

COMMUNICATIONS

CACM.ACM.ORG

OF THE

ACM

05/2013 VOL.56 NO.05



Collaboration with a Robotic Scrub Nurse

Vanishing Electronics

Discrimination in
Online Ad Delivery

Fair Use in Europe

Consumer Technologies
in Emerging Markets

SENSE THE TRANSFORMATION



CALL FOR SUBMISSIONS

Digital Art | CG Research | Animation | Technological Innovations
| Post Production | Interactive Applications | Mobile Graphics | Industry Trends

Present your creative achievements, innovations and stellar technical research.
Review program guidelines, plan your schedule and submit your best work.

Technical Papers	14 May
Courses	6 June
Symposium on Mobile Graphics and Interactive Applications	6 June
Art Gallery	13 June
Emerging Technologies	13 June
Computer Animation Festival	2 July
Posters	9 July
Technical Briefs	9 July

For complete details, visit sa2013.siggraph.org/submitters.



SIGGRAPH
ASIA 2013
HONG KONG

CONFERENCE 19 NOV - 22 NOV
EXHIBITION 20 NOV - 22 NOV

**HONG KONG CONVENTION
AND EXHIBITION CENTRE**

SA2013.SIGGRAPH.ORG

LEAD SPONSOR



SPONSORED BY





15th International Conference on Human-Computer
Interaction with Mobile Devices and Services

MOBILEHCI 2013

Munich, Germany

August 27-30

@ <http://www.mobilehci2013.org>

 <http://www.facebook.com/MobileHCI2013>

 @MobileHCI2013



Departments

- 5 **Editor's Letter**
Fricative Computing
Let's bring friction back into computing.
By Moshe Y. Vardi
-
- 7 **From the President**
ACM President's Salary Increased by 300%!
By Vinton G. Cerf
-
- 9 **Publisher's Corner**
A Few Good Reasons to Publish in *Communications*
By Scott E. Delman
-
- 12 **Letters to the Editor**
Try Old Boys Security Network
-
- 14 **BLOG@CACM**
Encouraging IT Usage In Future Healthcare, Quality in CS Education
Jeannette M. Wing considers how technology acts as a change agent for healthcare, while Mark Guzdial ponders ways to measure quality in computer science education.
-
- 41 **Calendar**
-
- 102 **Careers**

Last Byte

- 104 **Puzzled**
Ant Alice's Adventures
By Peter Winkler

News



- 17 **Proving Grounds**
Researchers are making headway with one of quantum computing's major theoretical problems: multi-prover interactive proofs.
By Alex Wright
-
- 20 **Vanishing Electronics**
Engineers are reinventing electronics by building safe devices that dissolve in the body or within the environment. The technology could redefine everything from medicine to computing.
By Sam Greengard
-
- 23 **'Small Data' Enabled Prediction Of Obama's Win, Say Economists**
"Big data" from crowdsourcing resulted in more complex predictions.
By Paul Hyman

Viewpoints

- 26 **Law and Technology**
Fair Use in Europe
Examining the mismatch between copyright law and technology-influenced evolving social norms in the European Union.
By P. Bernt Hugenholtz
-
- 29 **Historical Reflections**
Max Newman: Forgotten Man of Early British Computing
Reflections on a significant, yet often overlooked, computing pioneer.
By David Anderson
-
- 32 **Education**
Human-Centered Computing: A New Degree for Licklider's World
Combining computing and psychology, J.C.R. Licklider's prescient ideas are being applied in contemporary educational settings.
By Mark Guzdial
-
- 35 **Viewpoint**
The Science in Computer Science
Computer science is in a period of renaissance as it rediscovers its science roots.
By Peter J. Denning
-
- 39 **Viewpoint**
Moving from Petaflops to Petadata
The race to build ever-faster supercomputers is on, with more contenders than ever before. However, the current goals set for this race may not lead to the fastest computation for particular applications.

Practice

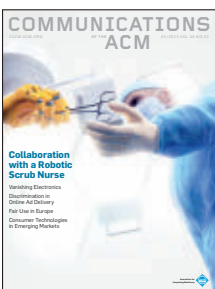


- 44 **Discrimination in Online Ad Delivery**
Google ads, black names and white names, racial discrimination, and click advertising.
By Latanya Sweeney

- 55 **Eventual Consistency Today: Limitations, Extensions, and Beyond**
How can applications be built on eventually consistent infrastructure given no guarantee of safety?
By Peter Bailis and Ali Ghodsi

- 64 **A File System All Its Own**
Flash memory has come a long way and it is time for software to catch up.
By Adam H. Leventhal

Q Articles' development led by acmqueue.queue.acm.org



About the Cover:
An operating room-based robotic arm capable of passing key surgical instruments to a doctor during surgery may soon free OR technicians to perform other concurrent tasks. A firsthand account of experiments with a robotic scrub nurse is detailed in this month's cover story, beginning on page 68. Cover illustration by Peter Crowther Associates.

Contributed Articles



- 68 **Collaboration with a Robotic Scrub Nurse**
Surgeons use hand gestures and/or voice commands without interrupting the natural flow of a procedure.
By Mithun George Jacob, Yu-Ting Li, George A. Akingba, and Juan P. Wachs

- 76 **Strategies for Tomorrow's 'Winners-Take-Some' Digital Goods Markets**
Markets characterized by multiple competing digital standards have room for more than one winner, unlike traditional analog markets.
By Chris F. Kemerer, Charles Zhechao Liu, and Michael D. Smith

Review Articles

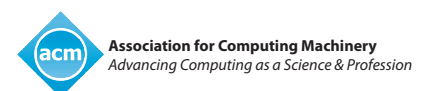


- 84 **The Promise of Consumer Technologies in Emerging Markets**
Employees in emerging markets find their own IT devices vital to job productivity and innovation.
By Iris Junglas and Jeanne Harris

Research Highlights

- 92 **Technical Perspective**
The Ray-Tracing Engine that Could
By Matt Pharr

- 93 **GPU Ray Tracing**
By Steven G. Parker, Heiko Friedrich, David Luebke, Keith Morley, James Bigler, Jared Hoberock, David McAllister, Austin Robison, Andreas Dietrich, Greg Humphreys, Morgan McGuire, and Martin Stich





ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.

Executive Director and CEO

John White
Deputy Executive Director and COO
 Patricia Ryan
Director, Office of Information Systems
 Wayne Graves
Director, Office of Financial Services
 Russell Harris
Director, Office of SIG Services
 Donna Cappel
Director, Office of Publications
 Bernard Rous
Director, Office of Group Publishing
 Scott E. Delman

ACM COUNCIL

President
 Vinton G. Cerf
Vice-President
 Alexander L. Wolf
Secretary/Treasurer
 Vicki L. Hanson
Past President
 Alain Chesnais
Chair, SGB Board
 Erik Altman
Co-Chairs, Publications Board
 Ronald Boisvert and Jack Davidson
Members-at-Large
 Eric Allman; Ricardo Baeza-Yates;
 Radia Perlman; Mary Lou Soffa;
 Eugene Spafford
SGB Council Representatives
 Brent Hailpern; Joseph Konstan;
 Andrew Sears

BOARD CHAIRS

Education Board
 Andrew McGettrick
Practitioners Board
 Stephen Bourne

REGIONAL COUNCIL CHAIRS

ACM Europe Council
 Fabrizio Gagliardi
ACM India Council
 Anand S. Deshpande, PJ Narayanan
ACM China Council
 Jianguang Sun

PUBLICATIONS BOARD

Co-Chairs
 Ronald F. Boisvert; Jack Davidson
Board Members
 Marie-Paule Cani; Nikil Dutt; Carol Hutchins;
 Joseph A. Konstan; Ee-Peng Lim;
 Catherine McGeoch; M. Tamer Ozsu;
 Vincent Shen; Mary Lou Soffa

ACM U.S. Public Policy Office

Cameron Wilson, Director
 1828 L Street, N.W., Suite 800
 Washington, DC 20036 USA
 T (202) 659-9711; F (202) 667-1066

Computer Science Teachers Association

Chris Stephenson,
 Executive Director

COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

Communications of the ACM is the leading monthly print and online magazine for the computing and information technology fields. *Communications* is recognized as the most trusted and knowledgeable source of industry information for today's computing professional. *Communications* brings its readership in-depth coverage of emerging areas of computer science, new trends in information technology, and practical applications. Industry leaders use *Communications* as a platform to present and debate various technology implications, public policies, engineering challenges, and market trends. The prestige and unmatched reputation that *Communications of the ACM* enjoys today is built upon a 50-year commitment to high-quality editorial content and a steadfast dedication to advancing the arts, sciences, and applications of information technology.

STAFF

DIRECTOR OF GROUP PUBLISHING

Scott E. Delman
 publisher@cacm.acm.org

Executive Editor

Diane Crawford

Managing Editor

Thomas E. Lambert

Senior Editor

Andrew Rosenbloom

Senior Editor/News

Larry Fisher

Web Editor

David Roman

Editorial Assistant

Zarina Strakhan

Rights and Permissions

Deborah Cotton

Art Director

Andrij Borys

Associate Art Director

Margaret Gray

Assistant Art Directors

Mia Angelica Balaquiot

Brian Greenberg

Production Manager

Lynn D'Addesio

Director of Media Sales

Jennifer Ruzicka

Public Relations Coordinator

Virginia Gold

Publications Assistant

Emily Williams

Columnists

Alok Aggarwal; Phillip G. Armour;
 Martin Campbell-Kelly;
 Michael Cusumano; Peter J. Denning;
 Shane Greenstein; Mark Guzdial;
 Peter Harsha; Leah Hoffmann;
 Mari Sako; Pamela Samuelson;
 Gene Spafford; Cameron Wilson

CONTACT POINTS

Copyright permission
 permissions@cacm.acm.org

Calendar items
 calendar@cacm.acm.org

Change of address
 acmhq@cacm.acm.org

Letters to the Editor
 letters@cacm.acm.org

WEBSITE
 http://cacm.acm.org

AUTHOR GUIDELINES
 http://cacm.acm.org/guidelines

WEBSITE

http://cacm.acm.org

AUTHOR GUIDELINES

http://cacm.acm.org/guidelines

ACM ADVERTISING DEPARTMENT

2 Penn Plaza, Suite 701, New York, NY
 10121-0701
 T (212) 626-0686
 F (212) 869-0481

Director of Media Sales

Jennifer Ruzicka
 jen.ruzicka@hq.acm.org

Media Kit acmm mediasales@acm.org

Association for Computing Machinery (ACM)

2 Penn Plaza, Suite 701
 New York, NY 10121-0701 USA
 T (212) 869-7440; F (212) 869-0481

EDITORIAL BOARD

EDITOR-IN-CHIEF

Moshe Y. Vardi
 eic@cacm.acm.org

NEWS

Co-Chairs

Marc Najork and Prabhakar Raghavan

Board Members

Hsiao-Wuen Hon; Mei Kobayashi;
 William Pulleyblank; Rajeev Rastogi

VIEWPOINTS

Co-Chairs

Susanne E. Hambrusch; John Leslie King;
 J Strother Moore

Board Members

William Aspray; Stefan Bechtold; Judith
 Bishop; Stuart I. Feldman;
 Peter Freeman; Seymour Goodman;
 Mark Guzdial; Richard Heeks;
 Rachele Hollander; Richard Ladner;
 Susan Landau; Carlos Jose Pereira de Lucena;
 Beng Chin Ooi; Loren Terveen;
 Jeannette Wing

PRACTICE

Chair

Stephen Bourne

Board Members

Eric Allman; Charles Beeler; Bryan Cantrill;
 Terry Coatta; Stuart Feldman; Benjamin Fried;
 Pat Hanrahan; Tom Limoncelli;
 Marshall Kirk McKusick; Erik Meijer;
 George Neville-Neil; Theo Schlossnagle;
 Jim Waldo

The Practice section of the CACM

Editorial Board also serves as
 the Editorial Board of *COMMUNIQUE*.

CONTRIBUTED ARTICLES

Co-Chairs

Al Aho and Georg Gottlob

Board Members

William Aiello; Robert Austin; Elisa Bertino;
 Gilles Brassard; Kim Bruce; Alan Bundy;
 Peter Buneman; Erran Carmel;
 Andrew Chien; Peter Druschel; Carlo Ghezzi;
 Carl Gutwin; James Larus; Igor Markov;
 Gail C. Murphy; Shree Nayar; Bernhard
 Nebel; Lionel M. Ni; Sriram Rajamani;
 Marie-Christine Rousset; Avi Rubin;
 Krishan Sabnani; Fred B. Schneider;
 Abigail Sellen; Ron Shamir; Yoav Shoham;
 Marc Snir; Larry Snyder; Manuela Veloso;
 Michael Vitale; Wolfgang Wahlster;
 Hannes Werthner; Andy Chi-Chih Yao

RESEARCH HIGHLIGHTS

Co-Chairs

Stuart J. Russell and Gregory Morrisett

Board Members

Martin Abadi; Sanjeev Arora; Dan Boneh;
 Andrei Broder; Stuart K. Card; Jon Crowcroft;
 Alon Halevy; Monika Henzinger;
 Maurice Herlihy; Norm Jouppi;
 Andrew B. Kahng; Xavier Leroy;
 Mendel Rosenblum; David Salesin;
 Guy Steele, Jr.; David Wagner;
 Alexander L. Wolf; Margaret H. Wright

WEB

Chair

James Landay

Board Members

Gene Golovchinsky; Marti Hearst;
 Jason I. Hong; Jeff Johnson; Wendy E. MacKay



ACM Copyright Notice

Copyright © 2013 by Association for Computing Machinery, Inc. (ACM). Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or fee. Request permission to publish from permissions@acm.org or fax (212) 869-0481.

For other copying of articles that carry a code at the bottom of the first or last page or screen display, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center; www.copyright.com.

Subscriptions

An annual subscription cost is included in ACM member dues of \$99 (\$40 of which is allocated to a subscription to *Communications*); for students, cost is included in \$42 dues (\$20 of which is allocated to a *Communications* subscription). A nonmember annual subscription is \$100.

ACM Media Advertising Policy

Communications of the ACM and other ACM Media publications accept advertising in both print and electronic formats. All advertising in ACM Media publications is at the discretion of ACM and is intended to provide financial support for the various activities and services for ACM members. Current Advertising Rates can be found by visiting <http://www.acm-media.org> or by contacting ACM Media Sales at (212) 626-0686.

Single Copies

Single copies of *Communications of the ACM* are available for purchase. Please contact acmhq@acm.org.

COMMUNICATIONS OF THE ACM

(ISSN 0001-0782) is published monthly by ACM Media, 2 Penn Plaza, Suite 701, New York, NY 10121-0701. Periodicals postage paid at New York, NY 10001, and other mailing offices.

POSTMASTER

Please send address changes to *Communications of the ACM*
 2 Penn Plaza, Suite 701
 New York, NY 10121-0701 USA



Association for Computing Machinery



Printed in the U.S.A.



Moshe Y. Vardi

DOI: 10.1145/2447976.2447977

Fricative Computing

Let's bring friction back into computing.

MARIE DE RABUTIN-CHANTAL, marquise de Sévigné, was a 17th century French Parisian aristocrat, remembered for her 30-year-long correspondence with her Provence-residing daughter. Over a thousand of their letters have been preserved and published in the 18th century. They are considered a treasure of French literature. The time gap—two weeks—between the mother's letters and her daughter's replies intensified the mother's worries, longings, and anxieties, which are vividly reflected in her letters.

It is difficult to imagine such correspondence today. The marquise and her daughter would be communicating frequently via email and text messages. The frequent exchanges would likely be quotidian and mundane, lacking the frisson that enriches the 17th century letters. The emotional depth of these letters resulted from the difficulty of communication between mother and daughter. Eliminate that difficulty, and the emotions are eliminated as well. It is quite unlikely that future generations will cherish personal correspondence from the 21st century.

Our discipline is dedicated to reducing friction. Latency must be eliminated, bandwidth must increase, and ubiquity should be universal. Our goal is to reduce the friction of computing and communication as much as possible. Facebook's CEO Mark Zuckerberg speaks of "frictionless sharing" as a goal. This reduction of friction has enabled the amazing world of the Internet and the Web we have created over the past 50 years, but should zero friction really be our goal?

One may dismiss my concern about letters-not-written as sentimental and anachronistic, but the effects of fric-

tionless computing are quite serious. On May 6, 2010, at 2:45 P.M., the U.S. stock market declined steeply, with the Dow Jones Industrial Average plunging about 600 points in five minutes, following an earlier decline of more than 300 points on worries about the Greek debt crisis. As described in a just-published book by Neil Irwin, *The Alchemists: Three Central Bankers and a World on Fire*, the "flash crash" deeply shook the confidence of central bankers and had a dramatic impact on their decision making.

It was only a few months later that a joint report issued by the U.S. Securities and Exchange Commission and the Commodity Futures Trading Commission identified high-frequency trading as the cause of the crash. It turns out that much of the trading taking place today on financial markets is algorithmic, with software tools entering trading orders, using speed and frequency that cannot be matched by humans. Proponents of high-frequency trading say it helps make the markets more "liquid," but Thomas Peterffy, a high-frequency trading pioneer, recently argued that "Today's drive for speed has absolutely no social value."

Imagine a mechanical engineer who declares that her goal is to eliminate friction, period. We would view this

The world cannot function without friction.

as insane. The world cannot function without friction. The goal should be to have the right amount of friction, in the right place, in the right time. Yet our discipline seems committed to the total elimination of friction in computing.

The adverse effects of frictionless computing are all around us. Email is the first example that comes to mind. It is simply too easy to send email messages, so we all send too many (I am famously guilty of this) and receive too many. It is also too easy to add recipients. We are simply drowning in email; with many articles bemoaning the "email tsunami" and "Pandora's inbox." An invention that was meant to free us from the overhead of paper communication ended up enslaving us electronically.

In almost every area touched by computing, we can see the symptoms of reduced friction. In a recent book by Dan Slater, *Love in the Time of Algorithms*, the author laments how online romance is threatening monogamy. "What if online dating makes it too easy to meet someone new?" he asks. "What if it raises the bar for a good relationship too high?" In essence, has online dating over-reduced the friction of dating?

The Greek philosopher Aristotle said "Anybody can become angry—that is easy, but to be angry with the right person, to the right degree, at the right time, for the right purpose, and in the right way—that is not easy." I feel the same about friction in computing. Reducing friction is easy, but having the right amount of friction, for the right application, in the right context, that is not easy, and is today a major challenge of computing.

Moshe Y. Vardi, EDITOR-IN-CHIEF

Autonomous Car Driving Simulator

- Vehicle Motion Platform -

www.vehiclemotion.net/ www.motionlab.com/
www.youtube.com/user/motionlabLLC

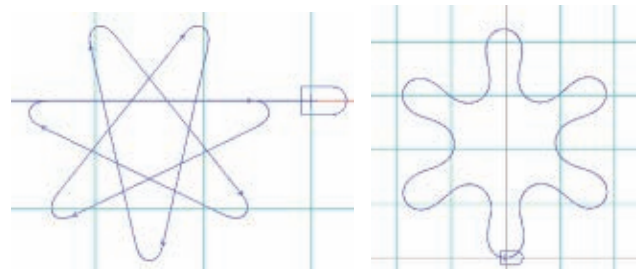
Unmanned car driving is one of the hottest topics in the robotics/AI/CS community and in the automotive industry; to approach this paradigm, we first must discover the principle of human drivers' steeringwheel-maneuvering skills. A single driver is capable of operating a number of different types of vehicles with distinct dimensions, weights, and powertrains. This observation led MotionLab to formalize the *hardware-independent* human driving skill, which it calls *abstract vehicle motion*.

MotionLab's *Vehicle Motion Platform* is a Java®-based simulator that installs all abstract vehicle motions developed so far. Its license can be purchased for \$399 from www.vehiclemotion.net/. The *Platform* also has a programmable capacity so that a user can write his or her own program to test and run a new abstract vehicle motion. The simulator can also be viewed at www.motionlab.com/cms.php?menuId=38.



Car-Like Robotics since 1975
MotionLab LLC; CA, U.S.A

The *Vehicle Motion Bible* is a PDF document that contains all abstract vehicle motions installed in the *Platform*, the *MotionMind Robot-Control Language Manual*, and an introduction to *Symmetric Geometry*. The Bible can be downloaded for free from www.vehiclemotion.net/. *Symmetric Geometry* is the rock-solid theoretical foundation of *Vehicle Motion Science*. (Download the *Platform* with confidence; a standalone Java® application never becomes a source of security threat.)



World-Renowned Journals from ACM

ACM publishes over 50 magazines and journals that cover an array of established as well as emerging areas of the computing field. IT professionals worldwide depend on ACM's publications to keep them abreast of the latest technological developments and industry news in a timely, comprehensive manner of the highest quality and integrity. For a complete listing of ACM's leading magazines & journals, including our renowned Transaction Series, please visit the ACM publications homepage: www.acm.org/pubs.

ACM Transactions on Interactive Intelligent Systems



ACM Transactions on Interactive Intelligent Systems (TIIS). This quarterly journal publishes papers on research encompassing the design, realization, or evaluation of interactive systems incorporating some form of machine intelligence.

ACM Transactions on Computation Theory



ACM Transactions on Computation Theory (ToCT). This quarterly peer-reviewed journal has an emphasis on computational complexity, foundations of cryptography and other computation-based topics in theoretical computer science.

PLEASE CONTACT ACM MEMBER SERVICES TO PLACE AN ORDER
Phone: 1.800.342.6626 (U.S. and Canada)
+1.212.626.0500 (Global)
Fax: +1.212.944.1318
(Hours: 8:30am–4:30pm, Eastern Time)
Email: acmhelp@acm.org
Mail: ACM Member Services
General Post Office
PO Box 30777
New York, NY 10087-0777 USA



Association for
Computing Machinery

Advancing Computing as a Science & Profession

www.acm.org/pubs



Vinton G. Cerf

DOI: 10.1145/2447976.2447978

ACM President's Salary Increased by 300%!

I HOPE THIS headline got your attention! Of course, as with all voluntary positions at ACM, the salary is \$0.00/year. However, you gain incalculably in satisfaction when you volunteer your time in any of the myriad opportunities afforded by the international ACM. Here's what ACM Past President Alain Chesnais says on his personal home page (<http://www.alainchesnais.com>): "I have been involved as a volunteer at ACM/SIGGRAPH for over 20 years. I started out by becoming a member of the Paris SIGGRAPH chapter in 1987, then volunteered to help set up a mailing list. In 1991 I was elected chair of the chapter and was appointed to the SIGGRAPH local groups steering committee. I've continuously held positions at SIGGRAPH and ACM ever since, including ACM SIGGRAPH president from 2002 through 2005. I currently serve as ACM past president."

I want to draw your attention to Chesnais' compelling essay on volunteerism (<http://dl.acm.org/citation.cfm?id=1831408>) and to the importance and value of volunteer work for ACM. The options range from leading special interest groups and chapters, to helping edit and review publications, to service on the ACM Council or other councils sponsored by ACM. These positions represent opportunities to serve the computer science community, to engage in substantive ways with colleagues, to learn and exercise leadership, and to shape our discipline and its image in the public's mind.

A good example is USACM Council, which serves as an advisory panel on policy matters that concern computer science and the industry it has spawned. The head of the USACM Council is appointed by the President of

ACM and has been led by Eugene Spaford ("Spaf") for many years. I think Spaf would say this volunteer effort has been remarkably rewarding. USACM has given ACM a platform for expressing its views on U.S. technology policy issues vital to our discipline. It has provided opportunities for ACM staff and volunteers to contribute broadly.

I also want to emphasize how valuable it has been to have staff support at ACM to reinforce the leadership of the volunteer members. For virtually all of the ACM volunteer leaders there is a corresponding person or group within ACM HQ to support the volunteers. For example, ACM has a policy office in Washington, D.C., led by Cameron Wilson and includes David Bruggeman and Renee Dopplick. They faithfully track legislation, research activities and policy issues, and many of their notices are triggers for USACM reaction.

There are many opportunities for ACM members to join volunteer posts. Some of these posts are elective, others are appointed, and many just take personal initiative. Starting a student or local chapter or a local SIG is one example where initiative counts. I had the pleasure of serving as the chair of the Los Angeles SIGART chapter for a couple of years in the mid-1960s. There are some ACM members with lists of volunteer leadership efforts as long as my arm. Others have served in particular capacities for long periods of time. I think of Kelly Gotlieb, who has served as chair or co-chair of the ACM Awards Committee for two decades, and the late Jim Horning who served as co-chair with Gotlieb for a decade, as well as the new co-chair, Cherri Pancake.

Volunteering takes time and not everyone has a lot of it available for

such work. Still, a remarkable number of our 100,000-plus members have found the time to take on additional responsibility on behalf of the organization. I think of people like Moshe Vardi and his collection of volunteer editors and the complementary staff at ACM, like Diane Crawford, executive editor of *Communications*, who corresponds regularly and particularly vigorously, especially when I am late with my monthly contribution to our flagship publication. This points out another key factor in the volunteer/staff relationship. ACM is deeply dependent on the synergy between the full-time staff and the volunteer leaders. Anyone who has ever had experience working as a volunteer can appreciate that the sinews of an organization are a critical part of binding the volunteer muscle of the organization into a functioning whole. I can't think of better examples of that than John White, the CEO of ACM, and Pat Ryan, the COO. They and the staff of ACM keep the trains running on time and help the volunteer talent to perform their responsibilities.

If you are interested in exploring some volunteer options, please visit ACM's website (<http://www.acm.org>) where you will find information about the Key People of the organization (<http://www.acm.org/key-people>) who can connect you with volunteers or staff who will help you discover such opportunities. See <http://www.acm.org/publications> for a list of volunteer opportunities for publications. In fact, I encourage everyone reading this column to explore the ACM website—it is a gold mine of information about your organization and opportunities for volunteer work.

Vinton G. Cerf, ACM PRESIDENT

The Ultimate Online Resource for Computing Professionals & Students

ACM DL DIGITAL LIBRARY

<http://www.acm.org/dl>



Association for Computing Machinery

Advancing Computing as a Science & Profession



DOI: 10.1145/2447976.2447979 Scott E. Delman

Communications remains today one of the top-cited publications in computing.

A Few Good Reasons to Publish in *Communications*

With the recent news that ACM authors now have the option to make their new articles Open Access via the ACM Digital Library in perpetuity by paying a one-time Article

Processing Charge (APC), I started thinking about the long-term impact this new model could have on ACM, its Digital Library, and ACM's publications. This naturally led me to thinking about the relationship ACM has with the computing community itself, why that relationship is so strong, and why the "business model" is ultimately less important for ACM

than the need to continue offering valuable services to you, the reader and member of this community. As the publisher of *Communications*, I thought it might be worthwhile to dedicate this column to highlight some of that value specifically related to ACM's flagship. formation that has consistently stood the test of time, not just over the years, but since the earliest days of the field itself. In addition to the data presented in this chart, I was able to count 52 ACM A.M. Turing Award recipients and 487 ACM Fellows who have thus far graced the pages of the magazine with some of their finest work. *Communications* remains today one of the top-cited publications in computing and ranks number one in terms of citations or impact factor in several of the computer science subject areas covered by Thomson Reuters' Journal Citation Reports.

Bibliometrics: Publication history.

Publication years	1958–2013
Publication count	11,257
Citation count	142,914
Available for download	11,256
Downloads (6 weeks)	351,159
Downloads (12 months)	2,203,902
Downloads (cumulative)	14,080,989
Average downloads per article	1,250.98
Average citations per article	12.70

formation that has consistently stood the test of time, not just over the years, but since the earliest days of the field itself. In addition to the data presented in this chart, I was able to count 52 ACM A.M. Turing Award recipients and 487 ACM Fellows who have thus far graced the pages of the magazine with some of their finest work. *Communications* remains today one of the top-cited publications in computing and ranks number one in terms of citations or impact factor in several of the computer science subject areas covered by Thomson Reuters' Journal Citation Reports.

Publishing in *Communications* is both an honor and privilege that many in the community even today covet as one of the crowning achievements of their careers in the same way scientists in other fields covet a publication credit in *Nature* or *Science*. When you publish in *Communications*, you know you are in good company and you know you will be read by the largest possible audience in computing, regardless of whether your article is Open Access or accessed via the thousands of institutions subscribing to the DL worldwide. When you look closely at some of the statistics generated by the ACM DL, I am sure you will agree.

One of the incredible features of the ACM DL is the ability to aggregate a wide range of data and provide some level of analysis at the publication level. The chart here provides a very compelling answer to the questions "Why publish in *Communications*?" and "What is the value we offer?"

In addition to serving as the "community's flagship," *Communications* is the computing community's go-to source for high-quality scholarly in-

Scott E. Delman, PUBLISHER

Association for Computing Machinery

Global Reach for Global Opportunities in Computing



Dear Colleague,

Today's computing professionals are at the forefront of the technologies that drive innovation across diverse disciplines and international boundaries with increasing speed. In this environment, ACM offers advantages to computing researchers, practitioners, educators and students who are committed to self-improvement and success in their chosen fields.

ACM members benefit from a broad spectrum of state-of-the-art resources. From Special Interest Group conferences to world-class publications and peer-reviewed journals, from online lifelong learning resources to mentoring opportunities, from recognition programs to leadership opportunities, ACM helps computing professionals stay connected with academic research, emerging trends, and the technology trailblazers who are leading the way. These benefits include:

Timely access to relevant information

- *Communications of the ACM* magazine
- *ACM Queue* website for practitioners
- Option to subscribe to the *ACM Digital Library*
- ACM's *50+ journals and magazines* at member-only rates
- *TechNews*, tri-weekly email digest
- *ACM SIG conference* proceedings and discounts

Resources to enhance your career

- **ACM Tech Packs**, exclusive annotated reading lists compiled by experts
- **Learning Center** books, courses, webinars and resources for lifelong learning
- Option to join **36 Special Interest Groups (SIGs)** and **hundreds of local chapters**
- **ACM Career & Job Center** for career-enhancing benefits
- *CareerNews*, email digest
- **Recognition of achievement** through Fellows and Distinguished Member Programs

As an ACM member, you gain access to ACM's worldwide network of more than 100,000 members from nearly 200 countries. ACM's global reach includes councils in Europe, India, and China to expand high-quality member activities and initiatives. By participating in ACM's multi-faceted global resources, you have the opportunity to develop friendships and relationships with colleagues and mentors that can advance your knowledge and skills in unforeseen ways.

ACM welcomes computing professionals and students from all backgrounds, interests, and pursuits. Please take a moment to consider the value of an ACM membership for your career and for your future in the dynamic computing profession.

Sincerely,

A handwritten signature in black ink, appearing to read 'Vint Cerf'. The signature is fluid and cursive, written over a white background.

Vint Cerf

President

Association for Computing Machinery



Association for
Computing Machinery

Advancing Computing as a Science & Profession



Association for
Computing Machinery

Advancing Computing as a Science & Profession

membership application & digital library order form

Priority Code: AD13

You can join ACM in several easy ways:

Online http://www.acm.org/join	Phone +1-800-342-6626 (US & Canada) +1-212-626-0500 (Global)	Fax +1-212-944-1318
--	---	-------------------------------

Or, complete this application and return with payment via postal mail

Special rates for residents of developing countries:

<http://www.acm.org/membership/L2-3/>

Special rates for members of sister societies:

<http://www.acm.org/membership/dues.html>

Please print clearly

Name _____

Address _____

City _____ State/Province _____ Postal code/Zip _____

Country _____ E-mail address _____

Area code & Daytime phone _____ Fax _____ Member number, if applicable _____

Purposes of ACM

ACM is dedicated to:

- 1) advancing the art, science, engineering, and application of information technology
- 2) fostering the open interchange of information to serve both professionals and the public
- 3) promoting the highest professional and ethics standards

I agree with the Purposes of ACM:

Signature _____

ACM Code of Ethics:

<http://www.acm.org/about/code-of-ethics>

choose one membership option:

PROFESSIONAL MEMBERSHIP:

- ACM Professional Membership: \$99 USD
- ACM Professional Membership plus the ACM Digital Library: \$198 USD (\$99 dues + \$99 DL)
- ACM Digital Library: \$99 USD (must be an ACM member)

STUDENT MEMBERSHIP:

- ACM Student Membership: \$19 USD
- ACM Student Membership plus the ACM Digital Library: \$42 USD
- ACM Student Membership PLUS Print CACM Magazine: \$42 USD
- ACM Student Membership w/Digital Library PLUS Print CACM Magazine: \$62 USD

All new professional members will receive an ACM membership card.
For more information, please visit us at www.acm.org

Professional membership dues include \$40 toward a subscription to *Communications of the ACM*. Student membership dues include \$15 toward a subscription to *XRDS*. Member dues, subscriptions, and optional contributions are tax-deductible under certain circumstances. Please consult with your tax advisor.

RETURN COMPLETED APPLICATION TO:

Association for Computing Machinery, Inc.
General Post Office
P.O. Box 30777
New York, NY 10087-0777

Questions? E-mail us at acmhelp@acm.org
Or call +1-800-342-6626 to speak to a live representative

Satisfaction Guaranteed!

payment:

Payment must accompany application. If paying by check or money order, make payable to ACM, Inc. in US dollars or foreign currency at current exchange rate.

- | | | |
|---|--|---|
| <input type="radio"/> Visa/MasterCard | <input type="radio"/> American Express | <input type="radio"/> Check/money order |
| <input type="radio"/> Professional Member Dues (\$99 or \$198) | \$ _____ | |
| <input type="radio"/> ACM Digital Library (\$99) | \$ _____ | |
| <input type="radio"/> Student Member Dues (\$19, \$42, or \$62) | \$ _____ | |
| Total Amount Due | \$ _____ | |

Card # _____ Expiration date _____

Signature _____

Try Old Boys Security Network

PAUL HYMAN'S COMPLAINT about the lack of adequate reporting of cybercrime statistics was well justified in his news story "Cybercrime: It's Serious, But Exactly How Serious?" (Mar. 2013). All we have, he acknowledged, are lower-bound data, writing, "This much but how much more is there?" Information security is open-ended, with real but unreported losses, vulnerabilities, and threats.

Trade and professional journals tell us how to achieve security solutions, but such advice is not supported by experience because experience itself must be kept confidential. The confidentiality needed to achieve security of security greatly inhibits valid research and adequate preparation. I have for 40 years advised victim enterprises to carefully evaluate the pros and cons of publicly reporting specifics of their security experience, as revealing them would be a violation of the very concept of security; they could lose more from reporting than from keeping the information confidential. Yet they have a moral, social, and possibly legal obligation to publicly report it. An SEC advisory letter to public corporations (*SEC Disclosure Guidance: Topic No. 2*, Oct. 13, 2011, <http://www.sec.gov/divisions/corpfin/guidance/cfguidance-topic2.htm>) requires publicly reporting cybersecurity risks to shareholders but also advised not to reveal information helpful to potential adversaries. How can they carry out such a contradictory dual mandate?

Security-information-sharing organizations (such as Infraguard, <http://www.infraguard.net>) in cooperation with the FBI and the inter-industry Information Sharing and Analysis Centers (<http://www.isaccouncil.org>) are helpful to a point. I suggest also using what I call the "old boys network" of informally sharing the most sensitive security information by developing mutual trust with fellow security practitioners in other enterprises, as has been the practice for a long time in industrial security.

Donn B. Parker, Los Altos, CA

Leapfrog Open Access Toward Open Research

Open access is a transitional publishing model limited by its historical context, preserving the constraints of print media (such as tying each published piece to its original time and content) while being transitional in its embrace of wide distribution through the Internet. Though ACM's view of its own approach to open-access publishing has evolved, as reflected in Ronald F. Boisvert's and Jack W. Davidson's "Letter from ACM Publications Board Co-Chairs," "Positioning ACM for an Open Access Future" (Feb. 2013), it may well be able to leapfrog open access toward a true model of open research that permits each article to evolve over time.

Elements of such a model would probably include: offering free online public access in a structured format to all research data used to support an article, so the data can be retested and (cautiously) combined with data from other research; publishing research results in source form under an open license (such as Creative Commons's Attribution-ShareAlike; <http://creativecommons.org/licenses/by-sa/2.5/>), letting it evolve through new contributions; and encouraging authors to publish early in order to address reviews from diverse sources that could help refine claims or simply express ideas more clearly.

The second and third elements represent known challenges; the integrity of an article requires its curators prevent or filter out low-quality and irrelevant changes. However, such an open-content model conforms better to how information is really created and exchanged than the current print model or the incremental advance represented by open access.

Andy Oram, Cambridge, MA

Relational Model Alive and Well, Thank You

Carl Hewitt's letter to the editor "Relational Model Obsolete" (Jan. 2013) betrayed a shallow and too-common

level of understanding of the relational model. Of the five specific claims he made regarding "limitations" of the model, none is valid:

Inexpressiveness. This was simply wrong. Negation and disjunction are easily—in fact, almost trivially—expressible. Type generalization and specialization are easily expressible, too, though, to some extent, this is more an issue for the accompanying type system than for the relational model as such.

Inconsistency non-robustness. This one was both wrong and confused. Suffice it to say that " p AND NOT p " is an inconsistency, but "Alice says p AND Bob says NOT p " is certainly not. Moreover, even if a database really does contain an inconsistency, the relational model would still function (so we are not talking about a problem with the model); rather, the problem is with the database and with the consequent fact one cannot trust the answers given by the system. Further, a query language based on logic that encourages logical contradictions is nonsense.

Information loss. Whether one's updates "lose information" is entirely up to how one uses the model; it has nothing to do with the relational model. One of us (Date) co-authored a book¹ 100% devoted to the use of the relational model to manage temporal data and thereby not "lose information." For the record, the relational model requires no "correction," no "extension," and, above all, no perversion, for this desirable aim to be realized.

Lack of provenance. This point has to do not with the model as such but with how it is used. Note "Alice says" and "Bob says" are provenance information. In fact, the relational model is ideally suited to recording such information, and even SQL DBMSs are widely used for this purpose.

Inadequate performance and modularity. Criticizing the relational model for having no concurrency abstraction is like criticizing a cat for not being a dog. (Hewitt said it is SQL that has no concurrency abstraction, but SQL and

the relational model are not the same thing; indeed, SQL has little to do with the relational model.) As for “a... type should be an interface that does not name its implementations,” and to the extent we even understand this remark, types in the relational model meet this criterion.

We would never publish a critique of (for example) Hewitt’s actor model without understanding it well; why then does he feel he can publish a critique of the relational model, when he demonstrably does not understand it?

C.J. Date, Healdsburg, CA, and

D. McGoveran, Deerfield Beach, FL

Reference

1. Date, C.J., Darwen, H., and Lorentzos, N.A. *Temporal Data and the Relational Model*. Morgan Kaufmann, San Francisco, 2003.

Relational Model Outgrown

Unfortunately, Date and McGoveran make no good arguments against the limitations of the relational model, as outlined in my letter (Jan. 2013), partly because we are using incommensurable terminology (such as “negation,” “disjunction,” “concurrency,” and “abstraction”); for details, see articles in *Proceedings of Inconsistency Robustness 2011* (<http://robust11.org>), especially regarding the requirement for inconsistency robustness. My point is not to dismiss relational databases as a failure but to build on their success and overcome their limitations.

Such an effort must meet several requirements:

Compatibility. All capabilities of current relational databases must continue to be implemented; in addition, incrementally integrating inconsistent information from multiple relational databases with incompatible schemas (something awkward in the relational model) must become standard practice; and

Natural language + gestures (as lingua franca for communication with computer systems). Semantics of natural language and coordinated gestures of multiple participants must be expressible. An important consequence is expressions and gestures must be able to mutually refer to each other, something awkward in the relational model; for example, if Alice points her finger at Bob and charges, “I accuse you of harassing

me,” and Bob retorts, “I deny it!,” then the mutual co-reference of language and gesture must be expressible.

To move beyond the relational model, I propose the actor model because it already meets these requirements, in addition to the ones I included in my earlier letter. I further propose ActorScript¹ as a more appropriate foundation than SQL for a family of languages for information integration, as it includes the following characteristics:

Security and safety. Applications cannot directly interfere with one another.

Excellent support for concurrency.

- ▶ Not restricted to just the transactional model of concurrency, as in SQL;
- ▶ Messages directly communicated without requiring indirection through channels, mailboxes, pipes, ports, or queues;

- ▶ Integration of functional, imperative, logic, and concurrent programming; and

- ▶ Low-level mechanisms (such as threads, tasks, locks, and cores) not exposed by programs.

Language extension. ActorScript has excellent meta-language capabilities for implementing extensions without interfering with existing programs.

Capabilities for extreme performance.

- ▶ No overhead imposed on implementation of actor systems; for example, message passing has essentially the same overhead as procedure calling and looping; and

- ▶ Concurrency dynamically adapted to available resources and current load.

Relational databases have been an outstanding success and become dominant in the commercial world. However, computing has changed dramatically in the decades since the relational model was developed. Consequently, the relational model and SQL have become obsolete due to the limitations I’ve outlined here, and innovations like the actor model and ActorScript are required to address current and future needs.

Carl Hewitt, Palo Alto, CA

Reference

1. Hewitt, C. *Tutorial for ActorScript*. arXiv, Cornell University, Ithaca, NY, Mar. 2013; <http://arxiv.org/abs/1008.2748>

Communications welcomes your opinion. To submit a Letter to the Editor, please limit yourself to 500 words or less, and send to letters@cacm.acm.org.

© 2013 ACM 0001-0782/13/05

ACM’s A.M. Turing Award Recipients: Shafi Goldwasser and Silvio Micali

Incentive and Rewards in Social Media

Consequential Analysis of Complex Events on the U.S.’s Critical Infrastructure

Content Recommendation on Web Portals

Access to the Internet is a Human Right

SimPL: An Algorithm for Placing VLSI Circuits

And the latest news on deep learning programs, flexible phones, and privacy in the age of augmented reality eyewear.

