IEEE SPECTRUM

FOR THE TECHNOLOGY INSIDER | 05.17

The Shiny Bits

Kees Immink's peculiar genius helped create the CD, and now he's applying it to DNA storage **P. 30**

HOLLYWOOD'S DIGITAL PROBLEM

Archiving has long been a sore spot for the movie studios. But now, digital filmmaking has turned a nuisance into a full-blown crisis

P. 24

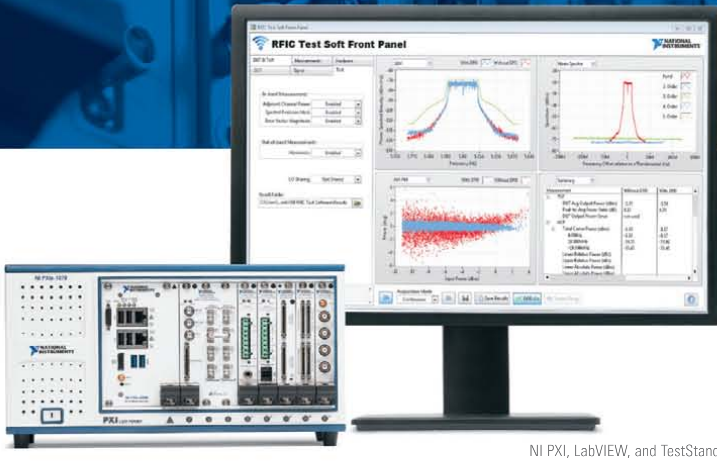
IEEE

FEATURES_05.17

IEEE SPECTRUM

On the cover **Illustration for *IEEE Spectrum* by Francesco Muzzi/StoryTK**

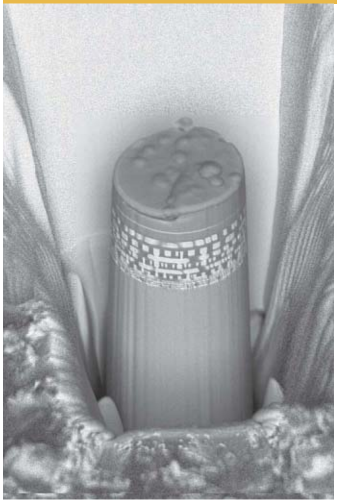
PHOTOGRAPH BY Ruud Baan

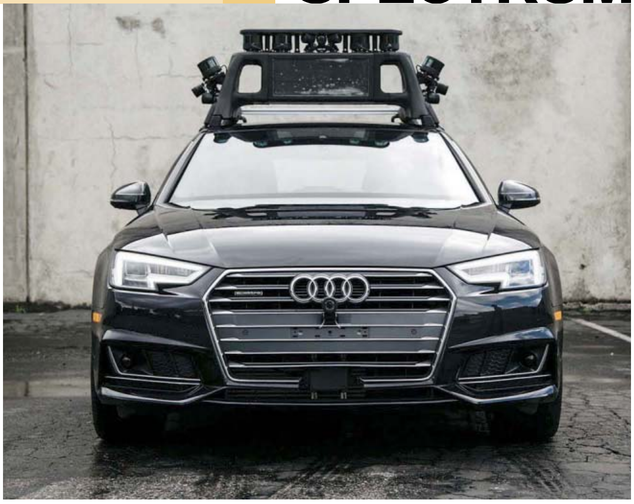


NI PXI, LabVIEW, and TestStand

DEPARTMENTS_05.17

IEEE SPECTRUM





Online

Spectrum.ieee.org

Deep Learning First
Drive.ai's autonomous vehicles use deep learning to turn past experience into better decisions, allowing its cars to navigate (mostly) autonomously—even in situations that are notoriously difficult for self-driving cars, such as at night or when it's raining. Watch it in action: http://spectrum.ieee.org/driveai0517

ADDITIONAL RESOURCES

Tech Insider / Webinars

Available at spectrum.ieee.org/webinar

- Phase-Shifter Technology and Applications—18 May

White Papers

Available at http://spectrum.ieee.org/whitepapers

- Your Design Problems, Answered
- Download 6 Essential Tips for Getting the Most Out of Your Oscilloscope
- Leveraging Natural-Language Processing in Requirements Analysis
- Introduction to EMC Compatibility
- Broadband Chokes for Bias Tee Applications: How to Successfully Apply a DC Bias Onto an RF Line

The Institute

Available at theinstitute.ieee.org

- **TECHNOLOGY AND ETHICS** This special report explores the ways in which IEEE is addressing the ethical implications of technology through its new IEEE TechEthics program, as well as a global initiative that focuses on the ethics behind AI and autonomous systems. You'll also find IEEE resources to help identify and resolve ethical issues.
- **HIGH-TECH SPY** Harold Lipset, a private detective and pioneer of electronic surveillance, invented such gadgets as a listening device that could be inserted into a martini olive.
- **HELPING THE HOMELESS** Two IEEE members developed HopeOneSource, a text-messaging system that notifies homeless people in Washington, D.C., about social services available to them, including medical centers and soup kitchens.

FROM LEFT: MIRKO HOLLER; MYCROFT AI; DRIVE.AI

IEEE SPECTRUM
(ISSN 0018-9235) is published monthly by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. © 2017 by The Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, New York, NY 10016-5997, U.S.A. Volume No. 54, issue No. 5, International edition. The editorial content of IEEE Spectrum magazine does not represent official positions of the IEEE or its organizational units. Canadian Post International Publications Mail (Canadian Distribution) Sales Agreement No. 40013087. Return undeliverable Canadian addresses to: Circulation Department, IEEE Spectrum, Box 1051, Fort Erie, ON L2A 6C7. Cable address: ITRIPLEE. Fax: +1 212 419 7570. INTERNET: spectrum@ieee.org. ANNUAL SUBSCRIPTIONS: IEEE Members: $21.40 included in dues. Libraries/institutions: $399. POSTMASTER: Please send address changes to IEEE Spectrum, c/o Coding Department, IEEE Service Center, 445 Hoes Lane, Box 1331, Piscataway, NJ 08855. Periodicals postage paid at New York, NY, and additional mailing offices. Canadian GST #125634188. Printed at 120 Donnelley Dr., Glasgow, KY 42141-1060, U.S.A. IEEE Spectrum circulation is audited by BPA Worldwide. IEEE Spectrum is a member of the Association of Business Information & Media Companies, the Association of Magazine Media, and Association Media & Publishing. IEEE prohibits discrimination, harassment, and bullying. For more information, visit http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

BACK STORY_

PINK RIBBON RED RIBBON

Fighting Stigma With High Tech

CELINA SCHOCKEN ISN'T THE KIND OF CEO who stays at her desk. As the head of Pink Ribbon Red Ribbon, a nonprofit dedicated to improving women's health in the developing world, Schocken [above right, with a pregnant woman in Uganda] spends almost half of her time in Africa or South America, visiting health clinics and meeting with local partners.

Soon, in service to her group's cervical-cancer-screening campaign, she'll also be showing off a new gadget. Her colleagues in Africa will begin field-testing an AI-enabled tool to help clinic workers spot the early warning signs of cancer [see "AI Medicine Comes to Africa's Rural Clinics," in this issue]. The tool is now being readied for deployment by artificial intelligence researchers at Intellectual Ventures Laboratory, an innovation hub in Bellevue, Wash.

Schocken cites estimates that only 5 percent of women in Africa have been screened for cervical cancer. Stigma is partially to blame, she says; some cultures believe that women who get cervical or breast cancer are being punished by the gods for unfaithfulness, while others blame witchcraft. Women often don't come in for screening until their symptoms can no longer be ignored, at which point their odds of survival are reduced. "It's similar to where we were with HIV about 15 years ago," Schocken says. "People believe that if you get it, you're going to die."

Schocken hopes that if early screening becomes the norm, survival rates will improve and the stigma around cancer will gradually decrease. Technology is the key to this transformation, she says. "Today's screening methods require too many health workers, and they're too hard to supervise," she says. "We need this new technology to go to scale." ■

CITING ARTICLES IN IEEE SPECTRUM *IEEE Spectrum* publishes an international and a North American edition, as indicated at the bottom of each page. Both have the same editorial content, but because of differences in advertising, page numbers may differ. In citations, you should include the issue designation. For example, Past Forward is in *IEEE Spectrum*, Vol. 54, no. 5 (INT), May 2017, p. 52, or in *IEEE Spectrum*, Vol. 54, no. 5 (NA), May 2017, p. 56.

CONTRIBUTORS_

Cary Champlin

Champlin is a principal engineer at Intellectual Ventures Laboratory, in Bellevue, Wash. In this issue, he, David Bell, and Celina Schocken describe their pioneering effort to improve cervical cancer detection tools with artificial intelligence [p. 40]. This type of project is what brought Champlin to the lab. "It sounds cheesy," he says with a laugh, "but I didn't want to just build toys or make packages ship faster. I wanted to make a difference in people's lives."

Francesco Muzzi

Muzzi is an award-winning graphic designer and illustrator at StoryTK, a storytelling and design studio in Oakland, Calif. He created the cover and opening illustration for "The Lost Picture Show" [p. 24]. "The cover depicts the expensive and seemingly endless cycle of digital conversions haunting our favorite films," Muzzi says. "Inside, the tape from the cover breaks apart and runs over the pages, suggesting the messy, unsolvable aspect of this business."

Marty Perlmutter

Perlmutter, who's based in Southern California, has worked in interactive video and new media for four decades. While giving a presentation on virtual reality at Sony Pictures last year, he first learned of the quiet crisis in digital film archiving, which he writes about in "The Lost Picture Show" [p. 24]. He found studio executives reluctant to talk publicly about the problem, a reticence he attributes to their desperation over the lack of a good solution.

Mark M. Tehranipoor

Tehranipoor is codirector of the University of Florida's Security and Assurance Laboratory, where he helps curate the lab's growing collection of cloned electronics products, like those addressed in "Invasion of the Hardware Snatchers" [p. 34]. He and coauthor Ujjwal Guin wrote the book *Counterfeit Integrated Circuits: Detection and Avoidance*. And with coauthor Swarup Bhunia, he recently started the *Journal of Hardware and Systems Security*.

W. Wayt Gibbs

Gibbs is a freelance writer and editorial director at Intellectual Ventures Laboratory, where he works with Nathan Myhrvold. In this issue, Gibbs describes creating voice-controlled gadgets using Amazon's Alexa and the open-source Mycroft [p. 18]. "These AI assistants can seem like magic," says Gibbs, "but the idea of turning over a lot of personal information to a giant tech company makes me uncomfortable. I'm happy to see alternatives like Mycroft."

IEEE SPECTRUM | Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page | Qmags THE WORLD'S NEWSSTAND®

Electronic License Plates for Drones

Should drones be required to broadcast an identifying code by radio?

In late 2015, mandatory drone registration went into effect in the United States. Since then, anyone who wants to fly a drone (or model aircraft) weighing over 0.55 pound (0.25 kilogram) must register with the U.S. Federal Aviation Administration to receive a unique identification number. This number needs to be placed on the drone, but there is no requirement for the tiny aircraft to broadcast signals to allow for remote identification. That might change in the future.

The FAA Extension, Safety, and Security Act of 2016 required the FAA administrator to "convene industry stakeholders to facilitate the development of consensus standards for remotely identifying operators and owners of unmanned aircraft systems and associated unmanned aircraft." Earlier this year, DJI, the world's largest commercial drone manufacturer, outlined a general scheme for doing just that.

The Chinese company's proposal attempts to balance the public's interest in being able to identify who is using a drone at a particular place and time with the privacy interests of the drone's owner or operator. As DJI points out, drone operators may want to maintain anonymity even if there are people around to witness their flights. Suppose, for example, a company is surveying land in anticipation of purchasing and developing it. That company might not want to clue in its competitors. Or perhaps the drone is being flown for the purposes of investigative journalism, in which case the journalists involved might not want others to know what they are looking into.

It's also easy to understand why law-enforcement or regulatory authorities would want to identify the owner or operator of a drone, say, if somebody felt the drone was invading their privacy or if a drone was being flown close to a nuclear power plant.

DJI's solution is to require drones to broadcast an identifying code by radio, by embedding that code in the telemetry or video transmissions. It would not include the name and address of the owner, but authorities would be able to use it to look up the information in a nonpublic database. Essentially, DJI is proposing electronic license plates for drones.

The requirement would not apply to all drones, though: Those at the small end of the spectrum would be exempt, just as they are from current FAA regulations requiring the registration of drones and model aircraft that exceed 0.25 kg. (By the way, DJI thinks that the 0.25-kg weight threshold is too low and that a more reasonable value should apply to both the current registration requirement and to any future requirement to broadcast an ID.)

Any new drone regulations are bound to be controversial, but DJI's proposals sound pretty reasonable to me. I am skeptical, though, about how well drone owners would comply. Consider that the FAA estimated that about a million small drones would be sold during the holiday season of 2015, shortly after the registration rules were put in place. But by the following February, only 325,000 registrations had been received.

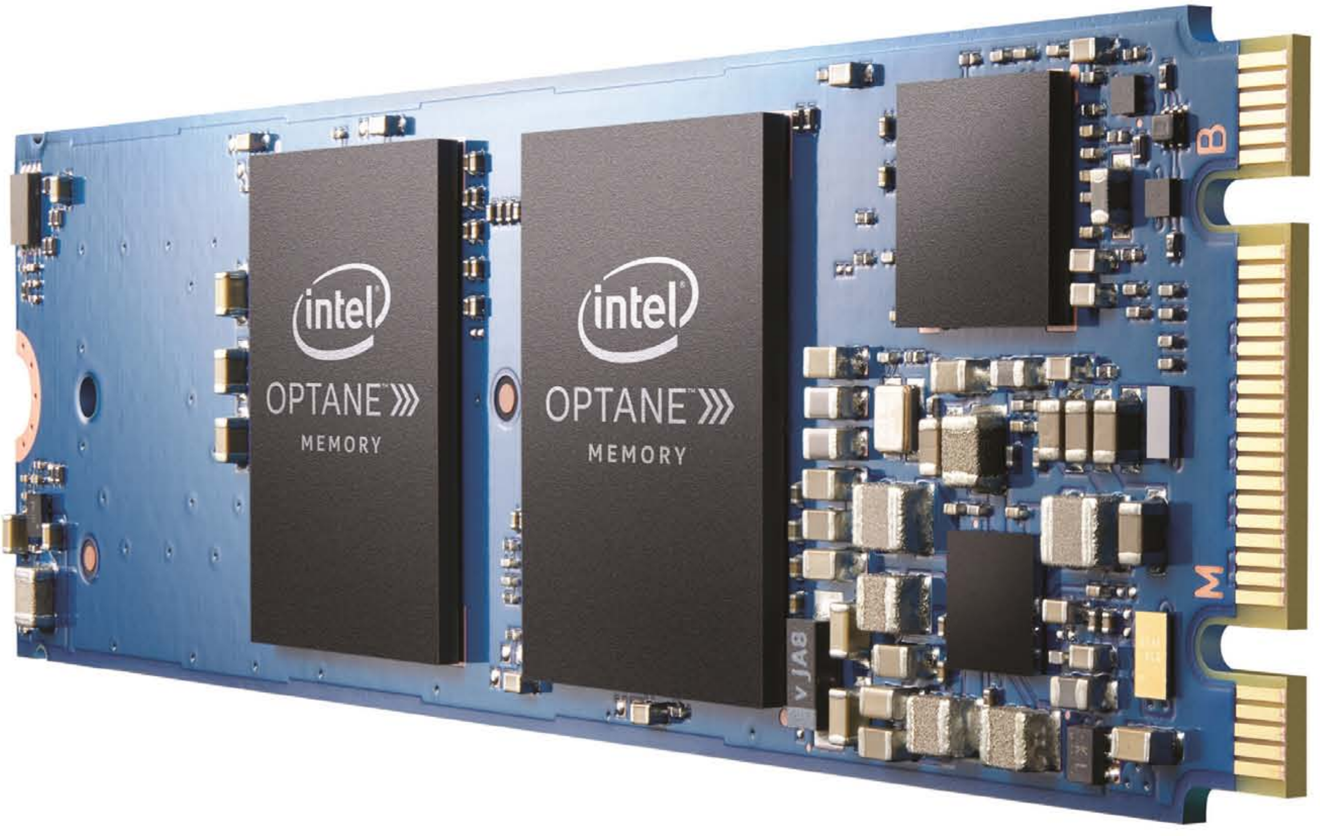
Now that doesn't necessarily mean that there were 600,000 or more scofflaw drones. Many drones sold as gifts could fall under the 0.25-kg limit. And one registration can cover multiple drones. Still, I wonder whether most owners would bother registering their drones.

My biggest concern is that the FAA would attempt to boost compliance by imposing harsh penalties for those who don't register. So far it hasn't happened—the FAA seems more keen to educate drone flyers than to fine them. But that could change, especially if hundreds of thousands of drones start broadcasting empty #0000000000 IDs. —DAVID SCHNEIDER

SAVERIO BLASI/GETTY IMAGES

POST YOUR COMMENTS at http://spectrum.ieee.org/spectrallines0517

NEWS



HAS INTEL CREATED A UNIVERSAL MEMORY TECHNOLOGY?

The company aims high with its mysterious XPoint memory in the Optane solid-state drive

LEARNING TO DRIVE: Intel's Optane storage device is a zippier solid-state drive. Memory modules are next.

INTEL

Today's computers shuttle data around a byzantine system of several different kinds of short- and long-term memory. No wonder, then, that engineers have long dreamed of one memory technology to rule them all, a universal memory that would simplify computing and streamline the path of data.

In March, Intel announced that it will sell to data centers a new kind of solid-state drive, called Optane, that it says could lead to this kind of simplification. Optane drives are nonvolatile, like flash memory, which means that they should use relatively little standby power and that they're fast, like DRAM. "It really starts to marry the worlds of memory and storage together," company CEO Brian Krzanich says in a promotional video, over the swells of heroic music. The technology "comes close to being the holy grail of memory," says »

DATA BYTE

Read latency in microseconds of optane drive versus solid-state drive when stressed by write demands:

20/700

Intel executive vice president William Holt in the same video.

Whether 3D XPoint, the mystery technology inside Optane, can live up to this promise is likely to depend on the performance it delivers as well as Intel's ability to scale up manufacturing using new materials and build out the right market. The 375-gigabyte Optane drive on offer now costs US $1,520, about three times the price of an equivalent solid-state drive.

This first product will enable data centers to do more with a smaller number of servers, says James Myers, who works on nonvolatile memory architecture at Intel. Myers gives an example of servers running a MySQL database, which, among other things, apps use to store instant messages. An equivalent flash drive can perform 1,400 such transactions per second; the Optane drive can perform over 16,000.

The Optane drive was announced with bombast, but the company is coy about the technology behind it. Myers says "3D" refers to the fact that the memory cells are stacked; "XPoint" alludes to the way the memory elements are arranged. While flash memory elements must be read and written in groups, XPoint elements—situated at the crossing point of interconnects—can be addressed individually. Myers says that this architecture, and something inherent to the storage materials themselves, makes 3D XPoint faster than flash memory.

Intel, which initially developed 3D XPoint in conjunction with Micron, won't say what the technology really is, but this doesn't seem to bother researchers or analysts. "Everyone seems to think it's phase-change memory," says semiconductor analyst Jim Handy. "I don't care." What matters to him—and, Intel hopes, its customers—is the performance.

The complexity of today's memory hierarchy—a combination that often includes magnetic disks and flash for storage and DRAM and static RAM for memory—is a necessary evil. Each technology has its own strengths, so they must be combined. Data are shuttled around from speedy but expensive SRAM caches—which are close to the processor and embedded within it—to slower, less expensive (but still pricey) DRAM. Finally, data are stored in slow but reliable flash or hard-disk drives, or both. Even if it's not possible to do it all in one memory technology, using only one for working memory close to the processor and one for longer-term storage would help simplify things. Intel says that XPoint memory could provide a speedier alternative to flash memory and magnetic hard disks. The company has also suggested it could supplement or supplant DRAM.

"DRAM is unique in its ability to waste power, so anything you can do to get rid of it is great," says Handy. For example, Google is thought to store the index of the entire Internet on several power-hungry, quick-access DRAM servers. If the company could switch this over to 3D XPoint—which Intel claims has 10 times the density of DRAM—Google could use fewer servers and thus save power and money, according to Steven Swanson, a computer scientist at the University of California, San Diego.

Intel is providing software that will enable computers to operate Optane as memory as opposed to storage, but it will be slow. Optane drive latencies max out at 7 or 8 microseconds—way faster than flash, which takes hundreds of microseconds, but not touching DRAM's low hundreds of nanoseconds. The Optane drives are fettered by the interface they use: They connect to the storage interface, not the memory interface.

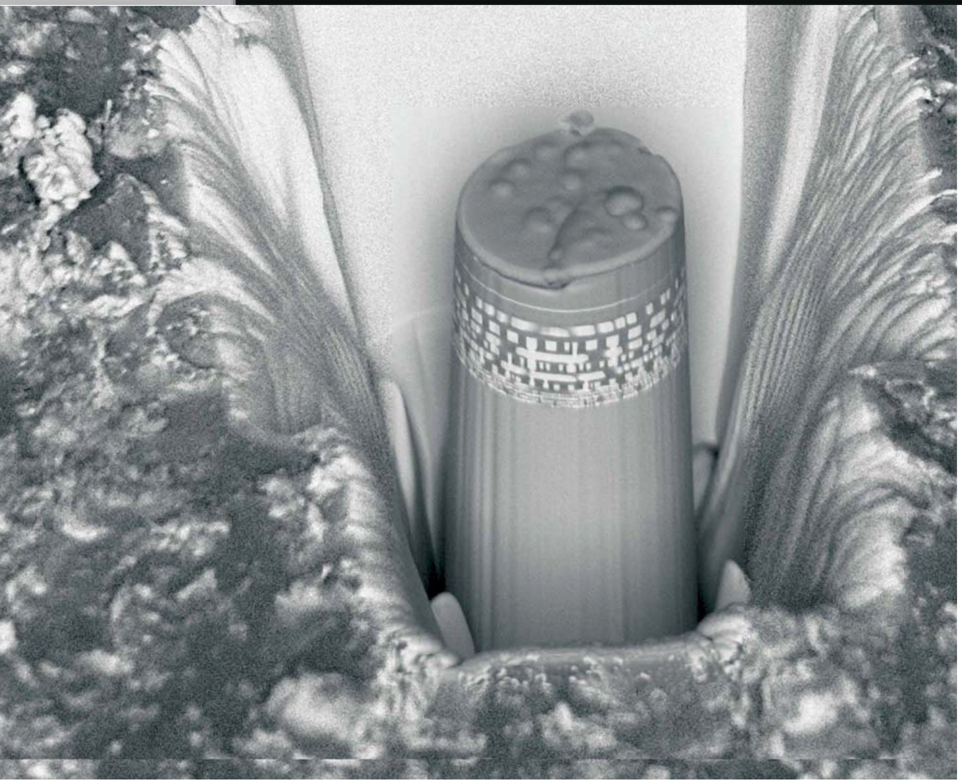
In the short term, Intel's drives are not likely to replace any existing memory technologies but will instead supplement them, says Swanson, who built a research drive based on the company's 3D XPoint technology in 2011. Swanson expects the path to memory and storage simplification to be complex because computing systems will have to be redesigned to route data in new ways.

