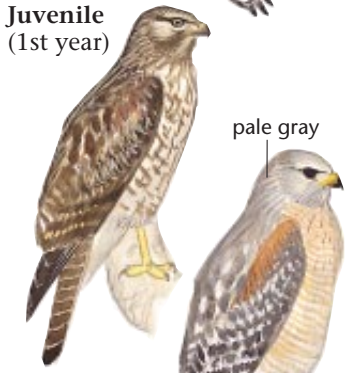
Juvenile



Adult

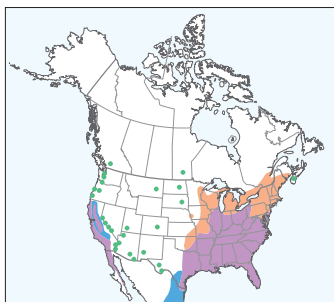


Juvenile (1st year)



Adult

Voice: Very vocal, with distinctive, far-carrying calls. Adult territorial call a high, clear, squealing *keeyuur keeyuur* . . . repeated steadily; often imitated by Blue Jay and Steller's Jay. Also a single or slowly repeated high, sharp *kilt*. Juvenile similar to adult. Calls of California may be a bit shorter, higher, and sharper than Eastern and Florida but very similar.



All eastern birds are similar, although southern Florida adult is much paler. California adult is more richly colored with solid orange breast lacking dark streaks, tibia feathers darker orange than belly, and fewer and broader white tail-bands. Juvenile is more like adult than eastern juveniles, with black and white wings and tail and dark rufous underwing coverts.

SAMPLE PAGE FROM *THE SIBLEY GUIDE TO BIRDS*

THE EDITORS' RECOMMEND

DAVID E. JONES'S *An Instinct for Dragons*. Routledge, New York, 2000 (\$24.95).



FROM AN INSTINCT FOR DRAGONS

Many societies have a concept of and a word for the dragon, even though the creature never existed. Why? Jones, professor of anthropology at the University of Central Florida, thinks

the concept derives from the experience of ancestral humans and prehumans with three kinds of predator: "Over millennia," he writes, "the raptor, big cat, and serpent began to form as a single construct—the dragon—in the brain/mind of our ancient primate ancestors." Jones got his idea from the behavior of vervet monkeys in Africa. They have three different alarm calls that provoke three different defensive responses: one for the leopard, one for the martial eagle and one for the python. Most of the 40 illustrations in the book portray dragons as different societies envisioned them. The common theme is that they look scary.

ROGER LEWIN, GARNISS CURTIS AND CARL SWISHER'S *Java Man: How Two Geologists' Dramatic Discoveries Changed Our Understanding of the Evolutionary Path to Modern Humans*. Scribner, New York, 2000 (\$27.50).

The two geologists are Curtis and Swisher of the University of California at Berkeley; Lewin is a science writer. They describe how Curtis and Swisher put the age of a fossil human skull found on the Indonesian island of Java at 1.7 million years. The date made resounding news because it pushed back by about a million years the time when humans migrating out of Africa first reached Eurasia. This finding by two geochronologists did not win unanimous agreement from paleoanthropologists, but the dating of skulls found recently in the Republic of Georgia appears to support the early arrival of humans in Eurasia. The authors then proceed to an absorbing discussion of the stages in human evolution. They argue that "unless humans are different from all other animals (for some unexplained reason), several species coexisted on Earth, perhaps sometimes in the same geographical region, until a little less than 30,000 years ago." *H. sapiens*, they say, remains alone because the other species lost out through economic competition. "Or the incoming population of modern humans might have hastened their extinction in ways that we know are all too human, that of violence."



FROM JAVA MAN

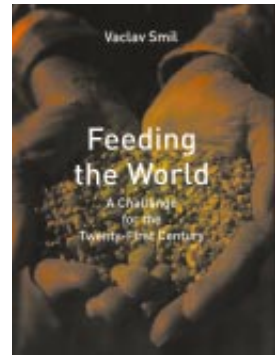
STUART A. KAUFFMAN'S *Investigations*. Oxford University Press, New York, 2000 (\$30).