The *Communications* Web site, <http://cacm.acm.org>, features more than a dozen bloggers in the **BLOG@CACM** community. In each issue of *Communications*, we'll publish selected posts or excerpts.

twitter

Follow us on Twitter at <http://twitter.com/blogCACM>

DOI:10.1145/2447976.2447981

<http://cacm.acm.org/blogs/blog-cacm>

Encouraging IT Usage In Future Healthcare, Quality in CS Education

Jeannette M. Wing considers how technology acts as a change agent for healthcare, while Mark Guzdial ponders ways to measure quality in computer science education.



Jeannette M. Wing
A Futuristic Health IT Scenario

<http://cacm.acm.org/blogs/blog-cacm/140646-a-futuristic-health-it-scenario/fulltext>

Nov. 8, 2011

The information technology trends supporting “cyber as a fifth dimension” are clear: Big Data, cell + cloud, wisdom of the crowds, co-robots, cyber-physical systems, Internet of Things, brain-machine interfaces, bio-molecular machines, nanocomputing on the one hand and exascale on the other, and quantum is still a teaser. Let’s project these trends onto a point of convergence in the future and consider the following scenario relevant to health and well being.

Imagine the day when an elderly woman in India feels ill. At birth, her genetic code had been entered into her medical record. Since birth, she has been able to record a complete history of time- and location-based

measurements of her physiological features (for example, temperature, blood pressure, height, and weight) and of her environment (for example, air and water quality, interactions with people). Ubiquitous sensor networks would collect this information. Today she might record this information using her cellphone and store it in the cloud. Today she might be illiterate but still be able to manage this information with speech input. These recordings are part of her personal medical record, which also includes past interactions with health and wellness professionals, such as diagnoses, interventions, treatments, and medical test results.

She contacts her doctor. They meet in cyberspace. Today this real-time communication could be in a virtual world through avatars or it could be through a wall-sized touch display in her home, projecting an image of the doctor. They share information visually. For example, she can demonstrate the pain she gets in moving her body

in certain ways. She can show the location and pattern of her rash. Her doctor can explain the meaning of a test result by zooming in on a medical image or by replaying a videograph. Tomorrow the doctor might be able to palpate the sore area in investigating the problem.

The doctor consults the world to help diagnose and treat her. Based on populations of people with similar genetic makeup and similar histories (including physiological paired with environment) who had similar symptoms and reported the effectiveness of their treatments, the doctor can determine the most appropriate treatment for her. A treatment for an elderly Indian woman will be based on populations more similar to her rather than on, say, middle-aged male Caucasians who grew up in the U.S. Today some of the data is already here to mine; tomorrow there will be more data, more ways to determine relevance, more ways to spot trends, and hopefully more ways to more quickly prescribe more effective treatments for the individual. We won’t need “special populations” for clinical trials anymore—we can take any subset of the world population as needed.

The doctor’s knowledge base also includes a model of the human, a multiresolution, multiscale, many-dimensional, highly parameterized computational model of a human body. It relies on exascale (or beyond) data and processing capability to simulate all systems of the human and their interactions, from the molecular level to the systems level. Any “What if?” is possible.

The doctor decides surgery is needed. The surgeon directs a robot with nanoscale or molecular-scale precision to aid in the operation, implanting an embedded programmable device for continuous monitoring or periodic drug activation. As her condition improves or as new discoveries in medical science are made, the device can be reprogrammed through unintrusive software updates.

She returns home from surgery. The doctor is able to monitor her recovery remotely and continuously. Her medical device communicates wirelessly to the cloud, adding entries to her personal medical record. Device readings are accessible by only the doctor except that alerts are also sent to emergency specialists. Her home co-robot makes sure she takes her medication. It helps fix her meals and cleans her house as she recuperates.

I probably missed the boat on some of these points, not just the computer science, but certainly the medical science. But for sure, our technology is the change agent for healthcare for the future.

I will close with two caveats. First, privacy. The medical profession upholds a principle of privacy that poses difficult technical challenges for computer scientists to tackle. How can we give the doctor access to a population's data and still preserve the privacy of the individuals in the population? How can we protect her personal medical record stored in the cloud? Privacy in healthcare is an emerging area of research in computer science. Second, ethics. As with any technology, just because we can does not mean we should. As computer scientists, we are responsible for explaining the benefits and limitations of informa-

**For sure,
our technology
is the change agent
for healthcare
for the future.**

tion technology and for participating in open debate on its ethical consequences. Technical solutions will not suffice; we will likely need new regulations and changes to social norms.



Mark Guzdial Enrollment and Quality: Does it Matter to Measure?

<http://cacm.acm.org/blogs/blog-cacm/143715-enrollment-and-quality-does-it-matter-to-measure/fulltext>

November 29, 2011

Lord Kelvin has been quoted as saying, “If you cannot measure it, you cannot improve it.” (But he also said, “There is nothing new to be discovered in physics now,” so what did he know?) In contrast, quality guru W.E. Deming wrote, “the most important figures that one needs for management are unknown or unknowable.” What can we measure in computing education, what can't we measure, and does it matter whether or not we can? I've been thinking about enrollment and quality—what can we measure, and what does it matter?

Enrollment: *Network World* declared computer science to be “the hottest major in campus” recently. Enrollment has risen dramatically at the top CS departments. But has it really risen nationally? Internationally? I recently visited Swinburne University in Melbourne, Australia, for their Melbourne Computing Education Conventicle. CS enrollments are down in the state of Victoria, and applications for next year are down 10%.

Most of what we know about CS enrollment in the U.S. we know from the Computing Research Association's Taulbee Report, which gathers data from Ph.D.-granting research institutions. There have been efforts to gather data more widely in the U.S. (called Taurus for “Taulbee for the Rest of Us”), but those have been small and not adequately funded. The U.S. Department of Education tracks undergraduate enrollment in their IPEDS database, but only for first-time and full-time students. Part-time students, and adults returning for more education, are not counted. In reality, we don't know how “hot” CS is as a major. Nobody has the broad view.

Is that a problem? Many were concerned about a lack of enrollment in

computer science. Some are now concerned about the rise in enrollment. We don't really know what the enrollment is, up or down, and maybe it doesn't really matter. We simply respond, and mostly invisible market forces will drive the students in ebbs and flows. If it is important to us (for example, to the IT industry, to those concerned about the economy), then we need to figure out a way to measure it.


Quality: I have argued in the past that we have only a few good instruments for measuring knowledge about computer science, and these aren't used often. We need these measures in order to figure out what works in computing education. I recently finished reading Richard DeMillo's new book, *From Abelard to Apple*. He talks about the challenges facing Universities today, from issues of cost, to issues of accessibility. The for-profit institutions threaten today's non-profit higher education institutions because they offer lower-cost and more flexible alternatives.

The argument is posed that the for-profits offer lower-quality offerings, that the non-profit colleges and universities offer better quality. Do they? How do we know? Rankings of colleges and universities are based on prestige and reputation, not on measures of learning outcomes. If a student wanted to choose an institution based on the one that could provide the most learning opportunities, how would she find that institution?

If learning quality matters, then we should try to measure it. But it might not matter. DeMillo recently pointed out (in a response to a blog post) that land lines offer higher quality phone calls, but cellphones won out because of the importance of flexibility and accessibility. The quality is good enough on cellphones. Does the added quality (if any, if measurable) of colleges and universities make the increased cost worthwhile? Or are all higher-education alternatives equally good enough, so choice is based on cost and accessibility? If quality matters, we should figure out how to measure it and demonstrate the value.

Jeannette M. Wing is a vice president of Microsoft and head of Microsoft Research International. **Mark Guzdial** is a professor at the Georgia Institute of Technology.

© 2013 ACM 0001-0782/13/05



Your **brain** will be **delighted**.
Both sides.

Find yourself among thousands of techno-enthusiasts as you engage in a dazzling array of mind-expanding programs, informative sessions, and blockbuster events showcasing the latest in computer graphics and interactive techniques.



SIGGRAPH2013
Left Brain + Right Brain

The **40th** International
Conference and **Exhibition**
on **Computer Graphics** and
Interactive Techniques

Conference 21–25 July 2013
Exhibition 23–25 July 2013
Anaheim Convention Center



Sponsored by ACM SIGGRAPH

www.siggraph.org/s2013



Proving Grounds

Researchers are making headway with one of quantum computing's major theoretical problems: multi-prover interactive proofs.

AS A YOUNG computer science master's student at the University of Paris, Orsay, Thomas Vidick began his first research project on the problem that has consumed him ever since: quantum entanglement.

Seven years and one Ph.D. later, Vidick—now a post-doctoral researcher at the Massachusetts Institute of Technology (MIT)—has finally published the culmination of that work. “It took us a while,” he says.

In October 2012, Vidick and co-author Tsuyoshi Ito, a researcher at NEC Labs in Princeton, N.J., revealed their long-sought-after result: demonstrating that an important theoretical security protocol, known as a multi-prover interactive proof, could work in a quantum computing environment.

The proof marks the closure of a major unsolved question in computer science—one with important implications for the study of computational complexity and for quantum computing in general.

“It’s an outstanding result,” says Professor John Watrous of the University of Waterloo, echoing the view of many theoreticians working in the quantum computing field. For the rest of us non-specialists, however, appreciating why this research matters requires a basic understanding of two conceptually



Multi-prover interactive proof systems limit the ability of independent provers to cheat.

challenging topics: interactive proof systems, and quantum mechanics.

For the past two decades, interactive proof systems have emerged as a well-established technique in modern cryptography and computer security, by providing a mechanism that allows one machine to confirm another’s identity. In the classic interactive proof, a “verifier” with limited computational abilities queries a “prover” that is assumed to have un-

limited computing power, but uncertain motives. The verifier issues a challenge to assess the prover’s trustworthiness by posing a series of questions, such as asking whether a particular formula can be satisfied. The verifier may then repeat the protocol as many times as necessary to accept or reject the proof.

Researchers have long known that systems with multiple respondents—so-called multi-prover systems—could

provide more reliable responses than single-prover ones, by forcing the verifier to solicit answers from two or more independent provers. In principle, such proofs should work more efficiently because the provers cannot communicate with each other, thus limiting their ability to cheat.

By way of analogy, consider the case of a married couple, in which one partner is a foreign citizen applying for legal residency in the U.S., while the other is an American citizen. In this scenario, an immigration officer might want to assess whether the couple is really in love or perpetrating a “green card marriage.” If the investigator interviews only one member of the couple, there is a reasonable chance that one person might get away with making misleading statements. However, if the investigator interviews both partners separately, then there is a much higher likelihood of spotting any deception—for example, by posing trick questions to catch them on inconsistencies. In much the same way, a proof involving multiple provers ought to yield more reliable results than one relying on a single prover.

In recent years, researchers have also started to explore whether interactive proof systems might work in a quantum computing environment that could, in theory, support exponentially faster algorithms.

Watrous played an important role in the study of quantum single-prover interactive proof systems, contributing to the landmark QIP=PSPACE result, which demonstrated that quantum systems operate in the same bounded space as classical proof systems. “The QIP = PSPACE result answered what was for me the biggest unanswered question about single-prover quantum interactive proof systems,” he says. “The multi-prover quantum interactive proof system model, on the other hand, is mostly wide open.”

In a quantum computing environment, multi-prover interactive proof systems come with an important wrinkle: the problem of entanglement, or what Albert Einstein once called “spooky action at a distance.” When quantum particles interact with each other, they enter a state of co-relation in which they will always behave as if connected, even if they are physically separated.

According to the laws of quantum mechanics, particles have no fixed properties until they are measured. As soon as an observer measures a particle, however, it assumes a fixed state. When quantum particles are entangled, that observation can influence the state of both particles. In theory, then, entangled provers could commingle their answers—posing a risk to the reliabil-

ity of the proof. What effect might those shared properties have on a multi-prover interactive proof? That is the problem that Ito and Vidick set out to address.

“If the provers can share entanglement, they can use it to generate outcomes that are correlated in a way that is much stronger than anything they could generate using classical means,” says Vidick. “The question is then whether they can use these correlations to cheat.”

From a quantum computing perspective, “cheating” refers not to any willful act of deception—quantum particles have no ulterior motives, after all—but rather to increasing the likelihood of winning beyond that of a classical proof.

From a mathematical point of view, establishing the effects of entanglement proved extremely challenging. Because quantum computing differs critically from classical computing in its intrinsic reliance on probabilities and interference, it requires a fundamentally new way of approaching computational problems.

“When particles are entangled, their probability distributions can’t be treated separately,” Vidick explains. “They’re really part of a single big distribution. But any mathematical description of that distribution supposes a bird’s-eye perspective that no respondent in a multi-prover proof would have. Finding a way to do justice to both the correlation be-

Milestones

Computer Science Honors, Awards

SLOAN FOUNDATION ANNOUNCES 2013 RESEARCH FELLOWS

The Alfred P. Sloan Foundation, which awards grants to support original research and broad-based education related to science, technology, and economic performance, and to improve the quality of American life, recently announced the recipients of the 2013 Sloan Research Fellowships.

These 126 early-career scientists and scholars “represent the very best that science has to offer,” according to the foundation.

Among the recipients are several who are active in computer science fields, including:

Nicholas Harvey, University of British Columbia

Bjorn Hartmann, University of California, Berkeley

Michael Lustig, University of California, Berkeley

David James Brumley, Carnegie Mellon University

Simha Sethumadhavan, Columbia University

Krzysztof Gajos, Harvard University

Derek W. Hoeim, University of Illinois, Urbana-Champaign

Svetlana Lazebnik, University of Illinois, Urbana-Champaign

Fei Sha, University of Southern California

Sachin R. Katti, Stanford University

Ryan Williams, Stanford University

Ruslan Salakhutdinov, University of Toronto

Bianca Schroeder, University of Toronto

Vinod Vaikuntanathan, University of Toronto

For a full listing of 2013 Sloan Research Fellows, see <http://www.sloan.org/sloan-research-fellowships/2013-sloan-research-fellows/>

ANITA JONES RECEIVES AAAS PHILIP HAUGE ABELSON AWARD

The American Association for the Advancement of Science recently selected Anita Jones, University Professor Emerita of Computer Science at the University of Virginia, to receive the 2012 Philip Hauge Abelson Award.

A specialist in computer security systems, Jones was honored for her scientific and technical achievements in computer science; contributions as a mentor, inspiration, and role model for other scientists and engineers; and her lifetime of public service to government, professional institutions, academia, and industry.

A member of the National Academy of Engineering (NAE), Jones’ previously has received the NAE’s Arthur M. Bueche Award, the Department of Defense Award for Distinguished Public Service, a Meritorious Civilian Award from the U.S. Air Force, and the IEEE Founders Medal.

tween the measurements and the separation of the measurers proved enormously difficult.”

Ito and Vidick’s proof relies on disguising the questioner’s intent by asking multiple questions, thus reducing the likelihood of cheating. That strategy stems from an earlier proof of the classical version of their result by Babai, Fortnow, and Lund in 1991. Their proof demonstrates that quantum entanglement would not allow multiple provers to trick the verifier with an incorrect answer.

Vidick feels the breakthrough insight was, as he puts it, to “stop trying to show that the provers will not be able to use their entanglement to cheat.” Instead, they focused their efforts on assessing the entangled provers as a whole. This process involved coming up with a complex set of tools that allowed them to analyze and perform reductions directly with entangled provers, rather than trying to extrapolate results by relating entangled provers to classical provers.

That result has proved an important conceptual breakthrough in the world of quantum interactive proofs. “A major consequence of our work is that it provides a technique to immunize proof systems against the use of entanglement,” says Vidick. “This is surprising because we know that sometimes entanglement allows for a large amount of cheating. So it wasn’t expected that any classical protocol could be generically resistant against entanglement-based cheating.”

Looking ahead, Vidick sees plenty of implications for his work in other research areas. He is particularly excited about the intersection of multi-prover interactive proofs with the world of mathematics, especially functional analysis. He points to Grothendieck’s inequality as one example of where he thinks his work could lead to deep extensions, as well as new approximation algorithms for classical problems in learning theory, such as principal component analysis.

While this proof opens up a number of promising new research avenues, the implications for practical computer science applications remain less clear.

“Even if we had quantum computers, it’s unlikely that anyone would be able to build one of these systems,”

Ito and Vidick’s proof relies on disguising the questioner’s intent by asking multiple questions, thus reducing the likelihood of cheating.

says Lance Fortnow, professor and chair of the School of Computer Science at Georgia Tech and author of *The Golden Ticket*, a newly available book on the well-known P vs. NP problem. The computational problems involved would simply be too daunting.

Fortnow believes, however, that the finding may have indirect applications for the larger world of quantum computing. “This result may give us new insights into the limits of quantum entanglement for the purpose of communication between two parties,” he says. “It (quantum entanglement) may be even less useful than we expected.”

Those implications aside, both quantum and classical interactive proof systems serve primarily as theoretical models for studying complexity theory, particularly around questions of computational efficiency and theoretical cryptography.

“One of the main reasons we study them is because, as theorists, we are drawn to models and notions that we consider to have fundamental importance from a mathematical viewpoint,” says Watrous. “The applications would be indirect.”

Vidick agrees that multi-prover interactive proofs hold interest primarily for theoretical researchers like himself, although he does see a bright future for more limited applications of quantum cryptography, such as quantum key distribution. “Quantum crypto is really much more powerful than classical crypto, and moreover typically requires only very simple quantum mechanical equipment—much easier to get than a universal quantum computer,” says Vidick.

The current generation of quantum protocols relies on single-prover interactive proofs, however—without the messy problem of entanglement. “Entanglement is hard to generate and even more to keep coherently,” says Vidick. While there are a few quantum key distribution protocols that rely on entanglement (for example, Ekert 1991), these protocols work by measuring two separated but entangled particles nearly instantaneously—currently a technical impossibility. “I would bet it will be done within 10 years,” he says, “but this doesn’t mean the use of entanglement will be practical.”

From Watrous’ perspective, the implications may be largely theoretical, but nonetheless potentially far-reaching. “We are still close to the beginning in terms of understanding this model.” ■

Further Reading

László Babai, Lance Fortnow, Carsten Lund. **Non-Deterministic Exponential Time Has Two-Prover Interactive Protocols.** *Computational Complexity*, 1, 1 (1991), 3–40.

A.K. Ekert **Quantum Cryptography Based on Bell’s Theorem.** *Physical Review Letters*, vol. 67, no. 6, 5 August 1991, pp. 661–663.

Tsuyoshi Ito and Thomas Vidick. **A multi-prover interactive proof for NEXP sound against entangled provers.** *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, September 2012. arXiv:1207.0550v2

Rahul Jain, Zhengfeng Ji, Sarvagya Upadhyay, and John Watrous. **2010. QIP = PSPACE.** In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC ’10)*. ACM, New York, NY, USA, 573–582. DOI=10.1145/1806689.1806768 <http://doi.acm.org/10.1145/1806689.1806768>

Julia Kempe, Hirofumi Kobayashi, Keiji Matsumoto, Ben Toner, and Thomas Vidick. **2011. Entangled Games Are Hard to Approximate.** *SIAM J. Comput.* 40, 3 (June 2011), 848–877. DOI=10.1137/090751293 <http://dx.doi.org/10.1137/090751293>

Assaf Naor, Oded Regev, Thomas Vidick. **2012. Efficient Rounding for the Noncommutative Grothendieck Inequality, to appear in Proceedings of the 45th ACM Symposium on Theory of Computing (STOC ’13).** arXiv:1210.7656v1

Alex Wright is a writer and information architect based in Brooklyn, NY.

Vanishing Electronics

Engineers are reinventing electronics by building safe devices that dissolve in the body or within the environment. The technology could redefine everything from medicine to computing.

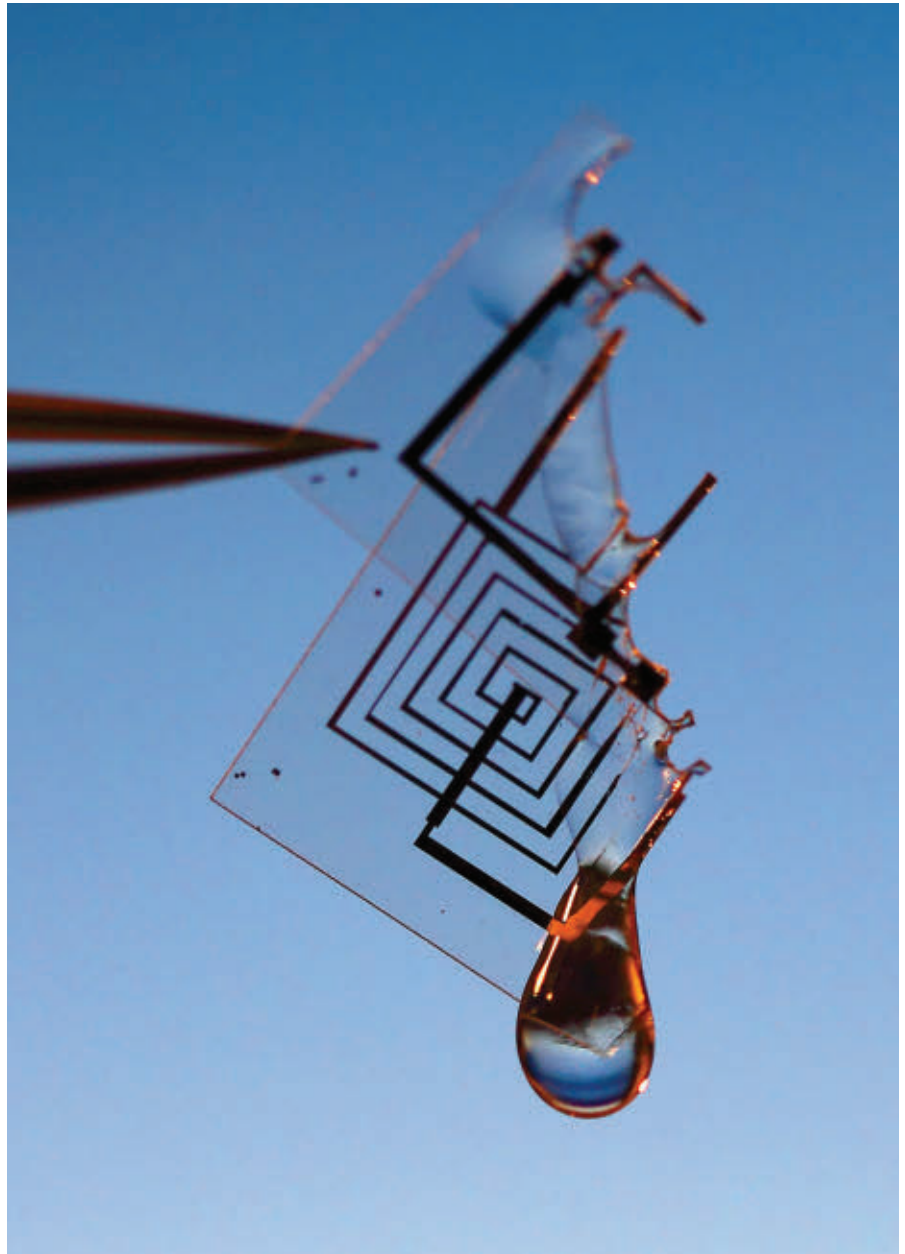
SINCE THE DAWN of the digital age, engineers and designers have struggled to make electronics more dependable. They have constructed microchips, circuit boards, and processors to better withstand the forces of nature—including dust, oxidation, and exposure to various other gasses and substances in the atmosphere. They have also experimented with materials that conduct electricity more effectively and fail less regularly.

Yet times, and electronics, change. Peer into the engineering lab at the University of Illinois Urbana-Champaign and you will view a new generation of electronic devices that will fundamentally change medicine, environmental science, and perhaps spying and warfare. John Rogers, a professor of material science and engineering, is constructing programmable devices that are designed to disintegrate over a period of weeks or months.

Within a few years, so-called transient electronic devices could deliver medication in a highly targeted way, monitor chemical and oil spills, and engage in clandestine tracking and listening for the military. “The disappearance of the device can occur at a physical level through dissolution or sublimation or it can take place at a chemical level as a result of a reaction event that’s built into the system,” Rogers says. “Transient electronics deliver a completely different design model that can be used in entirely different ways.”

Instability Matters

The origins of today’s electronic systems date back to the 1930s. That is when Austrian engineer Paul Eisler developed the first printed circuit boards. These systems—which allowed engineers to mass-produce circuits with electrical interconnects



A water droplet dissolves electronic components of a transient electronic device: transistors, diodes, inductors, capacitors and resistors, all on a thin silk substrate.

between specific components—were initially deployed by the military and used with proximity fuses that would detonate a bomb when it reached a specified distance from a target. By

the 1950s, circuit boards and other increasingly sophisticated electronic systems began to appear in televisions, radios, telephones, and various industrial machines.

Today, virtually every electronic device—from computer mice to microwave ovens and ceiling fans—incorporates integrated circuits, microchips, and other components. These machines can perform in far more complex ways and deliver much higher reliability thanks to these electronic brains. As a result, integrated circuits—typically constructed from silicon, laminates, metals, and resins—have become the building blocks for the digital age. They allow engineers to build an array of devices with computer-like capabilities.

Now, researchers and engineers are turning the concept upside down; they are reexamining the fundamental idea of building electronic systems designed for the long haul. “There is an expectation that if you buy a cellular phone it will last two or three years and if you buy a television it will work for at least 10 years,” states Yonggang Huang, a professor of mechanical engineering at Northwestern University in Chicago, Illinois. “The goal with transient electronics is to build a system with planned obsolescence.”

Consequently, researchers are now eyeing new materials, manufacturing schemes, device components, and theoretical design tools that support transient electronic systems. This includes ultrathin sheets of silicon and materials like silk, along with transient transistors, diodes, wireless power coils, temperature and strain sensors, photodetectors, solar cells, radio oscillators, antennas, and simple digital cameras that work inside tiny programmable and biocompatible devices. Rogers says these devices—manufactured from dissolvable magnesium and magnesium oxide (the latter substance is available as a vitamin tablet)—could eventually be the size of a grain of sand but possess the same attributes of a conventional microprocessor or the brains of an RFID tag. They would use inductive and radio frequency (RF) wireless technologies with dissolvable single-use batteries to transmit electrical signals.

Depending on the specific design and the way the device is encapsulated into silk, it could react to various chemicals, compounds, temperature, or pH in order to dissolve within minutes or last days, weeks, or potentially years.

“These devices offer robust performance using a bare minimum of electronics. They are able to dissolve or be reabsorbed within a specific timeline.”

“The technology is the exact opposite of conventional electronics systems, including integrated circuits, that are designed for long-term physical and electronic stability,” explains Fiorenzoomenetto, professor of biomedical engineering at Tufts University School of Engineering in Medford, MA. “These devices offer robust performance using a bare minimum of electronics. They are able to dissolve or be reabsorbed within a specific timeline.” Researchers are also exploring the use of bio-plastics and other fully degradable materials, he notes.

In fact, actual transient electronics devices are beginning to take shape. So far, Rogers, Huang, and Omenetto have teamed up to develop several functioning systems. One of them is a basic thermal device designed to monitor and prevent post-surgical infection. The unit, which measures only a few tens of nanometers in thickness and includes a 64-pixel digital camera, has been successfully tested on rats. The silk-based device—constructed from tiny circuits, including transistors and interconnects—dissolves in the body and is absorbed into tissue a couple of weeks after being exposed to internal fluids.

Power Plays

The uses for transient electronics are broad—and new possibilities will emerge as the technology takes hold. For example, these devices could usher in a new era of medical monitors—including devices that better measure specific heart or kidney performance, for example, and report back to doctors. They could also offer interven-

Milestones

Five Receive First Queen Elizabeth Prize

The first Queen Elizabeth Prize for Engineering, a global award given in recognition and celebration of “outstanding advances in engineering that have changed the world,” has been awarded jointly to five computing pioneers for their groundbreaking work that led to the Internet and World Wide Web.

The recipients, who will share a prize of 1 million pounds (\$1.5 million), are Louis Pouzin, Robert Kahn, Vinton Cerf, Tim Berners-Lee, and Marc Andreessen.

The Royal Academy of Engineering, which administers the Queen Elizabeth Prize for Engineering, explained that Pouzin, Kahn, and Cerf were chosen for having made “seminal contributions to the protocols (or standards) that together make up the fundamental architecture of the Internet.” Berners-Lee was selected for his creation of the World Wide Web (WWW), “which vastly extended the use of the Internet beyond email and file transfer,” while Andreessen was included because he “wrote the Mosaic browser that was widely distributed and which made the WWW accessible to everyone. His work triggered a huge number of applications unimagined by the early network pioneers.”

Kahn and Cerf jointly received the ACM A.M. Turing Award in 2004. Cerf is currently president of ACM.

The prize originally was envisioned as rewarding and celebrating an individual, or up to three individuals; the rules had to be changed to allow all five men to be included.

The five winners will meet Queen Elizabeth and receive their prizes in London in June.

—Larry Fisher

tional components or therapeutic devices that dispense optimal levels of medication to a specific location within the body. An implantable biocompatible device could be used to administer chemotherapy drugs or treat chronic pain.

In fact, this highly targeted approach to treatment could alleviate the side effects that frequently occur with conventional drugs, especially chemotherapy drugs that affect healthy tissue. Because these devices would monitor or micro-monitor specific areas of the body, they could adjust doses in real-time while providing a stream of data for doctors and other health professionals. In the end, physicians could address problems more promptly and understand how specific medications impact a condition.

But the possibilities for transient electronics do not stop there. Researchers are also developing environmental monitors that could be dropped at the site of chemical or oil spill or into the atmosphere in order to create an instant monitoring network. Scientists could monitor water or air dispersal patterns in order to obtain a more accurate assessment of how different approaches and tools combat the problem. These sensors would dissolve or compost when exposed to water, soil, or air. “They fall into the category of “eco-resorbable,” Huang says. “This eliminates the need for collection and recovery or any associated environmental issues.”

Finally, transient electronics could be used by the military for monitoring and spying. The U.S. Defense Advanced Research Projects Agency (DARPA), which currently funds research in this field, is exploring ways to use devices that would record video and audio but then disappear, so that there is no record of their existence. Likewise, transient electronics could allow the military or spies to dispense a toxin or medication without being detected. Says Rogers: “The ability to create devices that can disappear without a trace could redefine military and security applications.”

Switching on the Future

Although researchers are in the early stages of developing functional transient electronic devices, Rogers believes there is a little doubt they will play a ma-

One thing is certain: over the next decade, the field will continue to advance and transient electronics will unlock new opportunities and possibilities.

major role in defining the future of medicine and other fields. What’s more, as technology continues to advance and the ability to program smaller and smaller devices takes hold, the possibilities are likely to extend to new and different fields. For now, researchers are exploring the use of different materials. For instance, devices constructed with magnesium might also assist with the intake of minerals and vitamins that are essential for a healthy body.

Eventually, Rogers believes that transient electronics could replace conventional microprocessor and electronics designs. “At a certain point, you have to ask whether we might be able to build computers and cellphones with a specific lifespan and an ability to disintegrate into the environment. The ability to reduce overall waste and mitigate the hazardous waste stream could fundamentally change the way we build, consume, and discard an array of devices,” he explains.

Omenetto says that in addition to examining different combinations of transient materials, researchers are studying how they can assemble and combine all the different elements optimally. A key is executing software code in a way that provides the maximum possible results. “There is a huge focus on reliability and issues such as mean-time to failure. Obviously, if the device is used for medical purposes there are issues relating to approval by the Federal Drug Administration (FDA) and its counterparts in other countries.”

One thing is certain: over the next decade, the field will continue to ad-

vance and transient electronics will unlock new opportunities and possibilities. Although research is still in the early stages, it will gain momentum as multidisciplinary teams build rapidly on learning, Rogers says. “The actual electronics required for these systems is relatively modest. The challenges largely revolve around building systems that optimize the technology and are conducive to large-scale mass-production at a relatively low cost.”

In fact, Rogers expects to see an array of actual devices within a few years. “The technology is advancing rapidly. This is an area that offers incredible opportunities in both the medical arena and elsewhere,” he concludes. The ability to build transient electronic devices could fundamentally change medicine, environmental sciences and other fields. “It represents a fundamentally different approach to building electronic systems. It is an important part of the future of medicine, environmental sciences and other fields.” **□**

Further Reading

Hwang, S., Tao, H., Kim, D., Cheng, H., Song, J., Rill, E., Brenckle, M., Panilaitis, B., Won, S., Kim, Y., Song, Y., Yu, K., Ameen, A., Li, R., Su, Y., Yang, M., Kaplan D.L., Zakin, M.R., Slepian, M.J., Huang, Y., Omenetto, F.G., Rogers, J.A.

A Physically Transient Form of Silicon Electronics, *Science* 28, September 2012, Vol. 337 no. 6102 pp. 1640-1644 DOI: 10.1126/science.1226325 <http://www.sciencemag.org/content/337/6102/1640.abstract>

Kim, D., Viventi, J., Amsden, J.J., Xiao, J., Vigeland, L., Kim, Y., Blanco, J.A., Panilaitis, B., Frechette, E.S., Contreras, D., Kaplan, D.L., Omenetto, F.G., Huang, Y., Hwang, K., Zakin, M.R., Litt, B., Rogers, J.A.

Dissolvable Films of Silk Fibroin for Ultrathin Conformal Bio-integrated Electronics, *Nature Materials* 9, 511–517 (2010) doi:10.1038/nmat2745, April 2010. <http://www.nature.com/nmat/journal/v9/n6/full/nmat2745.html>

Sridharamurthy, S.S., Agarwal, A.K., Beebe, D.J., Jiang, H.

Dissolvable membranes as sensing elements for microfluidics based biological/chemical sensors. *Lab on a Chip, Issue 7, 2006. P. 840-842.*

Samuel Greengard is an author and journalist based in West Linn, OR.

‘Small Data’ Enabled Prediction Of Obama’s Win, Say Economists

“Big data” from crowdsourcing resulted in more complex predictions.

NINE MONTHS BEFORE the 2012 U.S. presidential election last November, economists knew the election’s outcome. They attribute their forecasting success not to social network-sourced “big data” as some in the media have suggested, but to “small data” they say could have just as accurately been figured with a pencil and paper as with a computer.

The electoral process in the U.S. is somewhat complex, which complicates the process of forecasting its outcome. The president and vice president are not elected directly by the voters; instead, the Electoral College, an institution made up of “electors” who are chosen by popular vote on a state-by-state basis, elects the nation’s top executives every four years. Each of the 50 U.S. states is entitled to as many electors as it has members of Congress (which includes two senators and a number of representatives proportional to each state’s population), and those electors are supposed to cast their votes for the winner of the popular vote in each state.

With all that in mind, economists say the 2012 presidential election was just too early to have relied on input from unproven data from sources like Facebook, Twitter, and search results. Instead, they used social media input to make more complex combination predictions—just as they anticipate doing for the next presidential election in 2016.

How data was used to predict the election’s results depended entirely on the forecasting methodology chosen, four basic types of which were employed by economists in 2012, says Justin Wolfers, a professor of economics and public policy at the University of Michigan.

He describes them this way:

► **The fundamentals method.** This



includes such factors as GDP growth, unemployment rate, and whether the candidate is an incumbent. “The model is a very simplistic, stripped-down one,” says Wolfers, “and that is its virtue. Because we only get one presidential election every four years, we have a limited number of observations from which to draw. A more complicated model would require additional observations we just don’t have.”

► **The polls method.** Two broad types include looking at a popular poll like the Gallup Poll, while a more sophisticated approach aggregates results from various pollsters. Statistician Nate Silver employed this latter method to correctly predict the winner in all 50 states and the District of Columbia. “It involves looking at every single poll,

thinking about each poll’s biases, and realizing that the average of many polls will do better than any individual poll,” says Wolfers. “Many regard this method as state of the art.”

► **The prediction market method.** Economists watch speculative markets in which current market prices can be interpreted as predictions of the election outcome.

► **The hybrid of polls and prediction market method.** “The latest twist in election forecasting in which, instead of asking people who they intend to vote for, they are asked who they think will win,” says Wolfers. “This method of polling—which, like a prediction market, aggregates the latent wisdom that exists in the broader population—worked so spectacularly well in 2012

that I expect it will play a much more prominent role in 2016.”

In Wolfers’ opinion, a year in advance of the election, the fundamentals approach works well while polls do not, because people have not started thinking about the election yet. Polls do a good job three months before the election, he says, but prediction markets do the best job regardless of when they are employed.

None of these methods involve what is commonly known as “big data,” says Patrick Hummel, currently a research scientist at Google who developed a model for forecasting elections with David Rothschild, an economist at Microsoft Research and a Fellow of the Applied Statistics Center at Columbia University, during their time at Yahoo! Research.

Hummel describes the way they utilized data in his 2012 presidential prediction as simple linear regression, first gathering from earlier elections historical data like economic indicators, presidential approval ratings, which party was in power and for how long, and biographical information about the candidates. Then, he and Rothschild compared how various pieces of data that were available nine months before the 2012 election correlated with the results of the earlier elections.

In February 2012, they predicted President Obama would win the Electoral College with 303 electoral votes to Romney’s 235, forecasting every state correctly except for Florida, where they predicted Obama would lose (in fact, Obama won Florida with

Hummel describes the way they utilized data in his 2012 presidential prediction as simple linear regression.

just 50.01% of the vote).

Hummel and Rothschild also accurately predicted the vote shares that President Obama would receive in all 50 states and, after the election, determined their median error in that prediction was 1.6 points.

“We are aware of 23 different polling organizations that made predictions of statewide vote shares in various states the day before the election,” Hummel says, “and of those 23, there was only one that ended up with an average error that was less than 1.6 points.”

Hummel and Rothschild’s dataset included hundreds of historical elections—the outcomes in 50 states for every year for the last several decades—that totaled approximately 100,000 unique pieces of data.

“I wouldn’t classify that as big data,” Hummel says, “which can involve as many as tens of billions of data points in one analysis. Our particular analysis, which could be done

with pencil and paper, doesn’t come anywhere close to that.”

While “big data” might not have been appropriate for predicting the presidency in 2012, it was what was needed for making complex, combination predictions, says David Pennock, principal researcher and assistant managing director at Microsoft Research New York City.

“If you’re just trying to predict the national winner, the small data model with just a few features—like economic variables, approval ratings, and so on—is the right way to go,” Pennock explains. “But you really do need computer science when you’re dealing with computations between things—like the chance that the same party wins both Ohio and Pennsylvania, or the chance that whoever wins Ohio will win the whole country. That’s when you’ve got not just 50 things to predict, but two to the 50th, which is something like one quadrillion, if you count up all the combinations.”

The 2012 election was the first in which economists were able to do complex combination predictions, due not only to the increase in available computational power, but also to algorithms that were developed just recently.

“For instance, we can ask what’s the likelihood of gas prices going up if Obama wins—or if Romney wins,” Pennock says. “And what is the likelihood of taxes rising. Or, if Candidate A wins, is it likely we’ll be involved in more wars ... or that the stock market will drop. These are the kinds of interesting predictions that help voters determine how the election’s outcome

At Press Time

Authors Accept ACM’s OA Options



ACM officially rolled out its anticipated publication

policy changes aimed to expand access to its magazines, journals, and conference proceedings on April 2. Reaction to these changes was evident within days of their debut—at press time, the first authors of recent manuscript submissions had chosen one of the new options to manage the

publication rights of their work.

As detailed by ACM Publications Board co-chairs Ronald F. Boisvert and Jack W. Davidson in their February 2013 editorial (<http://cacm.acm.org/magazines/2013/2/160170-positioning-acm-for-an-open-access-future/fulltext>), ACM authors now have three ways to manage their publication rights:

1. Authors who want ACM to manage the rights associated with their work

can use the standard ACM Copyright Transfer Agreement.

2. Authors who want to retain intellectual property rights for their work may choose an equivalent exclusive licensing agreement, which provides ACM with certain publication and distribution rights.

3. Authors may retain all rights to their work while making it openly accessible through the ACM Digital Library via an author-pays option.

ACM authors continue to

hold exclusive ownership of their patents and trademarks, as well as reuse rights for any portion of their own work without fee or permission from ACM. Major revisions created by an author continue to be owned by the author, and each author holds self-archiving rights for accepted versions of his/her own work in personal bibliographies and institutional repositories. For more information, see <http://authors.acm.org> as well as the back cover of this issue.

will affect them personally.”

Pennock admits being skeptical about the ability of social media to predict elections, especially when there are better ways to forecast a winner. Instead, he—and other economists—used social media during the 2012 election to understand more subjective information, like people’s sentiments or reactions—the sort of information, he says, that cannot easily be obtained from other sources.

That is why Rothschild is gathering “tons of data from social media like Twitter and Facebook and search results” that he is not applying immediately, but intends to use in forthcoming research in 2013 and 2014.

“I just didn’t have the confidence yet to use social media to answer in any meaningful way the questions people are asking,” he says. “It’s just too new and too complicated and there isn’t much of a historical track record for it.”

Rothschild says he is excited about the promise of the new social media data that, come the next presidential election, he foresees economists will be able to use to answer very complex questions—and to provide the results in real time.

“The current data doesn’t have the granularity to allow us to do that yet,” he says. “But the direction we’re moving in is to determine, say, what are the five things people want to know most—and then to be able to provide answers. Or, perhaps build a model where each person can input their own specific questions and then output the answers. That’s the promise of social media.”

Currently, Rothschild is working with teams at both Bing (Microsoft’s Web search engine) and Xbox 360 (Microsoft’s video-game console) to prepare for what he calls “the next generation of actively collecting data to answer questions.”