Swanson and Handy believe Intel started with storage to help smooth out some of the risk in launching a new memory technology. Making Optane a memory requires new circuit board designs and cooperation from programmers. To get those, Intel needs to show that there is a market for 3D XPoint and demonstrate its reliability. Intel says that a product using a memory interface will be out in 2018. Even using the interface, 3DXPoint still won't be quite as fast as DRAM, but Intel promises that it will be denser and less expensive.

The success of this new memory, then, will hinge on data centers taking it up in a less than ideal initial form while the company works on scaling up production. Even though the flaws in today's memory and storage hierarchy are universally acknowledged, trying to change it is a risky move. "Almost all memory companies have one or two potential competitors to this technology, and they're all waiting to see what happens before they jump into [a] big investment," says Swanson.

Memory enthusiasts disagree about whether a true universal memory is even physically possible. It is perhaps most useful as a goal to guide the computer industry forward. "The concept of the universal memory is attractive because the idea is to simplify," says Wei Lu, a computer scientist at the University of Michigan and chief scientist at Crossbar, a resistive RAM startup company. "We have a big-data problem, and today's computers are fundamentally not good at this." —KATHERINE BOURZAC

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/inteloptane0517



TOWER OF PTYCHOGRAPY: Scientists used X-ray ptychography to image the insides of this section of an Intel processor.

MIRKO HOLLER

3D X-RAY TECH FOR EASY REVERSE ENGINEERING OF ICs

Researchers map an Intel processor down to its transistors

A team of researchers based in Switzerland is on the way to laying bare much of the secret technology inside commercial processors. They pointed a beam of X-rays at a piece of an Intel processor and were able to reconstruct the chip's warren of transistors and wiring in three dimensions. In the future, the team says, this imaging technique could be extended to create high-resolution, large-scale images of the interiors of chips.

The technique is a significant departure from the way the chip industry currently looks inside finished chips to reverse engineer them or check that intellectual property hasn't been misused. Today, reverse-engineering outfits progressively remove layers of a processor and take electron microscope images of one small patch of the chip at a time.

But "all it takes is a few more years of this kind of work and you'll pop in your chip and out comes the schematic," says Anthony Levi, of the University of Southern California. "Total transparency in chip manufacturing is on the horizon. This is going to force a rethink of what computing is," he says, and what it means for a company to add value in the computing industry.

Even if this approach isn't widely adopted to tear down competitors' chips, it could find a use in other applications. One of those is verifying that a chip has only the features it is intended to have, and that a "hardware Trojan"–added circuitry that could be used for malicious purposes–hasn't been introduced.

The work was conducted at the Paul Scherrer Institute's Swiss Light Source. The facility is a synchrotron; it accelerates electrons to close to the speed of light in order to generate beams of X-rays.

When the Swiss team shines an X-ray beam through a 22-nanometer-generation Intel processor, the various circuit components–its copper wires and silicon transistors, and other features–scatter the light in different ways and cause constructive and destructive interference. The researchers pointed the beam at their sample from a number of different angles, and using a technique called X-ray ptychography, they could reconstruct a chip's internal structure from the resulting diffraction patterns.

The resolution for this technique in any one direction is 14.6 nm, which rendered only a fairly blurry image of the individual transistor components. The resolution can be improved, but thanks to Moore's Law it will be a race to keep up with the chip industry.

This is not the first time that researchers have attempted to use X-rays to image the interior of integrated circuits, says team member Gabriel Aeppli. But "the resolution is better than it's been in the past. The scale is also larger," he says. "It's a huge chunk of the chip compared to what you could do with any other technique."

NEWS

The imaging itself was no small feat. The sample has to remain stable, and interferometers must be used to continuously measure its position. It took about 24 hours to perform the X-ray measurements, and the data processing took about as long, says team leader Mirko Holler. But additional computers should easily speed the processing, he says. And improvements to X-ray sources as well as other parts of the experimental apparatus could improve the imaging speed by a factor of 1,000.

"What they've done is a proof of concept, and it's quite impressive at that level," says Dick James, an emeritus fellow at TechInsights, which strips chips down to the transistor level to examine how the devices are built and wired together.

James says there are practical limitations to the amount of circuitry that can be analyzed using those traditional tear-down techniques, which, depending on the packaging, sometimes start with a bath in sulfuric acid. But he notes that a lot can be gleaned by piecing together small electron microscope pictures of a chip: "You can get most of the cell library by looking at smaller areas."

Although X-ray ptychography does promise bigger, high-resolution views, it faces several obstacles, James says. For one, chip feature sizes at the cutting edge could prove a challenge for its resolution. "The industry is getting ahead of this technique already," he says. Another impediment is the need for a synchrotron source. These machines aren't likely to find their way into chipmaking fabs, although there are a number of such facilities around the world, and it is possible to rent time on them.

Because of these limitations, James says, the best application for this imaging technique could be on chips that are made using older manufacturing processes, and thus have larger features. This is the case for a number of chips used in military and space applications. "If you can look at the whole chip, then you can compare the chip with the original design," James says, and "do a direct comparison [to] see if there are any obvious faults or any extra circuitry that's been put in."

—RACHEL COURTLAND

The X-ray measurements took 24 hours to perform, but improvements to the X-ray sources and other parts of the apparatus could improve imaging speed by a factor of 10,000

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/xraychip0517

TRUMP'S PLANS COULD SCUTTLE NEW HYBRID CAR TECH IN U.S.

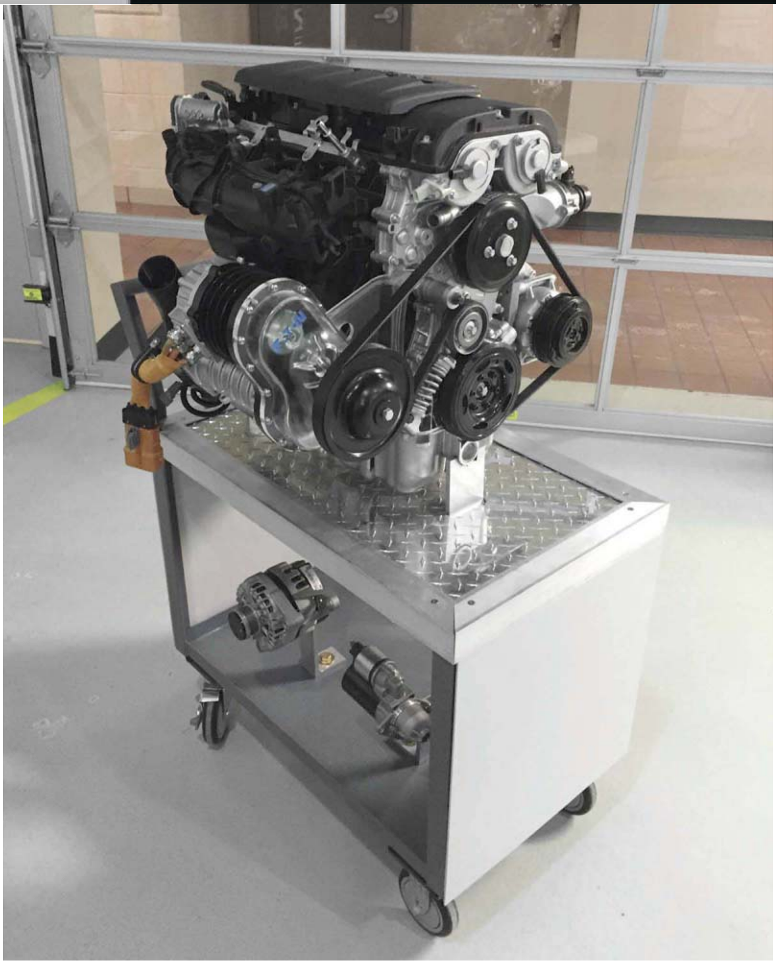
Weakened EPA rules could stall deployment of fuel-saving "mild hybrid" technology

Global policy efforts to boost fuel economy are progressively electrifying the automobile and expanding technology options to cut greenhouse gas emissions. One of the most exciting recent developments is, however, at risk of failure in the United States: a new generation of "mild hybrid" technology that could be undercut by President Donald Trump's attempts to stall U.S. climate policy.

In March, Trump ordered the Environmental Protection Agency to rethink fuel economy mandates set under President Obama. Those mandates require cars and light trucks sold in the United States between 2022 and 2025 to deliver an average of 6.5 liters per 100 kilometers (36 miles per gallon) in real-world driving– a 40 percent efficiency improvement over the average vehicle sold in 2016.

Mild hybrids rely on batteries, motors, and other components that are cheaper than the ones used in full hybrids, such as Chrysler's new Pacifica hybrid minivan [see "2017 Top Ten Tech Cars," *IEEE Spectrum*, April 2017]. By providing more targeted support to a vehicle's internal combustion engine, mild hybrids are both cheaper to produce and more cost-effective at cutting carbon dioxide emissions than full hybrids, according to vehicle experts. "You get about 70 percent of the fuel economy benefit of a full strong-hybrid system but at about 30 percent of the cost," says Sam Abuelsamid, a Detroit-based senior research analyst with energy consultancy Navigant Research. For the U.S. market, mild-hybrid tech would enable popular pickup trucks and SUVs to make do with cheaper four-cylinder engines instead of V-6s.

The technology itself is getting cheaper as lower-voltage designs replace their 100- to 120-volt prede-



MILD YET ELECTRIC: Eaton's Electrically Assisted Variable Speed supercharger uses braking energy and engine exhaust to power a mild-hybrid system.

EATON

cessors. Today's mild-hybrid standard is 48 volts, which allows for lighter batteries with fewer cells and requires far less electrocution shielding.

Most mild hybrids employ a motor-generator that absorbs braking energy to charge a lithium battery and then returns that energy to the engine's crankshaft during acceleration. An example is the electrified Silverado pickup that General Motors began offering in California last year and expanded in April to Hawaii, Oregon, Texas, and Washington. The technology cuts fuel use in city driving by up to 13 percent: It helps keep half of the V-8 engine's cylinders deactivated and lets it turn off at red lights.

Daimler plans to make 48-V motor-generators ubiquitous in Mercedes-Benz vehicles, starting with an electrically boosted S-Class sedan launching later this year. Besides its motor-generator, the mild hybrid's electrical system will supercharge the turbo engine, eliminating the traditional exhaust-driven turbo's torque lag.

Top-tier suppliers, meanwhile, have pushed the mild-hybrid concept further. Both Ohio-based Eaton and Ricardo, based in the United Kingdom, are testing mild hybrids that use engine exhaust to generate electricity, adding to the regenerative braking that charges the hybrid battery. Mihai Dorobantu, technology planning director for Eaton's vehicle group, says the company is showing off a demonstrator vehicle to automakers at its proving grounds in Marshall, Mich.

The challenge to selling such efficiency-boosting technology is that it increases a vehicle's sticker price. GM charges a US $500 premium on its hybrid Silverados. With gas prices in the United States currently averaging a low $2.30, the added system will take seven years to pay for itself. Most consumers—and the automakers serving them—do not make such investments unless mandates force their hands.

Mild hybrids were projected to make up 18 percent of cars and light trucks sold in 2025 in the United States, according to a joint analysis issued last year by the EPA, the National Highway Traffic Safety Administration, and California regulators. That number was 2.5 times as large as the combined share projected for full hybrids and battery EVs.

That projection looks optimistic under the Trump administration, according to Abuelsamid. He says Trump's EPA may retain Obama's standards but make them more flexible and lower penalties—in practice, rendering the rules voluntary. Abuelsamid says he had expected significant mild-hybrid growth from 2019 in the United States, but that could be delayed by four to five years.

Delay could hurt companies like Eaton that spent big on development. "Weakened or delayed standards would ultimately leave technology on the table," says Nic Lutsey, program director at the Washington, D.C.-based International Council on Clean Transportation (ICCT). A recent survey of suppliers by Ricardo found that 70 percent favored maintaining the EPA's rules. "They've made...those investments, and now they want to make sure there's still a market for them," Lutsey says.

That said, suppliers still have tougher mandates in markets such as China and Europe to rely on. European regulations, for example, limit gasoline-fueled cars to just 4.1 L/100 km (57 mpg) from 2021, and will require a further 17 to 28 percent reduction from 2025. Of the 5.6 million 48-V electric turbochargers that Navigant projects will be installed worldwide in 2025, 85 percent are for Europe and Asia.

As Dorobantu puts it, the "quest for higher and higher efficiency in vehicles" are "macro trends" that are independent of the EPA's standards. "It is those macro trends that are guiding our investments."
—PETER FAIRLEY

NEWS

A NEW SILICON PV RECORD

Japanese group pushes closer to the 29 percent efficiency limit

There's a new king of silicon solar cells. Researchers at Kaneka Corp., a resin and plastics manufacturer based in Osaka, have come up with a single-crystal heterojunction silicon solar cell that achieves a record-breaking 26.3 percent efficiency–a 0.7 percent increase over the previous record. That may not seem like a lot, but it's really a big step when you consider that the theoretical maximum efficiency for such cells–which make up about one-quarter of annual global production by gigawatts–is just over 29 percent.

Kaneka is a member of a project set up by the New Energy and Industrial Technology Development Organization (NEDO), a Japanese government entity established to help develop and promote new energy technologies.

In producing its new 180-square-centimeter prototype cell, Kaneka improved on several of the technologies promoted by NEDO. Chief among them is Kaneka's proprietary heterojunction technology. It reduces recombination, or resistive loss, where instead of exiting the device to produce electricity, positive and negative charges in the solar cell combine and produce heat.

In addition, Kaneka improved the energy-collection efficiency of the solar cell's interdigitated electrodes. What's more, the company moved the grid of electrodes from the front of the cell–the light-receiving area–to the back, boosting the amount of sunlight entering the cell.

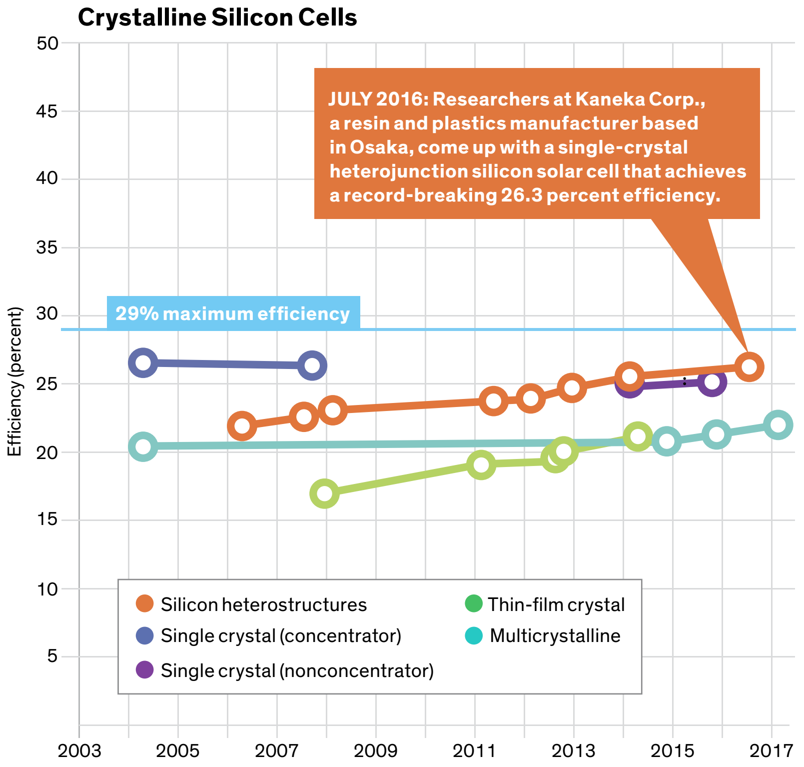
Panasonic, a member of the same NEDO project and holder of the previous energy-conversion-rate record of 25.6 percent for its 144-cm^2 solar cell set in 2014, employed the same key features in its device.

"But there are many types of materials, manufacturing processes, and architectures that can be selected," says Kunta Yoshikawa, a member of the Kaneka research team that worked on the new solar cell. "We achieved 26.3 percent efficiency by developing our [chemical vapor deposition] technology, optical management, and electrical-contact technology."

The biggest challenge the company faced in producing the prototype was obtaining a high degree of balance between the cell's lifetime and its optical characteristics while simultaneously reducing its internal resistance. "Although it is possible to obtain an outstanding value for a single characteristic, it is extremely difficult to balance all three properties to a high degree in one device," Yoshikawa says. For instance, if the prototype cell's lifetime had been shorter in relation to its optics and internal resistance, the cell's conversion efficiency, in theory, could have fallen to just 20 percent, he says.

"We overcame this challenge by designing a front-side architecture that produced excellent optical and lifetime properties," Yoshikawa says. "And at the same time, we worked to ensure that the rear-side architecture achieved a low resistance with a long lifetime."

As the company looks to the future, it hopes that by further improving the cell's main properties, it will eventually approach the technology's theoretical limits of just over 29 percent, Yoshikawa says. Check out the chart below to see where this development fits among other photovoltaics records. The data are drawn from "Interactive: Record-Breaking PV Cells" at *IEEE Spectrum*'s website. The interactive tool is a collaboration with the U.S. National Renewable Energy Laboratory, the National Institute of Advanced Industrial Science and Technology, in Japan, and Fraunhofer ISE, in Germany. –JOHN BOYD

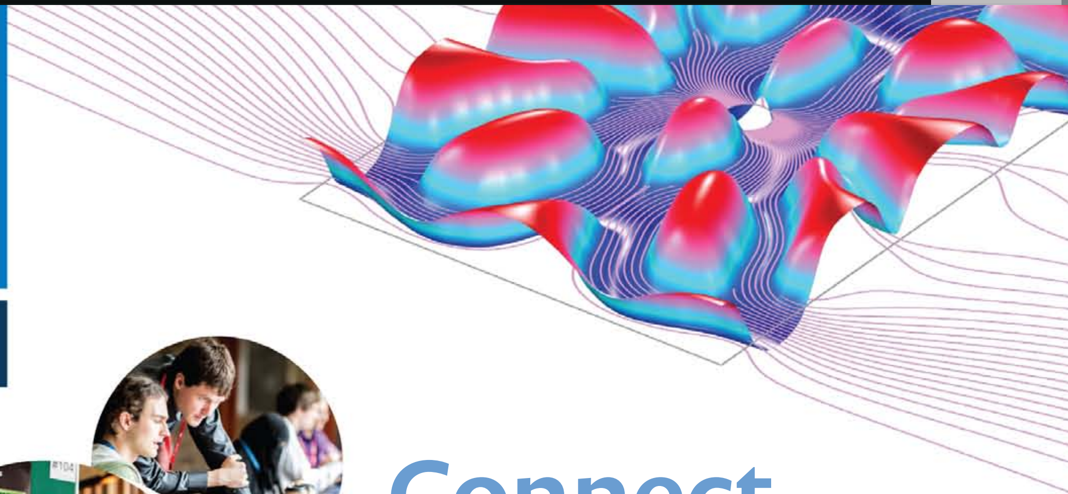


SOURCES: NREL, FRAUNHOFER ISE, AND NIAIST
GRAPH: ERIK VRIELINK

POST YOUR COMMENTS at http://spectrum.ieee.org/siliconrecord0517

COMSOL CONFERENCE 2017 BOSTON

OCTOBER 4–6
BOSTON MARRIOTT NEWTON



Connect. Learn. Innovate.

Discover how multiphysics modeling and simulation software spark innovation in your industry.