Kauffman's investigations concern nothing less than the nature of life. "It may be," he says, "that I have stumbled upon the proper definition of life itself." His deep and challenging argument runs as follows. Much of the order in living organisms is self-organized and spontaneous. "Self-organization mingles with natural selection in barely understood ways to yield the magnificence of our teeming biosphere. We must, therefore, expand evolution-

ary theory." The living organism, be it bacterial cell or human being, is a " 'propagating organization,' that is, that it literally constructs more of itself." This activity "has no statement in current physics or biology but constitutes that which constructs a biosphere." Kauffman, a founding member of the Santa Fe Institute, calls his actors autonomous agents and says we are on the verge of the capacity to create novel molecular autonomous agents. "When we do, or if we discover life on other planets and solar systems, science will enter a vast new phase in which we will create a 'general biology,' freed from the limitations of terrestrial biology."

VACLAV SMIL'S *Feeding the World: A Challenge for the Twenty-First Century*. MIT Press, Cambridge, Mass., 2000 (\$32.95).

Smil's message is that it can be done. He sees "no insurmountable biophysical reasons why we could not feed humanity in decades to come while at the same time easing the burden that modern agriculture puts on the biosphere." For one, "there is a very high probability that humanity will not double in number again, and that its 2050 total of around 10 billion people may be very close to (or perhaps even a bit above) its long-term maximum." Moreover, there is plenty of room for "tightening up the slack in the food system." Smil, professor of geography at the University of Manitoba, regards himself as a Malthusian in the sense expressed by Thomas Robert Malthus—usually seen as asserting that the population will eventually outrun the food supply—in the second, rarely read, edition (1803) of his essay on population: "Though our future prospects respecting the mitigation of the evils arising from the principle of population may not be so bright as we could wish, yet they are far from being entirely disheartening."



DEBORAH RUDACILLE'S *The Scalpel and the Butterfly: The War between Animal Research and Animal Protection*. Farrar, Straus and Giroux, New York, 2000 (\$25).

It's not possible to write a completely balanced account of the antivivisection movement, but Rudacille makes a brave attempt. The book starts with Victorian-era activists and researchers, pans through Nazi Germany and ends up with modern events such as the Silver Spring, Md., case (in which photographs of mutilated monkeys caused public outrage). The latter half of the book discusses current issues such as cloning, transplantation of animal organs into humans and the search for alternatives to animal research. The "scalpel" in the title is the rationality of science; given the profiles of animal researchers presented, it could just as well stand for the human male. The "butterfly" is the intuitionist to whom feelings and suffering are all important and—you guessed it—is female. The division makes sense because antivivisectionists have traditionally been women who were moved by compassion; it runs into trouble, though, with animal-rights philosophers such as Peter Singer, whose arguments, whether you like them or not, are very rational. The book itself probably would have benefited somewhat from a scalpel, for the writing starts out lively and entertaining but becomes more heavy-going two thirds of the way through.



Information Technology, 2500 B.C.

Philip and Phylis Morrison ponder the daily lives of those who first developed the written word some five millennia back

A year ago the world was bemused by three zeros. Those zeros won out, even as fear of practical confusions faded from our Y2K screens. We who live in a hurry were at least reminded of the long span of historical time. This New Year's we will consider another crucial millennium long ago, the one during which prehistory became history itself, and look back from the best-informed viewpoint of 2500 B.C.

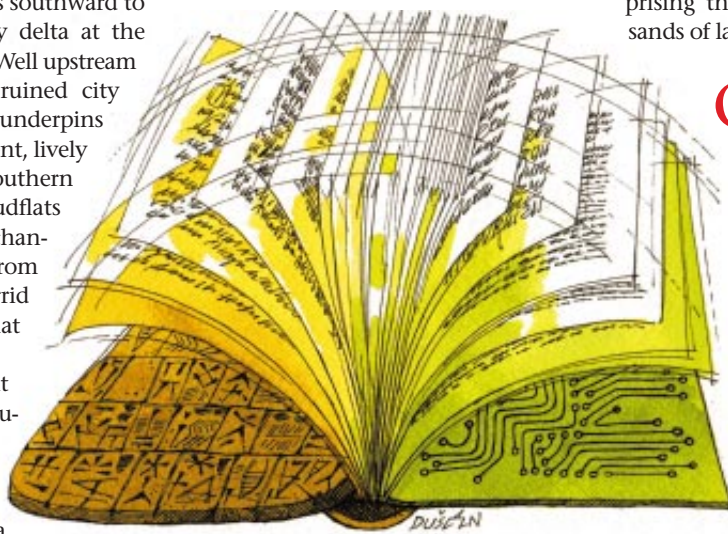
Our drama's stage is Mesopotamia. Across gently sloping lands the waters of two parallel rivers—the Euphrates to the west, the Tigris to the east—drain out the flanking Zagros Mountains southward to join into one wide, reedy delta at the shores of the Persian Gulf. Well upstream of their junction lies a ruined city whose aerial photograph underpins what we say about it. Ancient, lively and unique, Uruk was a southern lowlands city on broad mudflats traversed by many green channels, taking vital water from the Euphrates through torrid summers and winters that occasionally knew frost.