“People have been doing telephone polling for years,” he says, “but that is becoming more and more expensive, especially since many people don’t have standard phones anymore and tend not to answer their phones when pollsters call.”

Indeed, poll response rates are down from over 40% 20 years ago to less than 10% today.

Instead, he says, Xbox has a huge audience of engaged users who are

eager to supply information—and so he has been working on new methods of polling them, especially by making the process more enjoyable for them to participate.

In late September, he and his team launched a test run, averaging 20,000 interviews of five questions daily on Xbox, especially during the presidential debates.

“What we concluded was that we can make the product engaging so that people want to be in it and supply us information,” he says, “and that we can make the results meaningful, even today in 2012.”

Yet the real promise for forecasting in both the 2016 and 2020 Presidential elections, says Rothschild, is being able to answer new questions in a way that is both quick and, perhaps, more personalized.

Indeed, agrees Pennock, “there’ll be more complex predictions, since that seems to have captured the public’s imagination. We won’t just be talking about the horse race and who will win, but what will happen to each voter regardless of who wins. Which is actually more relevant, I think.” **C**

Further Reading

Rothschild, D. and Wilson C. Obama poised to win 2012 election with 303 electoral votes: The Signal Forecast, February 16, 2012; <http://news.yahoo.com/blogs/signal/obama-poised-win-2012-election-303-electoral-votes-202543583.html#OWqyiSA>

Rothschild, D. and Wolfers, J. Forecasting Elections: Voter Intentions versus Expectations, November 1, 2012; <http://www.brookings.edu/research/papers/2012/11/01-voter-expectations-wolfers>

Pennock, D. A toast to the number 303: A redemptive election night for science, and The Signal, November 10, 2012; <http://blog.odhead.com/2012/11/10/signal-redeemed/>

Hummel, P. and Rothschild, D. Fundamental Models for Forecasting Elections, http://researchdmr.com/HummelRothschild_FundamentalModel.pdf

Nate Silver: Forecasting the 2012 Election, April 12, 2012; <https://www.youtube.com/watch?v=P9dyDZsPPOE>

Pres. Obama on Top in First Presidential Election Forecast, July 13, 2012; https://www.youtube.com/watch?v=t_I2VpS2JMY

Paul Hyman is a science and technology writer based in Great Neck, NY.

© 2013 ACM 0001-0782/13/05

ACM Member News

ROBERT N.M. WATSON SEES CHALLENGES TO SECURITY IN FOCUS OF DESIGN TEAMS



Designing networks and computer systems that can provide robust security controls

continues to be a major challenge for developers. The way in which new security protocols are developed, and the teams that design them, must be altered in order to meet this challenge, says Robert N. M. Watson, Senior Research Associate in the Security Research Group at the University of Cambridge Computer Laboratory, and a Research Fellow at St. John’s College Cambridge.

Named Lecturer at Cambridge University in March, Watson spends much of his time looking at the ways in which computing systems have been designed, and has found that much of the work on security protocols has been segmented by discipline. As a result, Watson says, innovation is stifled because researchers in one discipline usually will only alter variables in their own field of expertise (such as hardware) to solve a problem, without considering alterations in other systems (such as software).

“The problem with this is that you’re exploring subsets of an overall solutions space, where you’re changing only one significant variable at a time,” Watson says. “We are interested in multivariable research, to really allow us to break sets of assumptions.”

Watson is assembling a team of researchers that spans a variety of disciplines, including hardware and software designers, networking specialists, and security researchers, with the goal of fostering a cross-disciplinary approach to solving large problems. He notes, “it’s not necessarily a set of people that would meet up and talk to each other outside of this context, but now we’re able to bring on projects that require a very broad view.”

—Keith Kirkpatrick



Law and Technology

Fair Use in Europe

Examining the mismatch between copyright law and technology-influenced evolving social norms in the European Union.

LIKE THE EURO, the law of copyright in the European Union seems to be in a state of perennial crisis. Copyright owners complain the law has left them defenseless against mass-scale infringement over digital networks, and call for enhanced copyright enforcement mechanisms. Users and consumers accuse the copyright industries of abusing copyright as an instrument to conserve monopoly power and outdated business models. Authors protest that the law does little to protect their claims to receive compensation—from the copyright industries and the users of their works alike.

A major cause of this crisis in copyright is the increasing gap between the rules of the law and the social norms that are shaped, at least in part, by the state of technology. Of course, technological development has *always* outpaced the process of lawmaking, but with the rapid and spectacular advances in information technology of recent years the law-norm gap in copyright has become so wide the system is now almost at a breaking point.

All this may look very familiar to U.S. readers well versed in the ongoing Great American Copyright De-

bate. However, in Europe the situation is much more complex, for at least two reasons. One is the intricacy of the EU lawmaking machinery, which requires up to 10 years for a harmonization directive to be adopted or revised. The other is the general lack of flexibility in the laws of copyright in the EU and its Member States, which—unlike the U.S.—do not permit “fair use” and thus allow little leeway for new technological uses not foreseen by the legislature.

As a consequence there is an increasing mismatch in the EU between

The EU legal framework leaves Member States little room to update or expand existing limitations and exceptions.

the copyright law and emerging social norms. For example, the law in many Member States fails to take account of current educational and scholarly practices, such as the use of copyright-protected content in PowerPoint presentations, in digital classrooms, on Blackboard sites, or in scholarly correspondence. By the same token, many European laws severely restrict the use of (parts of) copyright works for purposes such as data mining or documentary filmmaking. By obstructing these and other uses that many believe should remain outside the reach of copyright protection—and would likely be called *fair use* in the U.S.—the law in Europe impedes not only innovation, science, and cultural progress, but also undermines the social legitimacy of copyright law. Arguably, this ever-widening copyright law-norm gap is an important factor in explaining why illegal sharing of copyright content is more widespread in Europe than it is in the U.S.

Like in most countries of the world, copyright laws in the EU traditionally provide for “closed lists” of limitations and exceptions that enumerate the various uses that are permitted without authorization. Examples of such uses



are: private copying, quotation, parody, library archiving, classroom uses, and reporting by the news media. Statutory exceptions are usually detailed and connected to specific states of technology, and therefore easily outdated. To make matters worse, the EU legal framework leaves Member States little room to update or expand existing limitations and exceptions. The EU's Copyright in the Information Society Directive of 2001 lists 21 limitations and exceptions that Member States may provide for in their national laws, but does not allow exceptions beyond this "shopping list."

Search

A case in point, and a major cause of contention, is *search*. While operating a search engine in the U.S. would generally be considered fair use,^a courts across Europe are struggling to

accommodate information location tools in copyright laws that provide for extensive rights of reproduction but only narrow and outdated exceptions. This should come as no surprise given the widespread copying of copyright material that most search engine providers routinely engage in. Consider, for example, Google's web-crawler that makes digital snapshots of most content available on the Web at any time, or the Google cache that archives these copies for indexing purposes. For another example, consider the myriad thumbnails image search engines commonly return to users in response to queries. Dealing with these unauthorized copies in European legal systems that do not generally allow fair use is becoming increasingly problematic.

Court decisions in the Member States of the EU illustrate the current state of confusion regarding the copyright status of search. In a case brought against Google by Belgian newspaper publishers the Brussels Court of Appeals held that Google squarely infring-

es the rights of copyright owners in content cached by Google.^b By contrast, the Spanish Supreme Court held that the unauthorized copies in the Google cache were merely *de minimis*, and did not amount to infringement. In a case involving Google Image Search the German Federal Supreme Court took a middle-ground position by holding, on the one hand, that Google's unauthorized use of (thumbnail) images is not exempted by any existing statutory limitation. On the other hand, any author that makes his content available on the Web without blocking web crawlers, is deemed to have consented to the use of his content by search services.^c Other search-related cases decided by courts in France, Austria, and other countries point in yet other directions.⁷ Eventually, the European Court of Justice in Luxembourg will have the final say on

^b *Google Inc. v. Copiepresse*, Court of Appeal Brussels, May 5, 2011.

^c German Federal Supreme Court (*Bundesgerichtshof*), Judgment of April 29, 2010, Case I ZR 69/08, available in German at <http://www.bundesgerichtshof.de>.

^a See *Kelly v. Arriba Soft*, 336 F.3d 811 (9th Cir. 2003); *Field v. Google*, 412 F. Supp. 2d 1106 (D. Nev. 2006); *Perfect 10, Inc. v. Amazon, Inc.*, 487 F.3d 701 (9th Cir. 2007).

these questions. Will the Court allow search without permission? Or will the World Wide Web in Europe be searchable only on condition of a billion (or so) licenses?

To further complicate matters in Europe, the German government last year proposed special legislation that would require search engines and other online news aggregators to seek licenses from newspaper publishers for linking to news items.² Although the German parliament has recently watered down the bill to allow search engines to show news snippets, the bill has already set a dangerous precedent in other Member States, such as France, where newspaper publishers feel equally threatened by Google News and similar services.¹

User-Generated Content

Another area where a need for flexibility in copyright is evident is user-generated content. Whereas the social media have in recent years become essential tools of social and cultural communication, copyright law in most EU Member States leaves little or no room for sharing user-generated content that builds upon preexisting works. For example, a spoof video composed of materials taken from broadcast television and uploaded to YouTube or Facebook would be exempted only in the rare case that it would qualify as a quotation providing critical commentary, or as a parody or pastiche. As the European Commission already recognized in its 2008 *Green Paper on Copyright in the Knowledge Economy*, the absence in European copyright laws of an exemption permitting user-generated content “can be perceived as a barrier to innovation in that it blocks new, potentially valuable works from being disseminated.”⁴ The Commission’s suggestion to introduce a special user-generated content exception in EU copyright law has however as yet not materialized.

Good News

The good news is that the idea of introducing a measure of flexibility in the European system of copyright limitations and exceptions is now gradually taking shape. The Dutch government has in recent years repeatedly stated

Another area where a need for flexibility in copyright is evident is user-generated content.

its commitment to initiate a discussion at the European political level on a European-style fair use rule. In the U.K., the Hargreaves Report, a forward-looking government-commissioned study on copyright reform published in May 2011, recommends the U.K. government argue in Brussels for “an additional exception, designed to enable EU copyright law to accommodate future technological change where it does not threaten copyright owners.”⁵ The U.K. government’s official response to the *Review*⁶ highlights the need for more flexibility in EU copyright law. Most recently in Ireland the Copyright Review Committee advised the Irish government to consider the introduction of a general fair use rule, which would make Ireland part of a growing number of states—including Israel and Singapore—adopting the American model.³

Toward a Semi-Open Norm?

Clearly, the time is ripe for a critical review of the EU’s closed list of permitted limitations and exceptions to copyright. The Information Directive of 2001 that sought to deal with the early copyright challenges of the digital environment, is now well over 10 years old, but has never been properly reviewed by the European Commission. Opening up the Directive’s closed list to allow other fair uses that promote innovation and cultural development should feature high on the European Commission’s legislative agenda for the near future. A straightforward way to do this would be by allowing Member States to provide for *other* (that is, not specifically enumerated) limitations and exceptions permitting unauthorized uses, on the

condition these uses comply with the so-called three-step test. The three-step test, which is part of the WTO’s TRIPS Agreement and other international treaties that are binding upon the EU, is already incorporated in the Directive as an overarching norm preventing Member States from introducing overbroad copyright limitations. The test requires that exceptions: apply only in certain special cases; not conflict with the normal exploitation of copyright works; and not otherwise unreasonably prejudice the interests of rights holders. By combining the present system of circumscribed exceptions with an open norm that would allow other fair uses, a revised Directive would better serve the combined goals of copyright harmonization and promotion of culture and innovation.

A good example of such a semi-open norm can be found in the *European Copyright Code* that was released by a group of leading European copyright scholars (the Wittem Group) in 2010. If the EU legislature wishes to infuse a measure of flexibility and fair use in its currently defunct copyright system, it needs to look no further.⁸ **C**

References

1. A clash across Europe over the value of a click. *The New York Times* (Oct. 30, 2012); http://www.nytimes.com/2012/10/31/technology/european-newspapers-seeking-a-piece-of-google-ad-revenue.html?pagewanted=all&_r=0.
2. Assault on Google News: Berlin cabinet approves new Web copyright law. *Spiegel Online International* (Aug. 30, 2012); <http://www.spiegel.de/international/germany/german-lawmakers-propose-charging-fees-to-aggregators-like-google-a-852965.html>.
3. Copyright Review Committee. Copyright and innovation. A consultation paper. Dublin, 2012.
4. European Commission. *Green Paper on Copyright in the Knowledge Economy*. Brussels. COM(2008) 466/3 (16.07.2008), 19–20.
5. Hargreaves, I. *Digital Opportunity. A Review of Intellectual Property and Growth*. (May 2011), 5.
6. U.K. government response to the *Hargreaves Review of Intellectual Property and Growth* (Aug. 2011); <http://www.ipa.gov.uk/ipresponse-full.pdf>.
7. van der Noll, R. et al. *Flexible Copyright: The Law and Economics of Introducing an Open Norm in the Netherlands*. Study commissioned by the Dutch Ministry of Economic Affairs, Agriculture, and Innovation. SEO-rapport nr. 2012-60, Amsterdam, (Aug. 2012); http://www.ivir.nl/publications/vangompel/Flexible_Copyright.pdf.
8. Wittem Group. *European Copyright Code*; <http://www.copyrightcode.eu>

P. Bernt Hugenholtz (hughenoltz@uva.nl) is director of the Institute for Information Law at the University of Amsterdam (<http://www.ivir.nl/staff/hughenoltz.html>).

This Viewpoint is based on a study by P. Bernt Hugenholtz and Martin R.F. Senftleben, “Fair Use in Europe. In Search of Flexibilities”; <http://ssrn.com/abstract=1959554>.

Copyright held by author.

Historical Reflections

Max Newman: Forgotten Man of Early British Computing

Reflections on a significant, yet often overlooked, computing pioneer.

UNLESS YOU SPENT last year living in a cave, it was more or less impossible not to know 2012 was the Turing Centenary year. While it was gratifying to see so many exciting events arranged in Turing's honor, I could not help thinking we had moved more or less seamlessly from a position where Turing was largely ignored, to one where it sometimes felt as if no one else counted. In his editorial for the January 2013 *Communications*,⁵ Moshe Vardi similarly gives voice to a very reasonable concern that "we may have gone from celebration to hagiography."

Maxwell (Max) Herman Alexander Newman, whose own centenary passed relatively quietly in 1997,^a was closely associated with Turing,³ and from the mid-1930s onward played an important part in promoting and shaping Turing's career and promulgating as well as implementing his ideas on computing.^b

Despite being largely unknown in many computing circles, Max Newman



Maxwell (Max) Newman.

was, in his own right, one of the most significant figures in the early history of British computing. He was active in the field for more than 10 years initially at Cambridge before World War II, then at Bletchley Park during hostilities, and finally in the postwar setting of the University of Manchester in the

mid-late 1940s. Newman's very English habit of understating his personal contribution, combined with his preference for stressing the accomplishments of others may explain why historians of computing have generally paid little attention to this extraordinary man who is, in consequence, princi-

a See <http://www.cdpa.co.uk/Newman/MHAN/exhibition-panel.php?Title=Welcome%20to%20the%0Newman%20Exhibition&Picture=ExPoster1.jpg>

b For a much fuller account of Newman's contribution to the early development of computers in the U.K., see Anderson, D. "Was the Manchester 'Baby' conceived at Bletchley Park?," BCS eWIC, (2008); http://www.bcs.org/upload/pdf/ewic_tur04_paper3.pdf.

pally remembered as a mathematician working in the field of topology.

We have heard claims over the last year that modern computing began with Alan Turing. However, the roots of computing run much deeper, and any “starting point” is open to being contested. We can, for example, trace an interesting path from Turing back through Newman to a lecture delivered by David Hilbert at the Sorbonne at the turn of the 20th century, in which he proposed 23 “future problems.” These were taken up by the mathematics community and collectively formed the agenda for mathematics research in the century that followed. Hilbert’s 10th problem can be framed as asking if there could exist, in principle, a definite process involving a finite number of steps, by which the validity of any given first-order logic statement might be decided? Turing first met Hilbert’s so-called *Entscheidungsproblem* around the spring of 1935 when he was a student in Newman’s Part III Foundations of Mathematics course. Newman later recalled: “I believe it all started because he attended a lecture of mine on foundations of mathematics and logic in which I had mentioned...that what is meant by saying that the process is constructive is that it’s purely...mechanical...and I may even have said a machine can do it. But he took the notion...and did produce this extraordinary definition of a perfectly general,...computable, function thus giving the first idea really of a perfectly general computing machine.”²

In April 1936, Turing gave Newman

Had Newman done nothing else in his career, Colossus would have entitled him to a place in the pantheon of pioneers of early British computing.

a draft copy of his dazzlingly original answer to the *Entscheidungsproblem*, at the center of which was an idealized description of a person carrying out numerical computation that, following Church, we have come to call a “Turing machine.” All modern computers are instantiations of Turing machines in consequence of which Turing’s paper is often claimed to be the single most important in the history of computing.

Newman almost immediately took Turing under his wing, canvassing successfully for “On Computable Numbers” to be published by the London Mathematical Society. At the same time he enlisted Alonzo Church’s help in arranging that Turing should spend some time studying in Princeton.

Cambridge in the late 1930s and early 1940s was particularly fertile soil for computing pioneers, and Newman played a part in the education of most of them. In addition to Alan Turing and his contemporary Maurice Wilkes, other students of Newman’s included Tom Kilburn, Geoff Tootill, and David Rees.

In 1939, as war was spreading across Europe, Newman was awarded a fellowship of the Royal Society. However, he had little opportunity to use this fellowship to provide impetus for further work in mathematics, as the outbreak of hostilities took one colleague after another out of academic life into war work. Newman, whose young family had evacuated to the U.S., grew increasingly isolated and slowly became disillusioned with life at Cambridge. Acting at the suggestion of Patrick Blackett, he accepted a post at Bletchley Park. Neither of them could have had the least notion that Newman had thereby embarked on a course that was to completely alter the future direction of his career. Newman initially began work as a cryptanalyst working on “Fish”^c traffic as part of John Tiltman’s group. The type of transmission that attracted the greatest interest was known as “Tunny” and carried messages between the very highest ranks of the German command. Manual methods utilizing statistical techniques had been devised

for breaking into the code, but the sheer volume of traffic being intercepted was beginning to overwhelm the human resources available.

Newman was able to make a decisive contribution. He believed it was possible to mechanize the attack on Tunny and lobbied successfully to test his conviction by developing an electromechanical code-breaking machine that came to be known as the “Heath Robinson.” Newman was placed in charge of the development, and a cryptanalyst, two engineers, and 16 Wrens^d were placed at his disposal. He ran his section which, following the Bletchley Park tradition of naming sections after those in charge of them, was called “The Newmanry,” in a democratic spirit, encouraging his staff to speak up if they thought him mistaken. Democracy notwithstanding, Newman was later recalled by Donald Michie^e as leading with patriarchal authority and having something of the force of nature about him. Possessed of considerable intellectual daring, he demanded, and generally got, the impossible both of events and people. Michie further recalls that Newman proceeded with “vigor and certitude, seemingly as a vehicle without reverse gear.” Under Newman’s leadership originality flourished. The successes of his team, in which he took great pleasure, were not the result of detailed micromanagement but came about by finding people in whom he could place his trust and allowing them to work according to their own judgment. This allowed Newman to concentrate on the broader managerial and organizational issues in the service of which, in Michie’s words, he displayed an “unerring sense of direction in a broader-brush landscape of which [his staff] often had no inkling. Over time his persistence toward a perceived goal would fructify in a stunning coup.”

Although the Heath Robinson proved fairly unreliable, the results it achieved were sufficiently encouraging for approval to be granted to develop a

^c German High Command ciphers produced by Lorenz machines. These devices employed an encryption scheme more complicated than that used by Enigma.

^d The Women’s Royal Naval Service (WRNS; popularly and officially known as the Wrens) was the women’s branch of the Royal Navy.

^e Michie, D. Personal communication with David Anderson, Feb. 12, 2001. Unpublished email.

more sophisticated machine—the Colossus. A great deal has been written about this machine, the world’s first digital electronic computing device. Suffice it to note that, had Newman done nothing else in his career, Colossus would have entitled him to a place in the pantheon of pioneers of early British computing.

As it was, the Colossus profoundly affected Newman’s future career. He saw at once, as few others did, the impact computing would come to have on mathematics and he decided to establish a computer-building project as soon as the war was over. In Newman’s judgment, the mathematics department at Cambridge was not the right environment for this, and he began to look around for a more suitable setting. With Blackett’s help and encouragement, Newman was appointed to the Fielden Chair of Pure Mathematics at Manchester University, the post having become vacant when Louis Mordell moved in the opposite direction to take up the Sadleirian Chair. Newman had two clear goals in mind: To establish a first-rate department that could stand comparison with the best in the country and to build a computer. At Bletchley Park, Max had been surrounded by people who could help him achieve both objectives.

In a clear declaration of intent, Newman brought with him Jack Good and David Rees, both of whom had been at Cambridge before the war, and as part of Newman’s section at Bletchley Park, they had experience working on Colossus. With further assistance from Blackett, a substantial grant was obtained from the Royal Society. This was an innovative use of Royal Society funds and was the first award made for the purpose of developing a computer. The only piece of the puzzle that was missing was a lead engineer. Newman was not the only person looking for a top-flight engineer: the National Physical Laboratory (NPL) was also planning to build a computer and “good circuit” men, Newman wrote to von Neumann, were “both rare and not procurable when found.”⁴

Thus it was that, at war’s end, F.C. (Freddie) Williams found himself in the fortunate position of being a man much in demand. It must have been quite a disappointment to the NPL when, in

Newman’s influence on the first generation of British computer scientists was incalculable.

late 1946, Williams was offered and accepted the Edward Stocks Massey chair of Electro-Technics at the University of Manchester. Blackett and Newman, who had earlier discussed with I.J. Good the possibility of hiring Williams, were both on the appointments panel.


Having secured the support of the university, obtained funding from the Royal Society, and assembled a first-rate team of mathematicians and engineers, Newman now had all elements of his computer-building plan in place. Adopting the approach he had used so effectively at Bletchley Park, Newman set his people loose on the detailed work while he concentrated on orchestrating the endeavor. The result was success beyond all expectation. By the middle of 1948 the Small Scale Electronic Machine (SSEM) was up and running. Although little more than a proof of concept, it was still the world’s first working digital electronic stored program computer. The Manchester team had indisputably achieved a major coup.

Newman’s direct involvement with computing activity was, however, coming to an end. Like Blackett, Newman was opposed to the inevitable use of the Manchester computer in the development of nuclear weapons, and as the government took an ever-closer interest in the Manchester computer, Max stepped back to leave further development to the engineers.

Newman retired in 1964 but continued to be active in topology and two years later produced an engulfing theorem for topological manifolds. Until 1970, he taught in a succession of visiting professorships, mostly in the U.S. In May 1973, his wife Lynn died and later the same year he married Margaret Penrose. Together they enjoyed a happy and contented life surrounded by their children and occupied with

travel, music, and entertaining. Newman died in 1984 at age 89.

Newman was a deeply cultured man with an inquiring mind whose interests ranged over a broad spectrum. His influence on the first generation of British computer scientists was incalculable, and his early appreciation of the importance of computing was probably matched only by that of Alan Turing. The vision and leadership he showed at Bletchley Park during World War II and his single-minded determination to mechanize the British code-breaking efforts not only had an appreciable impact on the outcome of the conflict but also created a computing legacy that he was determined to carry into the postwar situation. Such was the deftness by which he accomplished the transfer of knowledge that some of those who gained most from his understanding were more or less unaware of the singular contribution made to their own success by this remarkable man.

I am content to leave the last word to Mary Cartwright, who in presenting Newman with the 1962 De Morgan Medal said of him: “Newman is a scholar of mathematics with a fine sense of the directions in which important advances are to be expected. It was, for instance, at least in part as a result of his interest and advocacy that Manchester University acquired one of the first general-purpose computers. He has also contributed greatly to mathematics as an expositor....Whether in writing or in speaking he has a distinction and clarity of language that few can rival and all must admire.”¹ 

References

1. Cartwright, M.L. Presentation of the De Morgan Medal to Professor M.H.A. Newman. *J. London Mathematical Society* 38 (1963), 130.
2. Evans, C.R. Interview with Maxwell Herman Alexander Newman. Unpublished interview (transcript by David P. Anderson). Science Museum/National Physical Laboratory, 1975.
3. Newman, W. Alan Turing remembered. *Commun. ACM* 55, 12 (Dec. 2012), 39–40; DOI: 10.1145/2380656.2380682.
4. Newman, M.H.A. Letter to John von Neumann. In Box 6, Folder 2, Item 2, The Newman Digital Archive, The Max Newman Digital Archive, the University of Portsmouth Future Proof Computing Group, and St. John’s College, Cambridge (Feb. 8, 1946) (unpublished letter).
5. Vardi, M.Y. Who begat computing? *Commun. ACM* 56, 1 (Jan. 2013), 5; DOI: 10.1145/2398356.2398357.

David Anderson (cdpa@btinternet.com) is the CiTECH Research Centre Director at the School of Creative Technologies, University of Portsmouth, U.K.

Copyright held by author.



Education

Human-Centered Computing: A New Degree for Licklider's World

Combining computing and psychology, J.C.R. Licklider's prescient ideas are being applied in contemporary educational settings.

IN THE 1960s, J.C.R. Licklider described his vision for the future of computing, which is remarkably like today's world. He saw computing as augmenting human intelligence,¹ and for communications among communities.² He foresaw cloud computing and the Semantic Web. Licklider's background was different than many of the early computer scientists. He was not an electrical engineer or primarily a mathematician—his degrees were mostly in psychology.

To predict today's world took a combination of computing and psychology. It is not surprising that understanding today's world of ubiquitous computing requires a blend of computing and social science. The phenomena of social computing are not primarily about technology. What is interesting about our modern computing milieu is the blend of technology, humans, and community. Human-centered computing is a new subdiscipline of computer science that prepares students for studying our socio-technical world.

Georgia Tech, Clemson University, and the University of Maryland, Baltimore County all offer graduate degrees in Human-Centered Computing (HCC). Students in Georgia Tech's HCC Ph.D. program work in areas like human-computer interaction (HCI), learning sciences and technologies, and cognitive science and AI. They use methods from social and behavioral sciences as well as engineering. While HCI focuses on the boundary (the interactions) be-



tween computing and humans, HCC places humans (as individuals and in societies) at the center of the research. HCC might lead to designs for new software, but it can also help us to *understand* what emerges from the world that Licklider predicted.

Georgia Tech's HCC degree program prepares students to design technology, to understand humans and societies, and to study what emerges

when that technology is ubiquitous. HCC at Georgia Tech has a core of three courses. The foundational course gives students theories that can be applied to understand human behavior with technology. The technology course ensures all HCC students can build prototypes of interactive systems to demonstrate their ideas. The third course ties together the threads to establish research themes. An annual seminar

engages students in discussion about recent literature that relates to HCC.

For me, HCC is an excellent way to prepare computing education researchers. People have always developed theories of how their worlds work, from ancient mythology to modern science. How are people explaining their computerized world to themselves? How can we help them develop a better and more useful understanding? What gets in the way of that understanding? Answering these kinds of questions requires knowledge of computer science (if only to recognize correct from incorrect understandings), but also of how people learn and how to study humans and their learning.

Not all HCC students address issues of computing education research, but even when that is not the explicit focus, HCC research often offers lessons about how people learn about computing. People always learn, and in a world filled with computing, that is often what they are learning about—though not always well, clearly, or efficiently. Here are stories of three HCC graduates whose dissertations inform us about how people learn about computing.

Reframing the Computing

Betsy DiSalvo (now an assistant professor at Georgia Tech) starts from an interesting observation. Many computer scientists (who are mostly white or Asian, and male) say they became interested in computing because of video games. No demographic group plays more video games than African-American and Hispanic teenagers and men. But few African-American and Hispanic males become computer scientists. Why was that?

DiSalvo explored her question with ethnographic methods. She observed African-American teen males playing video games and talked to them about how and why they played. She found they were playing video games differently than white teen males. Her participants never used “cheat codes” or modified their games in any way—they used video games like athletic competition. Manipulating the football or the field is cheating, so why would you change the video game? She used design research activities to explore how different ways of describing computing would make the technology more

Unlike science or mathematics, undergraduates often come to computer science with a poor understanding of what computer science is.

salient while still appealing to the audience she wanted to attract.

DiSalvo built the Glitch Game Testers project. Glitch successfully engaged African-American teen males in computer science by training and hiring them as game-testers. Game-testers must see video games as a technology with flaws. Glitch students learned computer science, motivated to become better testers. Glitch attracted students who loved video games, and kept them involved because it was a paying job. Most of her students went on to post-secondary computing education.

DiSalvo designed Glitch through a human-focused design process. She did not design a technology. She designed a new way for her students to think about computing. Her design research activities explored different ways of describing computer science with different lenses. Through that iterative design process, she found a reframing that could change who builds computing in the future.

Impact of Not Understanding Computing

Unlike science or mathematics, undergraduates often come to computer science with a poor understanding of what computer science is. Mike Hewner (now an assistant professor at Rose-Hulman Institute of Technology) wanted to know what the impact of that misunderstanding of computer science had on students who chose to major in computer science. Hewner interviewed 33 students at three different universities. He used a social science method called *grounded theory* to iden-

tify themes, create abstractions, and eventually come to a well-supported understanding of how CS majors make educational decisions.

Hewner found plenty of misunderstandings of computer science, like the two sophomore CS majors who told him computer graphics was the study of Photoshop. He also found more subtly different conceptions of computer science. There were the students who saw that computer science was the study of theory, and software engineering was “lower tier.” There were the students who saw that programming was “the end goal behind computer science,” and theory was in support of programming.

The biggest surprise in Hewner’s study was that these different conceptions did *not* significantly influence students’ educational choices. Few students that Hewner interviewed could even predict what would be in the next classes they took. Even when faced with choosing among specializations within the degree, students told Hewner the choice did not really matter, prompting the comment “I think I would have the same number of jobs available if I took either of them.” The students did not have a clear understanding of what jobs looked like in computer science, and consequently, they did not make choices in order to prepare themselves for a future job.

Hewner found his students used *enjoyment* as a proxy for *affinity* for a subject. Students would explore their interests through their classes. When they found something they really enjoyed, they chose that as a sign of their affinity for the subject. Hewner found students did not reason deeply about why they enjoyed a course (or not). A bad teaching assistant or an 8 A.M. lecture might lead the students to dislike a course, and that would be enough to cause students to switch majors to a course they enjoyed more. Once students committed to a major, they simply trusted the curriculum—and were willing to persist through difficult classes, once they had made the commitment.

Hewner drew on his deep understanding of computer science to make sense (and identify nonsense) in the students’ conceptualizations. His methods were drawn directly from social sciences. His findings help us to



Association for
Computing Machinery

ACM Conference Proceedings Now Available via Print-on-Demand!

Did you know that you can now order many popular ACM conference proceedings via print-on-demand?

Institutions, libraries and individuals can choose from more than 100 titles on a continually updated list through Amazon, Barnes & Noble, Baker & Taylor, Ingram and NACSCORP: CHI, KDD, Multimedia, SIGIR, SIGCOMM, SIGCSE, SIGMOD/PODS, and many more.

For available titles and ordering info, visit:
librarians.acm.org/pod



The modern computing world is not just designed. It emerges from people working with it.

better understand the process of preparing the students who will one day create future computing.

What Should Laypeople Understand About Computing?

Even the HCC students who study HCI topics can inform us about computing education. Erika Poole (now an assistant professor at Pennsylvania State University) is one of those. Poole studied how families attempting a variety of technology-related challenges, as a way of discovering how they sought out help.

The families in Poole's study did much worse than I might have guessed. For example, only two of 15 families were able to configure their wireless router. Some of the reasons were because of awful user interfaces (for example, a virtual keyboard that was missing some keys). But other tasks were challenging for more subtle reasons.

One of the challenges in Poole's study involved editing a Wikipedia page. When Poole's participants edited a Wikipedia page for the first time, without an account, they saw this warning message: "You are not currently logged in. Editing this way will cause your IP address to be recorded publicly in this page's edit history. If you create an account, you can conceal your IP address and be provided with many other benefits. Messages sent to your IP can be viewed on your talk page."

Poole's participants had to decide if "recording publicly" their "IP address" was a problem. One participant told Poole the process of editing the encyclopedia and reading this warning message made her "feel like a criminal." She canceled her changes. Another participant contacted his "computer savvy" friend to help interpret

the message. The friend warned him that having your IP address recorded was dangerous. This participant also gave up.

Editing Wikipedia is a natural act in Licklider's world. Should we expect people who edit Wikipedia to know what an IP address is? Is it a user interface mistake to expect Wikipedia contributors would understand the implications of publicly recording an IP address? While the Internet is in common use in the homes of all of Poole's participant families, the families (and even some of their "computer savvy" friends) clearly did not understand some of the basic terms and concepts of the technology they use.

Poole's study provides some concrete examples of what people understand, and misunderstand, about the computing in their lives. Certainly, she informs HCI designers, in thinking about expectations of user knowledge. On the other hand, an IP address is part of our modern world. Is it better to hide it, or explain it? Her study informs general education designers in thinking about what everyone needs to know about computing.

Understanding Licklider's World

J.C.R. Licklider was able to predict much of our modern computing world because he combined his understanding of computing with his understanding of people. To understand the world he predicted, we need researchers who understand computing and who can use the methods from Licklider's psychology (and sociology and other social sciences, too). The modern computing world is not just designed. It emerges from people working with it, trying to understand and use it (typically, in ways different than originally designed), and interacting with millions of others doing the same things. We have to study it as it is, not just as how we meant it to be. ■

References

1. Licklider, J.C.R. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics HFE-1*, (Mar. 1960), 4–11.
2. Licklider, J.C.R. and Taylor, R.W. The computer as a communications device. *Science and Technology* (Apr. 1968).

Mark Guzdial (guzdial@cc.gatech.edu) is a professor in the College of Computing at Georgia Institute of Technology in Atlanta, GA.

Copyright held by author.

Viewpoint

The Science in Computer Science

Computer science is in a period of renaissance as it rediscovers its science roots.

COMPUTER SCIENCE HAS for decades been ripped by an old saw: Any field that calls itself a science, cannot be science. The implied criticisms that we lack substance or hawk dubious results have been repeatedly refuted. And yet the criticism keeps coming up in contexts that matter to us.

It comes up in education in the debates about encouraging more student involvement in STEM (science, technology, engineering, and mathematics). Many critics see computer science mainly as technology or math. Will computer science be excluded because it is not seen as genuine science?

It comes up in research in debates about the predictive power of our analytic tools. In some subfields, such as storage management, performance prediction, and algorithms, experimental methods have led to reliable predictive models. In others, such as system safety and security, we lack predictive models and we can only speculate that experimental methods will lead to understanding. In his first ACM president's letter, Vint Cerf asks why software engineering does not rely more on experimental science (*Communications*, Oct. 2012). In so doing, he echoes a lament uncovered in a 1995 study of software engineering literature.¹⁰ Do enough of us know the experimental methods needed to do this consistently well?

In interdisciplinary collaboration, it comes up when teams are formed and



when credit is handed out. Why are computer scientists still often seen as professional coders rather than genuine collaborators?

My purpose here is to review the history of the question, “Is computing science?” and point to new answers that can help educators, researchers, and collaborators.

I use the term “computing” to refer to the set of related fields that deal with computation. These include computer science, computational science, information science, computer engineering,

and software engineering. Interestingly, I have encountered less skepticism to the claim that “computing is science” than to “computer science is science.”

A Short History of Science in Computing

Computing has been deeply involved in science since the beginning. A science vision pervaded the field through the 1950s, and then faded as technology development drew most of our energy through the 1980s. A science renaissance began in the 1990s, propelled by

ACM's Career & Job Center

Looking for your next IT job?

Need Career Advice?

Visit ACM's Career & Job Center at:

<http://jobs.acm.org>

Offering a host of career-enhancing benefits:

- A highly targeted focus on job opportunities in the computing industry
- Access to hundreds of corporate job postings
- Resume posting keeping you connected to the employment market while letting you maintain full control over your confidential information
- An advanced Job Alert system notifies you of new opportunities matching your criteria
- Career coaching and guidance from trained experts dedicated to your success
- A content library of the best career articles compiled from hundreds of sources, and much more!

The ACM Career & Job Center is the perfect place to begin searching for your next employment opportunity!

<http://jobs.acm.org>



Association for
Computing Machinery

Advancing Computing as a Science & Profession

computational science and the discovery of natural information processes. I will review each of these periods.

The pioneers who planned and built the first electronic computers were strongly motivated by visions of computers advancing science. The two most obvious ways were the numerical solution of mathematical models of physical processes, and the analysis of large datasets compiled from experiments. Computer science became a recognized academic field of study in 1962 with the founding of computer science departments at Purdue and Stanford. These departments maintained strong faculties in mathematical software, which directly supported science.

In 1967, Newell, Perlis, and Simon argued that the new field was a science concerned with all aspects of “phenomena surrounding computers.”¹² However, many traditional scientists disagreed with the science claim; they held that true science deals with phenomena that occur in nature (“natural processes”) whereas computers are man-made artifacts. Simon, a Nobel Laureate in economics, so strongly disagreed with the “natural interpretation” that he published a book *The Sciences of the Artificial* (MIT Press, 1969). He argued that economics and computer science met all the traditional criteria for science, and deserved to be called sciences even if, said Simon, their focal phenomena are “man-made as opposed to natural.”

In the initial years of the field, most computing people devoted their energy to building the systems that could realize the visionary dreams of the founders. By the late 1970s, the computing industry was recruiting system people so vigorously that university departments were experiencing a “brain drain” of systems-oriented faculty. ACM leadership was very concerned: this trend threatened experimental computer science. I was deeply involved as ACM president in arguing the importance of experimental methods for computing and in assisting the U.S. National Science Foundation (NSF) to support experimental computer scientists. I wrote in 1980 that the experimental method (that is, science) is essential in computer science,⁶ and in 1981 I cited the subfield of performance modeling and prediction as an exemplar of

the ideals of science.⁴ Despite these efforts, many university departments lost their experimentalists and the science vision faded into the background.

In the 1980s, science visionaries from many fields saw ways to employ high-performance computers to solve “grand challenge” problems in science. They said computing is not only a tool for science, but also a *new method of thought and discovery in science*. (Aha! Computational thinking!) They defined computational science as a new branch of science imbued with this idea. The leaders of biology, epitomized by 1975 Nobel Laureate David Baltimore, went further, saying biology had become an information science and that DNA translation is a natural information process. Another biologist, Roseanne Senson, attributed the efficiency of photosynthesis to a quantum algorithm embedded in the cellular structure of plant leaves (*Nature*, April 2007). Biologists have thus been leaders in driving nails into the coffin of the “natural science” argument about computing. Many other scientists have reached similar conclusions. They include physicists working with quantum computation and quantum cryptography, chemists working with materials, cognitive scientists working with brain processes, economists working with economic systems, and social scientists working with networks.⁹ All claimed to work with natural information processes. Stephen Wolfram went further, arguing that information processes underlie every natural process in the universe.¹³

Those two external factors—rise of computational science and discovery of natural information processes—have spawned a science renaissance in computing. Experimental methods have regained their stature because they are the only way to understand very complex systems and to discover the limits of heuristic problem solution methods.

Here is an example of an advance in algorithms obtained through an empirical approach. In May 2004, an international research group announced it had computed an optimal tour of 24,978 cities in Sweden (see <http://tsp.gatech.edu/sweden>). By iterating back and forth among several heuristic methods, they homed in on a provably optimal solution. Their computation took about one year on a bank of 96 parallel Intel Xeon

2.8GHz processors. With classical tour-enumeration algorithms, which are of order $O(n!)$, the running time would be well beyond the remaining age of the universe. With experimental methods, algorithm scientists quickly found optimal or near-optimal solutions.

New fields heavily based in experimental methods have opened up—network science, social network science, design science, data mining, and Bayesian inference, to name a few. The widening claims that information processes occur in nature have refuted the notion that computer science is not “natural” and have complemented Simon’s arguments that computing is a science of the artificial.

When Is a Field a Science?

This brief history suggests that computing began as science, morphed into engineering for 30 years while it developed technology, and then entered a science renaissance about 20 years ago. Although computing had subfields that demonstrated the ideals of science, computing as a whole has only recently begun to embrace those ideals. Some new subfields such as network science, network social science, design science, and Web science, are still struggling to establish their credibility as sciences.

What are the criteria for credibility as science? A few years ago I compiled a list that included all the traditional ideals of science:^{1,3}

- ▶ Organized to understand, exploit, and cope with a pervasive phenomenon.
- ▶ Encompasses natural and artificial

Although computing had subfields that demonstrated the ideals of science, computing as a whole has only recently begun to embrace those ideals.

processes of the phenomenon.

- ▶ Codified structured body of knowledge.
- ▶ Commitment to experimental methods for discovery and validation.
- ▶ Reproducibility of results.
- ▶ Falsifiability of hypotheses and models.
- ▶ Ability to make reliable predictions, some of which are surprising.

Computing’s original focal phenomenon was information processes generated by hardware and software. As computing discovered more and more natural information processes, the focus broadened to include “natural computation.”⁹ We can now say “computing is the study of information processes, artificial and natural.”¹¹

Computing is not alone in dealing with both natural and artificial processes. Biologists, for example, study artifacts including computational models of DNA translation, the design of organic memories, and genetically modified organisms (GMOs). All fields of science constantly face questions about whether knowledge gained from their artifacts carries over to their natural processes. Computing people face similar questions—for example, does studying a software model of a brain yield useful insights into brain processes? A great deal of careful experimental work is needed to answer such questions.