LEARN MORE AT comsol.com/conference

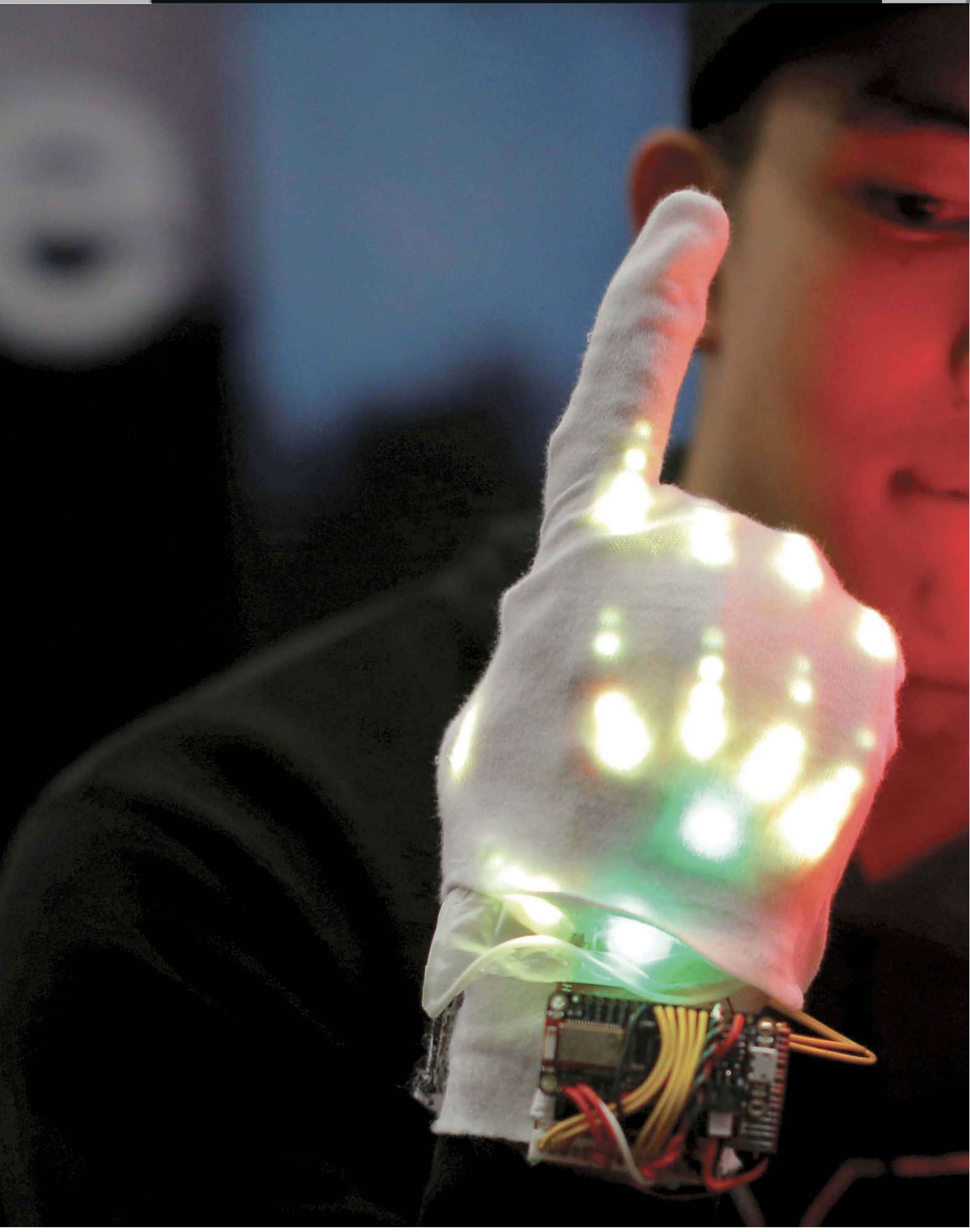
2,388 presentations in the past 5 years

Connect. Meet with colleagues and multiphysics simulation experts.

Learn. Expand your knowledge and discover new technologies through 30+ lectures and hands-on minicourses.

Innovate. See how numerical simulation is improving the product design workflow, leading to innovation.

COMSOL



PHOTOGRAPH BY Brian Snyder/Reuters

IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page

DANCE ENHANCER

THERE'S A REASON WHY terrible dancers are said to have two left feet. But feet aren't the only things that can make a dancer look clumsy–or cool. A team of engineers and designers from Tokyo has created the Groove, a glove that translates the movement of a dancer's hands and fingers into signals that cause the wearable device to light up. Each finger has bend sensors, and there's a pressure sensor and accelerometers in the palm. The Groove app uses data from the sensors to determine the pattern of light created by the 50 LEDs on each glove. The team showed off the device at the South by Southwest (SXSW) Festival in Austin, Texas, in March.

RESOURCES

7,899

THE NUMBER OF LINES OF CODE IN BAREMETAL-OS, A MODERN 64-BIT OPERATING SYSTEM WRITTEN ENTIRELY IN ASSEMBLER

RANDI KLETT

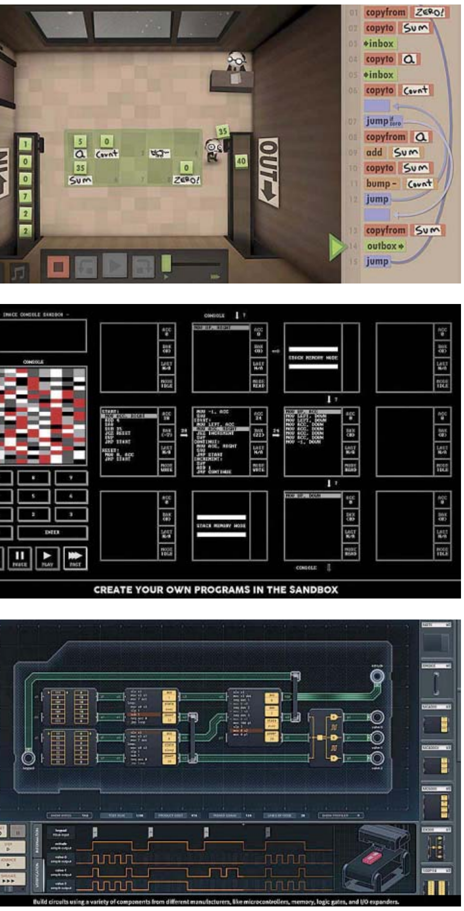
RESOURCES_GEEK LIFE

SOME ASSEMBLY (LANGUAGE) REQUIRED

THREE GAMES THAT MAKE LOW-LEVEL CODING FUN

Ah, assembly. Where all the pretense of high-level languages—the program structures, the data handling, the wealth of functions—gets stripped away. You get branches, bytes, and if you're lucky, a subtraction command. True, directly manipulating the state of a computer can be powerful, but few people code in assembly by choice. ● So I was surprised to find not one but three polished games that do a surprisingly good job of making coding in assembly language fun. To be clear, none of these titles involve writing assembly for real hardware. They all use virtual systems with minimal instruction sets. Still, they do capture the essence of assembly coding, with complex behaviors squeezed out of simple commands. ● The first game is *Human Resource Machine*, originally released in 2015 by Tomorrow Corp. and now available for Windows, Mac, Linux, and the new Nintendo Switch. In this game the player takes on the role of an office worker who must handle numbers and letters arriving on an "in" conveyor belt and put the desired results on an "out" conveyor belt. As you begin, you're given just two instructions to work with. As you progress and face more complex challenges, more instructions are provided. Challenges range in difficulty from outputting the larger of a pair of input numbers to sorting variable-length sequences. ● Because *Human Resource Machine* is highly abstracted, someone could play it as a straight puzzle game and be little the wiser. But those in the know will recognize the office worker as a register, the temporary workspace on the office floor as random access memory, and many of the challenges as classic introductory computer science problems. Because the game starts out with so few commands, and the interface is so intuitive, it would ▸

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/geeklife0517



INSIDE THE MACHINE: *Human Resource Machine* [top] pretends to be an office. *TIS-100* emulates an '80s microcomputer interface [middle], while *Shenzhen I/O* combines coding and wiring [bottom].

FROM TOP: TOMORROW CORP.; ZACHTRONICS(2)

make a great way to initiate beginners into the inner workings of processors. However, the game's story line doesn't do all that much to motivate players to keep completing levels.

TIS-100, from Zachtronics, solves this problem, albeit by presupposing a more sophisticated player. Also originally released in 2015, *TIS-100* is now available for Windows, Mac, Linux, and the iPad. The iPad version (released in 2016 as *TIS-100P*) gives perhaps the clearest idea of who the game is targeted to, as a monochrome text display is joined by an onscreen keyboard that looks (and clacks) just like what you'd expect to find attached to an old-school micro. And indeed, the game is a nostalgic tour de force, complete with a downloadable manual that will provide a pang of recognition for anyone who was programming in the 1980s: It's designed to look exactly like a second- or third-generation photocopy of the most important bits from the actual manual.

The story line for the game is that you've inherited a TIS-100 computer from your Uncle Randy. By solving problems in assembly, you unscramble corrupted portions of the computer's memory, providing hints as to the machine's original purpose. The TIS-100 has an unusual parallel architecture, being composed of computational nodes that pass messages to each other. The puzzles escalate fairly quickly in complexity, but all are doable, and if you're looking for something to take you back to a romanticized version of what it was like to code back in the day, *TIS-100* can't be beat.

Zachtronics, though, also brings us bang up to the present with its latest title, *Shenzhen I/O*, released last November and available for Windows, Mac, and Linux. In this wry game, you're a Western electrical engineer who's decided to go where all the actual stuff is getting made these days—Shenzhen, China. *Shenzhen I/O* shares key game mechanics with *TIS-100*—programming interconnecting modules to accomplish tasks—but it introduces a wide array of different modules (complete with faux datasheets) such as radios and displays. In *Shenzhen I/O*, the name of the game is getting the timing of signals just right so as to pass input/output test suites for the various electronic devices you've been hired to build. Many of the products are amusing, and EEs will find much to recognize, right down to the occasional judicious deployment of a NOP (no operation) instruction to get a timing cycle spot on. **—STEPHEN CASS**

BUILD YOUR OWN AMAZON ECHO

TURN A PI INTO A VOICE-CONTROLLED GADGET

As a young man, I yearned for a machine like the ship's computer on Star Trek: a gadget that can listen and obey a human voice, and answer in kind. Fifteen years ago, after reading about university researchers who had gotten voice-controlled artificial intelligence systems working, I taught myself Linux and set up a server in my attic in the hope that the technology had arrived to let me build such a thing myself. It had not.

But now voice control has come to the masses. Amazon's Echo smart speaker was the hot holiday gift of 2016, and last year the company released the smaller Dot and Tap gadgets. Like the Echo, these tie into Amazon's intelligent personal assistant, Alexa.

It is Alexa, running in the cloud, that converts your speech into text, interprets the text, and responds verbally, musically, or by passing commands to some other smart gadget such as a Wi-Fi–enabled lightbulb.

Alexa isn't the only player in this game, of course: Apple and Google have their own speech-driven AIs. But unlike those companies, Amazon has placed a heavy emphasis on inviting tinkerers and developers to expand the uses for Alexa, in two ways.

First, the company showed programmers how to create new "skills" (voice-controlled apps) that Alexa can invoke, and it set up a section of its online store to distribute them. Within months, the store contained almost 10,000 skills, with hundreds being added each week. (Currently, all skills are free.)

Second, and more interesting to me, Amazon released programming interfaces for Alexa and uploaded free source code and

POST YOUR COMMENTS at http://spectrum.ieee.org/handson0517

SPEAKERS THAT LISTEN: A Raspberry Pi and a sound processing board are mounted on the back of these speakers [bottom], creating a voice-controlled jukebox.

W. WAYT GIBBS(2)

tutorials to Github. Anyone can use these to make their own Echo-like gadget on hardware as inexpensive as a US $40 Raspberry Pi 3 equipped with a cheap USB microphone and speaker.

I resolved to build an Alexa Pi that could do all that an Echo can but also play music in stereo through better speakers. And for extra credit, I wanted to try to use the same hardware to make an intelligent speaker that doesn't rely on Amazon at all.

A quick survey of user forums turned up a problem with my plan to use my cheap USB microphone: Alexa needs cleaner audio input than one microphone can provide. The Echo uses seven microphones and sophisticated noise-cancellation circuitry to discern voice commands from across the room, even when music is playing.

Fortunately, audio-and-voice-tech company Conexant recently released the AudioSmart development kit, a board that includes two adjustable microphones, noise-cancellation hardware, and firmware preprogrammed to listen for the "Alexa" wake word. When the board hears the wake word, it sends a trigger signal to the Pi's general purpose input/output port to let the Pi know that it should start listening to a verbal command. Although the kit is aimed at development engineers (and pricey at $300), it can be reprogrammed to respond reliably to any wake word, unlike Amazon's Echo and Dot, which offer you only a choice of "Alexa," "Echo," "Amazon," or "Computer" (the latter proving that Amazon engineers watch Star Trek, too).

Following Amazon's tutorial on Github, I had the AudioSmart connected to the Pi and the system responding to verbal commands in a day. I linked it to the Alexa app on my iPhone, chose some skills from the online store, and soon had it turning lights on and off in the bedroom and queuing up TV shows on my Plex media server.

The effect was pretty magical—except for one glaring weakness. My setup required a monitor and keyboard to run: By default, Amazon forces a user to authenticate the device with its servers by manually logging into an Amazon Web page, which then passes a "token" (a long string of characters) to a graphical-user-interface program running on the Pi. The token expires after a few hours.

Clearly I wasn't going to set up a monitor and keyboard in the kitchen just to turn on the lights. There has to be a better way, I thought.

There is, but it turned out to be devilishly complicated. You can use a special Android app to generate a reusable token for the Alexa gadget that works even after rebooting. Amazon provides sample code for the app, but you have to configure, build, and run it yourself, using Android Studio. The documentation is sketchy and out of date. Many hours of work went into getting the app to run and communicate successfully with the Pi, and to then configure the Pi so that all the necessary pieces of software run in the right order at boot time.

At last however, I was able to unplug the monitor and keyboard, boot up the Pi, and say "Alexa, tell me a joke."

"What did the dog say after a long day at work?" Alexa responded: "Today was ruff." You're telling me.

Searching for a simpler route, I came across Mycroft AI, a startup in Lawrence, Kan., that has created a completely open-source alternative to Alexa. Even the hardware designs for its Echo-like product, called Mark 1, are free to download and build or modify yourself. I grabbed the Picroft disk image and copied it to a microSD card, which I stuck into my Pi. The Pi booted right up and started running the AI. (I did need to change a few files to make the system work with the AudioSmart board.)

Mycroft's system is still in its early days and has only a fraction of the skills that Alexa offers. It is much more flexible, however: You can set it to use IBM's Watson to convert your verbal commands to text and use Google Voice to talk to you, for example. Creating a new skill is as easy as writing a few dozen lines of Python code. The vast universe of open-source Linux software is there for the combining—so that your AI can boldly go where no AI has gone before. **—W. WAYT GIBBS**

PROFILE: ENVIRO POWER

THIS COMPANY IS BRINGING MICROCOGENERATION TO THE U.S.

BOXY BUT GOOD: Michael Cocuzza wants to replace traditional water heaters with turbines that also generate electricity.

ENVIRO POWER

Cogeneration, or combined heat and power (CHP), is the simultaneous production of electricity and useful heat. The global market for cogeneration equipment could hit US $43.8 billion by 2020, according to market research firm Global Industry Analysts. Cogeneration is often associated with large power plants in North America, and few there have modified the technology for small-scale commercial use. And nobody is selling a microcogeneration unit in the United States that is affordable enough for the average single-family house.

That's where mechanical engineer Michael Cocuzza, founder and CEO of Enviro Power, hopes to break in. He got started by creating a prototype for his own home's hot water boiler, capturing the steam generated by heating water to create electricity.

Enviro Power has a patent for a 10-kilowatt microcogeneration unit (mCHP) and is seeking financing to bring the product to market. Its mCHP uses a micro steam turbine, powered by propane or natural gas, to produce both heat and electricity. The company claims this cuts electricity consumption by up to 30 percent, reducing greenhouse gases and saving money. In addition, larger cogeneration units use an internal combustion engine that needs maintenance, while Cocuzza expects that a micro steam turbine engine will be maintenance free for 10 years.

On the strength of the energy savings demonstrated by his prototype, Cocuzza raised angel funding in 2015 that helped him set up in an incubator space. A graduate of the University of Connecticut, Cocuzza has staffed his company by hiring interns and employees recommended by UConn faculty and through the local startup ecosystem.

In partnership with a small municipal utility in Massachusetts, he plans to test his mCHP model in the "light commercial" market—apartment complexes, community swimming pools, nursing homes, and the like—in the early summer of 2017.

If it finds success there, Enviro Power plans to invest in making smaller units for residential customers. Initially these will be 5-kW units, but eventually the company hopes to bring 1.5-kW systems to market, which will be suitable for a standard 185-square-meter home.

"It's a proven market in Asia; it's growing in Europe," Cocuzza says. "We're at the very entry level of what appears to be a very large market."

About half the homes in the United States are heated by natural gas, so the market potential is strong. Comparable units in Europe are selling for two or three times the cost of a standard hot water boiler; Cocuzza's challenge is to be able to price his unit for significantly less. If Enviro Power can make the price low enough that one of its mCHPs will pay for itself through energy savings within three years, Cocuzza believes he'll succeed.

Enviro Power has a few competitors. Marathon Engine Systems, in East Troy, Wisc., has sold nearly 100 of its 4.4-kW ecopower units in the United States and Canada, including one installed in a 540-square-meter home, according to Marathon. Yanmar, a Japanese company, is trying to expand into the U.S. market with its 5-kW and 10-kW internal combustion engine mCHP systems: The company says it is selling the 5-kW unit into multifamily housing and large single-family houses.

Mike Cocking, of consulting firm MicroCogen Partners, says market conditions are currently ideal because gas prices are low and electricity costs are rising. "The competition is pretty meager," Cocking says. "On paper, [Enviro Power] measures very well."

Cocking approves of Cocuzza's plan to avoid direct residential sales for now: By starting with light commercial customers that need hot water all year, the return on investment is likely to be faster than in a single-family home. Worldwide, there are about 300,000 10-kW mCHP units in the market, but only 900 of them are in the United States, according to Cocking.

"It's a clever idea," Cocking says. "If he can make it viable, somebody will buy it."

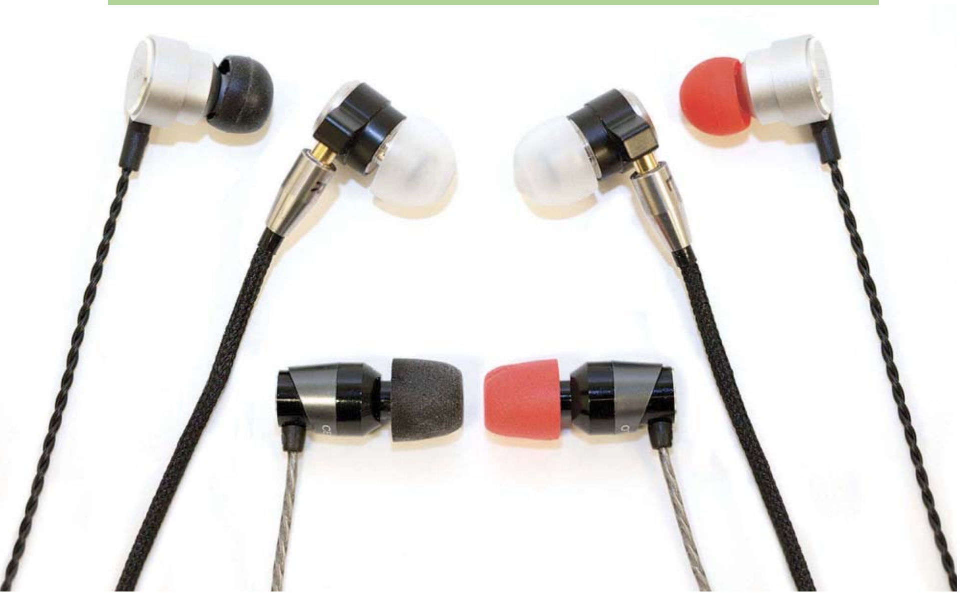
—THERESA SULLIVAN BARGER

Founded: 2013 **Location:** Mansfield Center, Conn. **Employees:** 7 **Total funding to date:** $450,000

POST YOUR COMMENTS at http://spectrum.ieee.org/startups0517

RESOURCES_REVIEW

EARPHONE SHOWDOWN, PART 2
THE DQSM D2, LKER I8, AND CRAZY CELLO



THE EAR DRUMMERS: From the outside in: the LKER i8, the DQSM D2, and the Crazy Cello. All provide good, if not perfect, sound at low cost.

More than a year ago I noticed I was spending most of my music-listening time playing songs on my smartphone. That realization sent me on a quest to find the most inexpensive earphone-based setup that would give me high-end, audiophile sound. I listened to dozens of earphones, half a dozen headphone amplifiers, and a like number of digital-to-analog converter-headphone amplifiers (DAC-headphone amplifiers). I found a few gems, such as the DragonFly Red DAC-headphone amplifier from Audioquest, in Irvine, Calif.

In my previous dispatch in April's issue of *IEEE Spectrum*, I described a few noteworthy models of earphones I've come across recently, including the Etymotic ER-4 and the LKER i1. In this follow-up, I'll consider a few more earphones toward the lower end of the market, all of them Chinese-made units I obtained via the AliExpress retail site. After listening to these models, I'm happy to report that it is indeed possible to get true audiophile sound out of a pocket-size setup.

The earphones I auditioned for this round were the DQSM D2 (US $168), the Tennmak Crazy Cello ($50), and the LKER i8 ($37, still available on eBay). As I did for the previous earphone comparison, I listened using the three different earphones to a few songs spanning widely different musical genres, over and over again, all stored in the Free Lossless Audio Codec (FLAC) format.

I started out particularly intrigued by the DQSMs, which have three sound-producing components—"drivers"—per ear; two are balanced-armature types, and the third is a dynamic driver. (For a quick tutorial on the nature of these driver types, check out the April article.)

Balanced armatures are prized more for their precise sound than for bass prowess, so the DQSMs promised the best of both worlds. Alas, they did not fully live up to that promise. They are very good earphones, and thanks to their balanced-armature drivers they were easily the best of the three in their ability to resolve fine detail. But when the DQSMs were coupled with my DragonFly DAC, the treble was sometimes a bit too forward, which made some passages sound harsh or strident.

As a side experiment, I tried them with my FiiO QOGIR E09k headphone amplifier, fed by a JDS Labs ODAC digital-to-analog converter. This setup played more to the DQSMs' strengths, and made me yearn to hear these earphones being fed by a single-ended tube headphone amplifier. But whatever its strengths, that combo wouldn't be portable, so I carried on with the Crazy Cello and the LKER i8 fed by my Samsung smartphone and the DragonFly Red.

The Crazy Cello earphones have a single titanium-coated dynamic driver for each ear, while the i8s have two dynamic drivers per ear ("dual dynamic"). Despite that difference, it proved a challenge to discriminate between these two models. Several rock and country songs sounded quite fine with both. More challenging fare was needed.