Bounding the settlement was a strong wall. A few thousand courtyard houses of sun-hardened clay bricks once sheltered 10,000 to 20,000 people around a high ceremonial mound and its ziggrats. There stood the public structures, a complex of large temples to the gods, where dwelt the priests who tended to divine and human needs on Uruk's lands and where worked the scribes whose accounts maintained the economic order for high and low. Such ancient cities are marked by three essentials: a wall, public buildings and a flow of worldly surplus, largely agrarian products but with bronze and silver, too, enabling steady growth and widespread trade.

Uruk (Erech, it is called in Genesis) had for two millennia been only one among hundreds of similar farming and pastoral villages in this land of Sumer. From their

images we know the Sumerians as a robust, stocky, black-haired people and from their written word as ones with a language all their own. Sumer was followed by later peoples with distinct languages, who nonetheless became for nearly three millennia cultural heirs to Sumer.

As power shifted far up the Tigris, the Sumerians leave our historical horizon, while newcomers densely peopled the larger surrounding area. These were speakers of Semitic tongues, first Babylonian, then Assyrian, and up to a dozen more, including scriptural Hebrew and Aramaic, all written in the Sumerian way. Almost



the entire populous valley is today within Iraq, where they use another Semitic language and write in alphabetic Arabic. Sumer is alive now only to those who dig and scrutinize clay.

The reason we can say so much about a people whose descendants we cannot now even point out is that theirs was the first society to leave a vivid written record. The written legacy of the kingships and empires of greater Mesopotamia were meticulously preserved in a script of Sumerian invention, its complex wedge marks, called cuneiform, etched in damp clay tablets baked for permanence. After some 3,500 years, this lasting syllabic record finally gave way to alphabets.

Writing on clay gave Uruk enduring influence, and abundant texts bring the vanished city to life in diverse tongues and places. As many as half a million cuneiform tablets, hand size up to book-page size, are now stored in the museums of the learned, from Baghdad upriver out to Moscow and Berkeley. Surely many more are waiting to be found. Those samples are of every quality: once prized accounts and receipts, schoolboys' lessons, litigation profound or droll, literary essays, erotica, mathematics—and entire ancient epics, centuries older than Father Abraham's. A mostly unread treasury, comprising the equivalent of tens of thousands of large printed volumes.

On a cool evening of the year we would number as 2500 B.C., a serious young citizen of Uruk is walking home along the street. His wife has laid in a cumin-spiced loaf or two of malted barley, prepared either to bake as bread or instead to crumble in water to let ferment a while, to become a popular tonic, barley beer. Dates with cream may end supper. The husband's garments are linen and wool, his shoes are leather, and a silver wire spirals around his wrist, from

which he might clip off a bit as proto-coinage. The street is busy with strangers. Ox carts rumble by laden with grain, their clever planked wheels as familiar as the simpler plows other oxen pull to furrow and channel the grainfields both within and without the walls. (Horses and war chariots will arrive hundreds of years later.)

Like many a hardworking person in Uruk, our subject pursues a career in information technology. The young scribe has studied for years to become a fluent writer of cuneiform; he goes daily to the temple precincts to assist estimators and archivists. He knows the songs and stories of that half-divine hero Gilgamesh, who

Continued on page 111

appointed clerk to the closet of the dowager Princess Augusta (she who helped to turn Kew into Royal Botanical Gardens). Young's royal closet job became available, so to speak, on the timely demise of its previous incumbent: Anglican vicar Stephen Hales, of whom I have spoken here before, a trustee of the colony of Georgia. Plant physiology ("Why does sap rise?") was a thing of his. That, and experiments I won't describe, but which involved cutting off frog heads and noting that the body still twitched some minutes after, thus giving rise to the visceral controversy: "Is the soul in the spinal cord?" Last but not least, Hales developed a ventilator to take fetid air out of shut-in spots like prisons, hospitals and ships. His airy gizmo was tested in 1751 on the good ship *Earl of Halifax* by hydrographer, surveyor and mineralogist Henry Ellis, who had just come back from being scientific observer on yet another of those Northwest Passage explorations that got nowhere but Hudson Bay.