The question of “scienceness” of computing has always been complicated because of the strong presence of science, mathematics, and engineering in the roots and practice of the field.^{8,11} The science perspective focuses on increasing understanding through experimental methods. The engineering perspective focuses on designing and constructing ever-improved computing systems. The mathematics perspective focuses on what can be deduced from accepted statements.

The term “theory” illustrates the different interpretations that arise in computing because of these three perspectives. In pure math, theory means the set of valid deductions from a set of axioms. In computing, theory more often means the use of formalism to advance understanding or design.

Effects on the Education System

Unfortunately, our education system for young people has not caught

up with these realities. From 2001 to 2009, college enrollments in CS majors dropped 50% (and are now recovering). From early analyses, we could see that students were losing interest in computing in high schools, half of which had no computer course at all, and many of the others relegated their one computer course to literacy in keyboarding and word processing. Very few had courses in the principles of computing. Around 1998, the U.S. Educational Testing Service wanted to help by focusing the Computer Science Advanced Placement (AP) curriculum on object-oriented programming. Unfortunately, the new AP curriculum did not help. Fewer than one-third of high schools actually used the CS AP curriculum and many teachers did not understand enough about object-oriented programming to teach it effectively.

Leaders in most of the STEM fields reported enrollment declines in the same period. Stimulating more student interest in STEM fields has become an international concern.

The science renaissance in computing has led to an explosion of new content on the principles of computing that is beginning to reach into high schools. With support from the U.S. National Science Foundation, a coalition of universities has defined a computer science principles introductory course and created prototypes (see <http://csprinciples.org>). The Educational Testing Service has embarked on a closely related project to redefine the AP curriculum around computing principles. Over the past two decades, Tim Bell of the University of Canterbury, New Zealand, has designed exercises and games for children 12–15 years old, allowing them to experience computing principles without using computers (see <http://csunplugged.org>). With my colleagues I have put together a presentation of all computer science principles (see <http://greatprinciples.org>).^{2,5}

The dream articulated by Newell, Perlis, and Simon 50 years ago has come true. It endured many skeptical antagonists and weathered many storms along the way. Computing is now accepted as science. Some of us even believe computing is so pervasive that it qualifies as a new domain of science alongside the traditional domains of physical, life, and social

We can now say computing is the study of information processes, artificial and natural.

sciences.⁷ Educators are finding innovative ways to teach computing science to young people, who are now being infected with the magic, joy, and beauty of the field.

Let Us Discuss

I am editor-in-chief of ACM's *Ubiquity*, an online peer-reviewed magazine about the future of computing and the people who are creating it. The *Ubiquity* editors put together a symposium of essays from 14 authors discussing various aspects of the question "Is computing science?" The authors include an ACM president, an ACM past president, two ACM A.M. Turing Award recipients, an NSF program manager, a journalist, six educators, and four interdisciplinary researchers. We drew five conclusions from the symposium.

First, the question of whether computing is science is as old as the field. It arose because traditional scientists did not recognize computational processes as natural processes. Even during the engineering years, when much of the energy of the field was devoted to building systems and understanding their theoretical limits, the field developed two important scientific theories. The theory of locality studied memory usage patterns of computations, and the theory of performance evaluation validated queueing network models for reliable performance predictions of computer systems.

Second, there is a growing consensus today that many of the issues we are studying are so complex that only an experimental approach will lead to understanding. The symposium documents advances in algorithmics, biology, social networking, software engineering, and cognitive science that use empirical methods to answer important questions.

Third, scientists in many fields now

recognize the existence of natural information processes. This dismisses an early perception that CS deals solely with artificial information processes. Computing is not constrained to be a "science of the artificial." Computing is indeed a full science.

Fourth, because information processes are pervasive in all fields of science, computing is necessarily involved in all fields, and computational thinking has been accepted as a widely applicable problem-solving approach. Many students are now selecting computer science majors because it preserves their flexibility in choosing a career field later.

Fifth, computing presented as science is very engaging to middle and high school students. The science perspective expands well beyond the unfortunate and prevalent notion that computer science equals programming. A growing number of STEM teachers are embracing these new methods.

I invite you to look in at the full symposium and see for yourself what these people have said (see <http://ubiquity.acm.org>), and then weigh in with your own observations. **C**

References

- Denning, P. Computing is a natural science. *Commun. ACM* 50, 7 (July 2007), 13–18.
- Denning, P. Great principles of computing. *Commun. ACM* 46, 11 (Nov. 2003), 16–20.
- Denning, P. Is computer science? *Commun. ACM* 48, 4 (Apr. 2005), 27–31.
- Denning, P. Performance analysis: Experimental computer science at its best. *Commun. ACM* 24, 11 (Nov. 1981), 725–727; <http://doi.acm.org/10.1145/358790.358791>.
- Denning, P. The great principles of computing. *American Scientist* 98 (Sept.–Oct. 2010), 369–372.
- Denning, P. What is experimental computer science? *Commun. ACM* 23, 10 (Oct. 1980), 543–544; <http://doi.acm.org/10.1145/359015.359016>.
- Denning, P. and Rosenbloom, P. Computing: The fourth great domain of science. *Commun. ACM* 52, 9 (Sept. 2009), 27–29.
- Gonzalo, G. Is computer science truly scientific? *Commun. ACM* 53, 7 (July 2010), 37–39. <http://doi.acm.org/10.1145/1785414.1785431>.
- Kari, L. and Rozenberg, G. The many facets of natural computing. *Commun. ACM* 51, 10 (Oct. 2008), 72–83; <http://doi.acm.org/10.1145/1400181.1400200>.
- Lukowicz, P., Tichy, W., Pechelt, L., and Heinz, E. Experimental evaluation in computer science: A quantitative study. *Journal of Systems and Software* 28, 1 (Jan. 1995), 9–18.
- Morrison, C. and Snodgrass, R.T. Computer science can use more science. *Commun. ACM* 54, 6 (June 2011), 36–38; <http://doi.acm.org/10.1145/1953122.1953139>.
- Newell, A., Perlis, A., and Simon, H. Computer science. *Science* 157 (1967), 1373–1374.
- Wolfram, S. *A New Kind of Science*. Wolfram Media, 2002.

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at the Naval Postgraduate School in Monterey, CA, is Editor of *ACM Ubiquity*, and is a past president of ACM.

Copyright held by author.

Viewpoint

Moving from Petaflops to Petadata

The race to build ever-faster supercomputers is on, with more contenders than ever before. However, the current goals set for this race may not lead to the fastest computation for particular applications.

THE SUPERCOMPUTER COMMUNITY is now facing an interesting situation: Systems exist that, for some sophisticated applications and some relevant performance measures, demonstrate an order of magnitude higher performance^{11,14,24} compared to the top systems from the TOP500 supercomputers list,² but are not on that list. Most of the TOP500 machines reach more than 80% of efficiency if they run LINPACK, on the other hand, if these machines run real engineering applications, they reach significantly less (~5%), due to the non-optimal manipulation with matrices, or due to the need for execution of non-numerical operations.

A creator of the TOP500 supercomputers list rightfully claimed the list sheds light on only one dimension of modern supercomputing,⁶ which is a relatively narrow one. This Viewpoint is intended to induce thinking about alternative performance measures for ranking, possibly ones with a much wider scope.²⁰ This Viewpoint is not offering a solution; it is offering a theme for brainstorming.

To demonstrate this need for such thinking, we will use the example of a particular type of systems, based on a kind of dataflow approach. Namely, we will focus on the solutions developed by Maxeler Technologies.¹² Typical applications of such systems include: geomechanical simulations,¹¹ financial stochastic PDEs,²⁴ and seismic



Oak Ridge National Laboratory's Titan supercomputer.

modeling in the oil and gas industry.¹⁴ There are several other efficient solutions with even more application-specific approaches, for example, the Anton machine for the calculation of interparticle forces in molecular dynamics simulation.¹⁹

Our perspective is that the performance metric should become multidimensional—measuring more than just FLOPS, for example, performance per watt, performance per cubic foot, or

performance per monetary unit (dollar, yen, yuan, euro, and so forth).

Here, we concentrate on the following issues: rationales (what are the evolutionary achievements that may justify a possible paradigm shift in the ranking domain); justification (what are the numerical measurements that require rethinking); suggestions (what are the possible avenues leading to potential improvements of the ranking paradigm). We conclude by specifying

ACM Transactions on Accessible Computing



This quarterly publication is a quarterly journal that publishes refereed articles addressing issues of computing as it impacts the lives of people with disabilities. The journal will be of particular interest to SIGACCESS members and delegates to its affiliated conference (i.e., ASSETS), as well as other international accessibility conferences.

www.acm.org/taccess
www.acm.org/subscribe



Association for
Computing Machinery

to whom all this might be most beneficial and opening possible directions for future research.

Rationales

The current era of supercomputing is referred to as the petascale era. The next big HPC challenge is to break the exascale barrier. However, due to technological limitations,^{16,23} there is growing agreement that reaching this goal will require a substantial shift toward hardware/software co-design.^{3,7,10,18} The driving idea behind the custom dataflow supercomputers (like the Maxeler solution), falls into this category: To implement the computational dataflow in a custom hardware accelerator. In order to achieve maximum performance, the kernel of the application is compiled into a dataflow engine. The resulting array structure can be hundreds to thousands of pipeline stages deep. Ideally, in the static dataflow form, data can enter and exit each stage of the pipeline in every cycle. It cannot be precisely itemized what portion of the improved performance is due to the dataflow concept, and what portion is due to customization; this is because the dataflow concept is used as a vehicle that provides customization in hardware.

For these dataflow systems, utilization of a relatively slow clock is typical, while the entire dataflow is completed more efficiently. This slow clock is not a problem for big data computations, since the speed of computation depends on pin throughput and local memory size/bandwidth inside the computational chip. Even when the dataflow is implemented using FPGA chips, and thus the general-purpose connections in FPGA chips bring a clock slowdown, this does not affect the performance: pin throughput and local memory size/bandwidth are the bottleneck. The sheer magnitude of the dataflow parallelism can be used to overcome the initial speed disadvantage. Therefore, if counting is oriented to performance measures correlated with clock speed, these systems perform poorly. However, if counting is oriented to performance measures sensitive to the amount of data processed, these systems may perform richly. This is the first issue

The current era of supercomputing is referred to as the petascale era. The next big HPC challenge is to break the exascale barrier.

of importance.

The second important issue is related to the fact that, due to their lower clock speed, systems based on this kind of a dataflow approach consume less power, less space, and less money compared to systems driven by a fast clock. Weston²⁴ shows the measured speedups (31x and 37x) were achieved while reducing the power consumption of a 1U compute node. Combining power and performance measures is a challenge that is already starting to be addressed by the Green 500 list. However, evaluating radically different models of computation such as dataflow remains yet to be addressed, especially in the context of total cost of ownership.

In addition to the aforementioned issues, the third issue of importance is that systems based on a kind of dataflow approach perform poorly on relatively simple benchmarks, which are typically not rich in the amount and variety of data structures. However, they perform fairly well on relatively sophisticated benchmarks, rich in the amount and variety of data structures.

Justification

Performance of an HPC system depends on the adaption of a computational algorithm to the problem, discretization of the problem, mapping onto data structures and representable numbers, the dataset size, and the suitability of the underlying architecture compared to all other choices in the spectrum of design options. In light of all these choices, how does one evaluate a computer system's suitability for a particular task such as climate modeling or genetic sequencing?

If we examine the Top500 list (based on LINPACK, a relatively simple benchmark dealing with LU decomposition), the top is dominated by traditional, control-flow based systems. One would expect these systems to offer the highest performance. However, if we turn to a relatively data-intensive workload (for example, order of gigabytes) used in banking environments, we see a system that shows a speedup of over 30 times compared to a traditional control-flow driven system.²⁴ On a highly data-intensive workload (for example, order of terabytes) used by geophysicists, a speedup of 70 times has been demonstrated.¹¹ On an extremely data-intensive workload (for example, order of petabytes) used by petroleum companies the same dataflow system shows an even greater speedup, close to 200 times.¹⁴

Of course, these results are obtained by creating a custom dataflow architecture for the specified problems. The question may arise: Could not a LINPACK implementation reveal the same potential? Indeed, custom hardware implementations have been shown to yield speedups. However, these are on a much lesser scale, 2–6 times.^{17,22} Furthermore, we believe all-out efforts to create LINPACK-targeting (as a benchmark) custom machines, and the informativeness of such results would not be highly useful, especially since even the implications of LINPACK results produced by some more general-purpose systems are already questioned.¹³ An additional confirmation of our opinion can be found in the fact that the TOP500 does not accept or publish results from systems with LINPACK custom hardware.

Suggestions

Taking all of these issues into account leads to the following two statements and suggestions.

1. The FLOPS count does not, on its own, sufficiently cover all aspects of HPC systems. To an extent it does provide estimates of HPC performance; however, it does not do so equally effectively for different types of systems.

This statement is not novel to the HPC community, as indicated by: Faluk,⁴ Pancake,¹⁵ and Wolter.²⁵ In fact, assessing the productivity of HPC sys-

tems has been one of the emphases of the Defense Advanced Research Projects Agency (DARPA) High Productivity Computing Systems (HPCS) program, with the *International Journal of High Performance Computing Applications* devoting a special issue just to this topic.⁹ Yet, the FLOPS count seems to persist as the dominate measure of performance of HPC systems.

We are not suggesting that the FLOPS count be eliminated, but rather that a data-centric measure could shed some more light on other aspects of HPC systems.

One idea is to look at result data generation rate per second (petabytes per second), per cubic foot, and per watt, for a particular algorithm and dataset size. The initial justification for expanding the FLOPS count can be found in the following fact: The floating-point arithmetic no longer dominates the execution time, even in computationally intensive workloads, and even on conventional architectures.⁴ Furthermore, some of the unconventional systems (like the Maxeler systems) have relatively large amounts of on-chip memory and avoid some types of instructions altogether, which further blurs the image obtained from looking at FLOPS. Consequently, for data processing applications, the rate of producing results is the logical measure, regardless of the type and number of operations required to generate that result.

Of course, financial considerations play a major role in computing. However, it is unreasonable to include non-transparent and ever-negotiated pricing information into an engineering measure. We know the cost of computer systems is dictated by the cost of the chips and the cost of the chips is a function of the regularity of the design, the VLSI process, chip area, and most importantly, volume. Encapsulating all these components in a measure remains a challenge.

2. LINPACK, the workload used to create the TOP500 Supercomputer list, is, on its own, not a sufficient predictor of performance.

Again, this point is not novel to the HPC community, as indicated in Anderson,¹ Gahvari,⁵ Geller,⁶ and in Singh²⁰ by a creator of the TOP500 list. Alternatives and expansions have been suggested, some of the most notable

Calendar of Events

May 18–26

35th International Conference on Software Engineering, San Francisco, CA, Sponsored: SIGSOFT, Contact: David Notkin, Email: notkin@cs.washington.edu

May 18–19

International Conference on Software and System Process, San Francisco, CA, Contact: Jurgen Munch, Email: j.muench@computer.org Phone: +358-50-3175318

May 19–22

SIGSIM Principles of Advanced Discrete Simulation Montreal, Canada, Sponsored: SIGSIM, Contact: Margaret L. Loper, Email: margaret.loper@gtri.gatech.edu Phone: 404-894-4663

May 20–24

IEEE International Parallel and Distributed Processing Symposium, Cambridge, MA, Contact: Charles Weerns, Email: weerns@cs.umass.edu

May 20–24

The 2013 International Conference on Collaboration Technologies and Systems, San Diego, CA, Contact: Waleed W. Smari, Email: smari@arys.org Phone: 937-681-0098

May 21–24

The Fourth International Conference on Future Energy Systems, Berkeley, CA, Contact: Catherine Rosenberg, Email: cath@ece.uwaterloo.ca Phone: 519-888-4510

May 28–30

International Conference on Application and Theory of Automation in Command and Control Systems, Naples, Italy, Contact: Francisco Saez, Email: franciscojavier.saez@upm.es

ones being the Graph 500 and the HPC Challenge. Both of them aim to include a wider set of measures that substantially contribute to the performance of HPC systems running real-world HPC applications. When put together, these benchmarks provide a more holistic picture of an HPC system. However, they are still focused only on control flow computing, rather than on a more data-centric view that could scale the relevance of the included measures to large product applications.

This Viewpoint offers an alternate road to consider. Again, we do not suggest LINPACK, Graph 500, or HPC Challenge to be abandoned altogether, but supplemented with another type of benchmarks: Performance of the systems when used to solve real-life problems, rather than generic benchmarks. Of course, the question is how to choose these problems. One option may be to analyze the TOP500 and/or Graph 500 and/or a list of the most expensive HPC systems, for example, and to extract a number of problems that top-ranking systems have most commonly been used for. Such a ranking would also be of use to HPC customers, as they could look at the list for whatever problem is the most similar to the problem of their own.

Finally, this type of ranking would naturally evolve with both the HPC technology and the demands presented to HPC systems by periodically updating the list. A generic benchmark, on the other hand, must be designed either by looking at a current HPC system and its bottlenecks, or typical demands of current HPC problems, or both. As these two change in time, so must the benchmarks.

Conclusion

The findings in this Viewpoint are pertinent to those supercomputing users who wish to minimize not only the purchase costs, but also the maintenance costs, for a given performance requirement. Also to those manufacturers of supercomputing-oriented systems who are able to deliver more for less, but are using unconventional architectures.²¹

Topics for future research include the ways to incorporate the price/complexity issues and also the satisfaction/profile issues. The ability issues (availability, reliability, extensibility, partitioning ability, programmability, portabil-

As we direct efforts to break the exascale barrier, we must ensure the scale itself is appropriate.

ity, and so forth) are also of importance for any future ranking efforts.

Whenever a paradigm shift happens in computer technology, computer architecture, or computer applications, a new approach has to be introduced. The same type of thinking happened at the time when GaAs technology was introduced for high-radiation environments, and had to be compared with silicon technology, for a new set of relevant architectural issues. Solutions that ranked high until that moment suddenly obtained new and relatively low-ranking positions.⁸

As we direct efforts to break the exascale barrier, we must ensure the scale itself is appropriate. A scale is needed that can offer as much meaning as possible and can translate to real, usable, performance, to the highest possible degree. Such a scale should also feature the same two properties, even when applied to unconventional computational approaches. □

References

- Anderson, M. Better benchmarking for supercomputers. *IEEE Spectrum* 48, 1 (Jan. 2011), 12–14.
- Dongarra, J., Meuer, H., and Strohmaier, E. TOP500 supercomputer sites; <http://www.netlib.org/benchmark/top500.html>.
- Dosanji, S. et al. Achieving exascale computing through hardware/software co-design. In *Proceedings of the 18th European MPI Users' Group Conference on Recent Advances in the Message Passing Interface (EuroMPI'11)*, Springer-Verlag, Berlin, Heidelberg, (2011), 5–7.
- Faulk, S. et al. L. Measuring high performance computing productivity. *International Journal of High Performance Computing Applications* 18 (Winter 2004), 459–473; DOI: 10.1177/1094342004048539.
- Gahvari, H. et al. Benchmarking sparse matrix-vector multiply in five minutes. In *Proceedings of the SPEC Benchmark Workshop* (Jan. 2007).
- Geller, T. Supercomputing's exaflop target. *Commun. ACM* 54, 8 (Aug. 2011), 16–18; DOI: 10.1145/1978542.1978549.
- Gioiosa, R. Towards sustainable exascale computing. In *Proceedings of the VLSI System on Chip Conference (VLSI-SoC)*, 18th IEEE/IFIP (2010), 270–275.
- Helbig, W. and Milutinovic, V. The RCA's DCFL E/D MESFET GaAs 32-bit experimental RISC machine. *IEEE Transactions on Computers* 36, 2 (Feb. 1989), 263–274.
- Kepner, J. HPC productivity: An overarching view. *International Journal of High Performance Computing*

- Applications* 18 (Winter 2004), 393–397; DOI: 10.1177/1094342004048533
- Kramer, W. and Skinner, D. An exascale approach to software and hardware design. *Int. J. High Perform. Comput. Appl.* 23, 4 (Nov. 2009), 389–391.
- Lindtjorn, O. et al. Beyond traditional microprocessors for geoscience high-performance computing applications. *IEEE Micro* 31, 2 (Mar./Apr. 2011).
- Maxeler Technologies (Oct. 20, 2011); <http://www.maxeler.com/content/frontpage/>
- Mims, C. Why China's new supercomputer is only technically the world's fastest. *Technology Review* (Nov. 2010).
- Oriato, D. et al. Finite difference modeling beyond 70Hz with FPGA acceleration. In *Proceedings of the SEG 2010, HPC Workshop*, Denver, (Oct. 2010).
- Pancake, C. Those who live by the flop may die by the flop. Keynote Address, 41st International Cray User Group Conference (Minneapolis, MN, May 24–28 1999).
- Patt, Y. Future microprocessors: What must we do differently if we are to effectively utilize multi-core and many-core chips? *Transactions on Internet Research* 5, 1 (Jan. 2009), 5–10.
- Ramalho, E. The LINPACK benchmark on a multi-core multi-FPGA system. University of Toronto, 2008.
- Shalf, J. et al. Exascale computing technology challenges. VECPAR (2010), 1–25; <https://www.nersc.gov/assets/NERSC-Staff-Publications/2010/ShalfVecpar2010.pdf>.
- Shaw, D.E. et al. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* 51, 7 (July 2008), 91–97; DOI: 10.1145/1364782.1364802.
- Singh, S. Computing without processors. *Commun. ACM* 54, 8 (Aug. 2011), 46–54; DOI: 10.1145/1978542.1978558.
- Stojanovic, S. et al. A comparative study of selected hybrid and reconfigurable architectures. In *Proceedings of the IEEE ICIT Conference*, (Kos, Greece, Mar. 2012).
- Turkington, K. et al. FPGA-based acceleration of the LINPACK benchmark: A high level code transformation approach. In *Proceedings of the IEEE International Conference on Field Programmable Logic and Applications* (Madrid, Spain, Aug. 2006), 375–380.
- Vardi, M.Y. Is Moore's party over? *Commun. ACM* 54, 11 (Nov. 2011); DOI: 10.1145/2018396.2018397.
- Weston, S. et al. Rapid computation of value and risk for derivatives portfolio. *Concurrency and Computation: Practice and Experience, Special Issue* (July 2011); DOI: 10.1002/cpe.1778.
- Wolter, N. et al. What's working in HPC: Investigating HPC user behavior and productivity. *CTWatch Quarterly* (Nov. 2006).

Michael J. Flynn (flynn@ee.stanford.edu) is a professor of electrical engineering at Stanford University, CA.

Oskar Mencer (o.mencer@imperial.ac.uk) is a senior lecturer in the department of computing at Imperial College, London, U.K.

Veljko Milutinovic (vm@etf.rs) is a professor in the department of computer engineering at the University of Belgrade, Serbia.

Goran Rakocevic (grakocevic@gmail.com) is a research assistant at the Mathematical Institute of the Serbian Academy of Sciences and Arts in Belgrade, Serbia.

Per Stenstrom (pers@chalmers.se) is a professor of computer engineering at Chalmers University of Technology, Sweden.

Roman Trobec (roman.trobec@ijs.si) is an associate professor at the Jožef Stefan Institute, Slovenia.

Mateo Valero (mateo.valero@bsc.es) is the director of the Barcelona Supercomputing Centre, Spain.

This research was supported by discussions at the Barcelona Supercomputing Centre, during the FP7 EESI Final Project Meeting. The strategic framework for this work was inspired by Robert Madelin and Mario Campolargo of the EC, and was presented in the keynote of the EESI Final Project Meeting. The work of V. Milutinovic and G. Rakocevic was partially supported by the iii44006 grant of the Serbian Ministry of Science.

Copyright held by author.

interactions

EXPERIENCES | PEOPLE | TECHNOLOGY



interactions' website interactions.acm.org, is designed to capture the influential voice of its print component in covering the fields that envelop the study of people and computers.

The site offers a rich history of the conversations, collaborations, and discoveries from issues past, present, and future.

Check out the current issue, follow our bloggers, look up a past prototype, or discuss an upcoming trend in the communities of design and human-computer interaction.

FEATURES

BLOGS

FORUMS

DOWNLOADS

interactions.acm.org

Association for
Computing Machinery



Article development led by [acmqueue](http://acmqueue.queue.acm.org)
queue.acm.org

Google ads, black names and white names, racial discrimination, and click advertising.

BY LATANYA SWEENEY

Discrimination in Online Ad Delivery

DO ONLINE ADS suggestive of arrest records appear more often with searches of black-sounding names than white-sounding names? What is a black-sounding name or white-sounding name, anyway? How do you design technology to reason about societal consequences like structural racism? Let's take a scientific dive into online ad delivery to find answers.

"Have you ever been arrested?" Imagine this question appearing whenever someone enters your name in a search engine. Perhaps you are in competition for an award or a new job, or maybe you are in a position of trust, such as a professor or a volunteer. Perhaps you are dating or engaged in any one of hundreds of circumstances for which someone wants to learn more about you online. Appearing alongside your accomplishments is an advertisement implying you may have a criminal record, whether you actually have one or not. Worse, the ads may not appear for your competitors.

Employers frequently ask whether applicants have ever been arrested or charged with a crime, but if an employer disqualifies a job applicant based solely upon information indicating an arrest record, the company may face legal consequences. The U.S. Equal Employment Opportunity Commission (EEOC) is the federal agency charged with enforcing Title VII of the Civil Rights Act of 1964, a law that applies to most employers, prohibiting employment discrimination based on race, color, religion, sex, or national origin, and extended to those having criminal records.^{5,11} Title VII does not prohibit employers from obtaining criminal background information, but a blanket policy of excluding applicants based solely upon information indicating an arrest record can result in a charge of discrimination.

To make a determination, the EEOC uses an adverse impact test that measures whether certain practices, intentional or not, have a disproportionate effect on a group of people whose defining characteristics are covered by Title VII. To decide, you calculate the percentage of people affected in each group and then divide the smaller value by the larger to get the ratio and compare the result to 80. If the ratio is less than 80, then the EEOC considers the effect disproportionate and may hold the employer responsible for discrimination.⁶

What about online ads suggesting someone with your name has an arrest record? Title VII only applies if you have an arrest record and can prove the employer inappropriately used the ads.

Are the ads commercial free speech—a constitutional right to display the ad associated with your name? The First Amendment of the U.S. Constitution protects advertising, but the U.S. Supreme Court set out a test for assessing restrictions on commercial speech, which begins by determining whether the speech is misleading.³ Are online ads suggesting the existence of an arrest record misleading if no one by that name has an arrest record?



72

69

66

63

80

70

60

cat interrupts
Timer Inter.
2319 ARC VUAR
6728 eth0
1409 105244
1470 1052

locks
dbadv

$L_{\alpha}^{T,S}$

$GAD_{\alpha}(x')$

$L_{\alpha}^{T,S}$

$GAD_{\alpha}(x')$

HEAVY

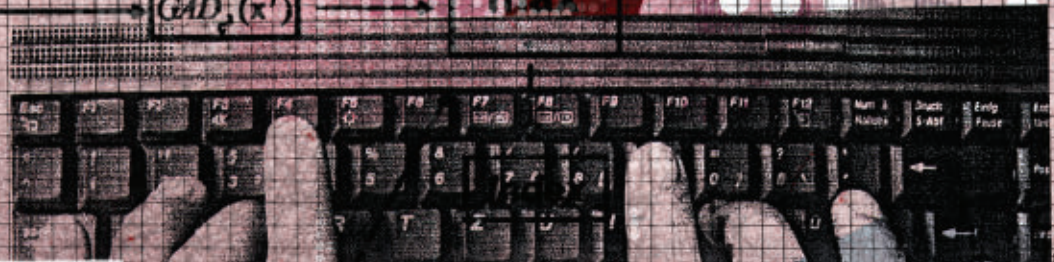


Figure 1. Ads from a Google search of three different names beginning with first name "Latanya."

Ads related to latanya farrell

[Latanya Farrell, Arrested?](#)
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

[Latanya Farrell](#)
www.publicrecords.com/
Public Records Found For: Latanya Farrell. View Now.

(a)

instantcheckmate DASHBOARD EDIT ACCOUNT INFO LOGOUT

LATANYA FARRELL
40 Lexington Htz
West Hartford, CT 06119
DOB: Jun 03, 1972 (40 years old)

Personal
Name, aliases, birthdate, phone numbers, etc.

Location
Detailed address history and related data, maps, etc.

Related Persons
Known family members, business associates, roommates, etc.

Marriage / Divorce
Marriage and divorce records and file...

Criminal History
Arrest records, speeding tickets, mugshots, etc.

Licenses
FAA Licenses, DEA Licenses, Other Licenses, etc.

Sex Offenders
Sex offenders living near Latanya Farrell's primary location.

Criminal History Rate This Content: ☆☆☆☆☆
This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.
We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Latanya Farrell has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.

Possible Matching Arrest Records

Name	County and State	Offenses	View Details
No matching arrest records were found.			

(b)

Ads by Google

[Latanya Sweeney, Arrested?](#)
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

[Latanya Sweeney](#)
Public Records Found For: Latanya Sweeney. View Now.
www.publicrecords.com/

[La Trova](#)
Search for La Tanya Look Up Fast Results neel
www.asik.com/Le+Tanya

(c)

instantcheckmate DASHBOARD EDIT ACCOUNT INFO LOGOUT

LATANYA SWEENEY
1420 Centis Ave
Pittsburgh, PA 15216
DOB: Oct 27, 1969 (53 years old)

Personal
Name, aliases, birthdate, phone numbers, etc.

Location
Detailed address history and related data, maps, etc.

Related Persons
Known family members, business associates, roommates, etc.

Marriage / Divorce
Marriage and divorce records and file...

Criminal History
Arrest records, speeding tickets, mugshots, etc.

Licenses
FAA Licenses, DEA Licenses, Other Licenses, etc.

Sex Offenders
Sex offenders living near Latanya Sweeney's primary location.

Criminal History Rate This Content: ☆☆☆☆☆
This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.
We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Latanya Sweeney has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.

Possible Matching Arrest Records

Name	County and State	Offenses	View Details
No matching arrest records were found.			

(d)

Assume the ads are free speech: what happens when these ads appear more often for one racial group than another? Not everyone is being equally affected by free speech. Is that free speech or racial discrimination?

Racism, as defined by the U.S. Commission on Civil Rights, is “any attitude, action, or institutional structure which subordinates a person or group because of their color.”¹⁶ *Racial discrimination* results when a person or group of people is treated differently based on their racial origins, according to the Panel on Methods for Assessing Discrimination of the National Research Council.¹² Power is a necessary precondition, for it depends on the ability to give or withhold benefits, facilities, services, and opportunities from someone who should be entitled to them and is denied on the basis of race. *Institutional or structural racism*, as defined in *The Social Work Dictionary*, is a system of procedures/patterns whose effect is to foster discriminatory outcomes or give preferences to members of one group over another.¹

These considerations frame the relevant socio-legal landscape. Now we turn to whether online ads suggestive of arrest records appear more often for one racial group than another among a sample of racially associated names, and if so, how technology can solve the problem.

The Pattern

What is the suspected pattern of ad delivery? Here is an overview using real-world examples.

Earlier this year, a Google search for *Latanya Farrell*, *Latanya Sweeney*, and *Latanya Lockett* yielded ads and criminal reports like those shown in Figure 1. The ads appeared on Google.com (Figure 1a, 1c) and on a news website, Reuters.com, to which Google supplies ads (Figure 1c), All the ads in question linked to instantcheckmate.com (Figure 1b, 1d). The first ad implied *Latanya Farrell* might have been arrested. Was she? Clicking on the link and paying the requisite fee revealed the company had no arrest record for her or *Latanya Sweeney*, but there is a record for *Latanya Lockett*.

In comparison, searches for *Kristen Haring*, *Kristen Sparrow*, and *Kristen Lindquist* did not yield any instant-

checkmate.com ads, even though the company's database reported having records for all three names and arrest records for *Sparrow* and *Lindquist*.

Searches for *Jill Foley*, *Jill Schneider*, and *Jill James* displayed instantcheckmate.com ads with neutral copy; the word *arrest* did not appear in the ads even though arrest records for all three names appeared in the company's database. Figure 2 shows ads appearing on Google.com and Reuters.com and criminal reports from instantcheckmate.com for the first two names.

Finally, we considered a proxy for race associated with these names. Figure 3 shows racial distinction in Google image search results for *Latanya*, *Lati-sha*, *Kristen*, and *Jill*, respectively. The faces associated with *Latanya* and *Lati-sha* tend to be black, while white faces dominate the images of *Kristen* and *Jill*.

These handpicked examples describe the suspected pattern: ads suggesting arrest tend to appear with names associated with blacks, and neutral or no ads appear with names associated with whites, regardless of whether the company placing the ad has an arrest record associated with the name.

Google AdSense

Who generates the ad's text? Who decides when and where an ad will appear? What is the relationship among Google, a news website such as Reuters, and Instant Checkmate in the previous examples? An overview of Google AdSense, the program that delivered the ads, provides the answers.

In printed newspapers, everyone who reads the publication sees the same ad in the same space. Online ads can be tailored to the reader's search criteria, interests, geographical location, and so on. Any two readers (or even the same reader returning to the same website) might view different ads.

Google AdSense is the largest provider of dynamic online advertisements, placing ads for millions of sponsors on millions of websites.⁹ In the first quarter of 2011, Google earned \$2.43 billion through Google AdSense.¹⁰ Several different advertising arrangements exist, but for simplicity this article describes only those features of Google AdSense specific to the Instant Checkmate ads in question.

Figure 2. Ad from a search of three different names beginning with the first name "Jill."

(a) Ad related to Jill Schneider
Jill Schneider Art
 www.posters2prints.com/
 Custom Frame Prints and Canvas. Shop Now, SAVE Big + Free Shipping!

We Found Jill Schneider
 www.instantcheckmate.com/
 Current Phone, Address, Age & More. Instant & Accurate Jill Schneider
 10,256 people + find this page
 Reverse Lookup - Reverse Cell Phone Directory - Data Check - Property Records

Located: Jill Schneider
 www.instantcheckmate.com/
 Information found on Jill Schneider Jill Schneider found in database.

(b) **JILL SCHNEIDER**
 1707 75th St
 Kansas City, MO 64118
 DOB: Mar 31, 1963 (43 years old)

Criminal History
 Rate This Content: ☆☆☆☆☆
 This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.

Possible Matching Arrest Records

Name	County and State	Offenses	View Details
1 Jill E Schneider	WI Admin Office of Courts(DM) disposition	Criminal/Traffic	View Details
2 Jill E Schneider	WI Admin Office of Courts(DM)	Criminal/Traffic	View Details
3 Jill E Schneider	WI Admin Office of Courts(DM) disposition	Criminal/Traffic	View Details
4 Jill E Schneider	WI Admin Office of Courts(DM)	Criminal/Traffic	View Details

(c) Ad related to Jill James
Located: Jill James
 www.instantcheckmate.com/
 Information found on Jill James Jill James found in database.

(d) **JILL JAMES**
 105 Sandhede Ct
 Cary, NC 27513
 DOB: May 31, 1952 (54 years old)

Criminal History
 Rate This Content: ☆☆☆☆☆
 This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.

Possible Matching Arrest Records

Name	County and State	Offenses	View Details
1 Jill B James	NC Admin Office of Courts demographic criminal	Criminal/Traffic	View Details
2 Jill James	NC Admin Office of Courts demographic criminal	Criminal/Traffic	View Details
3 Jill James	Individual NC courts	Criminal/Traffic	View Details
4 Jill B James	Individual NC courts	Criminal/Traffic	View Details
5 Jill Pate James	Individual NC courts	Criminal/Traffic	View Details
6 Jill Pate James	NC Admin Office of Courts demographic criminal	Criminal/Traffic	View Details
7 Jill Kelly James	NC Admin Office of Courts demographic criminal	Criminal/Traffic	View Details
8 Jill Kelly James	Individual NC courts	Criminal/Traffic	View Details
9 Jill Rosamond James	NC Admin Office of Courts demographic infractions	Criminal/Traffic	View Details
10 Jill Rosamond James	NC Admin Office of Courts demographic criminal	Criminal/Traffic	View Details

When a reader enters search criteria in an enrolled website, Google AdSense embeds into the Web page of results ads believed to be relevant to the search. Figures 1 and 2 show ads delivered by Google AdSense in response to various *firstname lastname* searches.

An advertiser provides Google with search criteria, copies of possible ads to deliver, and a bid to pay if a reader clicks the delivered ad. (For convenience, this article conflates Google AdSense with the related Google Adwords.) Google operates a real-time auction across bids for the same search criteria based on a “quality score” for each bid. A quality score includes many factors such as the past performance of the ad and characteristics of the company’s website.¹⁰ The ad having the highest quality score appears first, the second-highest second, and so on, and Google may elect not to show any ad if it considers the bid too low or if showing the ad exceeds a threshold (For example, a maximum account total for the advertiser). The Instant Checkmate ads in figures 1 and 2 often appeared first among ads, implying Instant Checkmate ads had the highest quality scores.

A website owner wanting to “host” online ads enrolls in AdSense and modifies the website to send a user’s search criteria to Google and to display returning ads under a banner “Ads by Google” among search results. For example, Reuters.com hosts AdSense, and entering *Latanya Sweeney* in the

search bar generated a new Web page with ads under the banner “Ads by Google” (Figure 1c).

There is no cost for displaying an ad, but if the user actually clicks on the ad, the advertiser pays the auction price. This may be as little as a few pennies, and the amount is split between Google and the host. Clicking the *Latanya Sweeney* ad on Reuters.com (Figure 1c) would cause Instant Checkmate to pay its auction amount to Google, and Google would split the amount with Reuters.

Search Criteria

What search criteria did Instant Checkmate specify? Will ads be delivered for made-up names? Ads displayed on Google.com allow users to learn why a specific ad appeared. Clicking the circled “i” in the ad banner (for example, Figure 1c) leads to a Web page explaining the ads. Doing so for ads in figures 1 and 2 reveals that the ads appeared because the search criteria matched the exact first- and last-name combination searched.

So, the search criteria must consist of both first and last names; and the names should belong to real people because a company presumably bids on records it sells.

The next steps describe the systematic construction of a list of racially associated first and last names for real people to use as search criteria. Neither Instant Checkmate nor Google are presumed to have used such a list.

Rather, the list provides a qualified sample of names to use in testing ad-delivery systems.

Black- and White-Identifying Names

Black-identifying and white-identifying first names occur with sufficiently higher frequency in one race than the other.

In 2003 Marianne Bertrand and Sendhil Mullainathan of the National Bureau of Economic Research (NBER) conducted an experiment in which they provided resumes to job posts that were virtually identical, except some of the resumes had black-identifying names and others had white-identifying names. Results showed white names received 50% more interviews.²

The study used names given to black and white babies in Massachusetts between 1974 and 1979, defining black-identifying and white-identifying names as those that have the highest ratio of frequency in one racial group to frequency in the other racial group.

In the popular book *Freakonomics*, Steven Levitt and Stephen Dubner report the top 20 whitest- and blackest-identifying girl and boy names. The list comes from earlier work by Levitt and Roland Fryer, which shows a pattern change in the way blacks named their children starting in the 1970s.⁷ It was compiled from names given to black and white children recorded in California birth records from 1961–2000 (more than 16 million births).

To test ad delivery, I combined the lists from these prior studies and added two black female names, *Latanya* and *Latisha*. Table 1 lists the names used here, consisting of eight for each of the categories: white female, black female, white male, and black male from the Bertrand and Mullainathan study (first row in Table 1); and the first eight names for each category from the Fryer and Levitt work (second row in Table 1). Emily, a white female name, Ebony, a black female name, and Darnell, a black male name, appear in both rows. The third row includes the observation shown in Figure 3. Removing duplicates leaves a total of 63 distinct first names.

Full Names of Real People

Web searches provide a means of locating and harvesting a real person’s first and last name (full name) by sampling

Table 1. Black-identifying names and white-identifying first names.

	White Female	Black Female	White Male	Black Male
(a)	Allison	Aisha	Brad	Darnell
	Anne	Ebony	Brendan	Hakim
	Carrie	Keisha	Geoffrey	Jermaine
	Emily	Kenya	Greg	Kareem
	Jill	Latonya	Brett	Jamal
	Laurie	Lakisha	Jay	Leroy
	Kristen	Latoya	Matthew	Rasheed
	Meredith	Tamika	Neil	Tremayne
(b)	Molly	Imani	Jake	DeShawn
	Amy	Ebony*	Connor	DeAndre
	Claire	Shanice	Tanner	Marquis
	Emily*	Aaliyah	Wyatt	Darnell*
	Katie	Precious	Cody	Terrell
	Madeline	Nia	Dustin	Malik
	Katelyn	Deja	Luke	Trevon
	Emma	Diamond	Jack	Tyrone
(c)		Latanya		
		Latisha		

names of professionals appearing on the Web; and sampling names of people active on social media sites and blogs (netizens).

Professionals often have their own Web pages that list positions and describe prior accomplishments. Several professions have degree designations (for example, Ph.D., M.D., J.D., or MBA) associated with people in that profession. A Google search for a first name and a degree designation can yield lists of people having that first name and degree.

The next step is to visit the Web page associated with each full name, and if an image is discernible, record whether the person appears black, white, or other.

Here are two examples from my test. A Google search for *EbonyPhD* revealed links for real people having *Ebony* as a first name—specifically, *Ebony Bookman*, *Ebony Glover*, *Ebony Baylor*, and *Ebony Utley*. I harvested the full names appearing on the first three pages of search results, using searches with other degree designations to find at least 10 full names for *Ebony*. Clicking on the link associated with *Ebony Glover* displayed an image.⁸ The *Ebony Glover* in this study appeared black.