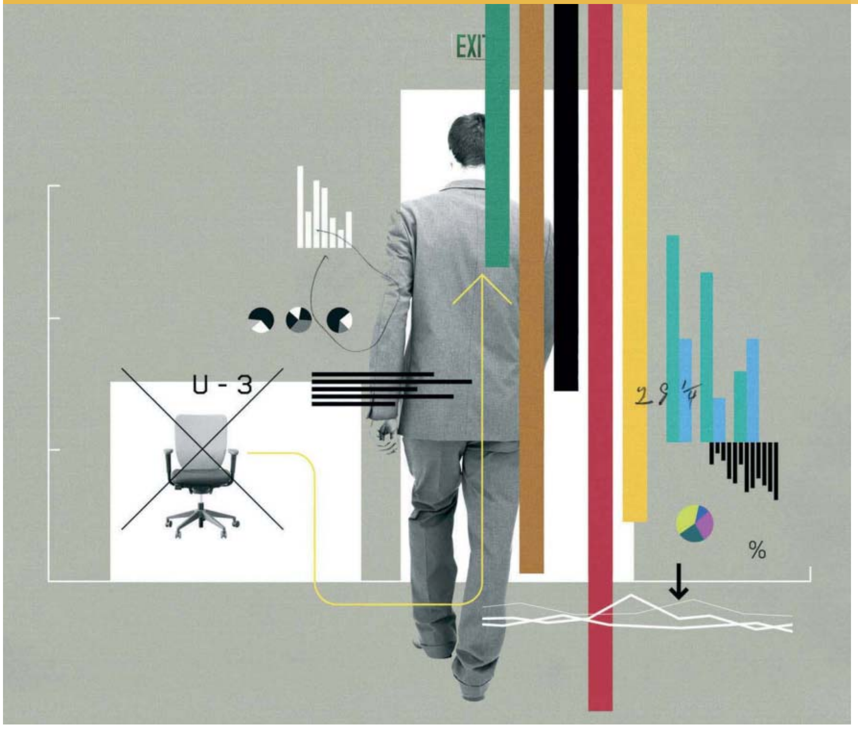
I selected the first movement of the string quartet by Maurice Ravel, in a digital recording by the Alban Berg Quartet. I was impressed by the ability of both models to render the gentle vibrato of the violins and the delicate interplay of the four instruments. At about 4 minutes into that movement, though, a sudden crescendo came across with a bit more finesse via the LKER i8.

That gave me an idea. The opening notes of Wolf Alice's song "Moaning Lisa Smile" go instantly from gentle strumming to towering power chords. The LKER i8 makes the most of them, providing more detail than the Crazy Cello.

Much to my surprise, the $37 i8 triumphed over rivals costing $50 and $168. Feed these earphones with a DragonFly Red ($199) and the USB Audio Player Pro app ($8), and you'll have a music system of audiophile quality that you can carry in your pocket, for under $250. It won't glow seductively or impress your friends with its imposing heft and size, but it will let you have marvelous musical experiences pretty much anywhere, and at any time. **—GLENN ZORPETTE**

RANDI KLETT

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/review0517



UNEMPLOYMENT: PICK A NUMBER

MANY ECONOMIC STATISTICS are notoriously unreliable, and the reason often has to do with what's included in the measurement and what's left out. Gross domestic product offers a good example of a measure that leaves out key environmental externalities, such as soil erosion, biodiversity loss, and effects of climate change. ● Measuring unemployment is also an exercise in exclusion. Casual consumers of U.S. economic news are familiar with only the official figure, which put the country's total unemployment at 4.8 percent at the beginning of 2017. But that is just one of six alternatives used by the U.S. Bureau of Labor Statistics to quantify "labor underutilization." ● Here they are, in ascending order (with rates, again, for January 2017). People unemployed 15 weeks or longer as a share of the total (civilian) labor force: 1.9 percent. People who lost jobs and who completed temporary jobs: 2.3 percent. Total unemployment (the official rate): 4.8 percent. All unemployed people plus discouraged workers, as a share of total and discouraged labor force: 5.1 percent. The previous category enlarged by all people only "marginally attached" to the labor force: 5.8 percent. And finally, the last category plus those who work only part time for economic reasons (that is, they would prefer to work full time): 9.4 percent. These six measures present quite a spread of values: The official unemployment rate (U-3) was only about half of the most encompassing rate (U-6), which was nearly five times as high as the narrowest measure (U-1). ● If you lose your job you count as unemployed only if you keep looking for a new one; otherwise, you never get counted again. That is why when trying to get closer to the "real" unemployment rate you must look at the labor force participation rate, which has recently been in decline. In early 2007, the U.S. labor force participation rate for people 16 years and over was 66.4 percent; by the end of 2008, a few months into the world's largest economic downturn since World War II, it was still at 65.8 percent. Its subsequent slide bottomed out in September 2015 at 62.4 percent, and by early 2017 the rate was marginally higher, at 62.9 percent. Nearly 95 million U.S. citizens above the age of 16 were not part of the labor force—a historic record.

This decline has come out of two countervailing trends: Male labor participation, now at 69 percent, has been falling ever since the early 1950s, when it peaked at about 86 percent. At the same time, female participation was rising until the year 2000 and has fallen since by less than 10 percent (it's now at about 57 percent).

As expected, the rates are much lower for the young (16 to 19 years old) in general (just 35 percent, compared with 60 percent in 1980) and for young African-Americans in particular (30 percent, compared with 41 percent in 2000). But even after we take into account the participation rates, we still do not get the full appraisal: Pure quantities do not capture essential qualities.

There has been some encouraging news on employment quality: The latest Gallup polling on U.S. jobs shows that 72 percent of U.S. employees are completely satisfied with their relations with coworkers and nearly as many feel physically safe in their workplaces. However, only 33 percent are satisfied with the money they earn and only 45 percent with the recognition they receive.

The complex reality of (un)employment can never be caught by aggregate numbers. Not a few people have left the labor force early and eagerly because they could afford to do so (hence they should be excluded when searching for the "real" unemployment rate). Others are trying desperately to get back into it. Many who are fully employed are unhappy with their lot but cannot change jobs easily or at all, because of their skills or family circumstances.

Numbers may not lie, but which truth do they convey? ■

PHOTO-ILLUSTRATION: STUART BRADFORD

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/unemployment0517

THE COUNTERINTUITIVE CLOUD

UNTIL RECENTLY, THE WORD *DATA* didn't require a modifier. But we passed a watershed moment when we started referring to *big* data. Apparently, that wasn't a sufficient description for some chunks of data, because people grasped for bolder terms, such as *humongous* data. Sadly, now, it appears that we have run out of appropriate adjectives. And yet data keeps getting bigger and bigger. • So instead of mentioning data, people have begun waving their hands and talking vaguely about the "cloud." This seems to be the perfect metaphor—a mystical vapor hanging over Earth, occasionally raining information on the parched recipients below. It is both unknowable and all-knowing. It answers all questions, if only we know how to interpret those answers. • This evolution brings to mind two images. The first is from the current scientific hypothesis that all of the information in a black hole resides in the event horizon that surrounds it. This is like the *idea* of the cloud, while on Earth below, the practical reality of the cloud manifests in proliferating server farms. These farms bring the second image to mind: Douglas Adams's city-size supercomputer, Deep Thought, from the classic novel (and radio play and TV show and movie) *The Hitchhiker's Guide to the Galaxy*. • With these imaginary end states in mind, I wonder: Where is all this headed? Will data increase indefinitely, or is there some point of diminishing returns? Is there such a thing as enough data—or possibly too much? • There is a popular saying that "data is the new oil." While I think this is an imperfect metaphor, it is true that both oil and data require refining to be useful. I'm mindful of the information pyramid described in T.S. Eliot's poem "The Rock": "Where is the wisdom we have lost in knowledge? / Where is the knowledge we have lost in information?"

For the purposes of our discussion, let's say that data is composed of 1s and 0s, information is the words and images encoded by data, and knowledge is what we glean or learn from that information. The critical refining is between information and knowledge. In the refining of oil, the ratio of the useful final product to the starting amount of crude is not a function of the amount of crude. Not so with information: The more crude information we have to deal with, the *less* knowledge we want to produce per bit. Otherwise, big data will simply overwhelm us as it continues to grow. What we want is the small knowledge that we obtain from the *big* information. As the data set gets bigger, the job gets harder. The catch, however, is that unless the big information is big enough, it may not contain the small signal that we are searching for.

Knowledge inevitably increases, so data has to increase even faster. Fortunately, storage technology seems capable of coping without turning Earth into a giant disk drive, but the crunch will be on the artificial intelligence and algorithms that turn data into knowledge. We have come a long way since Claude Shannon, in his classic paper on information theory, in 1948, could simply ignore the knowledge problem by writing: "Frequently the messages have meaning.... These semantic aspects of communication are irrelevant to the engineering problem."

I'm also mindful of the propensity of drawers, closets, and hard drives to eventually become filled with useless junk. I sometimes blame this on the second law of thermodynamics, which states that entropy—that is, disorder—always increases. Perhaps this will ultimately be true of the cloud. Old, useless information accumulates, and it's too much work to purge it. Moreover, who's to say what is useless and what is not? Everything is in there, but everything is too much. Entropy is maximized, and the data ultimately becomes, as Shakespeare put it, full of sound and fury, signifying nothing. ■

ILLUSTRATION: GREG MABLY

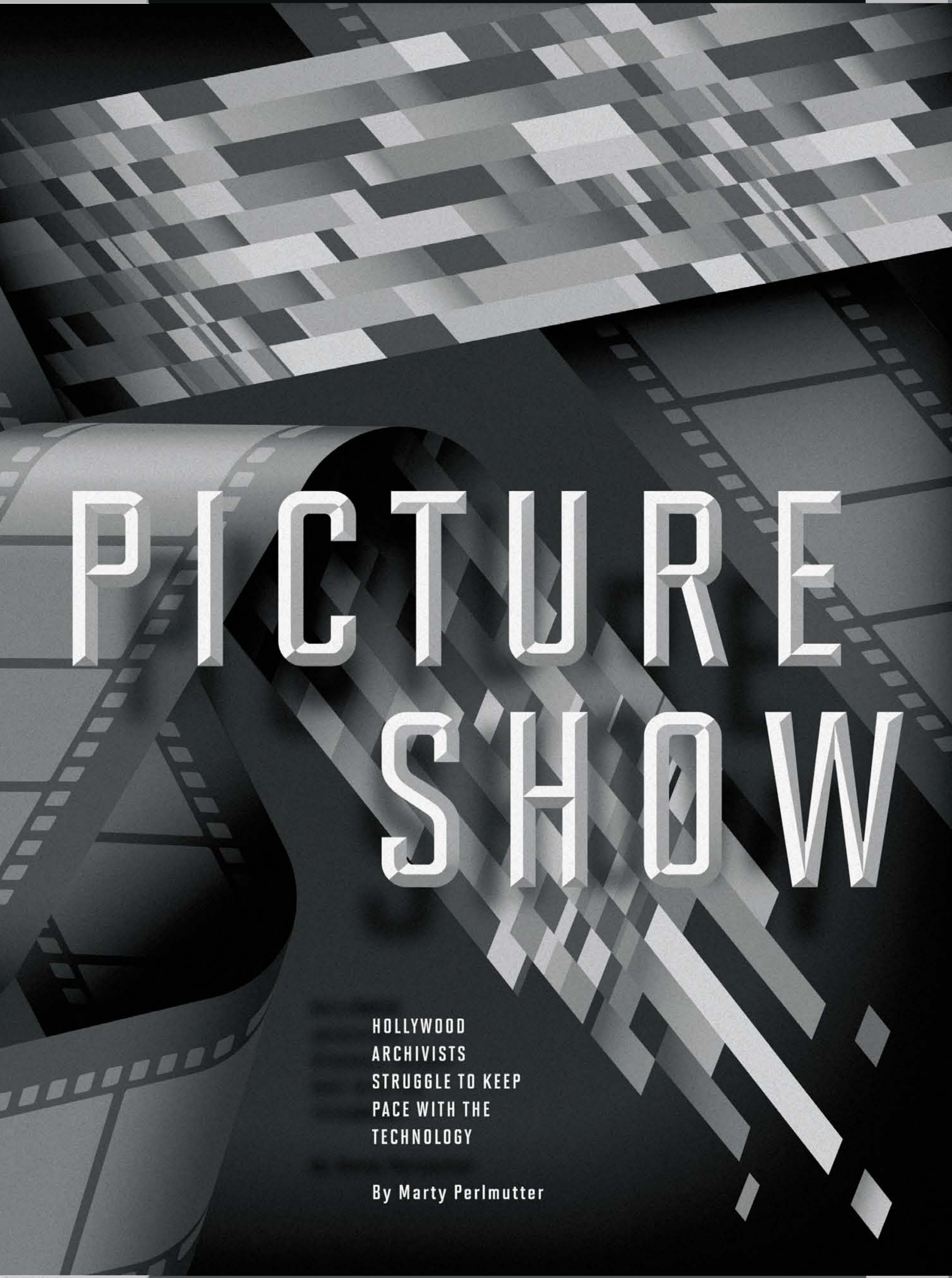
↗ POST YOUR COMMENTS at http://spectrum.ieee.org/reflections0517

THE LOST

AFTER EMBRACING MAGNETIC TAPE STORAGE,

ILLUSTRATION BY Francesco Muzzi/StoryTK

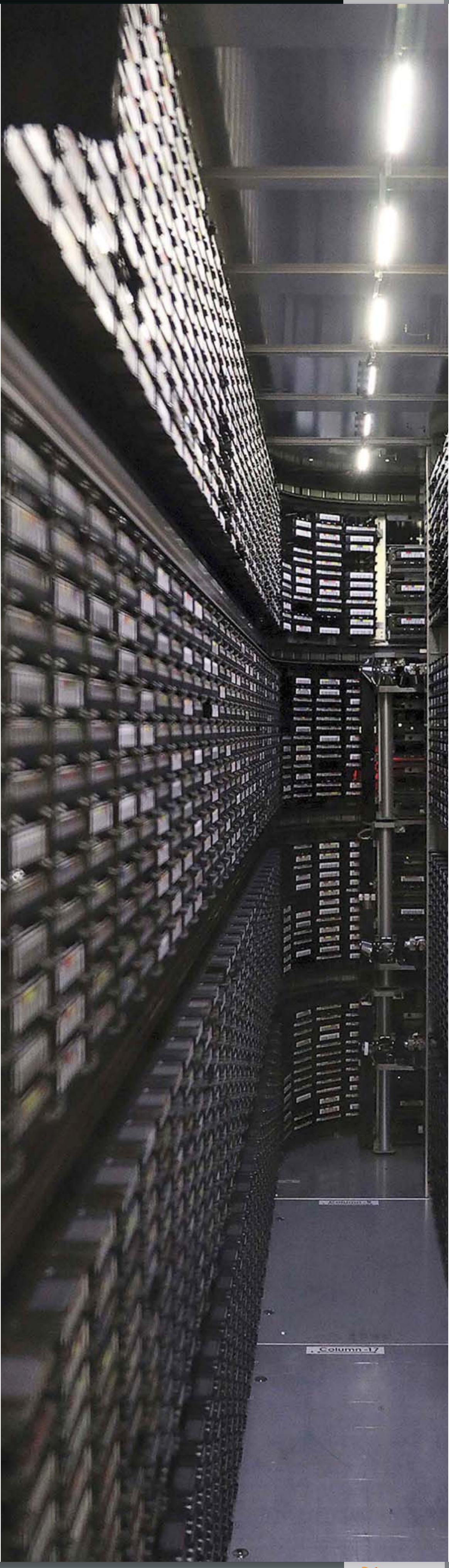
IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page



PICTURE SHOW

HOLLYWOOD ARCHIVISTS STRUGGLE TO KEEP PACE WITH THE TECHNOLOGY

By Marty Perlmutter



WHEN THE RENOWNED CINEMATOGRAPHER Emmanuel Lubezki began planning to shoot the wilderness drama *The Revenant*, he decided that to capture the stark, frozen beauty of a Canadian winter, he would use no artificial light, instead relying on sunlight, moonlight, and fire. He also planned to use traditional film cameras for most of the shooting, reserving digital cameras for low-light scenes. He quickly realized, though, that film "didn't have the sensitivity to capture the scenes we were trying to shoot, especially the things we shot at dawn and dusk," as he told an interviewer.

The digital footage, by contrast, had no noise or graininess, and the equipment held up much better in the extreme cold. The crew soon switched over to digital cameras exclusively. "I felt this was my divorce from film–finally," Lubezki said. The film, released in December 2015, earned him an Academy Award for cinematography two months later.

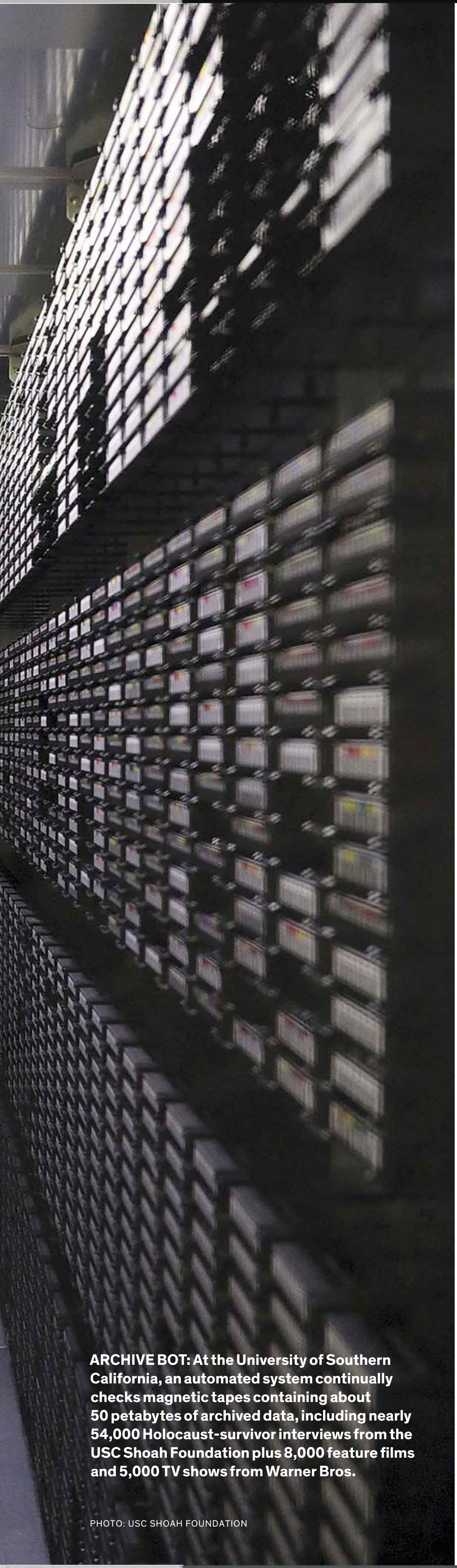
Lubezki's late-breaking discovery of digital is one that other filmmakers the world over have been making since the first digital cameras came to market in the late 1990s. Back then, digital moviemaking was virtually unheard of; according to the producer and popular film blogger Stephen Follows, none of the top-grossing U.S. films in 2000 were recorded digitally.

These days, nearly all of the films from all of the major studios are shot and edited digitally. Like Lubezki, filmmakers have switched to digital because it allows a far greater range of special effects, filming conditions, and editing techniques. Directors no longer have to wait for film stock to be chemically processed in order to view it, and digital can substantially bring down costs compared with traditional film. Distribution of films is likewise entirely digital, feeding not only the digital cinema projectors in movie theaters but also the streaming video services run by the likes of Netflix and Hulu. The industry's embrace of digital has been astonishingly rapid.

Digital technology has also radically altered the way that movies are preserved for posterity, but here the effect has been far less salutary. These days, the major studios and film archives largely rely on a magnetic tape storage technology known as LTO, or linear tape-open, to preserve motion pictures. When the format first emerged in the late 1990s, it seemed like a great solution. The first generation of cartridges held an impressive 100 gigabytes of uncompressed data; the latest, LTO-7, can hold 6 terabytes uncompressed and 15 TB compressed. Housed properly, the tapes can have a shelf life of 30 to 50 years. While LTO is not as long-lived as polyester film stock, which can last for a century or more in a cold, dry environment, it's still pretty good.

The problem with LTO is obsolescence. Since the beginning, the technology has been on a Moore's Law–like march that has resulted in a doubling in tape storage densities every 18 to 24 months. As each new generation of LTO comes to market, an older generation of LTO becomes obsolete. LTO manufacturers guarantee at most two generations of backward compatibility. What that means for film archivists with perhaps tens of thousands of LTO tapes on hand is that every few years they must invest millions of dollars in the latest format of tapes and drives and then migrate all the data on their older tapes–or risk losing access to the information altogether.

That costly, self-perpetuating cycle of data migration is why Dino Everett, film archivist for the University of Southern California, calls LTO



ARCHIVE BOT: At the University of Southern California, an automated system continually checks magnetic tapes containing about 50 petabytes of archived data, including nearly 54,000 Holocaust-survivor interviews from the USC Shoah Foundation plus 8,000 feature films and 5,000 TV shows from Warner Bros.

PHOTO: USC SHOAH FOUNDATION

"archive heroin–the first taste doesn't cost much, but once you start, you can't stop. And the habit is expensive." As a result, Everett adds, a great deal of film and TV content that was "born digital," even work that is only a few years old, now faces rapid extinction and, in the worst case, oblivion.

TO UNDERSTAND HOW THE MOVIE STUDIOS and archives got into this predicament, it helps to know a little about what came before LTO. Up until the early 1950s, filmmakers shot on nitrate film stock, which turned out to be not just unstable but highly flammable. Over the years, entire studio collections went up in flames, sometimes accidentally and sometimes on purpose, to avoid the costs of storage. According to the Film Foundation, a nonprofit founded by director Martin Scorsese to restore and preserve important films, about half of the U.S. films made before 1950 have been lost, including an astounding 90 percent of those made before 1929.

It wasn't just that film was difficult to preserve, however. Studios didn't see any revenue potential in their past work. They made money by selling movie tickets; absent the kind of follow-on markets that exist today, long-term archiving didn't make sense economically.

In the 1950s, nitrate film was eclipsed by more stable cellulose acetate "safety film" and polyester film, and it became practical for studios to start storing film reels. And so they did. The proliferation of television around the same time created a new market for film. Soon the studios came to view their archives not as an afterthought or a luxury but as a lucrative investment–and as an essential part of our collective cultural heritage, of course.