Ellis was considered enough of a transatlantic expert for his patron, George Montagu Dunk (*sic*), Earl of Halifax, to make him governor of Georgia, then governor of Nova Scotia (you're getting a feel for the old-boy network?), and then adviser to Dunk (love that name!) on matters American. Which were becoming decidedly lively by 1770, when Dunk became Lord Privy Seal in the government of his nephew, Lord North. Conciliation with the colonists was decidedly not North's cup of tea (especially after that 1773 Boston business), and things went rapidly down the tubes when the North entirely underestimated the colonists' will to resist (and the French will to help them). Mind you, so did nearly everybody else.

Except for the reformist parliamentarian Lord Shelburne, who lobbied vociferously in favor of the American rebels with the help of research notes provided by his librarian and traveling companion. The gent in question was himself so pro-Revolutionary (American *and* French) that he was eventually run out of town (and country) after a patriotic mob burned to ashes his house, lab and paperwork. But not before he had become the co-discoverer of oxygen, the inventor of soda water, and a grand scientific panjandrum. Joseph Priestley (for it was he) started out as a free-church preacher and in 1780 was running a Sunday school in Birmingham.

A teacher of his at the school was Rowland Hill's old man. SA

Wonders, continued from page 109

first built Uruk's city walls. (No written version of that tale appears until a few centuries later.) It is a fair conjecture that he might have formed some rough picture even of 1,000 years back, when the records of Uruk first became articulate. (Nowadays we can read 1,000 tablets and fragments stored from that early epoch, five out of six of them only number-filled accounting documents, as antique to our scribe as the comet of 1066 is to us.)

The recent brilliant unraveling of the evolution of cuneiform itself out of a long period of gamelike tokens and their impressions in clay is no part of this column, but compelling evidence centers on Uruk's place and time for the first copious records. Decipherment is active today; it grew largely out of the study of a monumental trilingual cuneiform text, carved on a rocky mountainside by ancient command of Darius of Persia, "King of Kings." In Victorian times that became the key to cuneiform, as the Rosetta stone had earlier served Egypt's hieroglyphics. Cuneiform texts reach us mainly out of buried or ruined libraries; Uruk's own lasted about 3,000 years, about as long as the first Imperial Archives of China. (No al-

phabetic collection so far has lasted even 1,000 years.) One poignant difference: when in our time a library burns, its books suffer even more than the building does, but clay tablets survived the structure's ruin, for tablets largely bake brick-red to new durability.

The recognition of our urban lives in old Uruk extends to its troubles. Millennia of irrigation led to the failure of the crops—even salt-tolerant barley—by salinization of the soil. Perhaps that is no great failure but simply part of the beginning of city life. Civilization may start where it is easy to begin, only later to seek sustainability. The peoples far upriver today still win good crops wherever variable floodwaters deposit renewing silt in their canal complex. The strong stratification of Uruk society is plain, too. In one document the young scribes report their fathers' occupations; these early IT professionals are sons, very rarely daughters, of the well-off. (Sumerian tavern keepers, though, were often women of means and status.) It was mainly war that brought ruin to walled Uruk and to all its royal successors. Whether we moderns will better manage our own overarmed world is far from a foregone conclusion. SA

www.sciam.com

This Month's Special Feature

MARTIAN WINDS

Armed with new data about the Red Planet, a few scientists now suggest that wind did some of the work previously credited to water

sciam.com delivers the latest headline news in science daily!

PAT RAWLINGS



The Open-Heart Open

Cardiac patients might want to consider the links—and not those of the sausage variety, says **Steve Mirsky**

Working too hard,” pop singer Billy Joel once warned, “can give you a heart attack-ack-ack-ack.” You ought to know by now, however, that playing too hard also may be hazardous to your health, especially for those with a history of cardiac problems. One of the challenges facing heart patients is to engage in physical activity strenuous enough to impart beneficial cardiovascular effects without being so vigorous as to impart death.

Fortunately, German researchers at the Center for Cardiovascular Diseases in Rottenberg and at the Institute for Sports Medicine at the University of Giessen have possibly pinpointed the proper pastime. As they reveal in last October’s issue of *Medicine and Science in Sports and Exercise*, it is, surprisingly enough, golf, which is called golf, it’s been said, only because the other four-letter words were already taken.