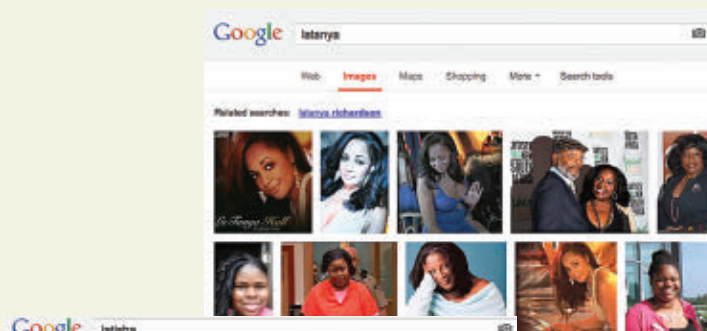
Similarly, search results for *JillPhD* listed professionals whose first name is *Jill*. Visiting links yielded Web pages with more information about each person. For example, *Jill Schneider*'s Web page had an image showing that she is white.¹⁴

PeekYou searches were used to harvest a sample of full names of netizens having racially associated first names. The website peekyou.com compiles online and offline information on individuals—thereby connecting residential information with Facebook and Twitter users, bloggers, and others—then assigns its own rating to reflect the size of each person's online footprint. Search results from peekyou.com list people having the highest score first, and include an image of the person.

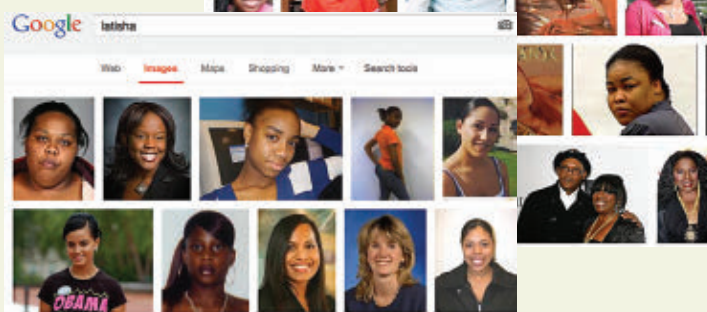
A PeekYou search of *Ebony* listed *Ebony Small*, *Ebony Cams*, *Ebony King*, *Ebony Springer*, and *Ebony Tan*. A PeekYou search for *Jill* listed *Jill Christopher*, *Jill Spivack*, *Jill English*, *Jill Pantozzi*, and *Jill Dobson*. After harvesting these and other full names, I reported the race of the person if discernible.

Figure 3. Image search results for first names Latanya, Latisha, Kirsten, and Jill.

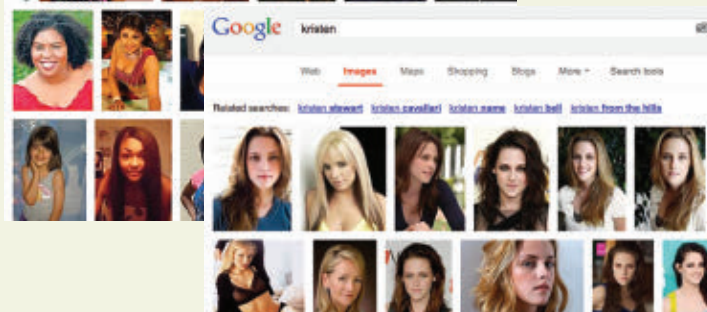
(a)



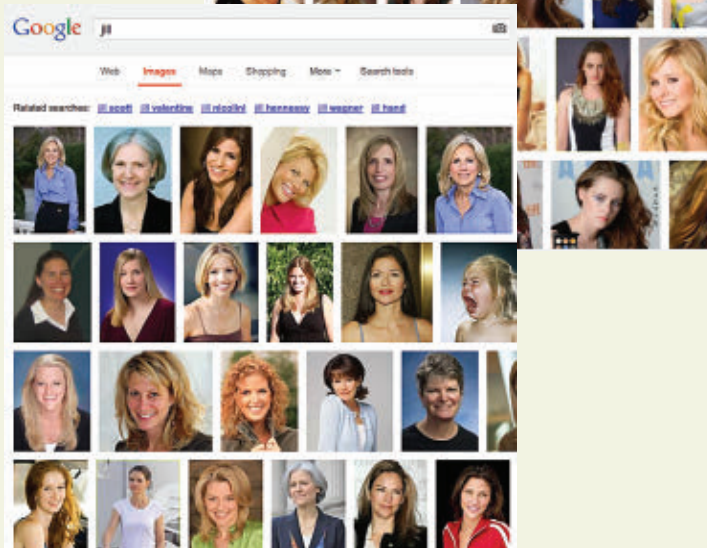
(b)



(c)



(d)



Armed with the approach just described, I harvested 2,184 racially associated full names of people with an online presence from September 24 through October 22, 2012. Most images associated with black-identifying names were of black people (88%),

and an even greater percentage of images associated with white-identifying names were of white people (96%).¹⁵

Google searches of first names and degree designations were not as productive as first name lookups on PeekYou. On Google, white male

names, *Cody*, *Connor*, *Tanner*, and *Wyatt* retrieved results with those as last names rather than first names; the black male name, *Kenya*, was confused with the country; and black names *Aaliyah*, *Deja*, *Diamond*, *Hakim*, *Malik*, *Marquis*, *Nia*, *Precious*, and *Rasheed* retrieved fewer than 10 full names. Only *Diamond* posed a problem with PeekYou searches, seemingly confused with other online entities. *Diamond* was therefore excluded from further consideration.

Some black first names had perfect predictions (100%): *Aaliyah*, *DeAndre*, *Imani*, *Jermaine*, *Lakisha*, *Latoya*, *Malik*, *Tamika*, and *Trevon*. The worst predictors of blacks were *Jamal* (48%) and *Leroy* (50%). Among white first names, 12 of 31 names made perfect predictions: *Brad*, *Brett*, *Cody*, *Dustin*, *Greg*, *Jill*, *Katelyn*, *Katie*, *Kristen*, *Matthew*, *Tanner*, and *Wyatt*; the worst predictors of whites were *Jay* (78%) and *Brendan* (83%). These findings strongly support the use of these names as racial indicators in this study.

Sixty-two full names appeared in the list twice even though the people were not necessarily the same. No name appeared more than twice. Overall, Google and PeekYou searches tended to yield different names.

Ad Delivery

With this list of names suggestive of race, I was ready to test which ads appear when these names are searched. To do this, I examined ads delivered on two sites, Google.com and Reuters.com, in response to searches of each full name, once at each site. The browser's cache and cookies were cleared before each search, and copies of Web pages received were preserved. Figures 1, 2, 5, and 6 provide examples.

From September 24 through October 23, 2012, I searched 2,184 full names on Google.com and Reuters.com. The searches took place at different times of day, different days of the week, with different IP and machine addresses operating in different parts of the United States using different browsers. I manually searched 1,373 of the names and used automated means¹⁷ for the remaining 812 names. Here are nine observations.

1. *Fewer ads appeared on Google.com than Reuters.com*—about five times



Of the more than 2,000 names searched, 78% had at least one ad for public records about the person being searched.



fewer. When ads did appear on Google.com, typically only one ad showed, compared with three ads routinely appearing on Reuters.com. This suggests Google may be sensitive to the number of ads appearing on Google.com.

2. *Of 5,337 ads captured, 78% were for government-collected information (public records) about the person whose name was searched.* Public records in the U.S. often include a person's address, phone number, and criminal history. Of the more than 2,000 names searched, 78% had at least one ad for public records about the person being searched.

3. *Four companies had more than half of all the ads captured.* These companies were Instant Checkmate, PublicRecords (which is owned by Intelius), PeopleSmart, and PeopleFinders, and all their ads were selling public records. Instant Checkmate ads appeared more than any other: 29% of all ads. Ad distribution was different on Google's site; Instant Checkmate still had the most ads (50%), but Intelius.com, while not in the top four overall, had the second most ads on Google.com. These companies dominate the advertising space for online ads selling public records.

4. *Ads for public records on a person appeared more often for those with black-associated names than white-associated names, regardless of company.* PeopleSmart ads appeared disproportionately higher for black-identifying names—41% as opposed to 29% for white names. PublicRecords ads appeared 10% more often for those with black first names than white. Instant Checkmate ads displayed only slightly more often for black-associated names (2% difference). This is an interesting finding and it spawns the question: Public records contain information on everyone, so why more ads for black-associated names?

5. *Instant Checkmate ads dominated the topmost ad position.* They occupied that spot in almost half of all searches on Reuters.com. This suggests Instant Checkmate offers Google more money or has higher quality scores than do its competitors.

6. *Instant Checkmate had the largest percentage of ads in virtually every first-name category, except for Kristen, Connor, and Tremayne.* For those names, Instant Checkmate had uncharacteristically fewer ads (less than 25%). Pub-

licRecords had ads for 80% of names beginning with *Tremayne*, and *Connor*, and 58% for *Kristen*, compared to 20% and less for Instant Checkmate. Why the underrepresentation in these first names? During a conference call with company's representatives, they asserted that Instant Checkmate gave the same ad text to Google for groups of last names (not first names).

7. *Almost all ads for public records included the name of the person, making each ad virtually unique, but beyond personalization, the ad templates showed little variability.* The only exception was Instant Checkmate. Almost all PeopleFinder ads appearing on Reuters.com used the same personalized template. PublicRecords used five templates and PeopleSmart seven, but Instant Checkmate used 18 different ad templates on Reuters.com. Figure 4 enumerates ad templates for frequencies of 10 or more for all four companies (replace fullname with the person's first and last name).

While Instant Checkmate's competitors also sell criminal history information, only Instant Checkmate ads used the word *arrest*.

8. *A greater percentage of Instant Checkmate ads using the word "arrest" appeared for black-identifying first names than for white first names.* More than 1,100 Instant Checkmate ads appeared on Reuters.com, with 488 having black-identifying first names; of these, 60% used *arrest* in the ad text. Of the 638 ads displayed with white-identifying names, 48% used *arrest*. This difference is statistically significant, with less than a 0.1% probability that the data can be explained by chance (chi-square test: $X^2(1)=14.32, p < 0.001$). The EEOC's and U.S. Department of Labor's adverse impact test for measuring discrimination is 77 in this case, so if this were an employment situation, a charge of discrimination might result. (The adverse impact test uses the ratio of neutral ads, or 100 minus the percentages given, to compute disparity: $100-60=40$ and $100-48=52$; dividing 40 by 52 equals 77.)

The highest percentage of neutral ads (where the word *arrest* does not appear in ad text) on Reuters.com were those for *Jill* (77%) and *Emma* (75%), both white-identifying names. Names receiving the highest percentage of ads with *arrest* in the text were *Darnell*

(84%), *Jermaine* (81%), and *DeShawn* (86%), all black-identifying first names. Some names appeared counter to this pattern: *Dustin*, a white-identifying name, generated *arrest* ads in 81% of searches; and *Imani*, a black-identifying name, resulted in neutral ads in 75% of searches.

9. *Discrimination results on Google's site were similar, but, interestingly, ad text and distributions were different.* While the same neutral and *arrest* ads having dominant appearances on Reuters.com also appeared frequently on Google.com, Instant Checkmate ads on Google included an additional 10 templates, all using the word *criminal* or *arrest*.

More than 400 Instant Checkmate ads appeared on Google, and 90% of these were suggestive of *arrest*, regardless of race. Still, a greater percentage of Instant Checkmate ads suggestive of *arrest* displayed for black-associated first names than for whites. Of the 366

ads that appeared for black-identifying names, 92% were suggestive of *arrest*. Far fewer ads displayed for white-identifying names (66 total), but 80% were suggestive of *arrest*. This difference in the ratios 92 and 80 is statistically significant, with less than a 1% probability that the data can be explained by chance (chi-square test: $X^2(1)=7.71, p < 0.01$). The EEOC's adverse impact test for measuring discrimination is 40%, so if this were employment, a charge of discrimination might result. (The adverse impact test gives $100-92=8$ and $100-80=20$; dividing 8 by 20 equals 40.)

A greater percentage of Instant Checkmate ads having the word *arrest* in ad text appeared for black-identifying first names than for white-identifying first names within professional and netizen subsets, too. On Reuters.com, which hosts Google AdSense ads, a black-identifying name was 25% more likely to generate an ad suggestive of an *arrest* record.

Figure 4. Template for ads for public records on Reuters for frequencies less than 10. Full list is available.¹⁵

instantcheckmate		Peoplesmart	
382	Located: fullname Information found on <i>fullname</i> <i>fullname</i> found in database.	87	We found: fullname 1) Get Aisha's Background Report 2) Current Contact Info—Try Free!
96	We found fullname Search Arrests, Address, Phone, etc. Search records for <i>fullname</i> .	105	We found: fullname 1) Contact <i>fullname</i> —Free Info! 2) Current Address, Phone & More.
40	Background of fullname Search Instant Checkmate for the Records of <i>fullname</i>	348	We found: fullname 1) Contact <i>fullname</i> —Free Info! 2) Current Phone, Address & More.
17	fullname's Records 1) Enter Name and State. 2) Access Full Background Checks Instantly.	Publicrecords	
195	fullname: Truth Arrests and Much More. Everything About <i>fullname</i>	570	fullname Public Records Found For: <i>fullname</i> . View now.
67	fullname Truth Looking for <i>fullname</i> ? Check <i>fullname</i> 's Arrests	128	fullname Public Records Found For: <i>fullname</i> . Search now.
176	fullname, Arrested? 1) Enter Name and State. 2) Access Full Background Checks Instantly.	13	Records: fullname Database of all lastname's in the Country. Search now.
55	fullname Located Background Check, Arrest Records, Phone, & Address. Instant, Accurate	56	fullname We have Public Records For: <i>fullname</i> . Search Now.
62	Looking for fullname? Comprehensive Background Report and More on <i>fullname</i>	Peoplefinders	
		523	We found fullname Current Address, Phone and Age. Find <i>fullname</i> . Anywhere.

Figure 5. Senator Claire McCaskill's campaign ad appeared next to an ad using the word "arrest."

Figure 6. An assortment of ads appearing for Latisha Smith.

These findings reject the hypothesis that no difference exists in the delivery of ads suggestive of an arrest record based on searches of racially associated names.

Additional Observations

The people behind the names used in this study are diverse. Political figures included Maryland State Representatives Aisha Braveboy (arrest ad) and Jay Jacobs (neutral ad); Jill Biden (neutral ad), wife of U.S. Vice President Joe Biden; and Claire McCaskill, whose campaign ad for the U.S. Sen-

ate in Missouri appeared alongside an Instant Checkmate ad using the word *arrest* (Figure 5). Names mined from academic websites included graduate students, staff, and accomplished academics, such as Amy Gutmann, president of the University of Pennsylvania. Dustin Hoffman (arrest ad) was among names of celebrities used. A smorgasbord of athletes appeared, from local to national fame (assorted neutral and arrest ads). The youngest person whose name was used in the study was a missing 11-year-old black girl.

More than 1,100 of the names harvested for this study were from PeekYou, with scores estimating the name's overall presence on the Web. As expected, celebrities get the highest scores of 10s and 9s. Only four names used here had a PeekYou score of 10, and 12 had a score of 9, including Dustin Hoffman. Only two ads appeared for these high-scoring names; an abundance of ads appeared across the remaining spectrum of PeekYou scores. We might presume that the bid price needed to display an ad is greater for more popular names with higher PeekYou scores. Knowing that very few high-scoring people were in the study and that ads appeared across the full spectrum of PeekYou scores reduces concern about variations in bid prices.

Different Instant Checkmate ads sometimes appeared for the same person. About 200 names had Instant Checkmate ads on both Reuters.com and Google.com, but only 42 of these names received the same ad. The other 82% of names received different ads across the two sites. At most, three distinct ads appeared across Reuters.com and Google.com for the same name. Figure 6 shows the assortment of ads appearing for *Latisha Smith*. Having different possible ad texts for a name reminds us that while Instant Checkmate provided the ad texts, Google's technology selected among the possible texts in deciding which to display. Figure 6 shows ads both suggestive of arrest and not, though more ads appear suggestive of arrest than not.

More About the Problem

Why is this discrimination occurring? Is Instant Checkmate, Google, or society to blame? We do not yet know. Google understands that an advertiser

may not know which ad copy will work best, so the advertiser may provide multiple templates for the same search string, and the "Google algorithm" learns over time which ad text gets the most clicks from viewers. It does this by assigning weights (or probabilities) based on the click history of each ad. At first, all possible ad texts are weighted the same and are equally likely to produce a click. Over time, as people tend to click one ad copy over others, the weights change, so the ad text getting the most clicks eventually displays more frequently.

Did Instant Checkmate provide ad templates suggestive of arrest disproportionately to black-identifying names? Or did Instant Checkmate provide roughly the same templates evenly across racially associated names but users clicked ads suggestive of arrest more often for black-identifying names? As mentioned earlier, during a conference call with the founders of Instant Checkmate and their lawyer, the company's representatives asserted that Instant Checkmate gave the same ad text to Google for groups of last names (not first names) in its database; they expressed no other criteria for name and ad selection.

This study is a start, but more research is needed. To preserve research opportunities, I captured additional results for 50 hits on 2,184 names across 30 Web sites serving Google Ads to learn the underlying distributions of ad occurrences per name. While analyzing the data may prove illuminating, in the end the basic message presented in this study does not change: there is discrimination in delivery of these ads.

Technical Solutions

How can technology solve this problem? One answer is to change the quality scores of ads to discount for unwanted bias. The idea is to measure real-time bias in an ad's delivery and then adjust the weight of the ad accordingly at auction. The general term for Google's technology is *ad exchange*. This approach generalizes to other ad exchanges (not just Google's); integrates seamlessly into the way ad exchanges operate, allowing minimal modifications to harmonize ad deliveries with societal norms; and, works regardless of the cause of the discrimi-

nation—advertiser bias in placing ads or society bias in selecting ads.

Discrimination, however, is at the heart of online advertising. Differential delivery is the very idea behind it. For example, if young women with children tend to purchase baby products and retired men with bass boats tend to purchase fishing supplies, and you know the viewer is one of these two types, then it is more efficient to offer ads for baby products to the young mother and fishing rods to the fisherman, not the other way around.


On the other hand, not all discrimination is desirable. Societies have identified groups of people to protect from specific forms of discrimination. Delivering ads suggestive of arrest much more often for searches of black-identifying names than for white-identifying names is an example of unwanted discrimination, according to American social and legal norms. This is especially true because the ads appear regardless of whether actual arrest records exist for the names in the company's database.

The good news is that we can use the mechanics and legal criteria described earlier to build technology that distinguishes between desirable and undesirable discrimination in ad delivery. Here I detail the four key components:


1. *Identifying Affected Groups.* A set of predicates can be defined to identify members of protected and comparison groups. Given an ad's search string and text, a predicate returns *true* if the ad can impact the group that is the subject of the predicate and returns *false* otherwise. Statistics of baby names can identify first names for constructing race and gender groups and last names for grouping some ethnicities. Special word lists or functions that report degree of membership may be helpful for other comparisons.

In this study, ads appeared on searches of full names for real people, and first names assigned to more black or white babies formed groups for testing. These *black* and *white* predicates evaluate to *true* or *false* based on the first name of the search string.

2. *Specifying the Scope of Ads to Assess.* The focus should be on those ads capable of impacting a protected group in a form of discrimination prohibited by law or social norm. Protec-



Discrimination is at the heart of online advertising. Differential delivery is the very idea behind it.



tion typically concerns the ability to give or withhold benefits, facilities, services, employment, or opportunities. Instead of lumping all ads together, it is better to use search strings, ad texts, products, or URLs that display with ads to decide which ads to assess.

This study assessed search strings of first and last names of real people, ads for public records, and ads having a specific display URL (instantcheckmate.com), the latter being the most informative because the adverse ads all had the same display URL.

Of course, the audience for the ads is not necessarily the people who are the subject of the ads. In this study, the audience is a person inquiring about the person whose name is the subject of the ad. This distinction is important when thinking about the identity of groups that might be impacted by an ad. Group membership is based on the ad's search string and text. The audience may resonate more with a distinctly positive or negative characterization of the group.

3. *Determining Ad Sentiment.* Originally associated with summarizing product and movie reviews, sentiment analysis is an area of computer science that uses natural-language processing and text analytics to determine the overall attitude of a writing.¹³ Sentiment analysis can measure whether an ad's search string and accompanying text has positive, negative, or neutral sentiment. A literature search does not find any prior application to online ads, but a lot of research has been done assessing sentiment in social media (sentiment140.com, for example, reports the sentiment of tweets, which like advertisements have limited words).

In this study, ads containing the word *arrest* or *criminal* were classified as having negative sentiment; ads without those words were classified as neutral.

4. *Testing for Adverse Impact.* Consider a table where columns are comparative groups, rows are sentiment, and values are the number of ad impressions (the number of times an ad appears, though the ad is not necessarily clicked). Ignore neutral ads. Comparing the percentage of ads having the same positive or negative sentiment across groups reveals the degree to which one group may be impacted more or less by the ad's sentiment.

Table 2. Negative and neutral sentiments of black and white groups.

	Black		White	
Negative	291	60%	308	48%
Neutral	197	40%	330	52%
Positive				
Totals	488		638	

A chi-square test can determine statistical significance, and the adverse impact test used by the EEOC and the U.S. Department of Labor can alert whether in some circumstances legal risks may result.

In this study the groups are black and white, and the sentiments are negative and neutral. Table 2 shows a summary chart. Of the 488 ads that appeared for the black group, 291 (or 60%) had negative sentiment. Of the 638 ads displayed for the white group, 308 (or 48%) had negative sentiment. The difference is statistically significant ($X^2(1)=14.32, p < 0.001$) and has an adverse impact measure of (40/52), or 77%.

An easy way of incorporating this analysis into an ad exchange is to decide which bias test is critical (for example, statistical significance or adverse impact test) and then factor the test result into the quality score for the ad at auction. For example, if we were to modify the ad exchange not to display any ad having an adverse impact score of less than 80, which is the EEOC standard, then arrest ads for blacks would sometimes appear, but would not be overly disproportionate to whites, regardless of advertiser or click bias.

Though this study served as an example throughout, the approach generalizes to many other forms of discrimination and combats other ways ad exchanges may foster discrimination.

Suppose female names tend to get neutral ads such as “Buy now,” while male names tend to get positive ads such as “Buy now. 50% off!” Or suppose black names tend to get neutral ads such as “Looking for Ebony Jones,” while white names tend to get positive ads such as “Meredith Jones. Fantastic!” Then the same analysis would suppress some occurrences of the positive ads so as not to foster a discriminatory effect.

This approach does not stop the appearance of negative ads for a store

placed by a disgruntled customer or ads placed by competitors on brand names of the competition, unless these are deemed to be protected groups.


Nonprotected marketing discrimination can continue even to protected groups. For example, suppose search terms associated with blacks tend to get neutral ads for some music artists, while those associated with whites tend to get neutral ads for other music artists. All ads would appear regardless of the disproportionate distribution because the ads are not subject to suppression.

As a final example, this approach allows everyone to be negatively impacted as long as the impact is approximately the same. Suppose all ads for public records on all names, regardless of race, were equally suggestive of arrest and had almost the same number of impressions; then no ads suggestive of arrest would be suppressed.

Computer scientist Cynthia Dwork and her colleagues have been working on algorithms that assure racial fairness.⁴ Their general notion is to ensure similar groups receive similar ads in proportions consistent with the population. Utility is the critical concern with this direction because not all forms of discrimination are bad, and unusual and outlier ads could be unnecessarily suppressed. Still, their research direction looks promising.

In conclusion, this study demonstrates that technology can foster discriminatory outcomes, but it also shows that technology can thwart unwanted discrimination.

Acknowledgments

The author thanks Ben Edelman, Claudine Gay, Gary King, Annie Lewis, and weekly Topics in Privacy participants (David Abrams, Micah Altman, Merce Crosas, Bob Gelman, Harry Lewis, Joe Pato, and Salil Vadhan) for discussions; Adam Tanner for first suspecting a pattern; Diane Lopez and Matthew Fox in Harvard’s Office of the General Counsel for making publication possible in the face of legal threats; and Sean Hooley for editorial suggestions. Data from this study is available at foreverdata.org and the IQSS Dataverse Network. Supported in part by NSF grant CNS-1237235 and a gift from Google, Inc. 

Related articles on queue.acm.org

Modeling People and Places with Internet Photo Collections

David Crandall, Noah Snaveley
<http://queue.acm.org/detail.cfm?id=2212756>

Interactive Dynamics for Visual Analysis

Jeffrey Heer, Ben Shneiderman
<http://queue.acm.org/detail.cfm?id=2146416>

Social Perception

James L. Crowley
<http://queue.acm.org/detail.cfm?id=1147531r>

References

- Barker R. *The Social Work Dictionary* (5th ed.). NASW Press, Washington, DC, ss, 2003.
- Bertrand, M. and Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. NBER Working Paper No. 9873, 2003; <http://www.nber.org/papers/w9873>.
- Central Hudson Gas & Electric Corp. v. Public Service Commission of New York*. Supreme Court of the United States, 447 U.S. 557, 1980.
- Dwork, C., Hardt, M., et al. 2011. Fairness through awareness. arXiv:1104.3913; <http://arxiv.org/abs/1104.3913>.
- Equal Employment Opportunity Commission. Consideration of arrest and conviction records in employment decisions under Title VII of the Civil Rights Act of 1964. Washington, DC, 915.002, 2012. http://www.eeoc.gov/laws/guidance/arrest_conviction.cfm.
- Equal Employment Opportunity Commission. Uniform guidelines on employee selection procedures. Washington, DC, 1978.
- Fryer, R. and Levitt, S. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics* 59, 3 (2004); <http://pricetheory.uchicago.edu/levitt/Papers/FryerLevitt2004.pdf>.
- Glover, E.; <http://www.physiology.emory.edu/FIRST/ebony2.htm> (archived at <http://foreverdata.org/onlineads>).
- Google AdSense; <http://google.com/adsense>.
- Google. Google announces first quarter 2011 financial results; http://investor.google.com/earnings/2011/Q1_google_earnings.html.
- Harris, P. and Keller, K. Ex-offenders need not apply: The criminal background check in hiring decisions. *Journal of Contemporary Criminal Justice* 21, 1 (2005), 6-30.
- Panel on Methods for Assessing Discrimination, National Research Council. Measuring racial discrimination. National Academy Press, Washington, DC, 2004.
- Pang, B. and Lee, L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004).
- Schneider, J. <http://www.lehigh.edu/bio/jill.html> (Archived at <http://foreverdata.org/onlineads>).
- Sweeney, L. Discrimination in online ad delivery (2013). (For details, see full technical report at <http://ssrn.com/abstract=2208240>. Data, including Web pages and ads, archived at <http://foreverdata.org/onlineads>).
- U.S. Commission on Civil Rights. Racism in America and how to combat it. Washington, DC, 1970.
- WebShot Command Line Server Edition. Version 1.9.1.1; <http://www.websitescreenshots.com/>.

Latanya Sweeney (latanya@fas.harvard.edu) is professor of government and technology in residence at Harvard University. She creates and uses technology to assess and solve societal, political, and governance problems and teaches others how to do the same. She is also founder and director of the Data Privacy Lab at Harvard.

How can applications be built on eventually consistent infrastructure given no guarantee of safety?

BY PETER BAILIS AND ALI GHODSI

Eventual Consistency Today: Limitations, Extensions, and Beyond

IN A JULY 2000 conference keynote, Eric Brewer, now VP of engineering at Google and a professor at the University of California, Berkeley, publicly postulated the CAP (consistency, availability, and partition tolerance) theorem, which would change the landscape of how distributed storage systems were architected.⁸

Brewer's conjecture—based on his experiences building infrastructure for some of the first Internet search engines at Inktomi—states that distributed systems requiring always-on, highly available operation cannot guarantee the illusion of coherent, consistent single-system operation in the presence of network partitions, which cut communication between active servers. Brewer's conjecture proved prescient: in the following decade, with the continued rise of large-scale Internet services, distributed-system architects frequently dropped “strong” guarantees in favor of weaker models—the most notable being *eventual consistency*.

Eventual consistency provides few guarantees. Informally, it guarantees

that, if no *additional* updates are made to a given data item, all reads to that item will eventually return the same value. This is a particularly weak model. At no given time can the user rule out the possibility of inconsistent behavior: the system can return *any* data and still be eventually consistent—as it might “converge” at some later point. The only guarantee is that, at some point in the future, something good will happen. Yet, given this apparent lack of useful guarantees, scores of usable applications and profitable businesses are built on top of eventually consistent infrastructure. How?

This article begins to answer this question by describing several notable developments in the theory and practice of eventual consistency, with

a focus on immediately applicable takeaways for practitioners running distributed systems in the wild. As production deployments have increasingly adopted weak consistency models such as eventual consistency, we have learned several lessons about how to reason about, program, and strengthen these weak models.

In summary, we will primarily focus on three questions and some preliminary answers:

How eventual is eventual consistency? If the scores of system architects advocating eventual consistency are any indication, eventual consistency seems to work “well enough” in practice. How is this possible when it provides such weak guarantees? *New prediction and measurement techniques allow system architects to quantify the behavior of real-world eventually consistent systems. When verified via measurement, these systems appear strongly consistent most of the time.*

How should one program under eventual consistency? How can system architects cope with the lack of guarantees provided by eventual consistency? How do they program without strong ordering guarantees? *New research enables system architects to deal with inconsistencies, either via external compensation outside of the system or by limiting themselves to data structures that avoid inconsistencies altogether.*

Is it possible to provide stronger guarantees than eventual consistency without losing its benefits? In addition to guaranteeing eventual consistency and high availability, what other guar-

antees can be provided? *Recent results show that it is possible to achieve the benefits of eventual consistency while providing substantially stronger guarantees, including causality and several ACID (atomicity, consistency, isolation, durability) properties from traditional database systems while still remaining highly available.*

This article is *not* intended as a formal survey of the literature surrounding eventual consistency. Rather, it is a pragmatic introduction to several developments on the cutting edge of our understanding of eventually consistent systems. The goal is to provide the necessary background for understanding both *how* and *why* eventually consistent systems are programmed, deployed, and have evolved, as well as where the systems of tomorrow are heading.

History and Concepts of Eventual Consistency

Brewer’s CAP theorem dictates it is impossible simultaneously to achieve always-on experience (*availability*) and to ensure users read the latest written version of a distributed database (*consistency*—as formally proven, a property known as “linearizability”¹¹) in the presence of partial failure (*partitions*).⁸ CAP pithily summarizes trade-offs inherent in decades of distributed-system designs (for example, RFC 677¹⁴ from 1975) and shows that maintaining an SSI (single-system image) in a distributed system has a cost.¹⁰ If two processes (or groups of processes) within a distributed system cannot communicate (are *parti-*

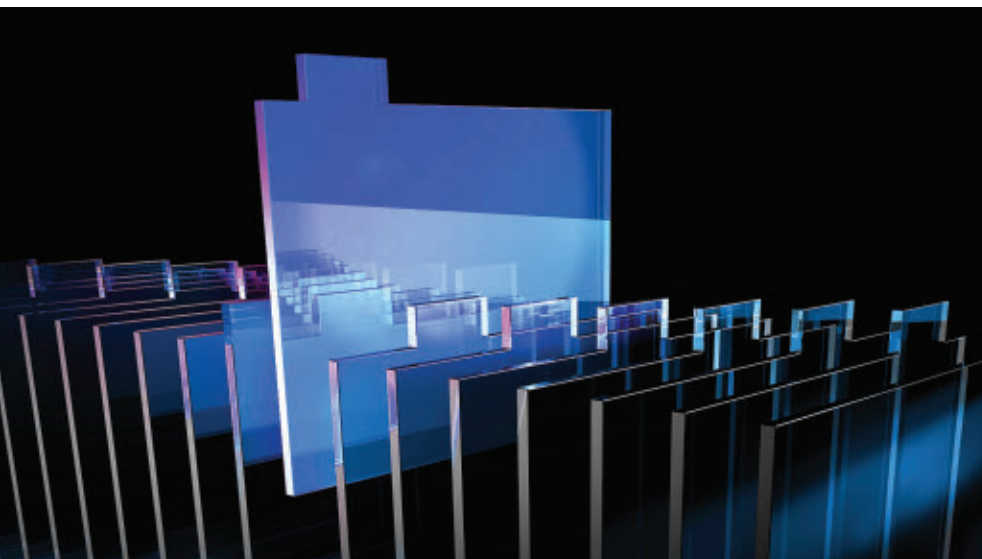
tioned)—either because of a network failure or the failure of one of the components—then updates cannot be synchronously propagated to all processes without blocking. Under partitions, an SSI system cannot safely complete updates and hence presents unavailability to some or all of its users. Moreover, even without partitions, a system that chooses availability over consistency enjoys benefits of low latency: if a server can safely respond to a user’s request when it is partitioned from all other servers, then it can *also* respond to a user’s request without contacting other servers even when it is able to do so.¹ (Note that you cannot “sacrifice” partition tolerance!¹² The choice is between consistency and availability.)

As services are increasingly replicated to provide fault tolerance (ensuring services remain online despite individual server failures) and capacity (to allow systems to scale with variable request rates), architects must face these consistency-availability and -latency trade-offs head on. In a dynamic, partitionable Internet, services requiring guaranteed low latency must often relax their expectations of data consistency.

Eventual consistency as an available alternative. Given the CAP impossibility result, distributed-database designers sought weaker consistency models that would enable both availability and high performance. While weak consistency has been studied and deployed in various forms since the 1970s,¹⁹ the eventual consistency model has become prominent, particularly among emerging, highly scalable NoSQL stores.

One of the earliest definitions of eventual consistency comes from a 1988 paper describing a group communication system¹⁵ not unlike a shared text editor such as Google Docs today: “...changes made to one copy eventually migrate to all. If all update activity stops, after a period of time all replicas of the database will converge to be logically equivalent: each copy of the database will contain, in a predictable order, the same documents; replicas of each document will contain the same fields.”

Under eventual consistency, all servers eventually “converge” to the same state; at some point in the future, servers are indistinguishable from one an-




other. This eventual convergence, however, does not provide SSI semantics. First, the “predictable order” will not necessarily correspond to an execution that could have arisen under SSI; eventual consistency does not specify which value is eventually chosen. Second, there is an unspecified window before convergence is reached, during which the system will not provide SSI semantics, but rather arbitrary values. As we will illustrate, this promise of eventual convergence is a rather weak property. Finally, a system with SSI provides eventual consistency—the “eventuality” is immediate—but not vice versa.


Why is eventual consistency useful? Pretend you are in charge of the data infrastructure at a social network where users post new status updates that are sent to their followers’ timelines, represented by separate lists—one per user. Because of large scale and frequent server failures, the database of timelines is stored across multiple physical servers. In the event of a partition between two servers, however, you cannot deliver each update to all timelines. What should you do? Should you tell the user he or she cannot post an update, or should you wait until the partition heals before providing a response? Both of these strategies choose consistency over availability, at the cost of user experience.

Instead, what if you propagate the update to the reachable set of followers’ timelines, return to the user, and delay delivering the update to the other followers until the partition heals? In choosing this option, you give up the guarantee that all users see the same set of updates at every point in time (and admit the possibility of timeline reordering as partitions heal), but you gain high availability and (arguably) a better user experience. Moreover, because updates are eventually delivered, all users eventually see the same timeline with all of the updates that users posted.

Implementing eventual consistency. A key benefit of eventual consistency is that it is fairly straightforward to implement. To ensure convergence, replicas must exchange information with one another about which writes they have seen. This information exchange is often called *anti-entropy*, a homage to the process of reversing entropy, or thermodynamic randomness, in a



Under eventual consistency, all servers eventually “converge” to the same state; at some point in the future, servers are indistinguishable from one another.



physical system.¹⁹ Protocols for achieving anti-entropy take a variety of forms; one simple solution is to use an asynchronous all-to-all broadcast: when a replica receives a write to a data item, it immediately responds to the user, then, in the background, sends the write to all other replicas, which in turn update their locally stored data items. In the event of concurrent writes to a given data item, replicas deterministically choose a “winning” value, often using a simple rule such as “last writer wins” (for example, via a clock value embedded in each write).²²

Suppose you want to make a single-node database into an eventually consistent distributed database. When you get a request, you route it to any server you can contact. When a server performs a write to its local key-value store, it can send the write to all other servers in the cluster. This write-forwarding becomes the anti-entropy process. Be careful, however, when sending the write to the other servers. If you wait for other servers to respond before acknowledging the local write, then, if another server is down or partitioned from you, the write request will hang indefinitely. Instead, you should send the request in the background; anti-entropy should be an asynchronous process. Implicitly, the model for eventual consistency assumes system partitions are eventually healed and updates are eventually propagated, or that partitioned nodes eventually die and the system ends up operating in a single partition.

The eventually consistent system has some great properties. It does not require writing difficult “corner-case” code to deal with complicated scenarios such as downed replicas or network partitions—anti-entropy will simply stall—or writing complex code for coordination such as master election. All operations complete locally, meaning latency will be bounded. In a geo-replicated scenario, with replicas located in different data centers, you do not have to endure long-haul wide-area network latencies on the order of hundreds of milliseconds on the request fast path. The mechanism just described, returning immediately on the local write, can put data durability at risk. An intermediate point in trading between durability and availability is to return after W


replicas have acknowledged the write, thus allowing the write to survive W-1 replica failures. Anti-entropy can be run as often or as rarely as desired without violating any guarantees. What's not to like?

Safety and liveness. While eventual consistency is easy to achieve, the current definition leaves some unfortunate holes. First, what is the eventual state of the database? A database always returning the value 42 is eventually consistent, even if 42 were never written. Amazon CTO Werner Vogels' preferred definition specifies that "eventually all accesses return the last updated value;" accordingly, the database cannot converge to an arbitrary value.²³ Even this new definition has another problem: what values can be returned before the eventual state of the database is reached? If replicas have not yet converged, what guarantees can be made on the data returned?


These questions stem from two kinds of properties possessed by all distributed systems: safety and liveness.² A *safety* property guarantees that "nothing bad happens;" for example, every value that is read was, at some point in time, written to the database. A *liveness* property guarantees that "something good eventually happens;" for example, all requests eventually receive a response.

The difficulty with eventual consistency is that it makes no safety guarantees—eventual consistency is purely a liveness property. Something good eventually happens—the replicas agree—but there are no guarantees with respect to what happens, and no behavior is ruled out in the meantime! For meaningful guarantees, safety and liveness properties need to be taken together: without one or the other, you can have trivial implementations that provide less-than-satisfactory results.

Virtually every other model that is stronger than eventual provides some form of safety guarantees. For almost all production systems, however, eventual consistency should be considered a bare-minimum requirement for data consistency. A system that does not guarantee replica convergence is remarkably difficult to reason about.



The difficulty with eventual consistency is that it makes no safety guarantees—eventual consistency is purely a liveness property.



How Eventual is Eventual Consistency?

Despite the lack of safety guarantees, eventually consistent data stores are widely deployed. Why? While eventually consistent stores do not promise safety, there is evidence that eventual consistency works well in practice. Eventual consistency is "good enough," given its latency and availability benefits. For the many stores that offer a choice between eventual consistency and stronger consistency models, scores of practitioners advocate eventual consistency.

The behavior of eventually consistent stores can be quantified. Just because eventual consistency does not promise safety does not mean safety is not often provided—and you can both measure and predict these properties of eventually consistent systems using a range of techniques that have been recently developed and are making their way to production stores. These techniques—which we discuss next—have surprisingly shown that eventual consistency often behaves like strong consistency in production stores.

Metrics and mechanisms. One common metric for eventual consistency is *time*: how long will it take for writes to become visible to readers? This captures the "window of consistency" measured according to the wall clock. Another metric is *versions*: how many versions old will a given read be? This information can be used to ensure readers never go back in time and observe progressively newer versions of the database. While time and versions are perhaps the most intuitive metrics, there are a range of others, such as numerical drift from the "true" value of each data item and combinations of each of these metrics.²⁵

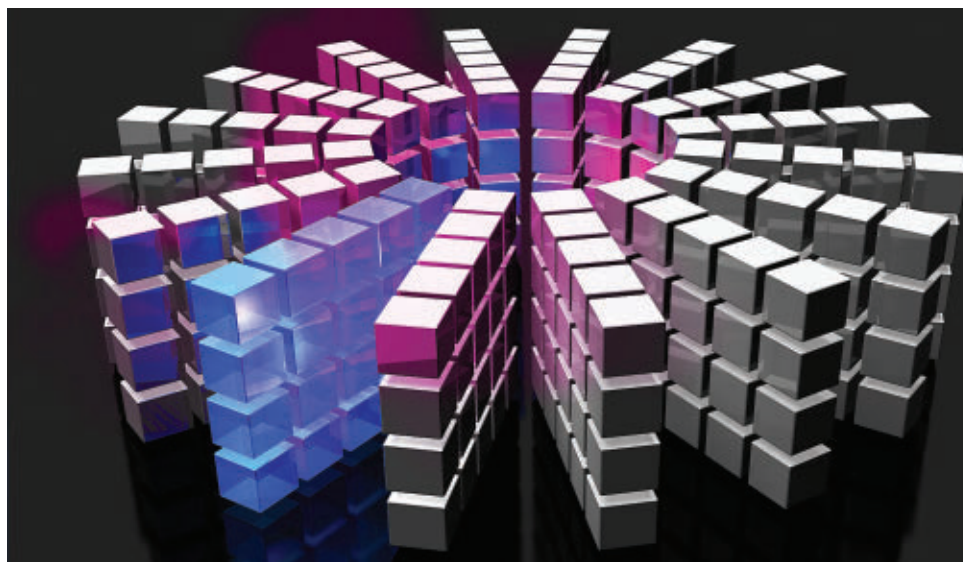
The two main kinds of mechanisms for quantifying for eventual consistency are measurement and prediction. *Measurement* answers the question, "How consistent is my store under my given workload right now?"¹⁸ while *prediction* answers the question, "How consistent will my store be under a given configuration and workload?"⁴ Measurement is useful for runtime monitoring and alerts or verifying compliance with service-level objectives (SLOs). Prediction is useful for probabilistic what-if analyses such as the effect of configuration

and workload changes and for dynamically tuning system behavior. Taken together, measurement and prediction form a useful toolkit.