The question then became: What's the best way to store a film? For decades, the studios took a "store and ignore" approach: Put the film reels on shelves, placed horizontally rather than vertically, at a constant cool temperature and 30 to 50 percent humidity. Ideally, they'd have redundant copies of each work in two or more of these climate-controlled vaults. Remarkably, the industry still uses film archiving, even for works that are born digital. A master copy of the finished piece will be rendered as yellow-cyan-magenta separations on black-and-white film and then preserved as traditional celluloid, on polyester film stock.

"We know how long film lasts," says the USC archivist Everett. "And archives were designed to store *things*. They're cool, they're dry, and they have shelves. Put the film on the shelf, and it will play in a hundred years."

One big problem with this approach is that to preserve the work, you must disturb it as little as possible. Dust, fingerprints, and scratches will obviously compromise the integrity of the film. Archive staff periodically check the stored masters for signs of degradation; occasionally, a master will be used to make a duplicate for public release, such as a showing at a repertory cinema or film festival. But otherwise, the archive remains pristine and off-limits. It's like having a museum where none of the art is ever on display.

Maintaining such a facility isn't cheap. And as chemical film stock becomes obsolete, along with the techniques used to create and manipulate it, relying on a film-based archive will only grow more difficult and more costly.

"The sad truth is that film images are ephemeral in nature, kept alive only by intensive effort," David Walsh, the head of digital collections at London's Imperial War Museum, has written. "Apart from anything else, if you are storing film in air-conditioned vaults or running digital mass-storage systems, your carbon footprint will be massive and may one day prove to be politically or practically unsustainable."

IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page

THE MOVIE INDUSTRY EXECUTIVES I interviewed would argue that the current system for digital archiving is already unsustainable. And yet when LTO storage first came along 20 years ago, it seemed to offer so much more than traditional film. Magnetic tape storage for computer data had been around since the 1950s, so it was considered a mature technology. LTO, as an open-standard alternative to proprietary magnetic tape storage, meant that companies wouldn't be locked into a single vendor's format; instead they could buy tape cartridges and tape drives from a variety of vendors, and the competition would keep costs down. Digital works could be kept in digital format. Tapes could be easily duplicated, and the data quickly accessed.

And manufacturers promised that the cartridges would last for 30 years or more. In an interview, Janet Lafleur, a product manager at Quantum Corp., which makes LTO cartridges and drives, said that LTO tape may still be "perfect" after 50 years. LTO came to be widely used for data backup in the corporate world, the sciences, and the military.

But the frequency of LTO upgrades has film archivists over a barrel. Already there have been seven generations of LTO in the 18 years of the product's existence, and the LTO Consortium, which includes Hewlett Packard Enterprise, IBM, and Quantum, has a road map that specifies generations 8, 9, and 10. Given the short period of backward compatibility–just two generations–an LTO-5 cartridge, which can still be read on an LTO-7 drive, won't be readable on an LTO-8 drive. So even if that tape is still free from defects in 30 or 50 years, all those gigabytes or terabytes of data will be worthless if you don't also have a drive upon which to play it.

Steven Anastasi, vice president of global media archives and preservation services at Warner Bros., therefore puts the practical lifetime of an LTO cartridge at approximately 7 years. Before that time elapses, you must migrate to a newer generation of LTO because, of course, it takes time to move the data from one format to the next. While LTO data capacities have been steadily doubling, tape speeds have not kept up. The first generation, LTO-1, had a maximum transfer rate of 20 megabytes per second; LTO-7's top rate is 750 MB/s. Then you need technicians to operate and troubleshoot the equipment and ensure that the migrated data is error free. Migrating a petabyte (a thousand terabytes) of data can take several months, says Anastasi.

And how much does it cost to migrate from one LTO format to the next? USC's Everett cited a recent project to restore the 1948 classic *The Red Shoes*. "It was archived on LTO-3," Everett says. "When LTO-5 came out, the quote was US $20,000 to $40,000 just to migrate it." Now that the film is on LTO-5, it will soon have to be migrated again, to LTO-7.

For a large film archive, data migration costs can easily run into the millions. A single LTO-7 cartridge goes for about $115, so an archive that needs 50,000 new cartridges will have to shell out $5.75 million, or perhaps a little less with volume discounts. LTO drives aren't cheap either. An autoloader for LTO-6 can be had for less than $3,000; an equivalent for LTO-7 is double that. And archivists are compelled to maintain and service each new generation of LTO drive along with preserving the LTO cartridges.

Lee Kline, technical director at Janus Films' Criterion Collection, regards data migration as an unavoidable hassle: "Nobody wants to do it, but you have to." Archivists like Kline at least have the budgets to maintain their digital films. Independent filmmakers, documentarians, and small TV producers don't. These days, an estimated 75 percent of the films shown in U.S. theaters are considered independent. From a preservation standpoint, those digital works might as well be stored on flammable nitrate film.

MEANWHILE, THE MOTION-PICTURE studios are churning out content at an ever-increasing rate. The head of digital archiving at one major studio, who asked not to be identified, told me that it costs about $20,000 a year to digitally store one feature film and related assets such as deleted scenes and trailers. All told, the digital components of a big-budget feature can total 350 TB. Storing a single episode of a high-end hour-long TV program can cost $12,000 per year. Major studios like Disney, NBCUniversal, Sony, and Warner each have archives of tens of thousands of TV episodes and features, and they're adding new titles all the time.

Meanwhile, the use of higher-resolution digital cameras and 3D cameras has caused the amount of potentially archivable material to skyrocket. "We went from standard definition to HD and then from HD to UHD," Peter Schade, NBCUniversal's vice president of content management, said in an interview. Pixel resolutions have gone from 2K to 4K and soon, 8K, he adds. Codecs–the software used to compress and decompress digital video files–keep changing, as do the hardware and software for playback. "And the rate of change has escalated," Schade says.

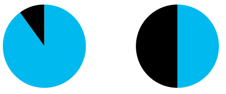
Computer-animation studios like Pixar have their own archiving issues. Part of the creative process in a feature-length animated film is developing the algorithms and other digital tools to render the images. It's impossible to preserve those software assets in a traditional film vault or even on LTO tape, and so animation and visual effects studios have had to develop their own archival methods. Even so, the sheer pace of technological advancement means those digital tools become obsolete quickly, too.

When Pixar wanted to release its 2003 film *Finding Nemo* for Blu-ray 3D in 2012, the studio had to rerender the film to produce the 3D effects. The studio by then was no longer using the same

PRESUMED LOST

Estimated lost U.S. films of the silent era **90%**

Estimated lost U.S. films prior to 1950 **50%**



Source: The Film Foundation

"Originally, we went all digital because it's so much cheaper. But is it? Really? We haven't solved the storage problem"

IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

animation software system, and it found that certain aspects of the original could not be emulated in its new software. The movement of seagrass, for instance, had been controlled by a random number generator, but there was no way to retrieve the original seed value for that generator. So animators manually replicated the plants' movements frame by frame, a laborious process. The fact that the studio had lost access to its own film after less than a decade is a sobering commentary on the challenges of archiving computer-generated work.

Another problem for archivists is that digital camera technology has allowed productions to shoot essentially everything. In the past, the ratio of what's shot to what's eventually used for a feature film was typically 10 to 1. These days, says Warner archive chief Anastasi, films can go as high as 200 to 1. "On some sets, they're simply not turning the camera off," he says.

All that material will typically get saved and stored for a while. But at some point, somebody will have to decide how much of that excess really needs to be preserved for posterity. Given the huge expense of film preservation, archivists are being ruthless about what they choose to store. "There's no way we can store it all," says USC archivist Everett. "We're just going to store the bare minimum."

At Warner, Anastasi has taken a triage approach. Four years ago, when he took over the studio's archives, he faced two distinct challenges: First he had to "stop the bleeding" by figuring out how to save those assets that were most vulnerable to being lost. Those on two-inch videotape, the medium of choice for network TV shows in the 1970s and 1980s, "were the most at risk. We captured that material on digital as uncompressed JPEG 2000 files." That part of the triage is now nearly complete.

The second challenge was finding a way to affordably maintain the studio's archive for more than a generation. He set the goal at "50-plus years." He also decided that rather than operating an in-house archive, the problem would be better handled by outsourcing it. And so in 2014, Warner signed a long-term contract with USC Libraries to maintain the studio's archives.

Sam Gustman, associate dean of the USC Libraries, says that the Warner archives are now part of 50 petabytes of archived data at USC, which also includes nearly 54,000 video interviews with Holocaust survivors gathered by the USC Shoah Foundation. For 20 years of storage, including power, supervision, and data migration every 3 years, USC charges $1,000 per terabyte, or $1,000,000 per petabyte. That works out to a relatively affordable $2.5 million per year for its current 50-PB holdings. It's not a money-making business, Gustman adds.

The USC archive maintains four copies of each tape: Two are held at USC, one at Clemson University in South Carolina, and one at Charles University in Prague. The aim is to "touch every tape" every six months, using an automated system, Gustman explains. A robotic arm selects a tape from a rack and loads it into a reader, which plays it back while a computer checks for aberrations. Any tape that isn't perfect is immediately trashed, and the archive makes a replacement from one of its remaining copies of the tape. The archive migrates to the latest version of LTO as it becomes available, so no tape is more than three years old.

Warner also began classifying its 8,000 feature films and 5,000 TV shows into two categories: those it will "manage"—that is, preserve for the long term—and those it deems "perishable." Managed assets include not just the finished work but also marketing materials and some deleted scenes. Perishable material may include dailies for features or unused footage; it will be stored for some time in the archive but may not be migrated. To decide what's perishable and what's not, the studio considers things like how successful the film has been, how popular its stars are, and whether the film could have enduring (or cult) appeal.

HIGH RES: Lasergraphics' Director 10K film scanner digitizes film stock at horizontal resolution of up to 10,000 pixels.

LASERGRAPHICS

The manage-or-perish scheme is by no means perfect, Anastasi admits, but he sees it as buying the studio a little time until a truly long-term digital storage technology comes along. If one ever does.

For now, he says, "We'll keep it, and there'll be time to rethink the strategy. But after 10 years, we can't guarantee access" to any of the material that hasn't been migrated to managed storage.

Everett says Warner's strategic thinking about digital archiving is pioneering. All of the studios, he notes, "are in a realm where there is no policy." Meanwhile, they're waiting for an archival technology that is better than LTO. "Originally, we went all digital because it's so much cheaper," Everett notes. "But is it? Really? We haven't solved the storage problem."

If technology companies don't come through with a long-term solution, it's possible that humanity could lose a generation's worth of filmmaking, or more. Here's what that would mean. Literally tens of thousands of motion pictures, TV shows, and other works would just quietly cease to exist at some point in the foreseeable future. The cultural loss would be incalculable because these works have significance beyond their aesthetics and entertainment value. They are major markers of the creative life of our time.

Most of the archivists I spoke with remain—officially at least—optimistic that a good, sound, post-LTO solution will eventually emerge. But not everyone shares that view. The most chilling prediction I heard came from a top technician at Technicolor.

"There's going to be a large dead period," he told me, "from the late '90s through 2020, where most media will be lost." ■

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/thelostpictureshow0517

THE PIT BOSS

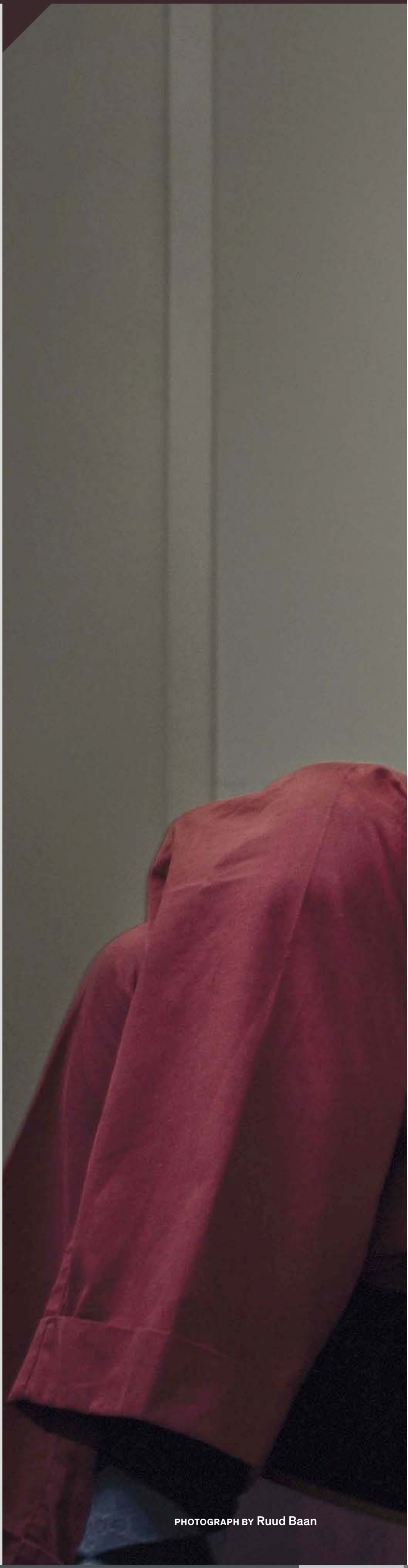
Kees Immink, this year's IEEE Medal of Honor recipient, put compact discs on track

By Tekla S. Perry

Remember vinyl records? More specifically, do you remember the way vinyl records skip when they're dusty or scratched?

Let me assume you're old enough to recall that annoyance, or perhaps you've experienced that vintage technology more recently. Now think back to when you got your first CD. Small and shiny, packing 74 minutes of music, it seemed magical, even more magical when you noticed that you could treat a disc pretty badly before physical damage affected the way it played.

A lot of different kinds of engineering, of course, went into figuring out how to put music on a CD and play it back so reliably. There's hardware, including a laser, optics to focus it, and mechanical systems to move the laser and turn the disc. And there's software–including pulse-code modulation, which turns regular samples of an analog signal into bits, and error-correcting codes, which make sure those bits don't get corrupted.



PHOTOGRAPH BY Ruud Baan

And there's one more piece of the CD puzzle: how to translate the digital bits into physical marks on the disc itself. This piece, known as the channel code, deserves much of the credit for making CDs skip resistant without sacrificing playing time. For this encoding scheme and related optical-recording technology, its creator, Kees Schouhamer Immink, will receive the 2017 IEEE Medal of Honor.

Immink was a young electrical engineer working in the laboratories of Philips, in Eindhoven, Netherlands, when the research effort that was to become compact-disc technology got under way in the mid-1970s. Initially, Immink had nothing to do with the project. Rather, he worked on control theory, assigned to a group of mostly optical engineers developing the analog laser videodisc. Philips introduced the LaserDisc in 1978, and it was a flop. This play-only system couldn't compete with video cassettes, which could be used to record and had a two-year head start in development. Come 1979, Immink was a research engineer without a project.

Meanwhile, another research group at Philips had developed a prototype of a laser-based digital audio disc system, while Sony engineers had demonstrated a similar technology. The two systems differed enough in the details to render them completely incompatible. But instead of racing separately to market and triggering a format war, in 1979 the two companies decided to merge their engineering talent and jointly settle on one design.

That's when Immink got involved. The team at Philips, he says, "needed someone to do measurements of the two systems, the quality, how they coped with scratches, how they coped with imperfections of the disc. My job with the LaserDisc was finished, so I said, 'Sure, I could do it.'"

Toward the end of the year, Sony sent engineer Hiroshi Ogawa over to Philips with a big box of electronics, Immink recalls. Ogawa hooked up his gear, Immink hooked up the Philips prototype, and the two, using experimental discs made with each company's coding technologies, started their tests. They messed up the plastic discs with scratches and dust and, eventually, crammed more bits onto the discs to see how long they could extend playing time before the discs became unreadable. And then the two engineers packed up their gear, went to Tokyo, and repeated the experiments—because, Immink notes facetiously, "Of course, the rules of physics are completely different in Tokyo than in Eindhoven." The whole process took several months.

> **The impact of Immink's EFM code didn't end with the CD. It went into the MiniDisc… the DVD…the Super Audio CD**

Although Immink was acting as a test engineer, he was a researcher at heart, so he had his eye out for parts of the design that could be improved. And he found one: the coding scheme that specifies how bits of data are translated into physical marks on the disc.

To understand what he came up with, first visualize the shiny surface of an optical disc. Just under that polycarbonate surface is a reflective layer of metal, marked with a pattern of pits and "lands," the places where it's not pitted. The laser hitting that metal reflects from the pits and lands differently, creating a variation in the light's intensity as it reflects back and hits an optical sensor. The system identifies a change in intensity as a 1; no change over a set distance represents a 0.

There's a problem, though, when too many pits or lands are bunched close together, because the servo system—the electronic and mechanical controls for the optical head—relies on the trail of pits and lands to stay on track. (A CD doesn't have grooves like a vinyl record to serve as a guide.) Immink explains this by referring to a classic fairy tale that goes by several names, among them "Little Thumb," in which children turned out into the forest drop stones along the way to find their way home (a scenario also played out in "Hansel and Gretel").

"If you do not drop a stone often enough, you lose your way," Immink says of the path created by the pits and lands on the disc. The problem is exacerbated by dust or scratches that hide spots on the disc, like a fallen leaf hiding a pebble.

Both Philips and Sony had come up with different rules for translating digital audio data to sequences of pits and lands, rules that took these considerations into account. But Immink thought both approaches were investing too much disc real estate in their efforts to keep the servo system on track. Instead, he thought, by designing better control mechanisms for the servo itself, he could place a few less "stones" along the path it needed to follow. He also could reduce the number of bits that separated each 8-bit block of data bits—so-called merging bits—as long as he selected them carefully in relation to the data bits around them. These two adjustments would allow 30 percent more data to fit into the same physical space without causing the optical head to skip.

The encoding system Immink devised came to be called Eight-to-Fourteen Modulation (EFM). Using a lookup table, it translates each 8-bit chunk of data into a series of 14 bits, with every binary 1 separated by at least two but no more than ten 0s. This ensures that no sequence of pits and lands is too long (problematic because the transitions between a pit and a land generate a signal used for timing recovery) or too short (problematic because short sequences are difficult to detect and prone to error). Three

DISC DRIVERS: Kees Immink [left] poses in 1984 with a videodisc while Philips Research Labs colleague Joost Kahlman holds a compact disc.

merging bits separate each sequence, for a total of 17 bits to represent each 8-bit chunk of data. (The initial Sony-proposed code translated 8 bits to 24 bits.)

In mid-1980, Immink presented his approach to the joint Philips and Sony development group. He demonstrated, he says, that the coding technique would cooperate with the servo controller as effectively as previous versions.

Then, Immink recalls, "a Sony engineer who was one of the major decision makers for the project told me my design for the decoder that would translate the recovered 14 bits into the 8 user bits was too complicated. 'Our design only requires five gates, and yours requires maybe 250 gates,' he said. 'If you can make it less than 70 gates–or maybe 100–then we will take it.' " Logic gates, at the time, were relatively costly.

Immink came back three weeks later with a version that contained only 52. "They had no choice but to accept it," he says.

Later that year, Sony and Philips jointly published the Red Book, the document that set the standard for CD audio. It included Immink's EFM technology. The first CD audio players hit the market in 1982, and by the early '90s, CD sales surpassed those of both vinyl records and cassette tapes–before being supplanted by digital downloads in the past couple of years (exactly which year depends on how you do the counting).

The CD is by no means a dead format; more than 1.5 billion discs were sold in the United States alone in 2015, according to the Recording Industry Association of America. And the impact of Immink's EFM code didn't end with the CD. It went into the MiniDisc, introduced in 1992, with a slightly more compact version, EFMPlus, going into the DVD, which was introduced in 1995. It also went into the Super Audio CD, introduced in 1999.

After the launch of CD technology, Immink, who had taught himself coding theory on the job, decided to get a formal credential in the topic and submitted a series of research papers to Eindhoven University of Technology, earning a Ph.D. degree. Meanwhile, as it was the custom at Philips for all researchers to change groups every five years, he left optical recording and moved to magnetic recording, developing the coding technology for what became the Digital Compact Cassette format, introduced in 1992 to little success. In the mid-1990s, still in the magnetic recording group, he developed a coding scheme for the MultiMedia CD, a Sony/Philips invention that both companies later dropped in favor of the Toshiba-led effort to develop the now ubiquitous DVD.

"My approach was 6 or 7 percent better in terms of storage capacity," Immink recalls. "But it would again have required a redesign of the servo systems."

Philips, along with its competitors, then turned its attention to the possibilities of a new kind of laser technology they expected would soon be available for commercial use: the blue laser. With a much shorter wavelength than the infrared lasers used in CD players and the red lasers used in DVD players, blue lasers would allow a dramatic increase in recording capacity, allowing high-resolution movies to be packed onto the same size disc as the CD and DVD.

Immink was not, however, assigned to this project, because he had rotated over to Philips's telecommunications group, where he was working on modems. And in 1998 he left the company.

"The official reason I give is that I didn't like it anymore. I just wanted to do something else," he says.

For someone in the Netherlands, that was an unusual move, says A.J. Han Vinck, who was a professor at Eindhoven University when Immink applied for his Ph.D. "It is said that once you start at Philips, you die at Philips. So when he left at age 50, a very successful researcher, it was a big surprise."