Of course, golf and medicine have a long history, but the connection has been for the most part related to an inability for patients to schedule Wednesday afternoon appointments. To gauge golf’s appropriateness as medicine, the researchers

put together a special event. “After written informed consent was obtained,” they report, “20 male golfers with cardiac diseases and 8 healthy controls participated in a golf tournament after their examination in the hospital,” which sounds like a typical weekend on the PGA Senior Tour. In fact, however, the intrepid golfers in the Infarct Invitational one-upped the geezer pros: the test subjects schlepped their own clubs on handcars, thus adding to their workout.

The researchers comment that suitable physical activity for heart patients should reduce cardiovascular risk factors, increase endurance and help to reintegrate patients socially. They suspected that golf could be ideal because the handicapping system allows players of all levels and ages to compete together and because the game poses physical challenges such as walking and other coordinated movements. The investigators also cryptically note that trauma risk is minor “unless the rules and etiquette are violated,” a possible veiled reference to those rare occasions when a golfer ignores the ball and swings instead at a gabby opponent.

In addition, substantial mental issues make golf more difficult than it might ap-

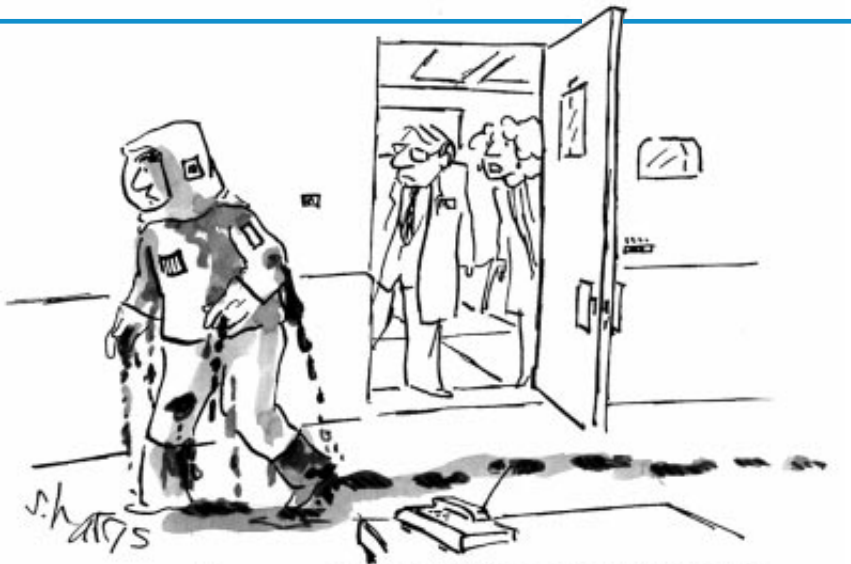
pear. Some golfers can’t tee off if anyone watches them, which pretty much rules out playing the big-money televised tournaments. Many develop the “yips,” a case of nerves that transforms a smooth putting stroke into what looks like a cross between water divination and an attempt to kill a rat with a shovel. Participants in the Cardiac Classic were faced with an additional psychological stress: “Before the competition,” the researchers explain, “the golfers were informed that the winners would receive valuable prizes,” which might ordinarily mean a new car but in this case could be nitroglycerin.

In another wrinkle that separated the heart study from more mundane golf outings, players had their heart rate recorded the entire time and their blood pressure measured after every three holes. Although continuous blood pressure monitoring would have been preferable, the study’s authors decided against it “in order not to disturb the golf swing,” thereby preventing incidents such as:

Pffft, pffft, pffft, pffft, pffft.
“Do you mind?”
“Sorry.”

Now, I don’t actually know much about golf, although I did once fade a 6-iron with tour sauce onto an elevated dance floor and drain the bird to take a nassau. Despite my ignorance, I was fascinated to learn about the German study, especially as it dovetailed with another recent finding—according to a report in the journal *Circulation*, patients who kept physically active after a first heart attack had a 60 percent lower risk of subsequent attacks than their sedentary counterparts.

Whether fit golfers benefit from the 18-hole hike remains unclear. But based on the data compiled during the tournament, golf indeed appears to have the potential to make coronary patients hearty, as in fewer ventricular arrhythmias, and hale, as in Irwin. For the cardiac-conscious, golf as exercise is thus much like the third bear bed sampled by Goldilocks: neither too hard nor too soft, but just right. SA



“Something must have gone wrong in the clean room.”