Probabilistically bounded staleness. As a brief deep dive into how to quantify eventually consistent behavior, we will discuss our experiences developing, deploying, and integrating state-of-the-art prediction techniques into Cassandra, a popular NoSQL. Probabilistically Bounded Staleness, or PBS, provides an *expectation* of recency for reads of data items.⁴ This allows us to measure how far an eventually consistent store's behavior deviates from that of a strongly consistent, linearizable (or regular) store. PBS enables metrics of the form: "100 milliseconds after a write completes, 99.9% of reads will return the most recent version," and "85% of reads will return a version that is within two of the most recent."

Building PBS. How does PBS work? Intuitively, the degree of inconsistency is determined by the rate of anti-entropy. If replicas constantly exchange their last written writes, then the window of inconsistency should be bounded by the network delay and local processing delay at each node. If replicas delay anti-entropy (possibly to save bandwidth or processing time), then this delay is added to the window of inconsistency; many systems (Amazon's Dynamo, for example) offer settings in the replication protocol to control these delays. Given the anti-entropy protocol, then—given the configured anti-entropy rate, the network delay, and local processing delay—you can calculate the expected consistency. In Cassandra, we piggyback timing information on top of the write distribution protocol (the primary source of anti-entropy) and maintain a running sample. When a user wants to know the effect of a given replication configuration, we use the collected sample in a Monte Carlo simulation of the protocol to return an expected value for the consistency of the data store, which closely matches consistency measurements on our Cassandra clusters at Berkeley.

PBS in the wild. Using our PBS consistency prediction tool, and with the help of several friends at LinkedIn and Yammer, we quantified the consistency of three eventually consistent stores running in production. PBS models



predicted that LinkedIn's data stores returned consistent data 99.9% of the time within 13.6ms; and on SSDs, within 1.63ms. These eventually consistent configurations were 16.5% and 59.5% faster than their strongly consistent counterparts at the 99.9th percentile. Yammer's data stores experienced a 99.9% inconsistency window of 202ms at 81.1% latency reduction. The results confirmed the anecdotal evidence: eventually consistent stores are often faster than their strongly consistent counterparts, and they are frequently consistent within tens or hundreds of milliseconds.

In order to make consistency prediction more accessible, with the help of the Cassandra community, we recently released support for PBS predictions in Cassandra 1.2.0. Cassandra users can now run predictions on their own production clusters to tune their consistency parameters and perform what-if analyses for normal-case, failure-free operation. For example, to explore the effect of adding solid-state drives (SSDs) to a set of servers, users can adjust the expected distribution of read and write speeds on the local node. These predictions are inexpensive; a JavaScript-based demonstration we have created⁴ completes tens of thousands of trials in less than a second.

Of course, prediction is not without faults: predictions are only as good as the underlying model and input data. As statistician George E.P. Box famously stated, "All models are wrong, but some are useful." Failure to account for an important aspect of the system or anti-entropy protocol may lead to

inaccurate predictions. Similarly, prediction works by assuming that past behavior is correlated with future behavior. If environmental conditions change, predictions may be of limited accuracy. These issues are fundamental to the problem at hand, and they are a reminder that prediction is best paired with measurement to ensure accuracy.

Eventual consistency is often strongly consistent. In addition to PBS, several recent projects have verified the consistency of real-world eventually consistent stores. One study found that Amazon SimpleDB's inconsistency window for eventually consistent reads was almost always less than 500ms,²⁴ while another study found that Amazon S3's inconsistency window lasted up to 12 seconds.⁷ Other recent work shows results similar to those presented for PBS, with Cassandra closing its inconsistency window within around 200ms.¹⁸

These results confirm the anecdotal evidence that eventual consistency is often "good enough" by providing quantitative metrics for system behavior. As techniques such as PBS and consistency measurement continue to make their way into more production infrastructure, reasoning about the behavior of eventual consistency across deployments, failures, and system configurations will be increasingly straightforward.

Programming Eventual Consistency

While users can verify and predict the consistency behavior of eventually consistent systems, these techniques do

not provide absolute guarantees against safety violations. What if an application requires that safety is always respected? There is a growing body of knowledge about how to program and reason about eventually consistent stores.

Compensation, costs, and benefits. Programming around consistency anomalies is similar to speculation: you do not know what the latest value of a given data item is, but you can proceed as if the value presented is the latest. When you have guessed wrong, you have to compensate for any incorrect actions taken in the interim. In effect, compensation is a way to achieve safety retroactively—to restore guarantees to users.¹³ Compensation ensures mistakes are eventually corrected but does not guarantee mistakes are not made.

As an example of speculation and compensation, consider running an ATM machine.^{8,13} Without strong con-

istency just as well as other errors such as data-entry mistakes.

An application designer deciding whether to use eventual consistency faces a choice. In effect, the designer needs to weigh the benefit of weak consistency B (in terms of high availability or low latency) against the cost C of each inconsistency anomaly multiplied by the rate of anomalies R :

maximize $B - RC$

This decision is, by necessity, application and deployment specific. The cost of anomalies is determined by the cost of compensation: too many overdrafts might cause customers to leave a bank, while propagation of status updates that is too slow might cause users to leave a social network. The rate of anomalies—as seen before—depends on the system architecture, configura-

when the cost of inconsistency is high, with tangible monetary consequences (for example, ATMs), compensation is more likely to be well thought out.

For some applications, however, the rate of anomalies may be low enough or the cost of inconsistency sufficiently small so that the application designer may choose to forgo including compensation entirely. If the chance of inconsistency is sufficiently low, users may experience anomalies in only a small number of cases. Anecdotally, many online services such as social networking largely operate with weakly consistent configurations: if a user's status update takes seconds or even minutes to propagate to followers, they are unlikely to notice or even care. The complexities of operating a strongly consistent service at scale may outweigh the benefit of, say, preventing an off-by-one error in Justin Bieber's follower count on Twitter.

Compensation by design. Compensation is error prone and laborious, and it exposes the programmer (and sometimes the application) to the effects of replication. What if you could program without it? Recent research has provided “compensation-free” programming for many eventually consistent applications.

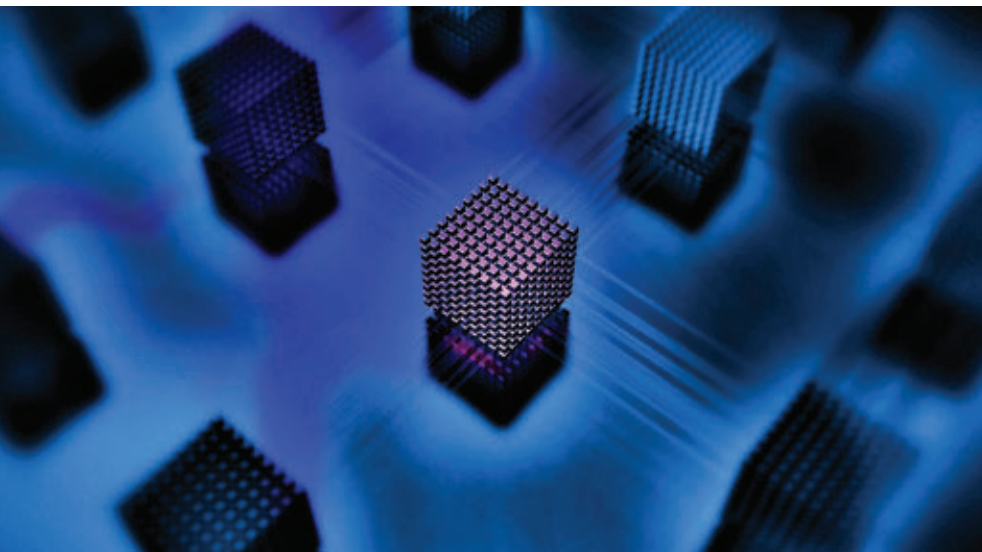
The formal underpinnings of eventually consistent programs that are consistent by design are captured by the CALM theorem, indicating which programs are safe under eventual consistency and also (conservatively) what is not.³ Formally, CALM means *consistency as logical monotonicity*; informally, it means programs that are monotonic, or compute an ever-growing set of facts (by for example, receiving new messages or performing operations on behalf of a client) and do not ever “retract” facts that they emit (that is, the basis for decisions it has already made do not change), can always be safely run on an eventually consistent store. (Full disclosure: CALM was developed by colleagues at UC Berkeley). Accordingly, CALM tells programmers which operations and programs can guarantee safety when used in an eventually consistent system. Any code that fails CALM tests is a candidate for stronger coordination mechanisms.

As a concrete example of this logical monotonicity, consider build-

sistency, two users might simultaneously withdraw money from an account and end up with more money than the account ever held. Would a bank ever want this behavior? In practice, yes. An ATM's ability to dispense money (availability) outweighs the cost of temporary inconsistency in the event that an ATM is partitioned from the master bank branch's servers. In the event of overdrawing an account, banks have a well-defined system of external compensating actions: for example, overdraft fees charged to the user. Banking software is often used to illustrate the need for strong consistency, but in practice the socio-technical system of the bank can deal with data inconsis-

tion, and deployment. Similarly, the benefit of weak consistency is itself possibly a compound term composed of factors such as the incidence of communication failures and communication latency.

Second, application designers actually must design for compensation. Writing corner-case compensation code is nontrivial. Determining the correct business application logic to handle each type of consistency anomaly is a difficult task. Carefully reasoning about each possible sequence of anomalies and the correct “apologies” to make to the user for each can become more onerous than designing a solution for strong consistency. In general,




ing a database for queries on stock trades. Once completed, trades cannot change, so any answers given that are based solely on the immutable *historical* data will remain true. However, if your database keeps track of the value of the *latest* trade, then new information—such as new stock prices—might retract old information, as new stock prices overwrite the latest ones in the database. Without coordination between replica copies, the second database may return inconsistent data.


By analyzing programs for monotonicity, you can “bless” monotonic programs as “safe” under eventual consistency and encourage the use of coordination protocols (such as strong consistency) in the presence of non-monotonicity. As a general rule, operations such as initializing variables, accumulating set members, and testing a threshold condition are monotonic. In contrast, operations such as variable overwrites, set deletion, counter resets, and negation (such as, “there does not exist a trade such that...”) are generally not logically monotonic.

CALM captures a wide space of design patterns sometimes referred to as ACID 2.0 (associativity, commutativity, idempotence, and distributed).¹³ *Associativity* means you can apply a function in any order: $f(a, f(b, c)) = f(f(a, b), c)$. *Commutativity* means a function’s arguments are order-insensitive: $f(a, b) = f(b, a)$. Commutative and associative programs are order-insensitive and can tolerate message reordering, as in eventual consistency. *Idempotence* means you can call a function on the same input any number of times and get the same result: $f(f(x)) = f(x)$ (for example, $\max(42, \max(42, 42)) = 42$). Idempotence allows the use of at-least-once message delivery, instead of at-most-once delivery (which is more expensive to guarantee). *Distributed* is primarily a placeholder for *D* in the acronym (!) but symbolizes the fact that ACID 2.0 is all about distributed systems. Carefully applying these design patterns can achieve logical monotonicity.

Recent work on CRDTs (commutative, replicated data types) embodies CALM and ACID 2.0 principles within a variety of standard data types, providing provably eventually consistent data structures including sets, graphs, and sequences.²⁰ Any program that correct-



The complexities of operating a strongly consistent service at scale may outweigh the benefit of, say, preventing an off-by-one error in Justin Bieber’s follower count on Twitter.



ly uses these predefined, well-specified data structures is guaranteed to never see any safety violations.

To understand CRDTs, consider building an increment-only counter that is replicated on two servers. We might implement the increment operation by first reading the counter’s value on one replica, incrementing the value by one, and writing the new value back every replica. If the counter is initially 0 and two different users simultaneously initiate increment operations at two separate, then both users may then read 0 and then distribute the value 1 to the replicas; the counter ends up with a value of 1 instead of the correct value of 2. Instead, we can use a G-counter CRDT, which relies on the fact that *increment* is a commutative operation—it does not matter in what order the two *increment* operations are applied, as long as they are both eventually applied at all sites. With a G-counter, the current counter status is represented as the count of distinct *increment* invocations, similar to how counting is introduced at the grade-school level: by making a tally mark for every increment then summing the total. In our example, instead of reading and writing counter *values*, each invocation distributes an increment *operation*. All replicas end up with two increment operations, which sum to the correct value of 2. This works because replicas understand the semantics of increment operations instead of providing general-purpose read/write operations, which are not commutative.

A key property of these advances is that they separate data store and application-level consistency concerns. While the underlying store may return inconsistent data at the level of reads and writes, CALM, ACID 2.0, and CRDT appeal to *higher-level* consistency criteria, typically in the form of application-level invariants that the application maintains. Instead of requiring that every read and write to and from the data store is strongly consistent, the application simply has to ensure a semantic guarantee (say, “the counter is strictly increasing”)—granting considerable leeway in how reads and writes are processed. This distinction between application-level and read/write consistency is often ambiguous

and poorly defined (for example, what does database ACID “consistency” have to do with “strong consistency”?). Fortunately, by identifying a large class of programs and data types that are tolerant of weak consistency, programmers can enjoy “strong” application consistency, while reaping the benefits of “weak” distributed read/write consistency.


Taken together, the CALM theorem and CRDTs make a powerful toolkit for achieving “consistency without concurrency control,” which is making its way into real-world systems. Our team’s work on the Bloom language³ embodies CALM principles. Bloom encourages the use of order-insensitive disorderly programming, which is key to architecting eventually consistent systems. Some of our recent work focuses on building custom eventually consistent data types whose correctness is grounded in formal mathematical lattice theory. Concurrently, several open source projects such as Statebox²¹ provide CRDT-like primitives as client-side extensions to eventually consistent stores, while one eventually consistent store—Riak—recently announced alpha support for CRDTs as a first-class server-side primitive.⁹

Stronger Than Eventual


While compensating actions and CALM/CRDTs provide a way around eventual consistency, they have shortcomings of their own. The former requires dealing with inconsistencies outside the system and the latter limits the operations that an application writer can employ. However, it turns out that it is possible to provide even stronger guarantees than eventual consistency—albeit weaker than SSI—for general-purpose operations while still providing availability.

The CAP theorem dictates that strong consistency (SSI) and availability are unachievable in the presence of partitions. But how weak does the consistency model have to be in order for it to be available? Clearly, eventual consistency, which simply provides a liveness guarantee, is available. Is it possible to strengthen eventual consistency by adding safety guarantees to it without losing its benefits?

Pushing the limits. A recent techni-



By analyzing programs for monotonicity, you can “bless” monotonic programs as “safe” under eventual consistency and encourage the use of coordination protocols in the presence of non-monotonicity.



cal report from the University of Texas at Austin claims no consistency model stronger than causal consistency is available in the presence of partitions.¹⁷ Causal consistency guarantees each process’s writes are seen in order writes follow reads (if a user reads a value $A=5$ and then writes $B=10$, then another user cannot read $B=10$ and subsequently read an older value of A than 5), and transitive data dependencies hold. This causal consistency is useful in ensuring, for example, comment threads are seen in the correct order, without dangling replies, and users’ privacy settings are applied to the appropriate data. The UT Austin report demonstrates that it is not possible to have a stronger model than causal consistency (that accepts fewer outcomes) without violating either high availability or ensuring that, if two servers communicate, they will agree on the same set of values for their data items. While many other available models are neither stronger nor weaker than causal consistency, this impossibility result is useful because it places an upper bound on a very familiar consistency model.

In light of the UT Austin result, several new data storage designs provide causal consistency. The COPS and Eiger systems¹⁶ developed by a team from Princeton, CMU, and Intel Research provide causal consistency without incurring high latencies across geographically distant datacenters or the loss of availability in the event of datacenter failures. These systems perform particularly well, at a near-negligible cost to performance when compared to eventual consistency; Eiger, which was prototyped within the Cassandra system, incurs less than 7% overhead for one of Facebook’s workloads. In our recent work, we demonstrated how existing data stores that are already deployed in production but provide eventual consistency can be augmented with causality as an added safety guarantee.⁶ Causality can be *bolted-on* without compromising high availability, enabling system designs in which safety and liveness are cleanly decomposed into separate architectural layers.

In addition to causality, we can consider the relationship between ACID transactions and the CAP theorem. While it is impossible to provide the

gold standard of ACID isolation—serializability, or SSI—it turns out many ACID databases provide a weaker form of isolation, such as read committed, often by default and, in some cases, as the maximum offered. Some of our recent results show many of these weaker models *can* be implemented in a distributed environment while providing high availability.⁵ Current databases providing these weak isolation models are unavailable, but this is only because they have been implemented with unavailable algorithms.


We—and several others—are developing transactional algorithms that show this need not be the case. By rethinking the concurrency-control mechanisms and re-architecting distributed databases from the ground up, we can provide safety guarantees in the form of transactional atomicity, ANSI SQL Read Committed and Repeatable Read, and causality between transactions—matching many existing ACID databases—without violating high availability. This is somewhat surprising, as many in the past have assumed that, in a highly available system, arbitrary multi-object transactions are out of the question.

Recognizing the limits. While these results push the limits of what is achievable with high availability, there are several properties that a weakly consistent system will never be able to provide; there is a fundamental cost to remaining highly available (and providing guaranteed low latency). The CAP theorem states that making staleness guarantees is impossible in a highly available system. Reads that specify a constraint on data recency (for example, “give me the latest value,” “give me the latest value as of 10 minutes ago”) are not generally available in the presence of long-lasting network partitions. Similarly, we cannot maintain arbitrary global correctness constraints over sets of data items such as uniqueness requirements (for example, “create bank account with ID 50 if the account does not exist”) and, in certain cases (for example, arbitrary reads and writes), even correctness constraints on individual data items are not achievable (for example, “the bank account balance should be non-negative”). These challenges are an inherent cost of choosing weak consistency—whether eventual or a stronger but still “weak” model.

Conclusion

By simplifying the design and operation of distributed services, eventual consistency improves availability and performance at the cost of semantic guarantees to applications. While eventual consistency is a particularly weak property, eventually consistent stores often deliver consistent data, and new techniques for measurement and prediction grant us insight into the behavior of eventually consistent stores. Concurrently, new research and prototypes for building eventually consistent data types and programs are easing the burden of reasoning about disorder in distributed systems. These techniques, coupled with new results pushing the boundaries of highly available systems—including causality and transactions—make a strong case for the continued adoption of weakly consistent systems. While eventual consistency and its weakly consistent cousins are not perfect for every task, their performance and availability implications will likely continue to accrue admirers and advocates in the future.

Acknowledgments

The authors would like to thank Peter Alvaro, Carlos Baquero, Neil Conway, Alan Fekete, Joe Hellerstein, Marc Shapiro, and Ion Stoica for feedback on earlier drafts of this article. 

Related articles on queue.acm.org

Eventually Consistent

Werner Vogels

<http://queue.acm.org/detail.cfm?id=1466448>

BASE: An Acid Alternative

Dan Pritchett

<http://queue.acm.org/detail.cfm?id=1394128>

Scalable SQL

Michael Rys

<http://queue.acm.org/detail.cfm?id=1971597>

References

1. Abadi, D. Consistency tradeoffs in modern distributed database system design: CAP is only part of the story. *IEEE Computer* (Feb. 2012).
2. Alpern, B. and Schneider, F.B. Defining liveness. *Information Processing Letters* 21 (Oct. 1985).
3. Alvaro, P., Conway, N., Hellerstein, J. and Marczak, W. 2011. Consistency analysis in Bloom: A CALM and collected approach. *Proceedings of the Conference on Innovative Data Systems Research* (2011).
4. Bailis, P., Venkataraman, S., Franklin, M., Hellerstein, J. and Stoica, I. Probabilistically bounded staleness

for practical partial quorums. In *Proceedings of Very Large Databases* (2012). (Demo from text: <http://pbs.cs.berkeley.edu/#demo>)

5. Bailis, P., Fekete, A., Ghodsi, A., Hellerstein, J. and Stoica, I. HAT, not CAP: Highly available transactions. arXiv:1302.0309 (Feb. 2013).
6. Bailis, P., Ghodsi, A., Hellerstein, J. and Stoica, I. Bolt-on Causal Consistency. In *Proceedings of ACM SIGMOD* (2013).
7. Bermbach, D. and Tai, S. Eventual consistency: how soon is eventual? An evaluation of Amazon S3's consistency behavior. In *Proceedings of Workshop on Middleware for Service-Oriented Computing* (2011).
8. Brewer, E. CAP twelve years later: How the “rules” have changed. *IEEE Computer* (Feb. 2012).
9. Brown, R. and Cribbs, S. Data structures in Riak (2012); <https://speakerdeck.com/basho/data-structures-in-riak>. RICON Conference.
10. Davidson, S., Garcia-Molina, H. and Skeen, D. Consistency in a partitioned network: A survey. *ACM Computing Surveys* 17, 3 (1985).
11. Gilbert, S. and Lynch, N. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *ACM SIGACT News* 33, 2 (June 2002).
12. Hale, C. You can't sacrifice partition tolerance (2010); <http://codahale.com/you-cant-sacrifice-partition-tolerance/>
13. Helland, P. and Campbell, D. Building on quicksand. In *Proceedings of the Conference on Innovative Data Systems Research* (2009).
14. Johnson, P. R., Thomas, R. H. Maintenance of duplicate databases; RFC 677 (1975); <http://www.faqs.org/rfcs/rfc677.html>.
15. Kawell Jr., L., Beckhardt, S., Halvorsen, T., Ozzie, R. and Greif, I. Replicated document management in a group communication system. In *Proceedings of the 1988 ACM Conference on Computer-supported Cooperative Work*: 395; <http://dl.acm.org/citation.cfm?id=1024798>.
16. Lloyd, W., Freedman, M., Kaminsky, M. and Andersen, D. Stronger semantics for low-latency geo-replicated storage. In *Proceedings of Networked Systems Design and Implementation* (2011).
17. Mahajan, P., Alvisi, L. and Dahlin, M. Consistency, availability, convergence. University of Texas at Austin TR-11-22 (May 2011).
18. Rahman, M., Golab, W., AuYoung, A., Keeton, K. and Wylie, J. Toward a principled framework for benchmarking consistency. *Workshop on Hot Topics in System Dependability* (2012).
19. Saito, Y. and Shapiro, M. Optimistic Replication. *ACM Computing Surveys* 37, 1 (Mar. 2005). <http://dl.acm.org/citation.cfm?id=1057980>
20. Shapiro, M., Preguiça, N., Baquero, C. and Zawirski, M. A comprehensive study of convergent and commutative replicated data types. INRIA Technical Report RR-7506 (Jan. 2011).
21. Statebox; <https://github.com/mochi/statebox>.
22. Terry, D., Theimer, M., Petersen, K., Demers, A., Spreitzer, M. and Hauser, C. Managing update conflicts in Bayou, a weakly connected replicated storage system. In *Proceedings on Symposium on Operating Systems Principles* (1995).
23. Vogels, W. Eventually consistent. *ACM Queue*, (2008).
24. Wada, H., Fekete, A., Zhao, L., Lee, K., A. and Liu, A. Data consistency and the tradeoffs in commercial cloud storage: the consumers' perspective. *Proceedings of the Conference on Innovative Data Systems Research* (2011).
25. Yu, H. and Vahdat, A. Design and evaluation of a conit-based continuous consistency model for replicated services. *ACM Trans. on Computer Systems* (2002).

Peter Bailis is a graduate student of computer science in the AMPLab and BOOM projects at UC Berkeley, where he works closely with Ali Ghodsi, Joe Hellerstein, and Ion Stoica. He currently studies distributed systems and databases, with a particular focus on distributed consistency models. He blogs at <http://bailis.org/blog> and tweets as @pbailis.

Ali Ghodsi (alig@cs.berkeley.edu) is an assistant professor at KTH/Royal Institute of Technology in Sweden and a Visiting Researcher at UC Berkeley since 2009. His general interests are in the broader areas of distributed systems, and networking. He received his Ph.D. in 2006 from KTH/Royal Institute of Technology in the area of distributed computing.

Flash memory has come a long way and it is time for software to catch up.

BY ADAM H. LEVENTHAL

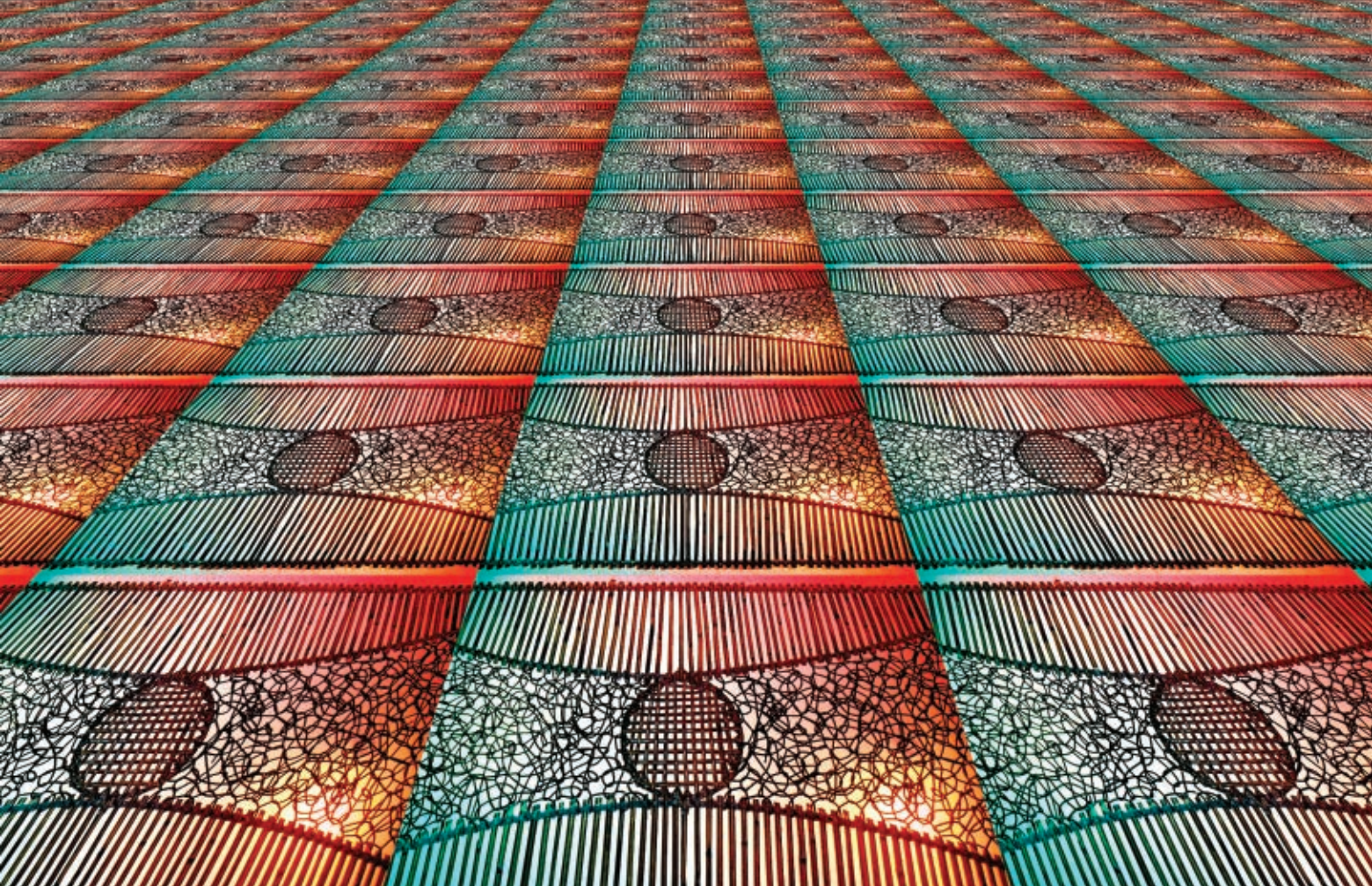
A File System All Its Own

IN THE PAST five years, flash memory has progressed from a promising accelerator,⁷ whose place in the data center was still uncertain, to an established enterprise component for storing performance-critical data.^{4,9} With solid-state devices (SSDs), flash arrived in a form optimized for compatibility—just replace a hard drive with an SSD for radically better performance. But the properties of the NAND flash memory used by SSDs differ significantly from those of the magnetic media in the hard drives they often displace.² While SSDs

have become more pervasive in a variety of uses, the industry has only just started to design storage systems that embrace the nuances of flash memory. As it escapes the confines of compatibility, significant improvements are possible in the areas of performance, reliability, and cost.

The native operations of NAND flash memory are quite different from those required of a traditional block device. The flash translation layer (FTL), as the name suggests, translates the block-device commands into operations on flash memory. This translation is by no means trivial; both the granularity and the fundamental operations differ. SSD controllers compete in subspecialties such as garbage collection, write amplification, wear leveling, and error correction.² The algorithms used by modern SSDs are growing increasingly sophisticated despite the seemingly simple block-read and block-write operations that they must support. A very common use of a block device is to host a file system. File systems, of course, perform their own type of translation: from file creations, opens, reads, and writes within a directory hierarchy to block reads and writes. There is nothing innate about file-system operations that make them well served by the block interface; it is just the dominant standard for persistent storage, and it has existed for decades.

Layering the file-system translation on top of the flash translation is inefficient and impedes performance. Sophisticated applications such as databases have long circumvented the file system—again, layers upon layers—to attain optimal performance. The information lost between abstraction layers impedes performance, longevity, and capacity. A file system may “know” that a file is being copied, but the FTL sees each copied block as discrete and unique. File systems also optimize for the physical realities of a spinning disk, but placing data on the sectors that spin the fastest does not



make sense when they do not spin at all. Volume managers, software that presents collections of disks as a block device, led to similar inefficiencies in disk-based storage, obscuring information from the file system.

Modern file systems such as Write Anywhere File Layout (WAFL)⁵ ZFS and B-tree file system (Btrfs)¹ integrated the responsibilities previously assigned to volume managers and reorganized the layers of abstraction. The resulting systems were more efficient and easier to manage. Poorly optimized software mattered when operations were measured in milliseconds; it matters much more on flash devices whose operations are measured in microseconds. To take full advantage of flash, users need software expressly designed for the native operations and capabilities of NAND flash.

The State of SSDs

For many years SSDs were almost exclusively built to seamlessly replace hard drives; they not only supported the same block-device interface, but also had the same form factor (for example, a 2.5- or 3.5-inch hard drive) and communicated using the same protocols

(for example, SATA, SAS, or FC). This is a bit like connecting an iPod to a car stereo using a tape adapter; now it seems that 30-pin iPod connectors are more common in new cars than tape decks are. Recently SSDs have started to break away from the old constraints on compatibility: some laptops now use a custom form-factor SSD for compactness, and many vendors produce PCI-attached SSDs for lower latency.

The majority of SSDs still emulate the block interface of hard drives: reading and writing an arbitrary series of sectors (512-byte or 4KB regions). The native operations of NAND flash memory are different enough to create some substantial challenges. Reads and writes happen at the granularity of a page (usually around 8KB) with the significant caveat that writes can occur only to erased pages, and pages are erased exclusively in blocks of 32–64 (256KB–512KB). While a detailed description of how an FTL presents a block interface from flash primitives is beyond the scope of this article, it is easy to get a sense of its complexity. Consider the case of a block in which all pages have been written, and the device receives an operation to logically

overwrite the contents of one page. The FTL could copy the block into memory, modify the page, erase the block, and rewrite it in its entirety, but this would be very slow—slower even than a hard drive! In addition, each write or erase operation wears out NAND flash. Chips are rated for a certain number of such operations—anywhere from 500–50,000 cycles today depending on the type and quality, and those numbers are shrinking as the chips themselves shrink. A native approach to block management would quickly wear down the media; and to compound the problem, a frequently overwritten region would wear out before other regions. For these reasons, FTLs use an indirection layer that allows data to be written at arbitrary locations and implements wear leveling, the process of distributing writes uniformly across the media.²

Bridging the Gap

The algorithms that make up an FTL are highly complex but no more than those of a modern file system. Indeed, the FTL and file system have much in common. Both track allocated versus free regions, both implement a logical to physical mapping, and both trans-

late one operation set to another. Newer FTLs even include facilities such as compression and deduplication—still marquee features for modern file systems. FTLs and file systems are usually built in isolation. The idea of a dramatic integration and reorganization of the responsibilities of the FTL and file system represents a classic conundrum: who will write software for non-existent hardware, and who will build hardware to enable heretofore-unwritten software?

Most SSD vendors are focused on a volume market where requiring a new file system on the host would be an impediment rather than an advantage. SSD vendors could enable the broader file-system developer community by providing different interfaces or opening up their firmware, but again—and without an obvious and compelling file system—there is little incentive. The exception was Indilinx's participation in the OpenSSD¹⁰ project, but the primary focus was FTL development and experimentation within conventional bounds. OpenSSD became effectively defunct when OCZ acquired Indilinx. There seems to be no momentum and only vague incentive for vendors to give developers the level of visibility and control they most want. Mainstream efforts to build flash awareness into file systems have led to more modest modifications to the interface between file system and SSD.

The most publicized interface between the file system and SSD is the

ATA TRIM command or its counterpart, the SCSI UNMAP command. TRIM and UNMAP convey the same meaning to a device: the given region is no longer in use. One of the challenges with an FTL is efficient space management; and the more space that is available, the easier it is to perform that task. As free space is exhausted, FTLs have less latitude to migrate data, and they need to keep data in an increasingly compact form; with lots of free space FTLs can be far sloppier.

For both performance and redundancy, almost all SSDs “overprovision.” They include more flash memory capacity than the advertised capacity of the SSD by anywhere from 10% to 100%. File systems have the notion of allocated and free blocks, but there is not a means—or a reason—to communicate that information to a hard drive. To let SSDs reap the benefits of free storage, modern file systems use the TRIM or UNMAP commands to indicate that logical regions are no longer in use. Some SSDs—particularly those designed for the consumer market—greatly benefit from file systems that support TRIM and UNMAP. Of course, for a file system whose steady state is close to full, TRIM and UNMAP have very little impact because there are not many free blocks.

Incremental Revolution

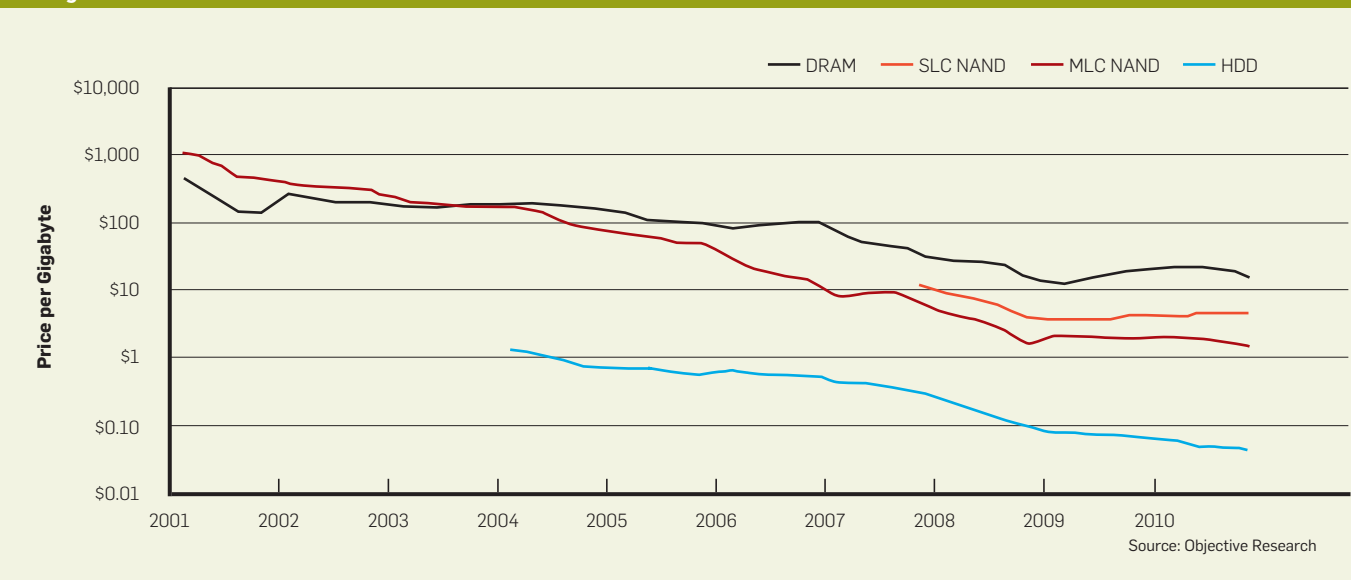
While many companies participate in incremental improvements, the most likely candidates to create a flash-opti-

mized file system are those that build both SSDs and software that runs on the host. The singular popularized example thus far is DirectFS⁶ from FusionIO. Here, the flash storage provides more expressive operations for the file system. Rather than solely using the legacy block interface, the DirectFS interacts with a virtualized flash storage layer. That layer manages the flash media much as a traditional FTL but offers greater visibility and an expanded set of operations to the file system above it.

DirectFS achieves significant performance improvements not by supplanting intelligence in the hardware controller, but by reorganizing responsibilities between the file system and flash controller. For example, FusionIO has proposed extensions to the SCSI standard that perform scattered reads and writes atomically.³ These are easily supported by the FTL, but dramatically simplify the logic required in a file system to ensure metadata consistency in the face of a power failure. DirectFS also relies on storage that provides a “sparse address space,” which effectively transfers allocation and block mapping responsibilities from the file system to the FTL, a task the FTL already must do. A 2010 article by William Josephson et al. states that “novel layers of abstraction specifically for flash memory can yield substantial benefits in software simplicity and system performance.”⁶

As with TRIM, incrementally adding expressiveness and functionality to the

Pricing trends.



existing storage interfaces allows file systems to take advantage of new facilities on devices that provide them. Storage system designers can choose whether to require devices that provide those interfaces or to implement a work-alike facility that they disable when it is not needed. Device vendors can decide whether supporting a richer interface represents a sufficient competitive advantage. Though this approach may never lead to an optimal state, it may allow the industry to navigate monotonically to a sufficient local maximum.

The Chicken and the Egg

There are still other ways to construct a storage system around flash. A more radical approach is to go further than DirectFS, assigning additional high-level responsibilities to the file system such as block management, wear leveling, read-disturb awareness, and error correction. This would allow for a complete reorganization of the software abstractions in the storage system, ensuring sufficient information for proper optimization where today's layers must cope with suboptimal information and communication. Again, this approach requires a vendor that can assert broad control over the whole system—from the file system to the interface, controller, and flash media. It is certainly tenable for closed proprietary systems—indeed, several vendors are pursuing this approach—but for it to gain traction as a new open standard would be difficult.

The SSDs that exist today for the volume market are cheap and fast, but they exhibit performance that is inconsistent and reliability that is insufficient. Higher-level software designed with full awareness of those shortcomings could turn that commodity iron into gold. Without redesigning part or all of the I/O interface, those same SSDs could form the basis of a high-performing and highly reliable storage system.

Rather than designing a file system around the properties of NAND flash, this approach would treat the commodity SSDs themselves as the elementary unit of raw storage. NAND flash memory already has complicated intrinsic properties; the emergent properties of an SSD are even more obscure and varied. A common pathology with SSDs, for example, is

variable performance when servicing concurrent or interleaved read and write operations. Understanding these pathologies sufficiently and creating higher-level software to accommodate them would represent the flash version of an existential software parable: enterprise quality from commodity components. It is a phenomenon that the storage world has seen before with disks; software such as ZFS from Sun has produced fast, reliable systems from cheap components.

The only easy part of this transmutation is finding the base material. Building such a software system given a single, unchanging SSD would already be complicated; doing it amid the changing diversity of the SSD market further complicates the task. The properties of flash differ between types and fabrication processes, but change happens at the rate of hardware evolution. SSDs change not only to accommodate the underlying media and controller hardware, but also at the speed of software, fixing bugs and improving algorithms. Still, some vendors are pursuing¹¹ this approach because, while it is more complex than designing for purpose-built hardware, it has the potential to produce superlative systems that ride the economic curve of volume SSDs.

Next for Flash

The lifespan of flash as a relevant technology is a topic of vigorous debate. While flash has ridden its price and density trends to a position of relevance, some experts anticipate fast-approaching limits to the physics of scaling NAND flash memory. Others foresee several decades of flash innovation. Whether it is flash or some other technology, nonvolatile solid-state memory will be a permanent part of the storage hierarchy, having filled the yawning gap between hard-drive and CPU speeds.⁸

The next evolutionary stage should see file systems designed explicitly for the properties of solid-state media rather than relying on an intermediate layer to translate. The various approaches are each imperfect. Incremental changes to the storage interface may never reach the true acme. Creating a new interface for flash might be untenable in the market. Treating

SSDs as the atomic unit of storage may be just another half-measure, and a technically difficult one at that.