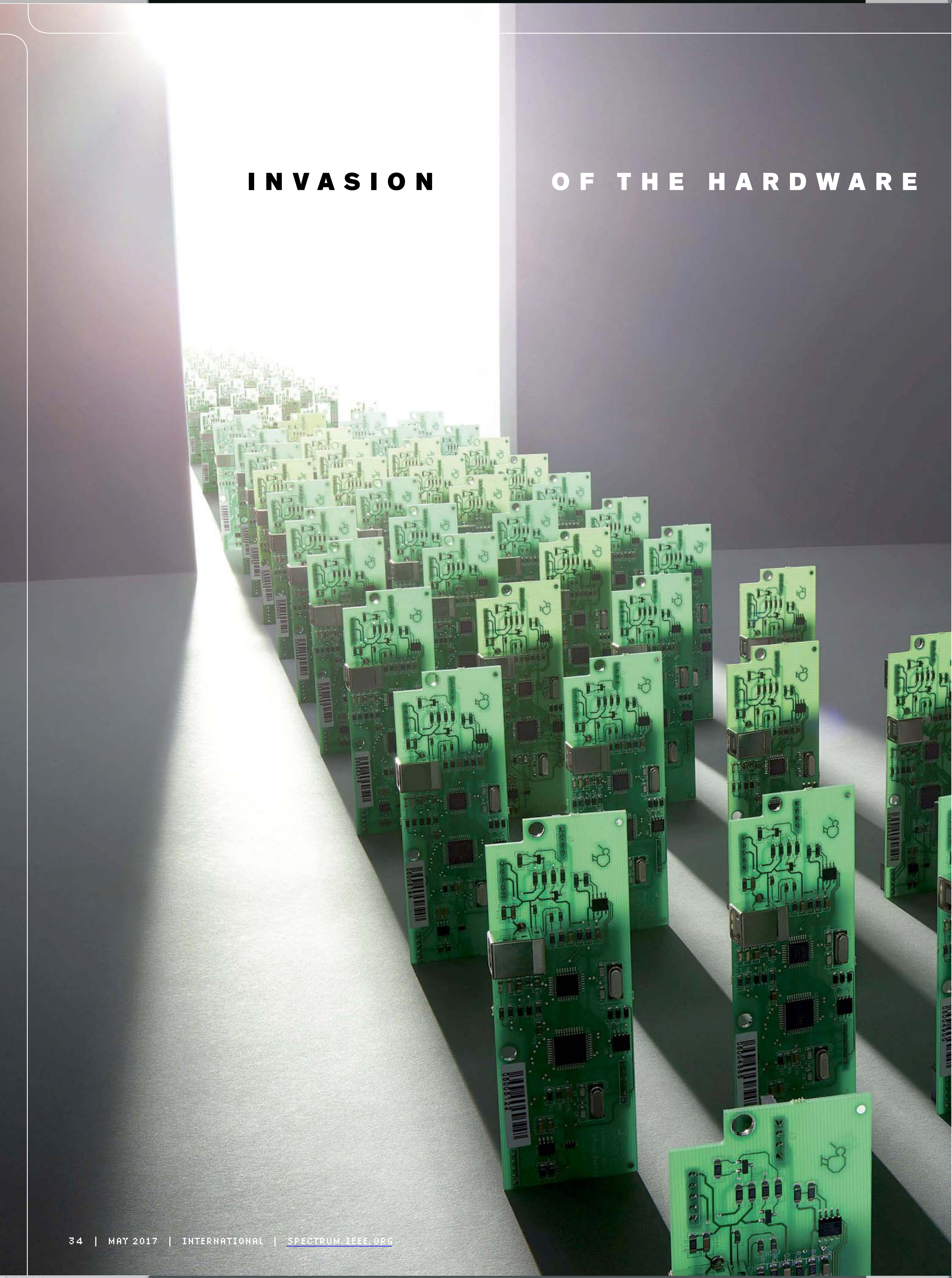
Generally, Vinck says, Immink does not do well accepting authority, recalling Immink's initial unwillingness to make revisions to the manuscript submitted for his Ph.D. thesis. And, Vinck points out, the Dutch have a tendency to "push down" anyone who begins to appear even a bit better than those around him or her, which was likely affecting Immink.

And, Immink recalls, the labs at Philips, like those at many large companies around the world, had | CONTINUED ON PAGE 47

KEES SCHOUHAMER IMMINK

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/immink0517

INVASION OF THE HARDWARE



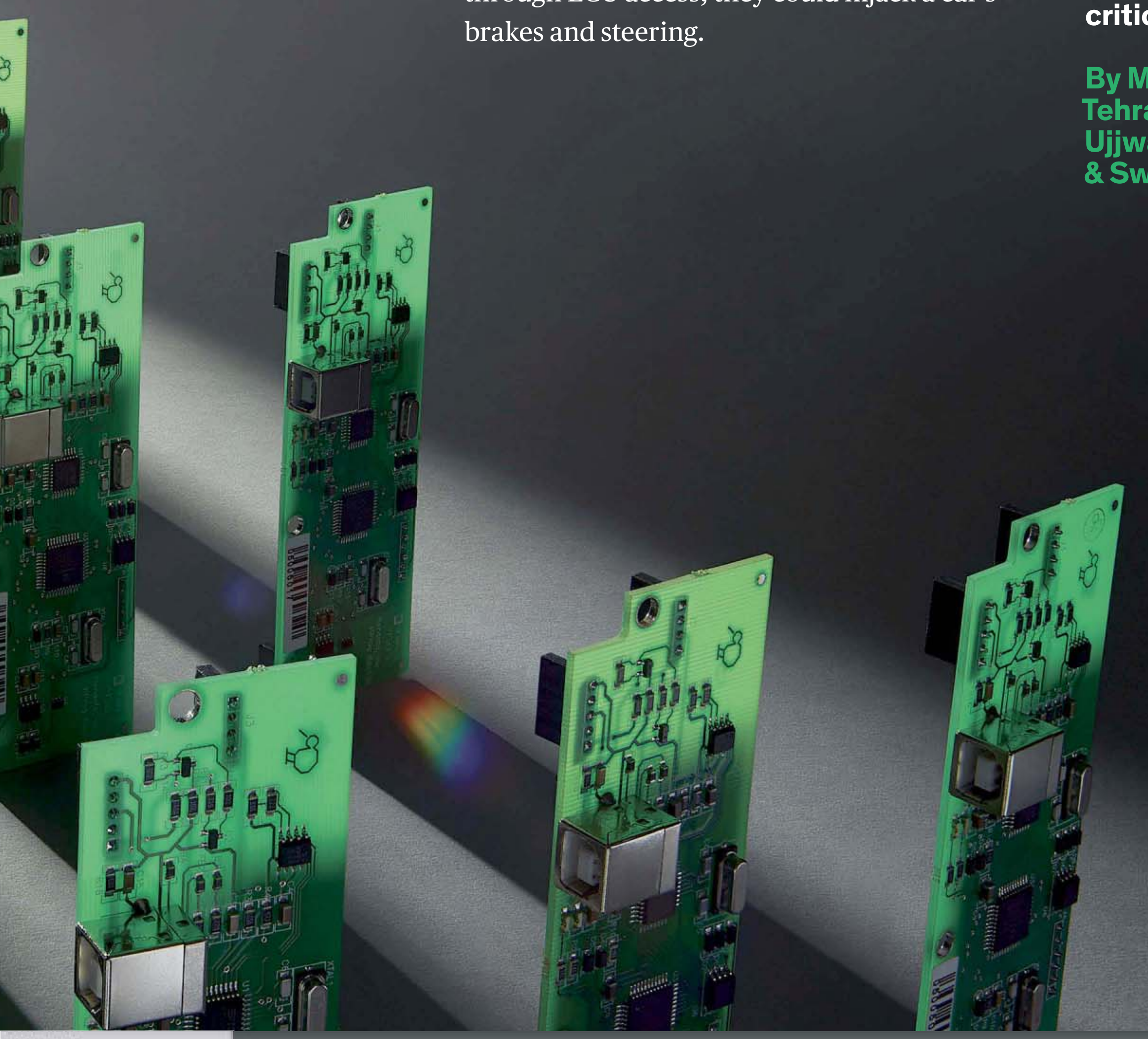
IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

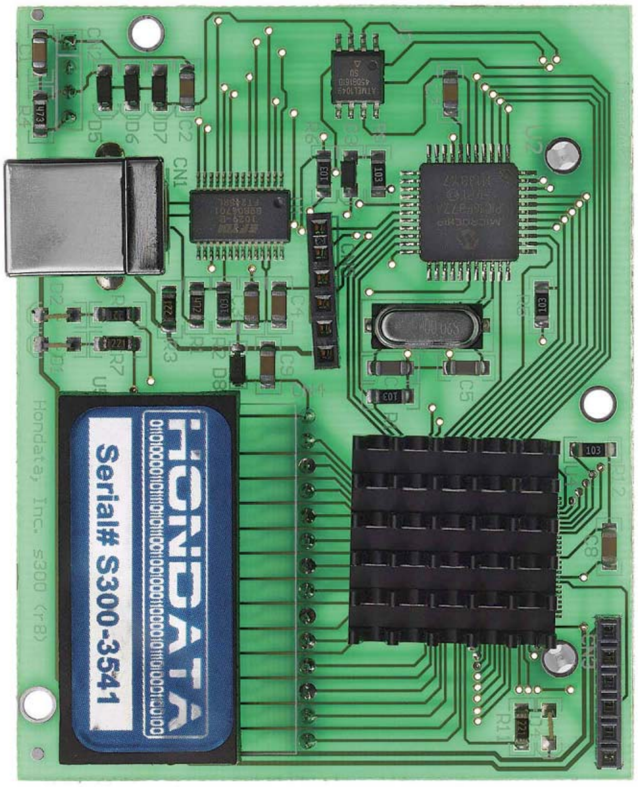
SNATCHERS

Fake hardware could open the door to malicious malware and critical failures

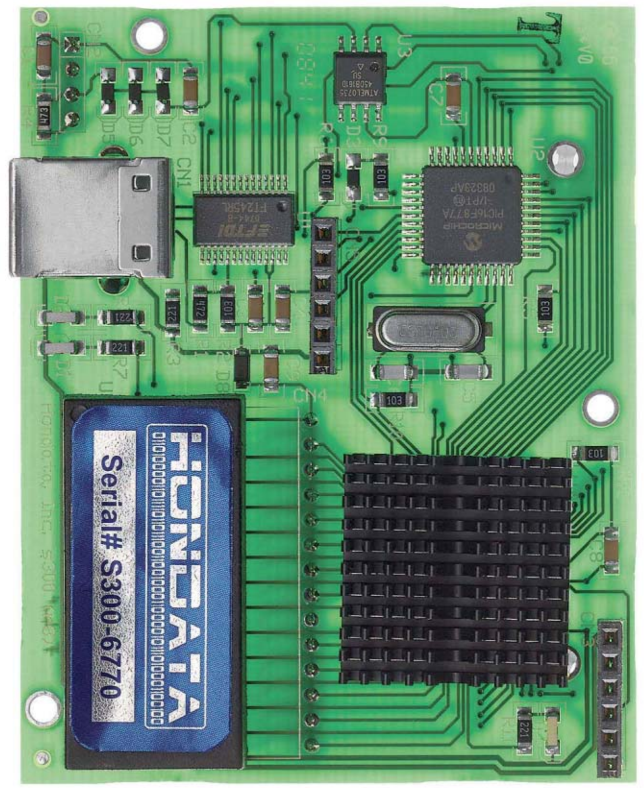
By Mark M. Tehranipoor, Ujjwal Guin, & Swarup Bhunia

IN FEBRUARY 2014, the FBI charged a Florida man, Marc Heera, with selling a cloned version of the Hondata s300, a plug-in module for the engine computer that reads data from sensors in Honda cars and automatically adjusts the air-fuel mixture, idle speed, and other factors to improve performance. The plug-in also allows users to monitor the engine via Bluetooth and make their own adjustments. The clones certainly looked like the genuine product, but in fact they contained circuit boards that had likely been built in China, according to designs Heera had obtained through reverse engineering. Honda warned that cars using the counterfeits exhibited a number of problems, including random limits on engine rpm and, occasionally, failure to start. Devices that connect to an engine control unit (ECU) present particular safety concerns; researchers have demonstrated that, through ECU access, they could hijack a car's brakes and steering.





FAKE



REAL

ALMOST TWINS: A close look at a clone of the Hondata s300 module alongside an authentic one reveals a few key differences—but not ones a consumer would be likely to spot. The s300 is a plug-in for car-engine computers.

It's not just car parts that are being cloned; network routers and parts for routers are also popular targets for cloners. That may not sound particularly scary until you consider that a hacker who has control of a cloned router can then intercept or redirect communications on the network. Look at the 2010 case of Saudi citizen Ehab Ashoor, who was convicted of purchasing cloned Cisco Systems gigabit interface converters with the intent to sell them to the U.S. Department of Defense. The devices were to be installed in Iraq in Marine Corps networks used for security systems and for transmitting troop movements and relaying intelligence from remote field operations to command centers.

While Ashoor appears to have been motivated by greed rather than any desire to do harm, the impact of ersatz equipment in critical electronic systems like a secure router or a car's engine can still be catastrophic, regardless of the supplier's intent.

And unlike counterfeit electronics of the past, modern clones are very sophisticated. Previously, counterfeiters would simply re-mark or repackage old or inferior components and then sell them as if they were new and top of the line; the main problem with these knockoffs was poor reliability. Cloned electronics these days are potentially more nefarious: The counterfeiters make their own components, boards, and systems from scratch and then package them into superficially similar products. The clones may be less reliable than the genuine product, having never undergone rigorous testing. But they may also host unwanted or even malicious software, firmware, or hardware—and the buyer may not know the difference, or even know what to look for.

Installing cloned hardware into networks, for instance, could open the door to hackers: They could launch man-in-the-middle attacks or secretly alter a secure communication path between two systems in order to bypass security mechanisms, like integrity verification, encryption, and end-point authentication. Software hidden in a router could allow an attacker to take control of other systems on the network, rerouting data to remote servers or even disrupting critical systems, such as the flow of electricity through a smart grid. A cloner who succeeds in embedding malicious software or hardware into a combat drone could shut it down or retarget it when it reached preset GPS coordinates.

Already, entire lines of consumer electronics have been cloned. Back in 2004, the Japanese electronics giant NEC Corp. received reports from Beijing and Hong Kong that pirates were selling keyboards, CDs, and DVDs bearing the company's logo. When NEC investigated further, it discovered the problem was far worse:

PREVIOUS PAGES: BOARDS FROM HONDATA

PREVIOUS PAGES AND THIS PAGE, PHOTOGRAPHY BY The Voorhes

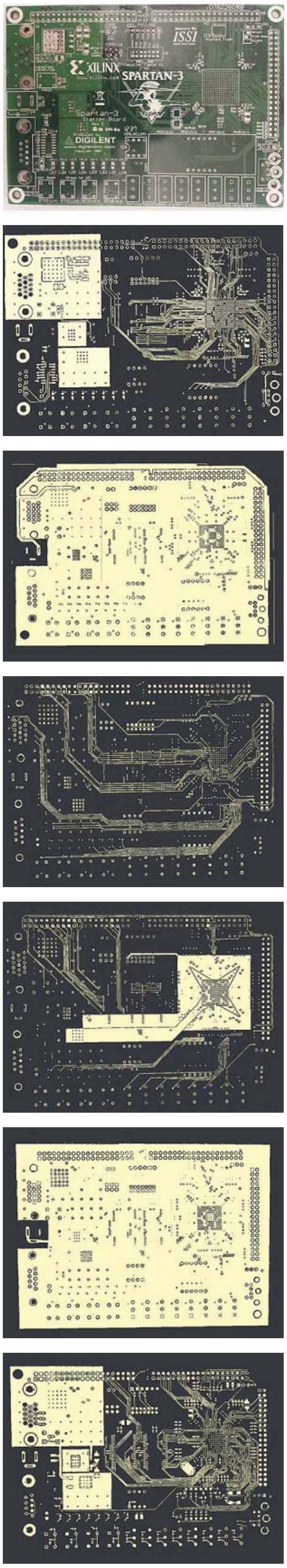
The cloners had developed a host of consumer electronic products–including home entertainment systems, MP3 players, batteries, microphones, and DVD players–and then sold them under the NEC label. The cloners even provided official-looking warranty documents to customers.

While there appear to be no published reports of injury or hacking related to this cloning, the risks are bigger today because more of the systems we interact with daily are connected to the Internet of Things. Cloned hardware may lack the security modules intended to protect such devices, and so it opens up the unsuspecting user to cyberattack.

Cloning spans all levels of electronics. Cloning a printed circuit board (PCB) can be relatively straightforward, particularly if the cloner is able to use off-the-shelf components. By contrast, cloning an integrated package of electronics such as a network router requires obtaining the details of all the parts as well as any embedded firmware, manufacturing or purchasing each part, and then assembling them into a functional product.

And cloning microchips, especially today's billion-transistor versions, is even more challenging. A chip becomes an attractive target for cloners only if the market demand is high enough to make it worth the effort. That can happen when a manufacturer stops producing a particular chip, thus forcing anyone who wants to use the discontinued chip to buy it through a distributor that still has stock. As time goes on, the chip will become scarce. Consider two Texas Instruments microprocessors with essentially the same specifications: The discontinued XOMAP3525BCBB lists for US $72, and the still-manufactured OMAP3515ECBBA lists for $52. As the price of the discontinued chip continues to climb, at some point cloners may be willing to make the investment in reverse engineering the original chip. Or, if they get their hands on a pirated design file, they can produce a clone without a big investment.

Nobody really knows the true scale of electronics cloning, thanks to the clandestine nature of the activity and the lack of adequate detection measures in the global supply chain. But we can infer the size of the market by extrapolating from the cloned products that are seized each year by police and customs authorities.



X-RAY VISION: These X-ray tomography images reveal, layer by layer, the layout of a commercial printed circuit board.

UNIVERSITY OF FLORIDA

Using this kind of extrapolation, the International Chamber of Commerce estimated that trade in counterfeit and cloned products–including nonelectronic products such as designer handbags–amounted to $650 billion globally in 2011. The ICC projected that the figure would nearly triple, to $1.7 trillion, by the end of 2015. (The organization does not track electronics separately.)

Estimates do exist for certain types of electronics. In 2005, for example, the FBI, U.S. Immigration and Customs Enforcement, and U.S. Customs and Border Protection kicked off Operation Network Raider, an international initiative to control the illegal distribution of fake network hardware. Between 2005 and 2010, the campaign led to 30 felony convictions and the seizure of nearly $143 million worth of fake network hardware; similar efforts leading to arrests and equipment seizures continued after that campaign ended, but the numbers haven't been aggregated.

Whatever the true size of the cloned electronics market, we are sure that it is growing, based on our work with SMT Corp., whose labs specialize in identifying cloned and counterfeited components in global supply chains. Among the trends contributing to the growth of cloning are more-sophisticated imaging and analysis tools–we'll discuss these in a moment–and the spread of contract manufacturing, a business model in which the companies that design chips and systems outsource their fabrication. As design files are passed back and forth between the designer and the contractor, cloners will exploit any crack in security to get those files. Once a cloner succeeds in fabricating credible copies, the ubiquity of online sales makes it easy for the sellers to hide their identities and attract bargain-minded customers.

So who exactly are the cloners? They could be just a couple of guys in a garage or a big state-funded organization, or something in between. State-sponsored cloning is thought to be common. Clon-

FIND THE FAKE: This encoder from Fairchild Semiconductor (now On Semiconductor) was discontinued in 2011, but parts—real and fake—are still being sold. Can you spot the fake? (It's the one on the top.)

SMT CORP.

ers in some countries argue that they don't trust U.S. manufacturers, so they clone U.S. chips to make sure their chips are free of deliberately implanted malicious circuits–sometimes called hardware Trojans. A side benefit is that the fake chips come with no licensing fees. These clones can and do get onto the international market.

Regardless of who's doing the cloning, they generally follow a two-step process: First, they copy a design, and second, they fabricate the product. The cloners may secretly buy the design from an employee at the target company, or they may hack into a computer containing that information.

A slightly more sophisticated approach is to reverse engineer the intended product. This kind of work has gotten much easier in the past two decades. The advent of better and cheaper imaging instruments and analysis tools have enabled reverse engineering of even the most complex microchips. Today's optical microscopes can produce 3D images of a chip with superfine resolution. Scanning electron microscopes and transmission electron microscopes can image the microchip's inner layers; high-resolution digital X-ray machines do the same for PCBs. Reverse-engineering companies like TechInsights (which recently acquired Chipworks) use this kind of imaging legitimately for the purposes of competitive analysis and patent research.

The other type of reverse engineering involves taking the product apart to understand how it's made. Using chemical etching, for instance, a cloner can remove each layer of a chip or PCB. Unlike the technique using high-end microscopes and X-ray machines mentioned above, this process requires only a low-power optical microscope to view the result. But the cloner might have to go through 50 or more chips or PCBs to get the design right.

In system-level cloning, such as that needed to reproduce routers and other network hardware, cloners may use some form of reverse engineering, or they may buy the design details from a friendly source at the manufacturer, along with stray parts or products that failed testing. But to complete the clone design, they also need to install firmware–the basic programs used to tell the system how to operate. Firmware for a router might include information such as what frequency the router communicates over or what type of security protocols it uses. On most PCBs, the system firmware is stored in nonvolatile memory–ROM, EEPROM, or flash. A cloner can copy it during the power-up cycle, by tapping a data bus when the processing unit loads system instructions from nonvolatile storage into active memory. Alternatively, the cloner can look directly at the memory itself, using a scanning electron microscope or infrared backside imaging; the latter takes advantage of the fact that semiconductor materials are transparent to certain wavelengths of light. Under a powerful microscope, the cloner can actually see the stored 1s and 0s and reconstruct the code.

Finally, to fabricate the fake, the cloner typically turns to an independent manufacturing facility, which may be similar to or even the same as that used to make the genuine product. Occasionally, cloners use their own production lines, but that's an expense few cloners can afford.

Fighting back against the cloners isn't easy, given the wide variety of electronic products that can be copied and the multitude of ways cloners use to get the job done. To date, the main defense has involved supply-chain security–essentially, giving each chip, PCB, and product a unique identification number to allow it to be tracked throughout the supply chain, from the manufacturer to the end user. That tactic, however, has proven of little value. The ID numbers are maintained in open databases so that companies can verify them. Cloners can simply access the databases, copy the ID numbers, and then place them on their counterfeit goods. Supply chains are so complicated that it's impossible to distinguish between a legitimate company making a database query and a cloner stealing IDs.

So researchers in academia, government, and industry have been working on other approaches. One tactic involves tagging chips and circuit boards with special materials, such as plant DNA. Clone fighters take botanical DNA sequences and scramble them, creating unique patterns that can be used as a signature for a batch of electronic parts. They then mix this DNA with selected fluorophores, which are chemicals that glow under specific wavelengths of light, and tag the electronics with this DNA ink.

Purchasers trying to confirm whether a chip is authentic will scan it for the fluorescent signature; if it's missing, it's a sure indication that something is wrong.

If the fluorescent mark exists, the purchaser will swab the spot to pull a sample of the DNA and then send it to a lab. Standard forensic DNA techniques are used to identify the plant sequences, which get checked against the product's database to confirm whether the label and the part matches.

According to Janice Meraglia at Applied DNA Sciences, one of the companies that offer DNA authentication, DNA tagging is pretty much clone proof. The DNA sequence data is held in a database that is accessible only to laboratory staff, unlike the open databases used for traditional parts IDs. Cloners also don't have access to an essential element of the DNA tagging process: the primer required to start the chain of DNA formation. Primers are small, custom-built sequences of DNA to which other specific sequences of DNA attach; they basically help DNA sequencing tools find the start of the DNA chain.

The downsides of DNA tagging are its expense and how long it takes for authentication. Right now, the list price of a single test is $250, according to Meraglia; this cost drops when an organization signs on for a regular supply-chain testing program. And, she says, the company is working on test equipment that could be used at a parts purchaser's location to reduce the time lag, turning tests around in under an hour.

These issues have kept the technique from becoming widely adopted, but some U.S. government agencies have already begun to tag critical electronic parts. Meraglia thinks the technology will soon move into the financial services industry to verify critical parts of its infrastructure, such as the routers that move data.

The latest promising countermeasure against electronics cloning is something called the physical unclonable function (PUF), which can potentially protect chips, PCBs, and even high-level products like routers. PUFs give each chip a unique "fingerprint." They rely on the physical variations among transistors or other components on a chip, like the width of metal traces, which in turn cause subtle differences in behavior.

The behavior most often exploited by PUF designers is the variation among switching speeds of different transistors. When many transistors are combined into a circuit, the differences in their switching speeds affect the signal propagation along a specific path, which can be measured and compared with the signal propagation of another, supposedly identical path on the same chip. From the two paths, the manufacturer can create a 1-bit signature for the chip. For example, if the switching speeds are faster along the first path than the second one, the manufacturer can assign the bit as "1." To create a longer signature—say, 16 or 64 bits—it typically compares more paths.

The PUF paths must be designed into the chip, either making use of existing features on the chip, such as test circuits or embedded memory, or by building in dedicated circuitry. Once a batch of chips is manufactured, the chipmaker checks how the PUF structures on each chip respond to specific inputs generated by external or internal test circuitry, and then registers the chip's unique signature in a database. A customer can query the database to see if a chip at any phase in the supply chain is authentic or fake, in much the same way a database containing biometric fingerprints can be checked to identify a person.

Because PUF fingerprints are determined during the manufacturing process, they are extremely difficult to replicate. But they suffer from several problems that have prevented their widespread adoption. The first is instability. The tiny variations in transistors that underlie the chip's digital signature can fluctuate along with the supply voltage or the ambient temperature. And as a transistor ages, its switching speed can slow down.

These problems can be addressed to some extent by increasing the number of signatures used for each chip. In this way, a chip can be recognized based on a less-than-perfect match over a large number of signatures; when it comes to detecting clones, a 90 percent match for 100 signatures may be as good as or better than a perfect match for just one signature. But nobody has yet figured out a way to avoid unstable signatures altogether.

Another problem with PUFs is their cost, which results from the additional design time, any new circuitry that needs to be included, and the signature-gathering from the finished chip. PUF technologies that rely on existing chip components—such as embedded memory or test structures to generate the signatures—will add little or no cost to the design.

A final problem is that cloners could use statistical modeling to predict the behavior of some PUFs. Researchers have demonstrated that signatures from some PUF technologies aren't as random as initially thought, and therefore chips with those types of PUFs aren't protected from being cloned.

Despite these concerns, some companies, including Microsemi Corp. and Xilinx, have started to use PUFs for chip identification. And researchers have been extending the concept to printed circuit boards. In this case, they use random variations inside the chips that go onto the PCB as well as variations in the metal traces that connect them. This approach is particularly exciting because, unlike chip fingerprints, PCB fingerprints can be checked remotely to verify the authenticity of a piece of equipment. So the technique could be used, for example, to ensure that critical infrastructure components, once installed, aren't later replaced with clones.

Research on advanced methods for clone detection is just beginning. Government and industry will need to pay more attention to this issue, as powerful reverse-engineering tools become cheaper and more accessible to would-be cloners.

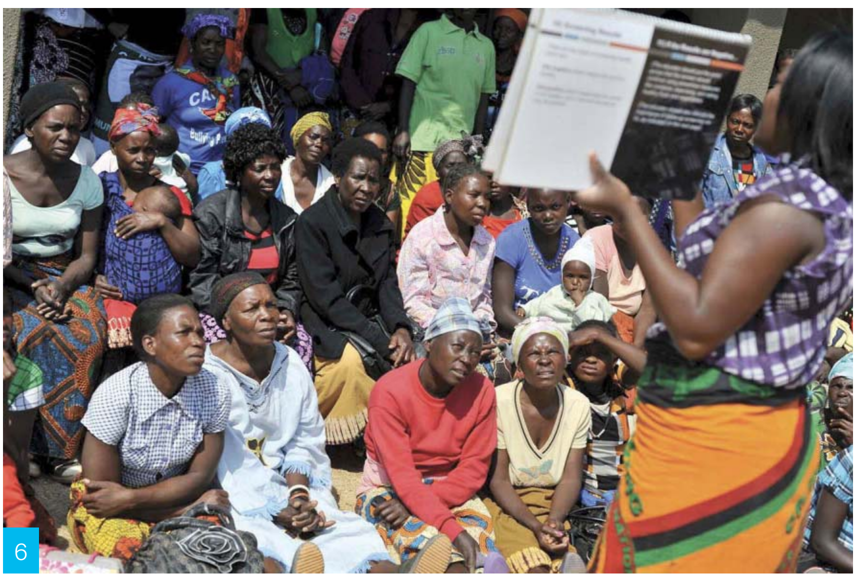
And as the number of Internet of Things devices grows, everyone should become more aware of the dangers that cloned products pose. Some industry projections estimate the IoT population could reach 30 billion by 2020. Imagine if just 1 percent of those connected devices were clones harboring malicious hardware or software. That would mean a potential army that's 300 million strong, waiting for the opportunity to launch an attack of the clones. ■

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/fakehardware0517


1


2


3


6

COME ONE, COME ALL: As part of a cervical cancer prevention campaign in six African countries, the nonprofit Pink Ribbon Red Ribbon works with local health clinics on special screening events. Here, women in Botswana [1] and Zambia [6] receive education about cervical cancer and are screened for signs of the disease. Women with positive results can be treated immediately.

EXPERT EYE: The EVA Scope [2] clips onto any Android smartphone and uses its camera. It acts as a cheap and user-friendly colposcope, the tool gynecologists use to view a magnified image of a woman's cervix. Soon this tool will integrate an artificial intelligence program to help health workers identify signs of cancer.

TRAINING DAY: The cervical cancer screening campaign relies on the EVA Scope. Here, nurses learn how to use the tool at a training camp in Kenya [3, 4, and 5].


5


4

IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

Smartphone-based diagnostic tools with an artificial intelligence upgrade get a tryout

BY CARY CHAMPLIN, DAVID BELL & CELINA SCHOCKEN

AI Medicine Comes to Africa's Rural Clinics

CLOCKWISE FROM TOP RIGHT: EMILY JOHNSON(3); RICCARDO GANGALE/PINK RIBBON RED RIBBON(2); MOBILEODT

In rural health clinics across Kenya, women have started showing up with a surprising request: They've come for their "cervical selfies." ¶ Their enthusiasm is a good omen for a public health campaign against cervical cancer now under way in six African countries. Using an optical accessory that snaps onto any Android smartphone and makes use of its camera, health workers are examining women and catching early signs of cancer, enabling them to get immediate treatment. And soon this diagnostic device will be better still. With the integration of artificial intelligence, this technology may serve as a model for smarter health care in Africa and beyond. ¶ The screening campaign relies on a tool developed by the Israeli company MobileODT—the acronym stands for "optical detection technologies." Health workers use a clip-on attachment, called the EVA (enhanced visual assessment) Scope, to turn a smartphone into a device similar to a colposcope, the tool gynecologists use to view a magnified image of a woman's cervix. With an associated phone app, the screeners can analyze the image, show it to the patient, and store the data in the cloud.

The campaign's organizers originally worried that women wouldn't be willing to be examined in such an intimate way—but in fact, many women have been not only willing but also quite interested in seeing their photos. Instead, the big challenge is ensuring that health workers make accurate diagnoses from these images. That's where AI comes into play.

At Global Good, the innovation hub where two of us (Champlin and Bell) work, we want to use today's ubiquitous mobile technologies to transform health care, particularly in parts of the world that lack medical infrastructure. As a test case, we partnered with MobileODT to integrate machine-learning technology into the EVA Scope. In late 2017 we'll begin field trials in Ethiopia.

This initiative fits the mission of Global Good, a collaborative effort between Bill Gates and the Bellevue, Wash.-based company Intellectual Ventures: to develop technologies that improve people's lives in poor parts of the world. In this case, we're drawing from seemingly esoteric research in machine learning and taking advantage of what are called convolutional neural networks (CNNs). Intellectual Ventures' founder Nathan Myhrvold pioneered the idea of applying these computer science tactics to medical diagnostics, arguing that we can use CNNs to transform mobile phones into supersmart diagnostic tools, and thus help save millions of lives. It may not be possible to send an expert doctor to every health clinic across Africa—but with AI, we can send their expertise.

There's good reason to focus on cervical cancer as a test case for this technology. About 270,000 women die from the disease every year, according to the World Health Organization, and 85 percent of those deaths occur in low-income countries. The disease strikes women in their prime adult years, when they're raising families and earning money. That's why the nonprofit Pink Ribbon Red Ribbon, which fights women's cancers in countries where the need is greatest, has partnered with groups around the world to provide cervical cancer screening at local health clinics. (One of us, Schocken, is Pink Ribbon Red Ribbon's CEO.)

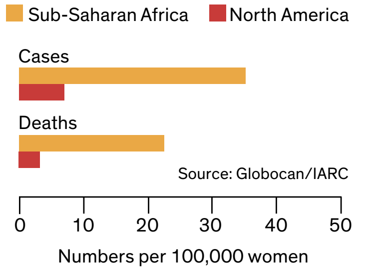
There's also good reason to think this campaign will make a real difference: Unlike so many forms of cancer, cervical cancer is largely preventable, treatable, and curable. Screening exams can reveal the early warning signs of the disease, which typically takes 10 to 15 years to progress to a dangerous stage. So health professionals have a tremendous opportunity to diagnose and treat this potential killer.

Up until now, though, the costs of wide-scale screening have been prohibitive in developing countries. Pink Ribbon Red Ribbon, based in Washington, D.C., estimates that only 5 percent of women in Africa have been checked.

For the traditional screening protocol that's been used around the world for decades, a health worker takes a sample of cervical cells (a Pap test), sends the sample to a lab for analysis, and then waits for results. This process is not only expensive; it can also take weeks in places with rough roads and few labs. Despite the slow-growing nature of cervical cancer, delay has serious consequences. Without same-day screening and treatment, many women never get the care they need. Women don't follow up for many reasons: They may not be able to travel to the clinic again, their husbands may raise objections, or they may not understand the need to return.

A lifesaving breakthrough came in the 1990s when researchers realized that applying acetic acid—the basis of simple household vinegar—to the cervix causes precancerous lesions to turn white. A health worker can then destroy those abnormal cells in much the same way dermatologists

Rates of cervical cancer cases and deaths for Sub-Saharan Africa and North America, 2012



Rates of cervical cancer cases and deaths worldwide, 2012

IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page

remove a wart, with either heat or cold. This fast and relatively painless treatment stops cancer from developing—for a total cost of less than US $20. (In wealthier countries, women may opt for long-term monitoring of precancerous lesions or a more thorough surgery, but these options aren't usually feasible in places like rural Kenya.)

Clinics in many developing countries are adopting visual screening programs, which have probably saved hundreds of thousands of lives already. However, the procedure isn't perfect. Frontline health workers need training and supervision to differentiate between lesions that are truly signs of impending cancer and the many suspicious-looking conditions that are actually benign. Screeners may also miss evidence of advanced cancer that requires referral to a specialist.

MAKING THE CALL: Kenyan nurses practice using the EVA Scope to differentiate between benign lesions and signs of cervical cancer.

EMILY JOHNSON

The team at Global Good saw this situation not as a medical challenge but as a software engineering challenge. Where human eyes needed help, we would bring computer vision backed up by artificial intelligence.

We started by reviewing images of the cervix obtained by colposcopes. We quickly realized that typical computer vision software couldn't handle this large and complex data set because the images had too many features with too much variability. We simply couldn't design algorithms with detailed and exhaustive procedures for distinguishing between a healthy cervix and one with signs of trouble.

The situation called for machine learning—the branch of computer science in which the computer is given an objective, a software framework, and a large training data set, and is then left to create its own solution for carrying out the task at hand.

A common type of machine learning relies on deep neural networks (DNNs), so named because the computing scheme loosely mimics the brain's interconnected neurons. Each computing node can be thought of as a neuron with many inputs. The artificial neuron performs some function based on those inputs and then outputs a single signal, which can serve as one of the inputs for other neurons. By arranging many layers of connected neurons, computer scientists enable these networks to handle tremendously complex tasks.

While the architecture of neural networks is inspired by the human brain, this brand of AI is far removed from human ways of thought. If somebody is explaining how to visually identify a bottle of beer in a store that also stocks bottles of wine, juice, and water, that person would likely describe its distinguishing features in terms of height, diameter, shape, texture, color, and patterns. Some descriptors might include analogies such as "satin finish" or "orange-peel texture." Every feature would be based on, and limited by, our human senses and perception. Yet that list wouldn't include all the factors that our brains use to distinguish one object from another, because much of the process is subconscious.

If we used a human list of features as the basis of an algorithm for recognizing beer bottles, we'd likely get poor results. So instead we'd feed a DNN thousands and thousands of highly variable images of bottles, with metadata indicating whether each image does in fact show a beer bottle. Through a complicated series of training runs, the network can eventually determine, on its own, the relevant distinguishing features. Many similar experiments have shown that neural networks can identify features quite unlike those any person would come up with. And their lists of salient features, cryptic as they are, often enable superb performance.

In tasks involving image processing and pattern recognition, the subtype of DNNs previously mentioned called convolutional neural networks (CNNs) are proving the most promising. This approach uses a few clever tricks to reduce the heavy computational task of making sense of an image.

Computer scientist Yann LeCun created some of the first CNNs in the 1980s at AT&T Bell Laboratories, using them in computer vision systems that could recognize handwriting, such as the zip codes on envelopes. (Today, LeCun is the director of AI research at Facebook.) But the potential of CNNs wasn't really explored until 2012, when University of Toronto graduate student Alex Krizhevsky and his colleagues used a CNN to win an image recognition challenge that involved 1.4 million photos of objects in 1,000 different categories. Their AlexNet program had an error rate more than 50 percent lower than previous winners. It handily recognized barometers, barbershops, bubbles, baseball players, bullet trains, bolo ties, burritos, bath towels, and Boston terriers, to name just a few of the categories under the letter *B*. (Krizhevsky and several of his teammates now work at Google.)

Since that game-changing demonstration, CNNs have taken off. Enabled by the availability of relatively cheap high-performance computing, which is needed for training these networks, CNNs have been adopted for many applications involving images.

In the medical field, the possibilities are exciting. For example, Jürgen Schmidhuber of the Swiss AI Lab IDSIA recently took images of breast cancer cells from real pathology reports and used a CNN to identify dividing cells that indicate an aggressive tumor—a detection task that's challenging even for trained experts. His demonstration was a proof of concept,

INSIDE A CONVOLUTIONAL NEURAL NETWORK

A CNN is a type of machine-learning program that's well suited for image classification tasks–like spotting the early signs of cervical cancer. Here's a simplified account of how a CNN works.

A. The original input for a CNN is a digital image with each pixel represented by number values for its red, green, and blue channels.

B. The CNN uses a multitude of filters (the individual "neurons" in the network) to examine the image in chunks. Here, it's looking at chunks of 3 by 3 pixels.

C. A filter performs simple calculations, multiplying its own number values by the pixel values in the first chunk. The filter looks for a particular feature in the image—perhaps an angled line or a certain gradation of color—as represented by the pixel values.

0	0	0	0	0	0
0	5	4	5	7	0
0	3	2	3	1	0
0	1	5	3	3	0
0	6	1	7	1	0
0	0	0	0	0	0

0	0	0	0	0	0
0	5	5	5	3	0
0	7	5	2	3	0
0	0	6	3	1	0
0	0	0	0	1	0
0	0	0	0	0	0

0	0	0	0	0	0
0	1	2	4	4	0
0	5	3	2	6	0
0	4	0	2	0	0
0	6	3	7	4	0
0	0	0	0	0	0

0	0	0
0	1	2
0	5	3

FILTER 1
Blue

-7	7	7
-5	-6	-3
1	-4	6

CONVOLUTIONAL LAYER 1

FILTER 1

Red

-6	7	1
3	-1	4
2	5	6

Green

0	-2	-5
-1	3	-4
6	7	1

Blue

-7	7	7
-5	-6	-3
1	-4	6

E. When the grids are filled, the totals for the red, green, and blue channel calculations are summed up along with a number called the weight, which represents the importance of this filter for the final output.

38	65	60	19
79	67	68	48
78	68	77	14
10	77	-6	23

+

49	69	45	37
-28	10	22	26
-63	-14	-19	1
-30	-27	-15	1

+

-14	-24	-15	-66
-34	2	-11	-44
26	10	22	9
-17	-83	-55	-73

+ *1* =

74	111	91	-9
18	80	80	31
42	65	81	25
-36	-32	-75	-48

FILTER 999

Red

4	-3	7
3	-2	-6
3	-6	-5

Green

3	1	5
-3	-1	2
-3	-5	-4

Blue

6	5	-3
0	1	2
3	-1	0

G. A typical CNN may have hundreds of filters in each layer. Each filter performs its simple calculations and feeds its feature map forward to all the filters in the next layer. Those filters examine the feature maps in the same chunk-by-chunk way.

-62	-41	-57	4
-36	-12	26	-3
-68	-21	-47	27
14	-16	15	22

+

-50	-64	-51	-39
9	-19	-13	-5
44	36	9	-6
30	21	28	9

+

0	22	19	4
6	23	32	56
14	58	14	59
32	35	25	16

+ *-2* =

-114	-85	-91	-33
-23	-10	43	46
-12	71	-26	78
74	38	66	45

but we're now reaching the critical point when research projects can be turned into tools that assist in clinical care.

In a CNN, many layers of artificial neurons perform remarkably basic calculations and feed the results forward in a simplified way. The power of CNNs—and their advantage over other neural networks—comes from the clever arrangement of these simple steps, which keeps the computational load within reasonable bounds.

Each neuron in a CNN can be thought of as a filter that scans an image for one particular feature. To make sense of any image—whether the CNN is trying to distinguish between a beer bottle and a soda bottle, or between a healthy and a cancerous cervix—it may use thousands of filters arranged in multiple layers, which collectively perform billions of calculations.

The first layer of filters looks at the digital image on the pixel scale, taking in each pixel's numerical values for its red, green, and blue channels. One filter may detect vertical lines, while another may look for a certain color. Every filter examines the image in small and manageable chunks, seeking its particular feature in each chunk, and then represents its findings as number values in a feature map. These uncomplicated maps are then fed forward, with each map being used as an input for the filters in the next layer. These next filters will respond to features that are slightly larger or more abstract, such as the edge of an object or the presence of a flesh tone.

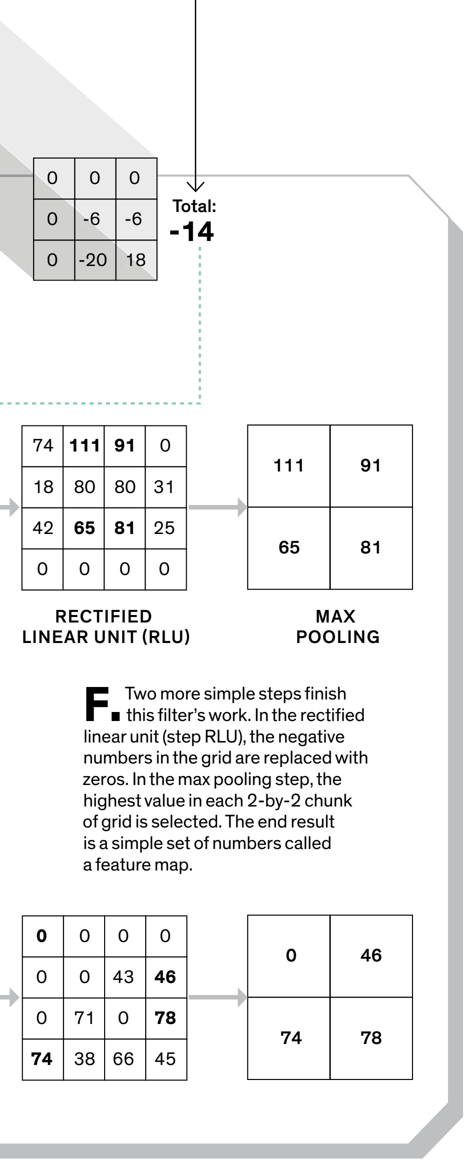
This routine continues, with each layer identifying increasingly complex forms or patterns. Finally, a layer sends its output to a "fully connected" layer that doesn't do a chunk-style scan of the feature maps but instead looks at them in their entirety. Typical CNNs finish with a couple of fully connected layers that look at the big picture (which is really just a lot of number values) and determine how well it matches the CNN's template for an object. That holistic view enables the final layer to declare: This is a beer bottle. Or, this is a healthy cervix. It can make this assessment even though human programmers never tell the network what filters to use—the AI has made those determinations itself during training.

HOW THE MACHINE LEARNS

To train a CNN to distinguish between a healthy cervix [top] and one showing signs of cancer [bottom], we give it thousands of labeled images of cervixes. For its first run through the image set, the CNN decides which filters to use and how much weight to give each filter. It then goes through the images and classifies each as either healthy or cancerous. When it's finished, it checks its accuracy; based on the results, it changes the filters and weights before its next run.