Some companies today are betting on the relevance of flash at least in the near term—some working within the confines of today's devices, others building, augmenting, or replacing the existing interfaces. The performance of flash memory has whetted the computer industry's appetite for faster and cheaper persistent storage. The experimentation phase is long over; it is time to build software for flash memory and embrace the specialization needed to realize its full potential. ■

Related articles on queue.acm.org

Anatomy of a Solid-state Drive

Michael Cornwell

<http://queue.acm.org/detail.cfm?id=2385276>

Enterprise SSDs

Mark Moshayedi, Patrick Wilkison

<http://queue.acm.org/detail.cfm?id=1413263>

Flash Disk Opportunity for Server Applications

Jim Gray, Bob Fitzgerald

<http://queue.acm.org/detail.cfm?id=1413261>

References

1. Btrfs wiki; https://btrfs.wiki.kernel.org/index.php/Main_Page
2. Cornwell, M. Anatomy of a solid-state drive. *ACM Queue* 10, 10 (2012); <http://queue.acm.org/detail.cfm?id=2385276>.
3. Elliott, R. and Batwara, A. Notes to T10 Technical Committee. 11-229r4 SBC-4 SPC-5 Atomic writes and reads; <http://www.t10.org/cgi-bin/ac.pl?t=d&f=11-229r4.pdf>; 12-086r2 SBC-4 SPC-5 Scattered writes, optionally atomic; <http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-086r2.pdf>; 12-087r2 SBC-4 SPC-5 Gathered reads—Optionally atomic; <http://www.t10.org/cgi-bin/ac.pl?t=d&f=12-087r2.pdf>
4. Gray, J. and Fitzgerald, B. Flash disk opportunity for server applications. *ACM Queue* 6, 4 (2008); <http://queue.acm.org/detail.cfm?id=1413261>
5. Hitz, D., Lau, J. and Malcolm, M. File system design for an NFS file server appliance. *WTEC '94 USENIX Winter 1994 Technical Conference*; <http://dl.acm.org/citation.cfm?id=1267093>
6. Josephson, W.K., Bongo, L.A., Li, K. and Flynn, D. DFS: A file system for virtualized flash storage. *ACM Transactions on Storage* 6, 3 (2010). <http://dl.acm.org/citation.cfm?id=1837922>
7. Leventhal, A. Flash storage today. *ACM Queue* 6, 4 (2008); <http://queue.acm.org/detail.cfm?id=1413262>
8. Leventhal, A. Triple-parity RAID and beyond. *ACM Queue* 7, 11 (2009); <http://queue.acm.org/detail.cfm?id=1670144>
9. Moshayedi, M. and Wilkison, P. Enterprise SSDs. *ACM Queue* 6, 4 (2008); <http://queue.acm.org/detail.cfm?id=1413263>
10. The OpenSSD Project; http://www.openssd-project.org/wiki/The_OpenSSD_Project
11. PureStorage FlashArray; <http://www.purestorage.com/flash-array/purity.html>

Adam H. Leventhal is the CTO at Delphix, a database virtualization company. Previously he served as Lead Flash Engineer for Sun and then Oracle where he designed flash integration in the ZFS Storage Appliance, Exadata, and other products.

© 2013 ACM 0001-0782/13/05

DOI:10.1145/2447976.2447993

Surgeons use hand gestures and/or voice commands without interrupting the natural flow of a procedure.

BY MITHUN GEORGE JACOB, YU-TING LI, GEORGE A. AKINGBA, AND JUAN P. WACHS

Collaboration with a Robotic Scrub Nurse

ERRORS IN THE delivery of medical care are the principal cause of inpatient mortality and morbidity (98,000 deaths annually in the U.S.).¹⁶ Ineffective team communication is often at the root of these errors.^{7,10,16} For example, in assessing verbal and nonverbal exchanges in the operating room (OR), Lingard et al.¹⁸ found frequent communication failure, with commands delayed, incomplete, or not received at all, as well as left unresolved. Firth-Cousins⁷ found 31% of all communications in the OR represent failures,⁷ with one-third of them having a negative effect on patient outcomes.⁷ And Halverson et al.¹⁰ found 36% of communication errors are related to equipment use.

Causes of errors include team instability (such as lack of familiarity between nurses and surgeons),⁵ lack of resources (such as minimal staffing), and distractions. Poor communication within a surgical team can result in greater likelihood of instrument-count discrepancies

among team members, possibly indicating retention of surgical instruments in a patient's body, with sponges and towels most common.⁶

Adding a robot to the operating theater as an assistant to a surgical team has the potential to reduce the number of miscommunications and their negative effects in two main ways: First, in the case of communication failure, a robotic scrub nurse (such as our Gestonurse) is able to deliver surgical instruments to the main surgeon communicating through hand gestures and speech recognition; timely, accurate surgical delivery to the surgeon can lead to decreased cognitive load, time, and effort for surgeons. And, second, the possibility of retained surgical instruments is avoided through accurate, thorough, timely tracking and monitoring of instruments used; retained instruments can puncture organs and cause internal bleeding. We have been developing Gestonurse at Purdue University for the past three years (see Figure 1).

The main use of robotics in surgery is not to replace the surgeon or surgical nurses but to work with them during surgery. In working side by side (see Figure 2), responsibility can be divided up like this: The robot passes instruments, sutures, and sponges during surgery and keeps an inventory of their use, while the surgical technician handles all remaining tasks (such as operating sterilizers, lights, suction machines, and electrosurgical units and diagnostic equipment and holding

» key insights

- Gestonurse is the first multimodal robotic scrub nurse to assist surgeons by passing and retrieving surgical instruments during simple procedures.
- Gestonurse recognizes both hand gestures and speech commands, mapping them to existing surgical instruments in a surgical tray.
- Gestonurse recognizes and tracks surgical instruments in use, retrieving them for the procedure, thus reducing the risk of retained instruments.



retractors and applying sponges to or suctioning the operative site). A robot controlled through hand gestures and speech commands is a natural alternative that does not affect the normal flow of surgery.

Related Work

Previous surgical robots include one used for object retrieval³ and others with haptic feedback (such as SOFIE²⁴ and the da Vinci surgical system¹¹ for minimally invasive procedures in endoscopic and laparoscopic surgeries).

Previous robotic scrub nurses include the voice-controlled “Penelope,”^{15,23} which localizes, recognizes,

and returns used instruments. Another voice-controlled robot,⁴ also uses computer-vision techniques to recognize, deliver, and retrieve surgical instruments. A problem with voice-only systems is performance degradation²² in noisy environments due to, say, the sound of drills, anesthesia machines, surgical staff side conversations, and operating equipment that can compromise patient safety; for example, a surgeon might say “50,000 units,” but the anesthetist hears “15,000 units.”¹ Errors can have dramatically adverse consequences for a patient’s well-being.

A voice-controlled robotic scrub nurse for laparoscopic surgery²⁷ uses

depth-based action recognition for instrument prediction; its 3D point-estimation method requires the surgeon wear markers of special reflective material for action recognition that could potentially compromise sterility.

Trauma Pod⁹ is a mobile-robotic-surgery effort sponsored by the U.S. Department of Defense intended to perform life-saving procedures on the battlefield; it responds only to voice commands, not physical gesture. Treat et al.²³ developed a robot that delivers instruments to the main surgeon following verbal requests, retrieving them as soon as they are no longer required. The instruments are identified through machine-vision algorithms, with decisions made through a cognitive architecture. In 2011, Jacob et al.¹²⁻¹⁴ and Wachs et al.²⁵ presented Gestonurse, the first surgical scrub nurse to understand nonverbal communication, including hand gestures.

In this article, we make two main contributions with respect to previous work: how verbal and nonverbal information can, when combined, improve the robustness of a robotic scrub nurse and how to assess the effectiveness of the interaction between surgeon and robot in an OR setting through a mock surgery—an abdominal incision and closure using a phantom simulator. Gestonurse gestures are recognized from the video/depth stream acquired by a Microsoft Xbox 360 Kinect sensor; a robotic arm then delivers the re-

Figure 1. Gestonurse robotic assistant.

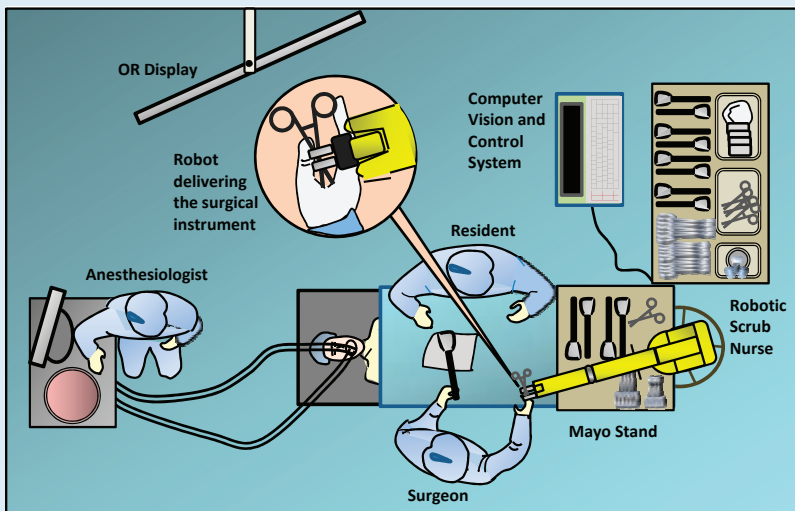
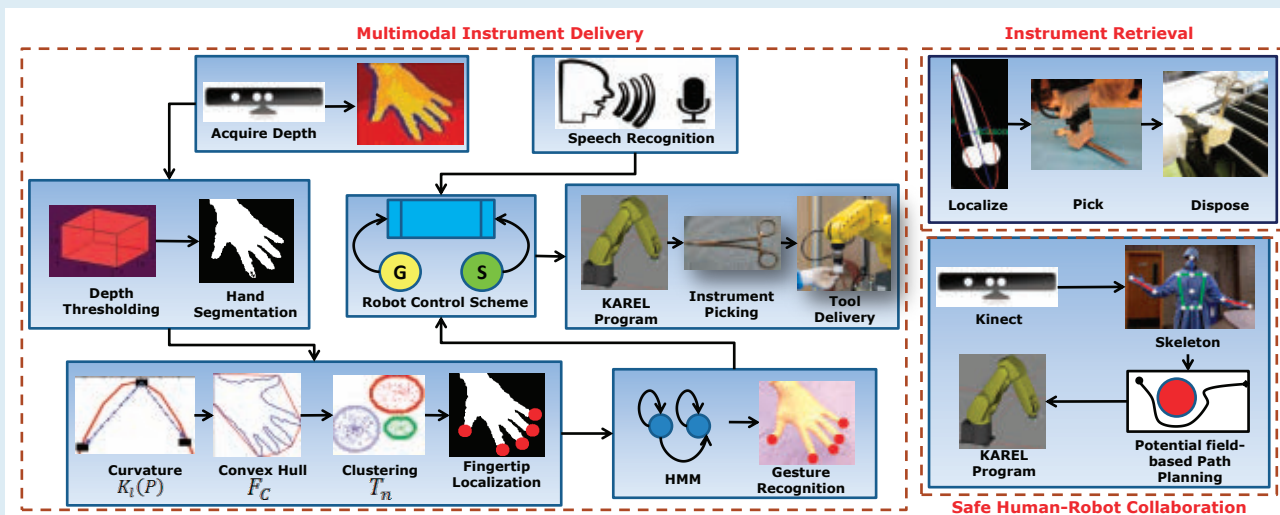


Figure 2. System overview.



requested instrument to the surgeon. A significant advantage of gesture-based communication is it requires no special training by the surgeon. Gesturing comes naturally to surgeons since their hands are already their main tools; moreover, hand signs are the standard method for requesting surgical instruments,^{8,19} and gestures are not affected by ambient noise in the OR. A multi-modal solution combining voice and gesture provides redundancy needed to assure proper instrument delivery.

Gestures for robotic control have been the focus of much research since the early 1980s. Early work was done with Richard A. Bolt's Put-That-There interface² followed by others using magnetic sensors or gloves to encode hand signs.^{20,21} Since then, gestures have been used in health care, military, and entertainment applications, as well as in the communication industry; see Wachs et al.²⁶ for a review of the state of the art.

System Architecture

Figure 2 outlines the Gestonurse system architecture. The streaming depth maps captured through the Kinect sensor are processed by the gesture-recognition module while a microphone concurrently captures voice commands interpreted by the speech-recognition module. Following recognition, a command is transmitted to the robot through an application that controls a Fanuc LR Mate 200iC robotic arm across the network through a Telnet interface. Gestonurse then delivers the required surgical instrument to the surgeon and awaits the next command. We also designed an instrument-retrieval-and-disposal system. A network camera monitors a specific region of the operating area, then, upon recognizing a surgical instrument, picks it up and delivers it to the surgeon. Meanwhile, the surgeon's hands are tracked to ensure robot and surgeon do not collide, ensuring safe human-robot collaboration.

Gesture recognition. To evoke a command, a member of the surgical staff places a hand on the patient's torso and gestures. The moment the hand is in the field of view, the gesture is captured by the Kinect sensor and segmented from the background through a depth-segmentation algo-



The surgeon requests scissors from Gestonurse using a hand signal recognized by the camera above the surgical bed.



Gestonurse safely hands off a surgical scissors to the surgeon requesting it.

Table 1. Gesture lexicon.











 (a) Scalpel	 (b) Bandage Scissors	 (c) Hook Retractor	 (d) Forceps	 (e) Hemostat
 (f) Needle	 (g) Speech On/Off	 (h) Scissors	 (i) Sleep/Wake	 (j) Retractor

Table 2. Confusion matrix at $\zeta = 0.99$ for the gesture lexicon.

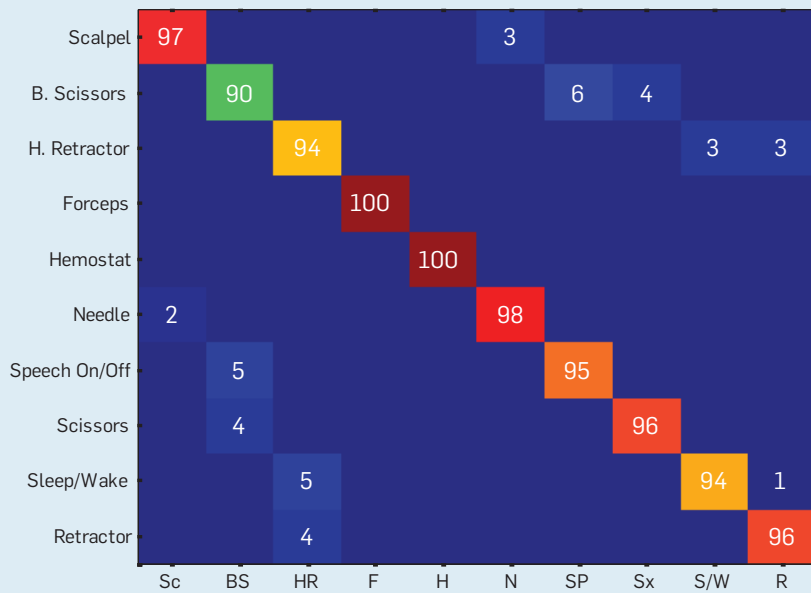
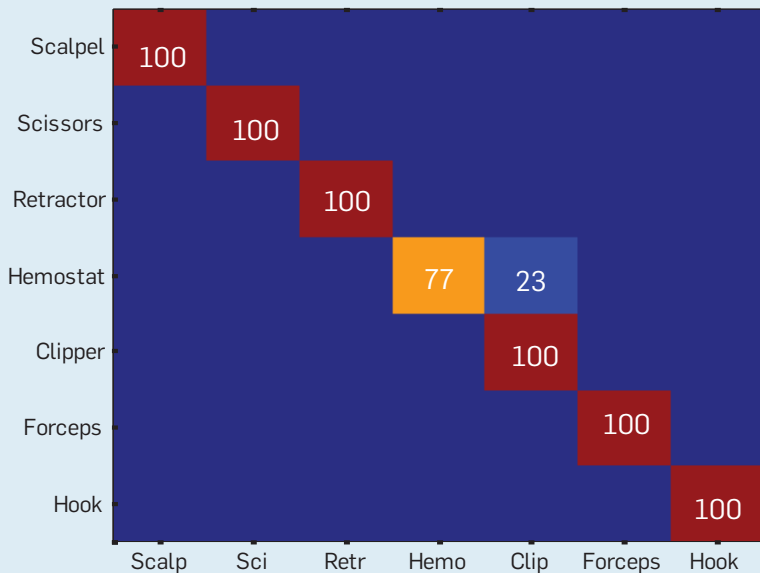


Table 3. Confusion matrix at $\eta = 0.5$.



rithm we developed.¹² A mask of the hand is obtained by thresholding the depth map and the area of the closest blob. The contour and convex hull of the blob identify the configuration of the fingers and associate the configuration to a hand in a static pose.¹³ A gesture is preceded by the hand in the static pose, terminating when the hand is not in the sensor's field of view. Hand-posture recognition uses this movement as a cue to temporally segment the gesture. The Gestonurse lexicon for surgical tasks (see Table 1) is based on standard OR gestures for requesting surgical instruments.^{8,19} The positions of the localized fingertips are recorded to obtain trajectories for each fingertip during the gesture. The trajectories are derived from the screen coordinates (in pixels) of the localized fingertip and the depth (in millimeters) of the fingertip with respect to the depth sensor; see the online Appendix for algorithmic details.

The system recognizes gestures to within 160ms (real time) from the time they were performed by the surgeon. The system's overall response time is slightly more than two seconds, including the time needed (160ms) to recognize the gesture, plus the overhead time the robotic arm needs to deliver the instruments (two seconds on average). The time the system needs includes for the surgeon to physically perform a gesture, approximately one to two seconds for experienced users and five to six seconds for novices.

Instrument localization and recognition. One of the many responsibilities of human surgical scrub techs is to remove instruments surgeons might place around the operating area during a procedure; this requires continuously monitoring the area to detect the presence of instruments. Detection means the instrument is localized, recognized, and finally removed. The camera monitoring the region is calibrated with respect to the robot so the coordinates of the instrument in the image plane can be converted to coordinates in the robot-frame view.

Robot control scheme. Delivering and retrieving instruments can be a risky activity due to potential harm to a patient and to the surgical team if potential collisions with the surgeon are not avoided. Imagine the robot passes

the scalpel and the surgeon moves a hand without noticing the scalpel in the way. In order to safely collaborate with a surgeon (on instrument retrieval and delivery), the robot must determine the position of the surgeon and plan a path to avoid a collision. We implemented a potential-field method to compute a safe path to that goal (fixed position for instrument delivery and variable position for instrument retrieval). Additionally, we use the skeleton-tracking ability of a Kinect sensor to track the position of the surgeon's hands. The camera is externally calibrated with the robot's coordinate system such that the position of the surgeon's hands is obtained in the robot's field of view.

Experiments. We conducted three main experiments to assess Gestonurse feasibility and robustness, evaluating gesture-recognition accuracy of its vision module, recognition of the instruments, and delivery and disposal of the surgical instruments. Finally we assessed users' learning ability, comparing the modalities used to communicate with the robot.

Gesture recognition. We assembled a database of 1,000 gestures from 10 users we asked to perform 10 gestures per instrument class, using 10-fold cross-validation to generate the set of receiver operating characteristic curves (see Figure 3) by varying a threshold parameter representing the strength of the recognition.

We normalized the log-likelihood scores from testing with the Viterbi algorithm such that the highest score was scaled to 1. We considered all instrument classes with normalized log-likelihood scores greater than ς to be

a positive result (a detection); $\varsigma \leq 0.99$ means an average gesture-recognition accuracy of 95.96%. Table 2 includes the confusion matrix for this operating point; we calculated the values as number of correct recognitions over the total number of instances presented.

Instrument recognition. We used a database of 700 color (RGB) images of 720x480 pixels each, including 10 images of the seven standard types of surgical instrument: scalpel, scissors, retractors, hemostats, clippers, forceps, and hooks. We segmented the instruments from the background by extracting nonstationary objects from the background buffer; we created and updated the background model using a Gaussian Mixture Model. We then used a support vector machine to classify the feature vectors of the segmented instrument. We used the fivefold cross-validation method to find the optimal support-vector parameter γ for

the radial basis function kernel function; Table 3 outlines the confusion matrix at the optimal $\gamma=0.5$. Though we assumed the instruments were clean (in a real-life scenario, they could be partially contaminated by blood or other secretions), the background model we used for instrument segmentation made detection independent of instrument color.

We found Gestonurse average instrument-picking accuracy to be 100%, with dropping accuracy of 92.38%. Some instrument classes perform much better than others, possibly due to variance in instrument shape.

Instrument picking and dropping. To measure performance, we placed instruments from the seven classes on a Mayo surgical tray, then recorded the robot's performance at picking and dropping per trial, conducting 15 trials total. We found instrument-picking accuracy of 100%, with av-

Figure 3. Receiver operating characteristic curve for Hidden Markov Model-based gesture recognition.

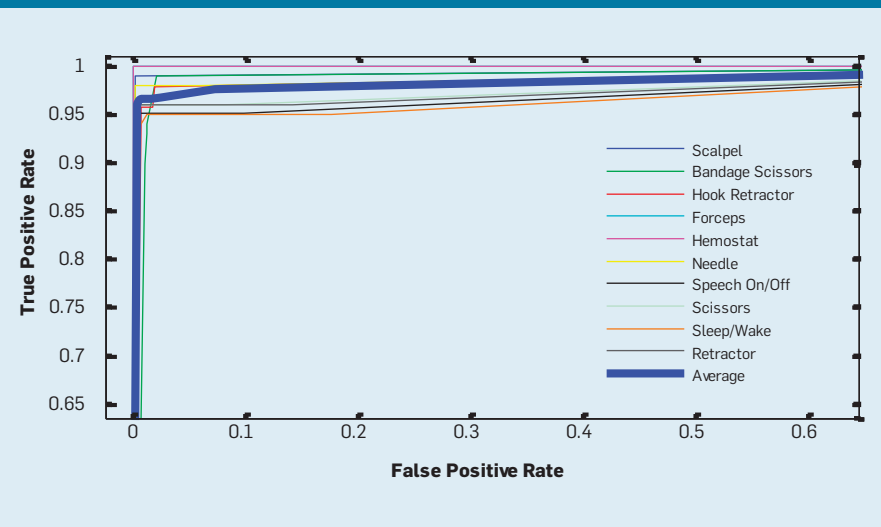
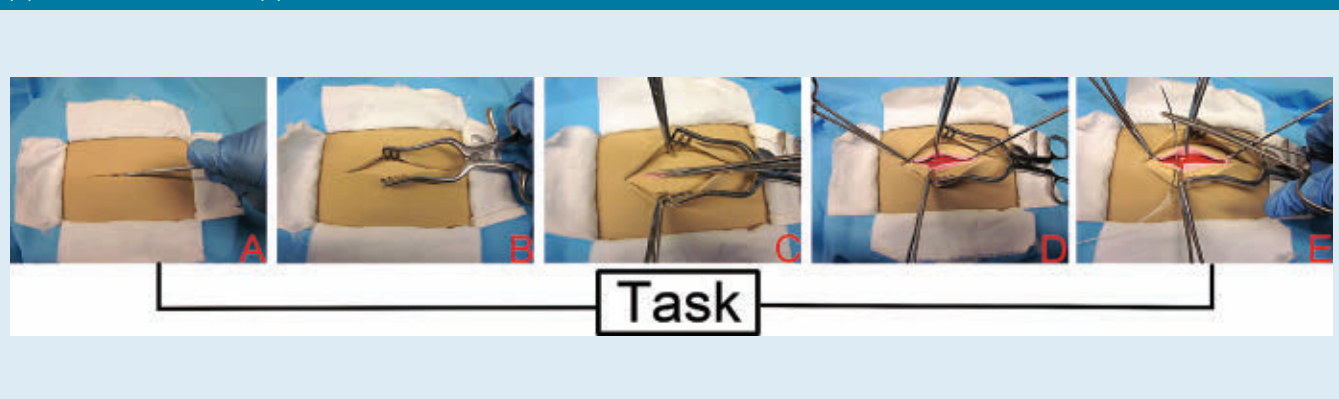


Figure 4. Stages of mock abdominal incision: (A) incision; (B) exposure of linea alba; (C) incision enlarged with scissors; (D) linea alba incised; and (E) incision closed.



erage dropping accuracy of 92.38%; per-class accuracy is included in the online Appendix.


Modalities compared. We recruited 12 graduate and undergraduate students, including eight males and four females, all 20 to 30 years old, to test the effectiveness of modality training on a mock surgical task simulating an abdominal incision and closure (see Figure 4), a task requiring five instrument classes: scalpel, scissors, needle, retractor, and four hemostats, a total of eight instruments.

We tested Gestonurse under three conditions: speech (S), gesture (G), and combined speech and gesture (SG). Note that in SG, we used the gestures and speech (see the online Appendix) to request the surgical instruments but not simultaneously. While Gestonurse can deal with simultaneous requests from multiple modalities, simultaneous requests using different modalities is not desirable during real-life surgeries. Surgeons are allowed to use only speech, only gestures, or speech and gestures one at a time during surgery.


We assigned the 12 subjects randomly to one of three test groups depending on whether they would be using speech, gestures, or both, each participating in two experiments. Subjects in the S and G groups could use only five commands to request five instruments from the robot for the mock procedure. We asked the SG test group to use speech to request half the required instruments and gestures for the rest.

Within each group, we trained two subjects to communicate with Gestonurse before performing the procedure, then asked them to repeat each command 15 times. We then read the name of the recognized instrument to the subject through a text-to-speech program, Microsoft SAM text-to-speech. We similarly conducted training for gesture recognition, with each test subject repeating each gesture 15 times and shown a bar graph with the log-likelihood score of the gesture for each gesture class.

Each subject performed the surgical task six times, while we recorded the task-completion times, which were determined mainly by type of surgical procedure, not the speed of the computer-vision algorithm; for



Another goal is to add the ability to predict the next likely surgical instrument according to the type of procedure (the context) instead of relying on a subjective, variable chain of verbal commands.



example, repairing an open abdominal aortic aneurysm can take up to eight hours.

Discussion

Having robotics support surgical performance promises shorter operating times, greater accuracy, and fewer risks to the patient compared with traditional, human-only surgery. Gestonurse assists the main surgeon by passing surgical instruments while freeing surgical technicians to perform other tasks. Such a system could potentially reduce miscommunication and compensate for understaffing by understanding nonverbal communication (hand gestures) and speech commands with recognition accuracy, as we measured it, over 97%. We validated the system in a mock surgery, an abdominal incision and closure. In it we computed learning rates of 73.16% and 73.09% for the test subjects with and without gesture training, indicating learning occurred at the same rate with and without gesture training, and that improvement of 75.44 seconds (12.92% less) in task completion time was due directly to the training provided to the test subjects prior to the six trials. This means the test subjects' skill was due to understanding and participating in the surgical task, rather than from learning to use hand gestures. Gesturing is presumably intuitive enough to be used by surgical staff with (almost) no training. Our informal discussions with surgical staff at Wishard Hospital, a public hospital affiliated with the Indiana University School of Medicine in Indianapolis, found surgeons excited about the possibility of using such a robot in a surgical setting.

The multimodal system is 55.95 seconds faster (14.9% less) than a speech-only system on average. However, we also found that gesture and voice together are no faster than gesture alone, performance that could be due to having to switch between modalities (and related additional cognitive load) that affects performance time. Our future work aims to address the kind of performance (in terms of functionality, usability, and accuracy) a robotic system must deliver to be a useful, cost-effective alternative to traditional human-only practice.

Conclusion

Gestonurse is a multimodal robotic scrub nurse we developed at Purdue and the Indiana University School of Medicine to reliably pass surgical instruments to surgeons and other members of a human surgical team, yielding gesture-recognition accuracy of 95.96% on average. The related Kinect-based robotic vision system we developed recognizes and picks instruments with a recognition rate of 92.38%. We also developed an instrument-picking system with 100% accuracy and a related disposal system with 92.83% accuracy (on average), as well as a field-path-planning algorithm to maximize safety in human-robot collaboration, implementing it in Gestonurse.

We conducted experiments on the effects of modality training for participants in a mock surgical procedure, calculating pre-task training delivered 12.92% reduced task-completion time on average across both speech and gesture. We also showed the system (following user training in both speech and gesture) is 14.9% faster than speech only, on average.

Future work on Gestonurse aims to fuse speech- and gesture-recognition data in a probabilistic fashion and transition to a real surgical setting involving animals at the Veterinary School at Purdue University in West Lafayette, IN. Another goal is to add the ability to predict the next surgical instrument likely to be needed according to the type of procedure (the context) instead of on a subjective, variable chain of verbal commands. We also aim to improve specific features of the system; for example, we assumed instruments were placed by surgical staff in fixed positions in a Mayo tray, with instrument coordinates saved as trajectory points in the teach pendant's (the robot's remote control) memory. Future versions will use the algorithm we developed to automatically detect and pick instruments, regardless of location. This will require minimal work since the current design already uses these techniques to retrieve instruments.

Acknowledgments

This project was funded, in part, by the Indiana Clinical and Translational Sciences Institute, by Grant Num-

ber Grant #TR000006 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. We also thank Dr. Steve Adams of Purdue University for his consultation and use of the Veterinary School OR, Dr. Rebecca Packer of Purdue for the surgical supplies we used in the experiment, and Hairong Jiang of Purdue for her support implementing the Gestonurse obstacle-avoidance module. **C**

References

1. Beyea, S.C. Noise: a distraction, interruption, and safety hazard. *AORN Journal* 86, 2 (2007), 281–285.
2. Bolt, R.A. Put-That-There: Voice and gesture at the graphics interface. *Commun. ACM* 14, 3 (1980), 262–270.
3. Borenstein, J. and Koren, Y. A mobile platform for nursing robots. *IEEE Transactions on Industrial Electronics* 2 (2007), 158–165.
4. Carpintero, E., Perez, C., Morales, R., Garcia, N., Candela, A., and Azorin, J. Development of a robotic scrub nurse for the operating theatre. In *Proceedings of the Third IEEE International Conference on Biomedical Robotics and Biomechatronics* (Elche, Spain, Sept. 26–29, 2010), 504–509.
5. Carthey J., de Laval, M.R., Wright, D.J. et al. Behavioral markers of surgical excellence. *Safety Science* 41, 5 (2003), 409–425.
6. Egorova, N.N., Moskowitz, A., Gelijns, A. et al. Managing the prevention of retained surgical instruments: What is the value of counting? *Annals of Surgery* 247, 1 (2008), 13–18.
7. Firth-Cozens, J. Why communication fails in the operating room. *Quality Safety Health Care* 13, 5 (Oct. 2004), 327.
8. Fulchiero, G.J., Vujevic, J.J., and Goldberg, L.H. Nonverbal hand signals: A tool for increasing patient comfort during dermatologic surgery. *Dermatological Surgery* 35, 5 (2009), 856–857.
9. Garcia, P., Rosen, J., Kapoor, C., Noakes, M., Elbert, G., Treat, M., Ganous, T., Hanson, M., Manak, J., Hasser, C., Rohler, D., and Satava, R. Trauma pod: A semi-automated telerobotic surgical system. *International Journal of Medical Robotics and Computer-Assisted Surgery* 5, 2 (2009), 136–146.
10. Halverson, A.L., Casey, J.T., Andersson, J., Anderson, K., Park, C., Rademaker, A.W., and Moorman, D. Communication failure in the operating room. *Surgery* 149, 3 (2010), 305–310.
11. Intuitive Surgical. da Vinci Surgical System; <http://www.intuitivesurgical.com/>
12. Jacob, M.G., Li, Y., Akingba, G., and Wachs, J.P. Gestonurse: A robotic surgical nurse for handling surgical instruments in the operating room. *Journal of Robotic Surgery* 6, 1, (Mar. 2012), 53–63.
13. Jacob, M.G., Li, Y., and Wachs, J.P. A gesture-driven robotic scrub nurse. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (Anchorage, AK, Oct. 9–12, 2011), 2039–2044.
14. Jacob, M., Li, Y.T., and Wachs, J.P. Gestonurse: A multimodal robotic scrub nurse. In *Proceedings of the Seventh ACM/IEEE International Conference on Human Robot Interaction* (Boston, MA, Mar. 5–8). ACM Press, New York, 2012, 153–154.
15. Kochan, A. Scalpel please, robot: Penelope's debut in the operating room. *Industrial Robot: An International Journal* 32, 6 (2005), 449–451.
16. Kohn, L.T., Corrigan, J., and Donaldson, M.S. *To Err Is Human: Building a Safer Health System, Volume 6*. Joseph Henry Press, 2000.
17. Li, Y.T., Jacob, M., Akingba, G., and Wachs, J.P. A cyber-physical management system for delivering and monitoring surgical instruments in the OR. *Surgical Innovation* (PubMed: 23037804); <http://sri.sagepub.com/content/early/2012/10/03/1553350612459109.abstract?rss=1>
18. Lingard L., Espin S., Whyte S. et al. Communication failures in the operating room: An observational classification of recurrent types and effects. *Quality*

- Safety Health Care* 13, 5 (2004), 330–334.
19. Phillips, N., Berry, E., and Kohn, M. *Berry & Kohn's Operating Room Technique*. Mosby/Elsevier, 2004.
20. Sturman, D.J. and Zeltzer, D. A survey of glove-based input. *IEEE Computer Graphics and Applications* 14, 1 (1994), 30–39.
21. Starner T. and Pentland A. Visual recognition of American Sign Language using Hidden Markov Models. In *Proceedings of the International Workshop on Automatic Face and Gesture Recognition* (Zurich, Switzerland, 1995),189–194.
22. Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H., and Shikano, K. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 4 (May 2009), 650–664.
23. Treat, M.R., Amory, S.E., Downey, P.E., and Taliaferro, D.A. Initial clinical experience with a partly autonomous robotic surgical instrument server. *Surgical Endoscopy* 20, 8 (2006), 1310–1314.
24. van den Bedem, L. *Realization of a Demonstrator Slave for Robotic Minimally Invasive Surgery*. Ph.D. dissertation, Technische Universiteit Eindhoven, Eindhoven, the Netherlands, 2010; <http://alexandria.tue.nl/extra2/684835.pdf>
25. Wachs, J.P., Jacob, M.G., and Li Y. Does a robotic scrub nurse improve economy of movements? In *Proceedings of the Image-Guided Procedures, Robotic Interventions, and Modeling Conference, SPIE Medical Imaging* (San Diego, Feb. 5–7, 2012).
26. Wachs, J.P., Kölsch, M., Stern, H., and Edan, Y. Vision-based hand-gesture applications: Challenges and innovations. *Commun. ACM* 54, 2 (Feb. 2011), 60–71.
27. Yoshimitsu, K., Miyawaki, F., Sadahiro, T., Ohnuma, K., Fukui, Y., Hashimoto, D., and Masamune, K. Development and evaluation of the second version of scrub nurse robot for endoscopic and laparoscopic surgery. In *Proceedings of the Intelligent Robots and Systems Conference* (Oct. 29–Nov. 2, 2007), 2288–2294.

Mithun George Jacob (mithunjacob@purdue.edu) is a Ph.D. student in the Department of Industrial Engineering at Purdue University, West Lafayette, IN.

Yu-Ting Li (yutingli@purdue.edu) is a Ph.D. student in the School of Industrial Engineering at Purdue University, West Lafayette, IN.

George A. Akingba (aakingba@iupui.edu) is an assistant professor of surgery and biomedical engineering in the Division of Vascular Surgery at the Indiana University School of Medicine, Indianapolis, IN.

Juan P. Wachs (jpwachs@purdue.edu) is an assistant professor in the Industrial Engineering School at Purdue University, director of the Intelligent Systems and Assistive Technologies Lab at Purdue, and affiliated with the Regenstrief Center for Healthcare Engineering, West Lafayette, IN.

DOI:10.1145/2447976.2447994

Markets characterized by multiple competing digital standards have room for more than one winner, unlike traditional analog markets.

**BY CHRIS F. KEMERER, CHARLES ZHECHAO LIU,
AND MICHAEL D. SMITH**

Strategies for Tomorrow's 'Winners-Take-Some' Digital Goods Markets

MOST MANAGERS (AND CONSUMERS) understand the key patterns of market evolution from earlier standards wars.¹⁰ History provides a number of examples that begin with two or more similar, but incompatible, information technologies introduced to address consumer market needs. Incompatibilities between the technologies mean users of one cannot enjoy the benefits of the other, in terms of either users to communicate with or of content to consume.

Vendors of both technologies, recognizing the network effects associated with adoption, start a “standards war,”

given their expectation that only one will win, and thus that firms must compete *for* the market before they compete *in* the market. This result is common in markets with networks of complementary goods (such as software for hardware, media for players, and games for game consoles), where the market desires a single, dominant standard, and consumers prefer to adopt the market leader, and may even withhold purchases until a dominant technology emerges.¹⁵

In order to win a standards war vendors may engage in competitive behavior, (such as subsidizing early adopters to increase network size and offsetting the lack of network benefits to early adopters), thereby causing the market to “tip” to their technology platform.^a When tipping occurs, the winning firm can extract economic rents from its dominant position in the market, and future-generation technologies must then offer significant improvements

a “Platform” is used here as defined by Eisenmann, et al.: “...products and services that bring together two distinct groups of user in a two-sided network.” Platforms tend to become standards, though not all standards are platforms. Our analysis here is of technology standards broadly, though some examples are platforms. Markets that tip to a single, dominant standard are commonly called winner-takes-all, despite the fact that the dominant market share may be less than 100%.

» key insights

- **Winner-takes-all markets, where products tend to “tip” toward a single standard, have been common in IT, as in VHS over Betamax, Blu-ray over HD-DVD, and Microsoft Windows and Office over multiple competitors.**
- **Changes in development and delivery of digital goods may portend a winners-take-some outcome that demands a switch from yesterday’s subsidization of early adopters and later reliance on network effects to drive the market toward tipping.**
- **Digital formats conversion across platforms increasingly allows coexistence of multiple winners, as with hardware devices, like flash memory cards, and with multiple digital audio, video, and image formats, allowing conversion across multiple standards.**



to overcome the network advantages of the incumbent platform. Examples of such standards wars include VHS over Betamax VCRs, DVD over Divx, Blu-ray over HD-DVD, and the XM-Sirius satellite merger;^{1,4,5} see the sidebar “Digital Winner-Takes-All Standards.”

These examples share a number of characteristics: Competing technologies were effectively substitutes; competitors’ formats were incompatible; complements (media, software, and content) were critical for consumer value; and, most important, technologies were not easily converted from one standard to another due to factors including the time and effort involved

in conversion, quality degradation inherent to conversion, technological restrictions or limitations, and digital rights management (DRM) restrictions.

Emerging Winners-Take-Some Market

Managers should note this established pattern of strategic interaction might become less relevant in the context of digital standards, where cheap and perfect conversion from one format to another is possible. In this setting our research suggests managers are more likely to face a winners-take-some outcome where multiple different standards can coexist.

Examples of this new competitive environment are appearing in a variety of contexts; for example, while the content-platform characteristic of flash memory cards and card converters may look similar to the competition between VHS and Betamax, flash memory cards have not yet seen a strong winner-takes-all outcome, where one dominant standard emerges. Instead, the flash memory card market has seen multiple formats—Compact Flash, Memory Stick, Secure Digital, Smart Media, xD Picture, and MultiMedia Card—with no obvious trend toward market consolidation (see Figure 1).