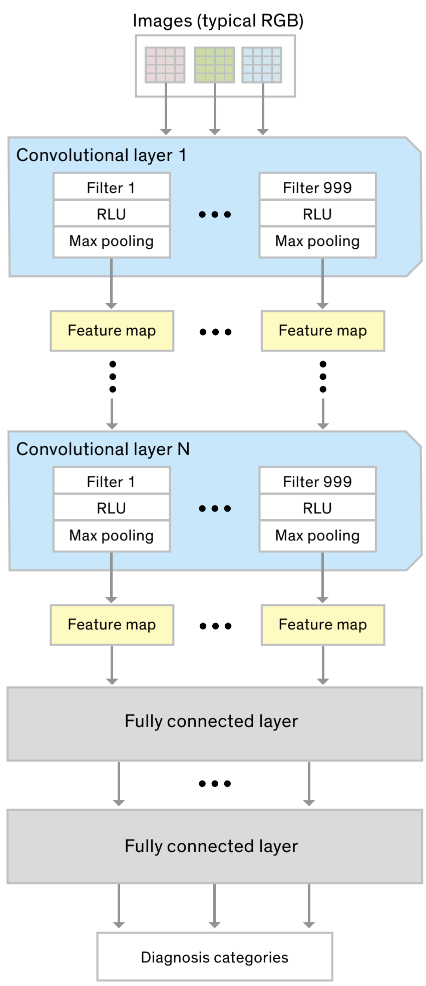
PHOTOS: GLOBAL GOOD

D. The resulting products for the first chunk are summed up, and the total is put down in one cell of a grid. Then the filter moves over one pixel to the right and looks at the next 3-by-3 chunk.



F. Two more simple steps finish this filter's work. In the rectified linear unit (step RLU), the negative numbers in the grid are replaced with zeros. In the max pooling step, the highest value in each 2-by-2 chunk of grid is selected. The end result is a simple set of numbers called a feature map.

H. For each digital image, a CNN uses many layers of convolutional filters. Finally, the last convolutional layer outputs all of its feature maps to a "fully connected" layer, which examines the maps in their entirety. The CNN uses several fully connected layers to make a final determination about the image's content.



We human programmers aren't mere observers, however. We establish the CNN's computing architecture by setting both the number of filters in each layer and the number of layers. We explore many combinations: Adding filters might make the CNN take note of more fine-grained details, or adding layers could cause a more gradual progression from raw image to abstract classification. We examine the results of each architecture, make some changes, and do another run. Equally important to the CNN's success is the set of images we provide for training: The images must represent the range that will be encountered in the real world and must be correctly sorted into labeled categories. A bad image set will yield bad results.

To train our CNN for cervical cancer screening, we're feeding in approximately 100,000 images of cervixes sorted into categories such as healthy tissue, benign inflammation, precancerous lesions, and suspected cancer. We're aided here by a partnership with the U.S. National Cancer Institute, which has given us access to its databases of high-quality, annotated, anonymized images. We're training our CNN with these "ideal" images before turning to the trickier images obtained from clinics.

We're using a software program called Caffe, which was developed at the University of California, Berkeley, as the training framework for our CNN. We first define the CNN architecture, and then Caffe runs our image set through the CNN using one set of filters. After that, it checks to see how well its classification system performed. Caffe then adjusts the CNN filters to try to improve the system's overall accuracy. It's like a black box with millions of knobs being automatically turned: We understand some of the features it focuses on, like colors and lines, but many are completely inscrutable. Caffe keeps turning knobs until the CNN reaches some plateau of performance or until it becomes obvious that it's a bad run. That's when we humans step back in to try a new architecture.

Once our CNN seems proficient, we'll challenge it to classify cervical images not used during its learning process. This is a crucial validation step, because CNNs can get great results when classifying familiar images but fail spectacularly when confronted with a new data set. When we get to this step later this year, we'll use a subset of images from the National Cancer Institute that we've kept apart.

CNN training requires plenty of computational power. When Caffe is putting our CNN through a run, it performs nearly a trillion simple mathematical computations, which would take weeks or even months on a high-end multicore CPU machine. But doing the calculations on the graphics processing units (GPUs) used for high-end video games and simulations can drastically speed up this number crunching. At Global Good, we use two high-performance computer clusters that are stuffed with GPUs to run through these iterations using images stored on a massive array of disks. It takes about 72 hours to do a single training session. Still, in the world of CNNs, a three-day run is considered quick!

The final challenge will be bringing our well-trained system out into the field, to see whether it can make sense of images captured in a wide range of conditions. A health worker using MobileODT's EVA Scope to conduct an exam employs a speculum and a light to view the cervix. Our AI will be trained to ignore the speculum and the light reflection from it. But it must also deal with variations that arise when you're taking a photo with a mobile device, such as irregularities in lighting, alignment, steadiness, and focus. Our system must be trained with images from these in-the-field exams to make sense of such inconsistencies. When we begin our trials in Ethiopia later this year, we'll assess our smart device's performance by checking its diagnoses against medical experts' assessments and pathology tests in the lab.

To make the technology practical for remote rural clinics, the smart EVA Scope won't require connection to the cloud to evaluate images. The heavy computations take place during the training of a CNN, but once the system is configured it doesn't require much processing power to evaluate an image. The EVA Scope will do that with an app on the smartphone to which it's attached.

We're already looking ahead to future generations of this technology. Currently, diagnostic tools based on machine learning are trained on a large initial data set prior to deployment, but they don't learn while they're being used in the field. Eventually, we hope to develop AI systems that can continue to improve their skills based on the cases and patients they encounter, adapting to changing conditions and improving their ability to make diagnoses. This adaptation wouldn't require the heavy computational power that's needed for basic training, since the CNN would be adding only a few images at a time to its data set. And if we can design AIs that are really good at making adjustments as they go, we can spend less time and energy on the initial training.

Our current effort may be just the beginning for machine-learning diagnostic tools. As many other medical exams rely on images such as X-rays and MRI scans, it's easy to imagine using other smart tools to classify images, finding patterns and outliers. We can also use CNNs to detect rare objects that are tough to spot. For example, our team is now working on a program to diagnose malaria, which requires examining a blood sample under a microscope to find minute malaria parasites. For people with low parasite levels, that's a challenge akin to finding a handful of marbles in a football field.

In many ways, machine learning is still in its infancy. A 3-year-old child has to look at only a few pictures of a cat before getting the concept, while a CNN has to look at millions. But in the medical field the comparison between humans and machines could shift in coming decades, as programs use features that are beyond our comprehension to scan medical images and draw conclusions. The machines may surpass doctors, but if they're doing it in the service of humanity, we'll take it. ■

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/aimedicine0517

THE PIT BOSS

CONTINUED FROM PAGE 33

started changing, with less and less room for "unfettered research."

"In the old days, the Philips research lab was an unregulated environment–good scientists and engineers got lots of freedom, and nice results came out of that," recalls Jos Weber, an associate professor at the Delft University of Technology who has known Immink for decades and collaborates with him today. "When Philips restructured its labs to run more as a business unit, with specific targets, he was not happy."

Immink's employment agreement with Philips prevented him from working for a year; he spent most of that time working on a book. Meanwhile, efforts to develop blue-laser optical disc technology, now referred to as Blu-ray Disc, continued, with more and more companies signing onto a group effort, including Korean upstart LG. The coding scheme then on the table, but not yet approved, was a version of Immink's EFM, called Enhanced EFM. Immink realized he had an idea for a different and better approach, one that started with 9 bits of data–not 8–before translating them into sequences of pits and lands. It was just a few percentage points better, he says, but it would allow more data to fit on a disc.

He filed patent applications for the technology, and then started working with a patent attorney to see about obtaining patents in at least the major markets. The process, he discovered to his dismay, would cost him hundreds of thousands of dollars and take years. And he'd have to file those applications within a year of having filed his application in the Netherlands.

On the advice of a Japanese friend, he contacted the Korean consumer electronics manufacturers, hoping to sell his work on the basis of the application and let the buyer complete the patent process.

The timing turned out to be perfect. LG, having just joined the group of companies defining the Blu-ray Disc format, wanted to submit a technology it owned for possible inclusion in the standard, but its team had nothing, Immink says. "At that moment, I knocked on their door with my technology, and they said, 'We want to buy it.' Just like that."

The technology did not, in the end, make it into the Blu-ray standard. But the sale made Immink independently wealthy.

Since then, he's stayed involved with research. Officially, he's heading up his own company, Turing Machines. That business, created to manage his patents, now holds his investments in startup companies. He spends his time researching on his own or with faculty and graduate students at Delft University and at the Singapore University of Technology and Design (SUTD). His rela-

DIGITAL HORIZONS: These days, Kees Immink continues to research technologies for digital data storage, working in the Netherlands and Singapore.

RUUD BAAN

tionship with Delft University is informal; at SUTD he's a visiting professor.

Immink's research still focuses on the storage of digital data, but he's turned from optical to solid-state drives as well as more far-out projects, like translating bits into DNA sequences for data archiving. "I love to come up with inventive methods to solve new problems," he says.

SUTD associate professor Cai Kui is one of the people laying out those problems. Immink spends time in Singapore when it's winter in Europe. During his stay, Kui says, "We meet three times a week, discussing new ideas. We'll take a day in between to think about what we discussed and do individual work on it. Then we might hand off two ideas to Ph.D. students or postdocs to analyze and compare.

"Even at 70," she says, "Immink continues doing hands-on research and can always come up with new ideas. His continued enthusiasm for doing research has inspired me."

In the Netherlands, Immink bikes about 15 kilometers from his home in Rotterdam to Delft University, where he meets Weber about once a month. "He is very creative," Weber says. "He's still the one that comes up with the ideas." The two regularly publish joint papers, which Immink drafts in the riverside studio he shares with his wife, an artist. Delft students seeking guidance sometimes visit him there.

When he's not thinking about digital storage, he rows. Immink is part of a mixed-gender group that trains on eight-person sculls on the Rotte river. And he listens to music. His tastes are eclectic, running from Strauss waltzes to country and western–and, of course, he listens to it on CDs. ■

IEEE SPECTRUM Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page

Applications are invited for faculty positions in the Department of Electrical and Computer Engineering for tenure, research, and teaching-tracks at multiple locations.

The Department of Electrical and Computer Engineering (ECE) invites applications for tenure, research, and teaching-track positions at its main campus in Pittsburgh (www.ece.cmu.edu), as well as Silicon Valley (www.cmu.edu/silicon-valley) and Rwanda (www.cmu.edu/rwanda) locations. Silicon Valley faculty will have a unique opportunity for research collaborations with industry and involvement in entrepreneurial activities emerging from research projects. Rwanda faculty will collaborate with regional and multi-national industry in technology innovation and entrepreneurship and deliver innovative, interdisciplinary graduate teaching and research programs within the African context.

We are strongly committed to all members of our community: students, faculty, and staff. Our vision is to be a creative driving force, within the university and worldwide, of the highest scholarly and entrepreneurial quality. Our mission is to inspire and educate engineers capable of pursuing fundamental scientific problems and important societal challenges. We strive to accomplish this with the highest commitment to quality, integrity, and respect for others. We are particularly interested in applicants who are committed and have passion for a culturally diverse environment in research and/or teaching, and demonstrate a willingness to nurture the uniquely inclusive Carnegie Mellon environment. We take pride and active steps in considering a diverse applicant pool in terms of gender, race, veteran status, and disability. Carnegie Mellon University seeks to meet the needs of dual-career couples and is a member of the Higher Education Recruitment Consortium (HERC) that assists with dual-career searches.

We are looking to hire in strategic thrust areas (www.ece.cmu.edu/research/index.html):

- Theoretical and Technological Foundations serve as a bedrock for our work;
- Systems and Technologies, connecting system level to physical and cyber levels (beyond CMOS, compute/storage systems, cyber-physical systems, data/network science systems, secure systems), with particular attention to nanoscale systems and integration of heterogeneous devices, sensors and materials; and
- Application Domains, where our fundamental and system level work makes significant impact (energy, healthcare & quality-of-life, mobile systems, smart infrastructure).

In Silicon Valley, our emphasis on experimental computing systems, platforms for mobile supercomputing, human, vehicle and environment sensing, and sensor signal processing and inference. In Rwanda, our focus is on software engineering, mobile computing, cloud computing, big data analytics, communications, information and cyber security, intelligent infrastructures, wireless networking, mHealth, energy systems, eLearning, and mobile applications.

- Tenure-track faculty carry a moderate teaching load that allows time for quality research and close involvement with students. We expect you to establish and grow a strong research program, contribute to our teaching mission, and show your passion for mentoring and advising students.
- Research-track faculty are not required to teach, but do so occasionally when of clear benefit to the faculty and the Department; you will be compensated for both teaching and advising Ph.D. students. You will typically focus on developing leadership within your area of research, developing research collaborations, and supervising Ph.D. students.
- Teaching-track faculty typically focus exclusively on teaching and service, but may do research as well. We will rely on you to help strengthen our teaching and mentoring mission.

For all tracks, we are seeking individuals who hold a Ph.D. in a relevant discipline and have demonstrated commitment to our core values: scientific truth, creativity, quality, innovation, and engineering solutions, all within a diverse and tight-knit community guided by respect and joy of doing. Faculty positions are primarily at the Assistant Professor level; however, appointments may be made at the rank of Associate Professor or Professor depending on the qualifications. Our Department and the College of Engineering are ranked among the top programs in the United States both at the undergraduate and graduate levels. We house and have ties to several multidisciplinary institutes and centers. We collaborate with colleagues around the world through a number of formal research and educational programs. We have extensive experimental and computing infrastructure, including state-of-the-art nanofabrication facilities.

Please submit an online application at www.ece.cmu.edu/faculty-staff/employment/index.html. We will begin evaluation of applications immediately and will continue throughout the academic year until positions are filled; we encourage you to submit early. Carnegie Mellon is an EEO/Affirmative Action Employer -- M/F/Disability/Veteran.

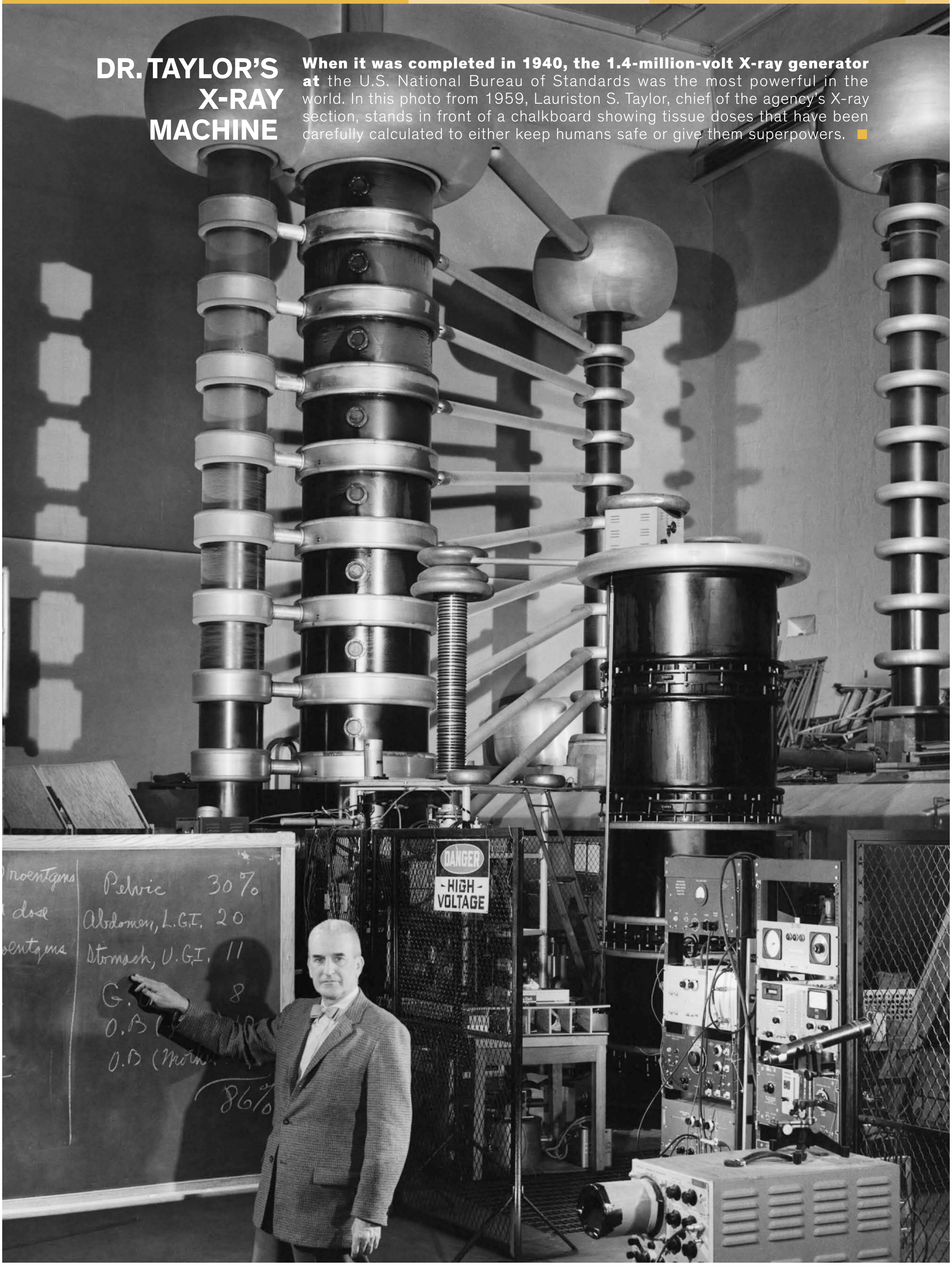
Electrical & Computer ENGINEERING

Carnegie Mellon University College of Engineering

PAST FORWARD_BY EVAN ACKERMAN

DR. TAYLOR'S X-RAY MACHINE

When it was completed in 1940, the 1.4-million-volt X-ray generator at the U.S. National Bureau of Standards was the most powerful in the world. In this photo from 1959, Lauriston S. Taylor, chief of the agency's X-ray section, stands in front of a chalkboard showing tissue doses that have been carefully calculated to either keep humans safe or give them superpowers. ■



NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

↗ POST YOUR COMMENTS at http://spectrum.ieee.org/pastforward0517


ieee.tv
Featured Videos
Videos by Series
Event Showcase
Most Popular
NEW
New Videos
Join IEEE
IEEE
About IEEE
About This App
tv
About IEEE.tv
Member Login
Share
IEEE

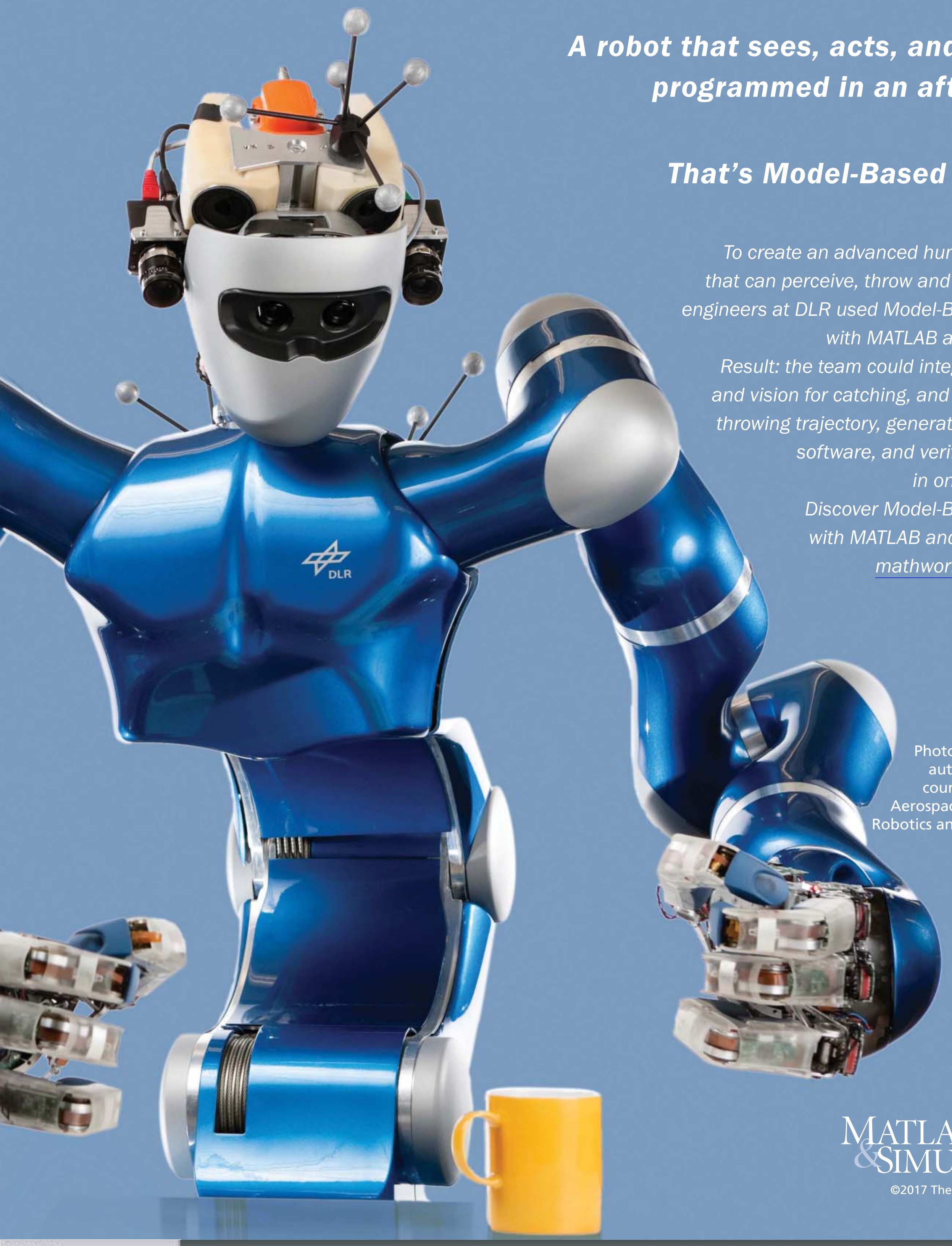


Photo of Agile Justin autonomous robot courtesy of German Aerospace Center (DLR), Robotics and Mechatronics Center