Figure 1. Flash memory card market share, January 2003–August 2006; data source: NPD Group¹¹

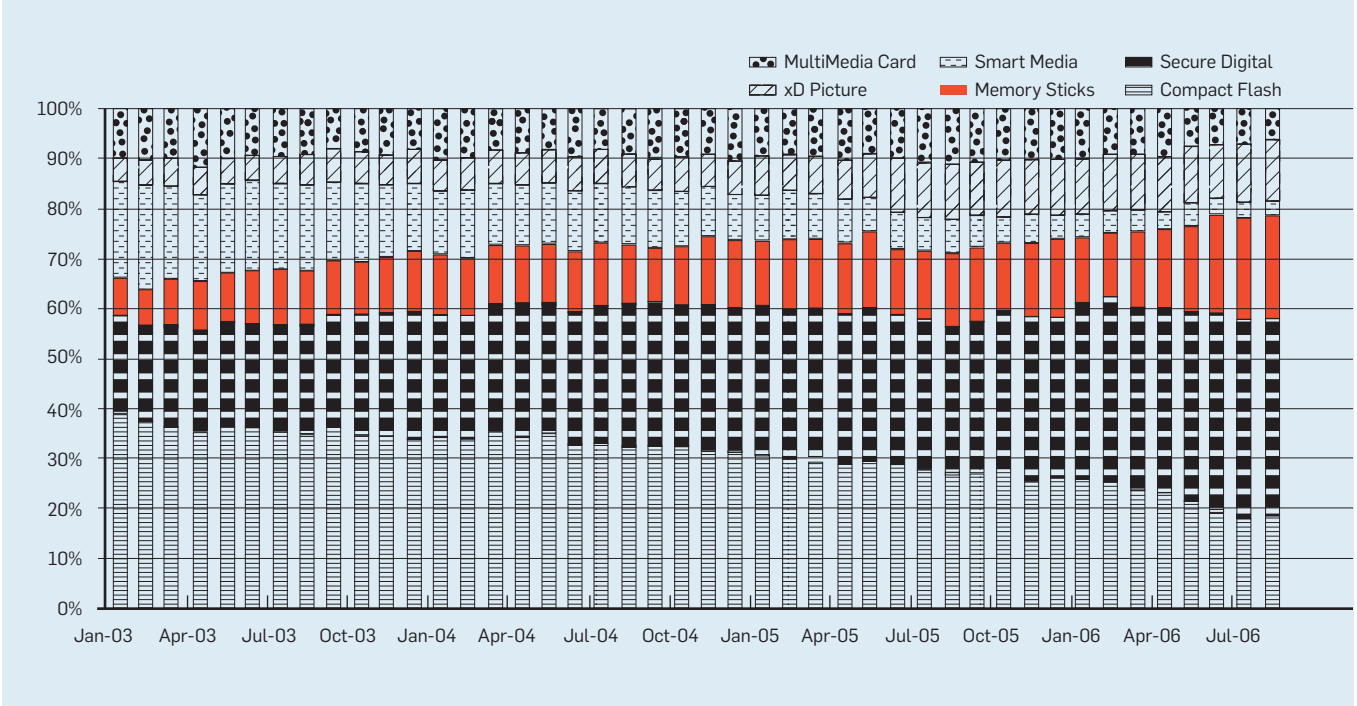
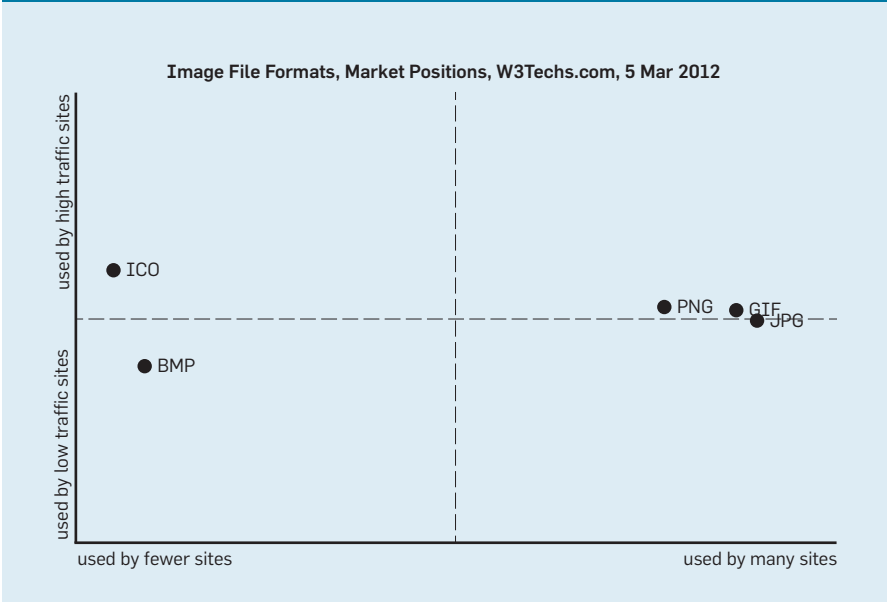


Figure 2. Adoption of digital image formats by market position and Web traffic; data source: http://w3techs.com/technologies/market/image_format



Digital-image formats use, November 2011–April 2012, top one million websites; data source: http://w3techs.com/technologies/history_overview/image_format/all

Digital Image Format Used	November 2011	April 2012
JPG	71.8%	72.4%
GIF	69.9%	67.3%
PNG	50.9%	55.6%
BMP	0.8%	0.7%
ICO	0.2%	0.2%
None	9.9%	9.4%

In the case of digital image formats several competing standards, including .jpg, .gif, .png, .bmp, and .tiff, have coexisted for years. Although some formats tend to be more popular than others, there is little tendency toward a winner-takes-all outcome, perhaps best illustrated in the adoption of various digital image formats by websites worldwide (see the table here).

It is clear that the majority of these websites adopt multiple formats to display images, but a dominant position has not led to a self-reinforcing growth path. The market shares of the three leading formats have been relatively stable over time, and, in fact, the market share of the third leading image format—.png—has shown some growth, a phenomenon that does not support what might be predicted by the classic theory of network effects. Moreover, the popularity of the leading formats is not driven by Web traffic, hence not by visitors' preference, suggesting compatibility among these formats is not a major factor in adopting a particular digital image format (see Figure 2).

Similar situations are seen in other digital media formats, as in audio (such as .wav, .aac, .mp3, .wav, .aac, .mp3, .wma, .flac, and Apple Lossless), video (such as .wmv, .mpg, .avi, .flv, and .mov), and file compression

(such as .arj, WinRAR, and WinZip). These examples suggest that first-mover advantage does not always translate into the persistent market power to be expected of a winner-takes-all outcome.^b

Several factors drive this trend: First, in a digital environment, a large number of essentially equivalent designs is possible, making an increased variety of independently produced formats more likely. In contrast, in an analog environment, natural laws tend to limit the design space. Additionally, digital forms are relatively easy to copy, encrypt, compress, and communicate than their analog counterparts, all of which reduce the overall cost of diversity.

However, this environment suggests only how multiple standards come into existence, not why they survive. One critical factor enabling the coexistence of competing digital standards is the presence of hardware- or software-based digital converters; for example, hardware-based flash memory converters allow users of one standard to easily transfer their content onto other devices through the ubiquitous USB interface. For digital file conversion, a computer with appropriate software can serve as a flexible universal converter, allowing for, say, straightforward conversion of a .jpg image file to a .gif image file. Likewise, video editors on a Macintosh platform easily convert .wma audio files created on Windows PCs to an iTunes-compatible .aac format, with little discernible loss in media quality.

In contrast, prior to the digital revolution, analog media “readers” were typically fixed in hardware and relatively inflexible. Conversion between two incompatible standards in this context was slow, and led to significant signal loss; for example, conversion from vinyl record albums to analog tape is costly in terms of both time and lost audio quality. Likewise, providing the ability to play two incompatible formats (such as VHS and Betamax videotapes) and write in at least one of them would nearly double the cost of the hardware.⁴ These significant costs

^b Not all analog markets become winner-takes-all, and not all digital markets become winners-take-some; however, when conversion is essentially lossless and costless, the likelihood of a multiple-winners outcome increases substantially.

Digital Winner-Takes-All Standards

But wait, aren't Blu-ray discs digital? And isn't satellite radio a digital signal? Why do these markets have a single winner?

Blu-ray/HD-DVD. It is important to note that while Blu-ray content is digital, it is encoded onto a fixed medium not easily converted between standards due to intellectual-property protections. Moreover, differences in the lasers used to encode Blu-ray and HD-DVD content mean “dual players” would cost nearly twice as much as a single player (similar to a dual VHS/Betamax player). As a consequence, the Blu-ray vs. HD-DVD “standards war” had a single winner, as with Beta-VHS.

XM and Sirius. XM and Sirius were established by the U.S. Federal Communications Commission as separate licenses to promote competition. However, bidding for content and subsidization of early adopters threatened to bankrupt both sides. Their merger was the solution to the failed regulatory enforcement of multi-vendor competition. A converter, in the form of a dual receiver, would still be required to receive both signals. For these reasons a single provider has survived.

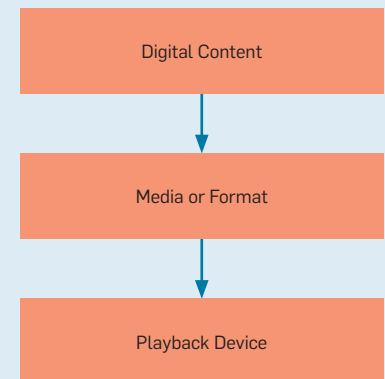
of multi-homing led consumers to choose a single standard, and a single winner generally emerged.^c

Digital Products Delivery Chain

A model is helpful for observing similarities among what otherwise might be seen as disparate products. Figure 3 highlights three essential elements in the digital-products delivery chain, starting with Digital Content. These information goods represent anything that can be encoded digitally, including data, images, music, and video. Producers of information goods must decide how they deliver these goods, represented as the second step—Media or Format—in the delivery chain. In the analog era, goods were delivered in fixed media (such as videocassette tapes and vinyl records). The first move toward full digitalization was digital information communicated on physical media (such as audio CDs and video DVDs); see the sidebar “Digital Winner-Takes-All Standards.” Digital goods today are increasingly delivered as a stream of bits following a standard format (such as one of the audio, image, and video formats outlined earlier). Formats can be seen as “contain-

^c Consistent with Eisenmann et al.,⁸ “multi-homing” is used here as “affiliating with multiple platforms”; for example, in the context of video games, demand-side multi-homing would involve a consumer with more than one video-game console, while supply-side multi-homing would involve a video-game creator developing versions for more than one console platform.

Figure 3. Digital-products delivery chain.



ers” that fulfill the role once filled by physical media.

Finally, the end consumer needs a playback device, or reader, to allow consumption of the information good. In the analog era these were single-purpose devices (such as VCRs and players), a model that persists today in single-purpose e-book readers. Increasingly, however, general-purpose devices with built-in converters serve the playback role for multiple media types; for example, e-books may be read on multipurpose devices (such as a PC, a smartphone, or tablets like Apple’s iPad⁶), reducing what would otherwise be multi-homing costs.

Future Standards Wars

How might these trends evolve in the future when even more products are digital? It seems likely that media quality will continue to be an impor-

tant aspect of consumer adoption decisions. The cost of digital conversion will continue to fall, given the prevalence of general-purpose computers and increasing reliance on media consumption through software-based devices and the Internet (such as Google Docs and other cloud-based services). It also seems likely that important technology markets will continue to have strong complementary goods relationships due to lower compatibility barriers. As a result, consumers will increasingly value product features over mere platform compatibility, and design features and functionality will be key dimensions of competition; see, for example, Apple's history with the iPhone and iPad.³

If these predictions hold we can expect an environment where technology vendors benefit from coordinating with other firms through cross-licensing agreements to increase their total effective market size.¹¹ In response, consumers will be more aggressive about early adoption of technologies, since the risk of being "stranded" on the wrong technology is reduced. This early consumer adoption should lead to a larger and more competitive market, more rapid technology innovation, and potentially more entrants in standards- and platform-based markets. Finally, we can expect more incremental technological changes relative to prior analog markets because there will be fewer installed-base barriers of the kind that might cause discontinuous, step-function technology

changes during platform-change windows (such as from analog tapes and records to compact discs, and from floppy drives to CD-ROMs) and the attendant rush to upgrade to the latest media format.

From Here to There

The Digital Markets Evolution Diamond (see Figure 4) outlines three potential paths that might be taken by technology vendors, as well as consumers, in the evolution from an analog winner-takes-all outcome to a digital winners-take-some market. The simplest, most direct path would be from Figure 4's Stage 1 to 3 via 2—Direct Digital Transition—where products move directly to a digital format (such as from analog TV signals to digital TV signals). However, such direct evolution may turn out to be a special case, and, perhaps more likely in the short run we will see two "detours" to the same end result. In the first—the left-hand path from Stage 1 to 2a to 3—the market evolves by undergoing a transition stage through fixed media. Products move to a digital future in two steps, the first a digitally based transitional form (such as from analog vinyl records to digital CDs), then a second (such as from digital CDs to pure digital downloads, or from analog video tapes through digital DVDs to digital downloads); see the sidebar "Netflix: A Missing Link."

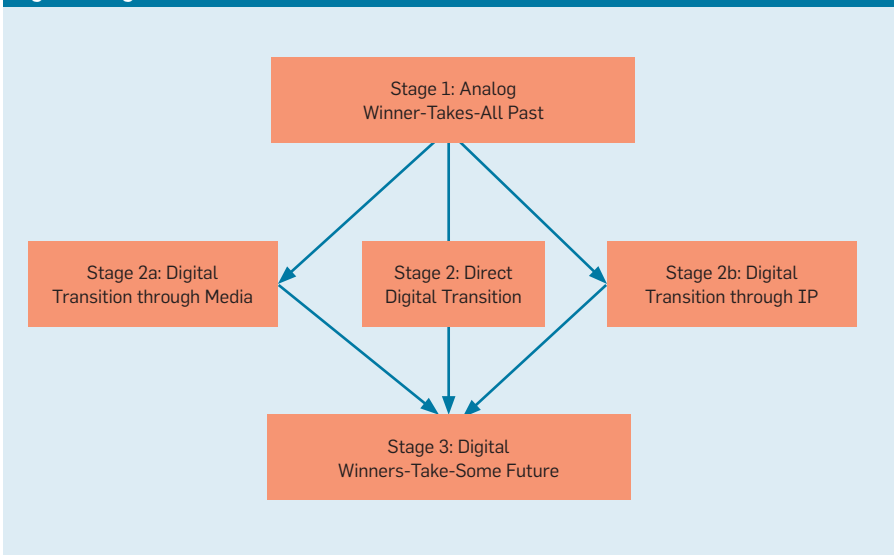
A second detour on the evolution to Stage 3 is a right-hand path from Stage 1 to 2b to 3 in Figure 4. The

product's first step is a set of multiple digital formats vendors guard with traditional intellectual property protections (such as patents and copyrights) while still imposing a winner-takes-all outcome supporting their technology. The earlier Stage 1 analog formats typically provided rational explanations for single winner-takes-all outcomes (such as the physical incompatibility between VHS and Betamax tapes). Similarly, the installed base of software may have created disincentives for multi-formats due to learning costs and the incentives for winner-takes-all outcomes through network effects due to platform creation and the benefit of a large number of complementary products. In Stage 2b vendors attempt to replicate winner-takes-all outcomes by creating proprietary digital formats protected by intellectual-property controls.

One example of Stage 2b intermediate migration is digital music from Apple's iTunes DRM to DRM-free downloads. As noted earlier, while inter-standards conversion is easy for most digital goods, products with DRM typically cannot be converted between formats. The record industry's mandated use of Apple-controlled DRM may have created a virtuous cycle for Apple, where customers who purchased content on the iTunes store were locked into using iPod media devices, and, out of convenience, most iPod users used the iTunes store to purchase music. This may have contributed to the early market dominance of the iTunes Music Store and the resulting market power Apple was able to exercise over music labels in pricing and marketing negotiations.¹³ Other observers have made similar comments about the market for e-books.¹⁴ In order to avoid such outcomes in the future producers of complementary content goods may have a strategic incentive to support multiple competing standards to reduce the likelihood of having to deal with a monopolist partner.⁹ With digital goods it may also be less costly for such a producer to convert its content to multiple formats.

However, such Stage 2B approaches may be short-lived, as the ease and quality of digital conversion makes it difficult to create advantages for pro-

Figure 4. Digital-markets evolution diamond.



proprietary formats (such as Sony's unsuccessful attempt to establish its Memory Stick format as the dominant flash memory standard) and given users' ability to defeat proprietary schemes to create constrained environments (such as so-called "jailbreaking" of iPhones and the defeat of copy-protection schemes in general).²

Moreover, in winner-takes-all markets vendors of digital technology have sought to establish their formats as a dominant standard and protect it from being copied. Sometimes, this took the form of not licensing their innovations to other firms, so as to retain sole manufacturing rights (such as Sony with Betamax and Apple with the Macintosh operating system). As these products could maintain higher margins they tended to command only niche markets. Therefore, these vendors moved to partner with other firms to co-produce devices or their complementary goods while still aiming to establish a single standard. In 1979 Sony successfully teamed with Philips to produce the audio CD standard, and video-game console manufacturers have contracted with video-game software producers to create entertainment systems that tend to produce generational, winner-takes-all results in video games (such as Nintendo's NES and Sony's PlayStation).⁷ However, the approach has seen notable failures as well; Sony, in particular, has created or backed a variety of unsuccessful efforts to standardize devices; see the sidebar "Sony Lessons Learned, Lessons Missed."

The market for digital technology may be seeing the emergence of an alternative strategy; for example, in flash memory, instead of attempting to promote a proprietary single standard, as Sony did with the Memory Stick, SanDisk sells a variety of flash memory formats. Likewise, Amazon provides converters allowing its users to read its Kindle DRM-protected titles on Kindle devices, as well as on other portable devices, including the iPad and iPhone.⁶ Finally, recent versions of Microsoft's Office productivity suite allow users to save their output in non-Microsoft formats (such as .pdf) and have made the file-format standard more accessible through .xml; see the sidebar "Microsoft Word and Adobe .pdf."

Netflix: A Missing Link

Netflix, Inc. is a subscription service that began as a "DVD rental by mail" service and has since begun offering streaming content over the Internet. At the end of 2011 Netflix had 23 million streaming subscribers and prior to that was mailing approximately two million DVDs on an average day. Netflix represents a classic transition path through fixed media. While it may ultimately provide only a direct digital download service, it began life by offering an alternative to making a trip to the video store. Sources: http://online.wsj.com/quotes/key_facts.html?mod=2_0470&symbol=NFLX&news-symbol=NFLX and <http://www.Netflix.com>

Sony Lessons Learned, Lessons Missed

Sony has been extremely successful in the consumer electronics market, including with its Walkman cassette player, audio CD standard (jointly with Philips), and PlayStation video-game system. However, less visible are a number of product attempts that have been relatively unsuccessful. Beyond Betamax, which, despite losing the standards war, went on to be successful as a commercial videotape standard, a variety of other Sony products failed to establish themselves in the marketplace. According to business author Steve Knopper, these include the Minidisc audio format, PressPlay music store, MusicClip (an SDMI-compliant digital music player based on the ATRAC DRM-protected standard), Connect music store (also ATRAC-based), and eXtended Copy Protection placed on music CDs via a software rootkit. See Knopper, *S. Appetite for Self-Destruction: The Spectacular Crash of the Record Industry in the Digital Age*. Free Press, New York, 2009.

Microsoft Word and Adobe .pdf

Two digital formats with significant worldwide installed bases are Microsoft Word (.doc and .docx) and Adobe Portable Document Format (.pdf). How do they fit within the trends described here? In the specific example of Word, high switching costs and complementary investment in learning or training played an important role in its early dominant market position. Further, converters (such as OpenOffice Writer) were introduced so late in the process that most potential users likely already invested in Word and its related learning and training. Moreover, converters to other formats are often imperfect, and, therefore, conversion is not lossless in the example of Word, meaning a winners-take-some outcome was less likely from the start.

The Adobe example involves similar technical "costs" to conversion but in this case through patents and proprietary standards. It is important to note that Adobe's initial strategy with .pdf was to exploit the two-sided nature of the reader/writer markets by giving away its reader software to consumers as a way to increase the utility of its .pdf writer software to publishers.⁷ However, to execute this strategy, Adobe needed to maintain its monopoly position in the writer market, through a combination of DRM (.pdf documents encoded with Adobe DRM can be read only by Adobe's reader software) and intellectual-property protections by holding patents that prevent other companies from developing competing ".pdf writer" software. This protection strategy was successful for Adobe for many years, though it was abandoned in 2008 when Adobe allowed its patents to be licensed royalty free and applied for ISO standardization of the .pdf standards.

These examples may reflect the first wave of a new strategy where platform rights holders choose to allow conversion in many cases. In the context of digital products a new equilibrium can emerge with technology vendors agreeing to provide converters at a suf-

ficiently low price to all consumers.¹¹ In this approach both the incumbent firm and potential new entrants are better off, since the possibility of conversion between formats provides multiple benefits: it helps adoption of both existing and new products, as consum-

ers need not wait on the sidelines for fear of being stranded by choosing the wrong standard; it reduces the need for price competition and subsidies to try to create a single winning standard; and it may even generate revenue through the sale of devices or software that perform the conversion. Moreover, users may benefit from being part of a larger network and generally having more opportunity to consume the new product.¹²

Our research in the market for flash memory shows that a variety of formats coexist in a winners-take-some outcome, rather than the traditional winner-takes-all outcome.¹² We find the existing network effects in flash memory use are moderated by the adoption of digital converters; specifically, digital converters provide a measurable reduction in the price premium of leading flash-card formats relative to that of formats with smaller market shares. These market dynamics imply that the provision of conversion technology increases the ability of new entrants to survive the standards competition, as converters tend to neutralize the impact of network effects. Our further analysis shows that market concentration in the flash memory market decreases as converters become more widely available, implying that adoption of converters fosters a more competitive market.

A variety of new and emerging products may fit this model; for example, there is intense competition in the e-book market among Amazon's Kindle, Apple's iPad, Barnes and Noble's Nook, and others.⁶ Given the digital nature of the content it seems probable that a winners-take-some result will emerge, with the ability for potential consumers to consume e-book content on multiple platforms, rather than a classic winner-takes-all outcome. This winners-take-some outcome is made possible by, in part, the fact that the cost for vendors of stocking multiple formats is much lower for digital goods than it was in, say, the Beta and VHS videocassette tape era, when significant quantities of physical inventory had to be kept in each supported format. With digitization and cheap, perfect copying a single master digital copy in each format is sufficient.

Caveat Manager

Predicting the future is, of course, a tricky business. While we expect to see the winner-takes-all phenomena replaced by the winners-take-some phenomena in many markets for digital goods, we also expect exceptions to emerge. What signs should a manager look for as advance warning that the market being pursued is unlikely to proceed to a winners-take-some outcome? We imagine three important conditions: First, especially early on, traditional market power may still prevail, with big vendors with deep pockets and strong distribution links in the marketplace choosing to follow the old rules and survive for an initial period of time. Eventually, though, as more examples of winners-take-some outcomes emerge, fewer technology vendors will take the risk. In addition, vendors that elect to try to follow the traditional path will be subject to increasing governmental antitrust oversight, as has been the case with many IT firms, including Google, Intel, and Microsoft.

A second exception may occur when a few collaborators in a consortium emerge to share in the financial returns, but also work to keep out others so as to keep sharing to a minimum. This is another market-power exception, but with an oligopoly instead of a monopoly outcome. These results are likely to be an initial transition point for market leaders that increasingly perceive the risks of a go-it-alone strategy.

Finally, we may still see winner-takes-all outcomes when governments dictate or otherwise greatly reward them. In some circumstances it may be appropriate, as when there are significant social and private costs of nonstandardization (such as HDTV and telecommunication standards) or when scale makes conversion a relatively expensive option. However, other circumstances will see less benign government intervention as when, say, regulators, under the influence of organizations with market power, or through a "fighting the last war" analysis of winner-takes-all markets, issue regulations that favor single winners. Managers are well advised to closely monitor emerging government policies in this regard.

Looking Ahead

The movement toward greater digitization will bring about an overall better marketplace for vendors and consumers alike, marked by quicker technology innovation, fewer consumer "dead-weight" losses due to technological stranding, more product choices, less vendor risk, and more interoperability. Managers should prepare to seize the related opportunities rather than fight the last war. **C**

References

1. Brynjolfsson, E. and Kemerer, C. Network externalities in microcomputer software: An econometric analysis of the spreadsheet market. *Management Science* 42, 12 (Dec. 1996), 1627–1647.
2. Cleary, P.J. The Apple Cat and the Fanboy Mouse: Unlocking the Apple iPhone. *North Carolina Journal of Law & Technology* 9, 2 (Spring 2008), 295–322.
3. Cusumano, M. The puzzle of Apple. *Commun. ACM* 51, 9 (Sept. 2008), 22–24.
4. Cusumano, M.A., Mylonadis, Y., and Rosenbloom, R.S. Strategic maneuvering and mass-market dynamics: The triumph of VHS over Beta. *Business History Review* 66, 1 (Spring 1992), 51–94.
5. Dranove, D. and Gandal, N. The DVD vs. DIVX standard war: Empirical evidence of network effects and preannouncement effects. *Journal of Economics and Management Strategy* 12, 3 (Fall 2003), 363–386.
6. Dunn, B.K. and Kemerer, C. *Barnes & Noble: Managing the eBook Revolution*. University of Pittsburgh teaching case, Aug. 2012.
7. Dunn, B.K. and Kemerer, C. *Class Warfare: A Re-Examination of Video Game Console Competitions*. University of Pittsburgh working paper, May 2012; <http://www.pitt.edu/~ckemerer/Research.htm>
8. Eisenmann, T., Parker, G., and Van Alstyne, M. Strategies for two-sided markets. *Harvard Business Review* 84, 10 (Oct. 2006), 92–101.
9. Farrell, J. and Katz, M.L. Innovation, rent extraction, and integration in systems markets. *Journal of Industrial Economics* 48, 4 (Dec. 2000), 413–432.
10. Gladwell, M. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, Little, Brown and Company, Boston, 2000.
11. Liu, C., Gal-Or, E., Kemerer, C., and Smith, M. Compatibility and proprietary standards: The impact of conversion technologies in IT markets with network effects. *Information Systems Research* 22, 1 (Mar. 2011), 188–207.
12. Liu, C., Kemerer, C., Slaughter, S., and Smith, M. Standards competition in the presence of digital conversion technology: An empirical analysis of the flash memory card market. *MIS Quarterly* 36, 3 (Sept. 2013), 921–942.
13. Stone, B. Want to copy iTunes music? Go ahead, Apple says. *New York Times* (Jan. 6, 2009); <http://www.nytimes.com/2009/01/07/technology/companies/07apple.html>
14. Stross, C.; <http://www.antipope.org/charlie/blog-static/2012/04/understanding-amazons-strategy.html>
15. Varian, H. and Shapiro, C. *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Review Press, Cambridge, MA, 1998.

Chris F. Kemerer (ckemerer@katz.pitt.edu) is the David M. Roderick Professor of Information Systems at the University of Pittsburgh and adjunct professor at the School of Computer Science at Carnegie Mellon University, Pittsburgh, PA.

Charles Zhechao Liu (charles.liu@utsa.edu) is an assistant professor of information systems at the University of Texas at San Antonio, San Antonio, TX.

Michael D. Smith (mds@andrew.cmu.edu) is a professor of information technology and marketing and co-director of the Initiative for Digital Entertainment Analytics at Carnegie Mellon University's Heinz College, Pittsburgh, PA.



THE ACM A. M. TURING AWARD

by the community ♦ from the community ♦ for the community



ACM, Intel, and Google congratulate

SHAFI GOLDWASSER and SILVIO MICALI

for transformative work that laid the complexity-theoretic foundations for the science of cryptography, and in the process pioneered new methods for efficient verification of mathematical proofs in complexity theory.



"The work of Goldwasser and Micali has expanded the cryptography field beyond confidentiality concerns," said Limor Fix, Director of the University Collaborative Research Group, Intel Labs. "Their innovations also led to techniques for message integrity checking and sender/receiver identity authentication as well as digital signatures used for software distribution, financial transactions, and other cases where it is important to detect forgery or tampering. They have added immeasurably to our ability to conduct communication and commerce over the Internet."

For more information see www.intel.com/research.



"Alfred Spector, Vice President of Research and Special Initiatives at Google Inc., said Goldwasser and Micali developed cryptographic algorithms that are designed around computational hardness assumptions, making such algorithms hard to break in practice. "In the computer era, these advances in cryptography have transcended the cryptography of Alan Turing's code-breaking era. They now have applications for ATM cards, computer passwords and electronic commerce as well as preserving the secrecy of participant data such as electronic voting. These are monumental achievements that have changed how we live and work."

For more information, see <http://www.google.com/corporate/index.html> and <http://research.google.com/>.



Financial support for the ACM A. M. Turing Award is provided by Intel Corporation and Google Inc.

DOI:10.1145/2447976.2447995

Employees in emerging markets find their own IT devices vital to job productivity and innovation.

BY IRIS JUNGLAS AND JEANNE HARRIS

The Promise of Consumer Technologies in Emerging Markets

ACROSS ASIA, EUROPE, and the Americas, consumer technologies are finding a home away from home. Employees increasingly use their own IT devices, such as tablets and smartphones, to solve problems within the enterprise that employs them. They use these resources in addition to—and sometimes in lieu of—tools supplied by the company.

Indeed, employees themselves are driving this IT consumerization trend. Influenced by their personal experiences with IT, employees perceive their own devices as more valuable as well as easier and more enjoyable to use than the tools provided by their organization. As part of our survey of more than 4,000 employees worldwide, we found that a majority (52%)

is already harnessing the capabilities of consumer gadgets for business purposes. More than 23% are using personal consumer devices at work

» key insights

- The usage of consumer devices and applications in the workplace is a global phenomenon, however, it is by no means a uniform one. Employees in Brazil, China, India, and Mexico show disproportionately higher consumer IT utilization rates in the workplace than any other markets studied.
- Employees in emerging markets believe that using consumer tools for work boosts their empowerment and, with it, their innovative nature—more so than employees in mature markets.
- More employees in emerging markets experience consumer IT as a productivity enhancer, making them more efficient and effective at work.



PHOTOGRAPH BY DIC LIEW/ISTOCKPHOTO

on a regular basis (see Table 1); for example, to stay connected with colleagues or customers. And 29% are using them at least once during an ordinary work week.

But IT consumerization does not stop there. In addition to consumer devices, employees are also using consumer software apps that are not part of the corporate enterprise IT portfolio, but are available for easy consumption through the cloud by home users and employees alike. Google Apps, Facebook, and Skype are just a few examples of consumer apps that employees are regularly using as work aids. From boosting personal productivity on the job, to strengthening social and business relationships, to empowering virtual communications among colleagues,

consumer applications are used daily by more than 20% of employees worldwide, according to our survey (see Table 1); 29% use such apps at least once a week.

Interviews we conducted at the outset of our research with more than 47 senior executives across multiple industries showed that while most had recognized IT consumerization as a pressing and even threatening topic, they had not attempted to formally quantify its organizational impact. When prompted about the benefits of IT consumerization for the organization, executives typically identified one (or more) of three themes: employee innovation, employee productivity, and employee satisfaction.

Employee innovation, particularly in business processes, was commonly

mentioned as a potential outcome that, over time, would also likely turn into cost savings. A CIO of a hospital, for instance, described a nurse's innovation regarding bandaging wounds. When nurses change patients' dressings, it often happens that the attending doctor arrives late, but still needs to inspect the wound. Instead of taking off a fresh bandage, the nurse had taken pictures of the wound with her phone and was able to show and document the healing progress. Later, the IT group helped to institutionalize a new process by providing a wizard for uploading and storing the image and linking it to a streaming viewer that allowed the doctor to securely retrieve the image. This example also illustrates another potential benefit: employee productivity.

Several executives mentioned the value to the organization of employees using the technology to productively link to resources while outside enterprise boundaries and normal working hours. Others mentioned the opportunity to enrich the interaction with customers—such as a pharmaceutical representative using a tablet computer and an interactive app or short video in a dialogue with a busy physician. The most commonly discussed benefits, however, related to employee satisfaction. Many interviewees recognized the changing characteristics of the generation of employees now entering the workforce, particularly their high comfort levels with, and expectations about, social networking and consumer technologies. IT consumerization was seen as a valuable tool in attracting and harnessing these new hires. Current employees, on the other hand, were perceived as valuing the independence and even the enjoyment that came from being able to choose their own tools as well as the convenience of working with up-to-date technology that paralleled what they used outside of work and better suited the characteristics of their job.

Distinctive Usage Patterns

While adopting consumer technologies for work purposes is a worldwide phenomenon, it is by no means a uniform one. Our subsequent survey of 4,097 full-time employees from large organizations in 16 different countries reveals the adoption of consumer tools and applications in the workplace is more prevalent in emerging markets than in mature markets. Brazil, China, India, and Mexico, for instance, show disproportionately higher consumer IT utilization in the workplace (see Table 1). Even when controlling for the type and number of devices employees receive from their employer (desktops, laptops, smartphones, tablet PCs), employees in those nations are twice as likely to use a consumer technology in the workplace as their counterparts in mature economies, such as Germany or France, whose usage rates are below 10%.

What explains these differences in consumer IT usage patterns? One possibility is that many people in emerging-market countries have leapfrogged earlier forms of communication technology, particularly landline phones and personal computers. Another is that rising income levels of a wide

middle class and mass-market pricing in emerging markets have made consumer technologies particularly affordable and appealing tools in these areas—for home and work. Consumer IT is small, cheap, modular, and ready to go. Not surprisingly, adoption rates of consumer IT in emerging economies are accelerating. While the overall usage of consumer technologies is increasing at an average of 5% for devices and 11% for applications around the globe, the fastest uptake in the near future, according to our study, will be in China (16% increase forecasted for devices; 35% for applications) and Mexico (11% forecasted increase for devices; 29% for applications, see Table 1). The size of the population using consumer technology is already much bigger in emerging markets and is growing much faster than in mature markets. In addition to these developments, differences in attitudes regarding the value of consumer IT in the workplace may shed further light on the usage patterns we found.

Contrasting Attitudes

Employees perceived the value of consumer IT in the workplace similar to those mentioned by our executives

Table 1. Consumer IT use in workplaces around the globe.

	Consumer IT devices		Consumer IT applications	
	Percentage of employees that use their own consumer IT devices for work often or very often	Percentage of employees that plan to use their own consumer IT devices for work more often from now on	Percentage of employees that use consumer IT applications for work purposes often or very often	Percentage of employees that plan to use consumer IT applications for work purposes more often from now on
Australia	21%	19%	18%	18%
Brazil	31%	37%	29%	43%
Canada	20%	19%	17%	24%
China	40%	56%	31%	67%
France	8%	15%	9%	15%
Germany	9%	7%	8%	11%
India	41%	48%	40%	56%
Italy	24%	30%	24%	32%
Japan	10%	12%	6%	11%
Mexico	39%	50%	30%	59%
Scandinavia	9%	12%	8%	12%
Singapore	34%	34%	28%	37%
South Korea	34%	46%	25%	48%
Spain	16%	24%	15%	29%
United Kingdom	16%	17%	11%	18%
United States	21%	20%	18%	20%
Sample Average	23%	28%	20%	31%

Note: Bold typeface marks top three for each category

interviewed for this study—namely as one of three things: innovation, productivity, and satisfaction. But unlike our interview data, our survey data of more than 4,000 full-time employees indicates fundamental differences exist regarding each of the three categories across the globe. Specifically, employees in emerging-market economies, more so than those in developed economies, see consumer IT as a tool that can drive innovation, increase personal productivity, and help companies attract and retain talent. (Note: The survey sample was drawn equally across industry and age groups, with approximately 250 employees from each of the 16 countries. All survey participants worked at companies with more than 100 employees. For demographics, refer to the appendix available online).

Driving innovation. Consumer technology's potential to drive innovation in the workplace is reflected in the variety of ways employees use such IT to approach, tackle, and solve problems. Employees in countries such as Brazil, China, India, and Mexico are more than twice as likely as those in mature markets to use or download consumer applications from the Internet to solve

a work problem, according to our study (see Table 2). They are also more inclined to create their own tech-based solutions or use their personal social contacts (gained through social media technology) to solve a problem. Instead of relying on established processes, for example, employees in emerging markets have the unique potential to create their own practices from scratch, with the help of consumer IT—and they happily do so.

Admittedly, improvisation (or creative problem solving) is an inherent characteristic of markets that, historically, have been resource deprived. Those cultures are also inherently more relational, that is, they utilize personal contacts to a greater extent in order to get business done. In addition, those countries typically invest less in IT when compared to mature markets. But even when taking all these explanations into account and considering our data was collected from large organizations—organizations with significant IT investments and formal IT governance structures—stark discrepancies in innovation perceptions between emerging and mature markets still exist. It almost seems that consumer IT is the

perfect tool that is able to harness those improvisation and connectedness tendencies of emerging markets, and bring them to fruition. Employees in emerging and developed countries alike also see a direct relationship between permitting the use of consumer technologies in the workplace and innovation. However, those in emerging-market economies believe more strongly that use of such technologies can make a significant difference in their innovativeness. When prompted to indicate whether they expect consumer IT to help them be more innovative in their job, employees in emerging economies rated their perception consistently higher—on average between 21% to 35%—than the world average (see Table 2).

Enhancing productivity. Our study also showed that employees in Brazil, China, India, and Mexico spend considerably more time (as much as 19% to 46% more) than those in other countries seeking out applications they think will make them more productive at work (see Table 3). When asked whether they believe consumer technology would lead to better use of their resources or higher-quality work, employees in emerging economies con-

Table 2. Viewing consumer IT as an innovation booster.

	Percentage of employees that agree or strongly agree with each statement				
	"I often spend time looking for suitable applications that make me better at work"	"I often use or download applications from the Internet to solve a problem at work"	"I often create technology-based solutions that help me solve a business problem"	"I often use my personal social network to find solutions to work problems"	"If I were allowed to choose my own hardware and software for work, I think I would be more innovative"
Australia	15%	17%	16%	8%	38%
Brazil	36%	40%	38%	26%	64%
Canada	17%	18%	18%	13%	42%
China	63%	60%	54%	38%	78%
France	8%	9%	14%	5%	33%
Germany	9%	10%	10%	7%	22%
India	55%	49%	47%	34%	70%
Italy	26%	30%	30%	21%	57%
Japan	13%	26%	9%	7%	40%
Mexico	48%	50%	48%	20%	76%
Scandinavia	10%	11%	9%	7%	31%
Singapore	34%	23%	26%	16%	64%
South Korea	24%	29%	18%	23%	57%
Spain	25%	22%	20%	12%	50%
United Kingdom	11%	15%	15%	7%	45%
United States	15%	15%	20%	12%	41%
Sample Average	26%	26%	24%	16%	50%

Note: Bold typeface marks top three for each category

sistently “agreed” or “strongly agreed” on such IT’s positive impact. More than 70% of our study participants from India and Mexico, and 82% of

our China respondents, showed strong agreement—compared to much lower percentages of participants from developed nations.

For emerging and mature economies, productivity and technology seem to perpetuate one another. But for emerging markets specifically, this effect seems remarkably pronounced. As those markets ramp up their economies, consumer technologies are becoming increasingly available to a broad spectrum of people. This, in turn, could fuel economic growth even further. The World Bank has found that with the addition of one mobile phone per 10 people, the per-capita GDP of a typical emerging nation will increase by approximately 0.8%.⁶

In this context, the availability of cloud-based data and applications accelerates opportunities for value creation, making workers more productive irrespective of time and location with lower investment upfront. Cloud computing provides tools in a more accessible, cheaper, and easier-to-use manner than technologies traditionally used by corporate IT. Indeed, compared to their Western equivalents, employees in emerging economies use the cloud almost twice as often in the workplace (see Table 3). This effect may stem from the lack of existing enterprise systems, the lack of IT policies, or both.

Capturing and keeping top talent. Lastly, our study revealed differences in attitudes regarding the power of consumer IT to attract and retain talent. As companies strive to compete by recruiting the best employees and keeping them on board, they are finding that offering these workers access to leading-edge technology can help. In particular, encouraging the use of consumer technologies sends a clear message that employees will have the most current tools available as well as maximum flexibility and freedom in carrying out their work. Yet again, this effect is remarkably more pronounced for consumer technologies in emerging-market economies than in mature ones.

For example, our survey respondents from emerging nations showed consistently higher levels of job satisfaction if allowed to use consumer technologies for work purposes (see Table 4). While the impact is noticeable around the globe (sample average is 59%), employees in emerging markets feel most influenced by the

Table 3. Perceiving consumer IT as productivity enhancer.

Percentage of employees that agree or strongly agree with each statement			
	“If I were allowed to choose my own hardware and software for work, I would be more resourceful”	“If I were allowed to choose my own hardware and software for work, I would be able to do higher quality of work”	“My work data gets stored with third party providers on the Internet”
Australia	38%	36%	11%
Brazil	56%	63%	23%
Canada	44%	41%	13%
China	82%	79%	36%
France	32%	37%	6%
Germany	23%	20%	4%
India	72%	72%	28%
Italy	54%	56%	19%
Japan	52%	46%	4%
Mexico	74%	72%	23%
Scandinavia	31%	35%	8%
Singapore	65%	67%	14%
South Korea	62%	54%	16%
Spain	48%	48%	15%
United Kingdom	44%	40%	7%
United States	41%	40%	12%
Sample Average	51%	50%	15%

Note: Bold typeface marks top three for each category

Table 4. Seeing consumer IT as a workforce-retention weapon.

Percentage of employees that agree or strongly agree with each statement		
	“By allowing employees to use their desired applications and devices for work-related purposes, an organization could increase satisfaction among employees”	“I would gladly pay for some of the applications and devices if only my organization would allow me to use them for work.”
Australia	57%	18%
Brazil	71%	36%
Canada	54%	26%
China	79%	37%
France	57%	14%
Germany	50%	13%
India	72%	51%
Italy	63%	31%
Japan	33%	18%
Mexico	76%	46%
Scandinavia	42%	10%
Singapore	69%	36%
South Korea	56%	31%
Spain	65%	28%
United Kingdom	50%	15%
United States	49%	20%
Sample Average	59%	27%

Note: Bold typeface marks top three for each category

type of tools available. With rates well above 70%, employees in countries such as Brazil, China, India, and Mexico assign more importance to the use of the latest consumer IT than their mature-markets counterparts. In fact, having the latest technologies at hand is so important to emerging-market employees that a large percentage is willing to pay for their own IT. And as their country's middle class grows, more of them can pay. More than 36% of our study participants from emerging economies have no objection to paying for some of their devices and applications for work, contrasted by mature-market countries, where employees are less inclined to shell out for such tools (see Table 4).

This figure is particularly interesting in the light of differences in purchasing power, which is 2 to 15 times lower in emerging markets than in mature markets.⁵ Despite these differences, the prospect of using a consumer device or application for work and home seems to hold more appeal for people in emerging economies. The novelty of these tools may be one reason. Another may be that IT policies are still nascent in emerging markets, owing to the lack of a preexisting infrastructure. Employees who have long been restricted to using certain mandated technologies may perceive consumer technologies as desirable must-haves.

Consumer Technologies and Economic Growth

The arrival of consumer-originated devices and applications into the workplace is empowering a second iteration of an employee-driven IT revolution. The first revolution, over 40 years ago, was the invasion of corporate offices by employees armed with personal computers. While this revolution was largely mitigated by IBM whose endorsement precipitated acceptance by information systems departments and, in time, policies for appropriate use, today no dominant vendor with the clout of the near-monopolistic IBM exists that could come to the rescue. The number of consumer devices and applications are too many, their functionalities are too diverse, and their price points are too low.



The adoption of consumer tools and applications in the workplace is more prevalent in emerging markets than in mature markets.



While the majority of companies are still in the contemplative, reactive, or experimentation phases, ranging between the extremes of “authority,” such as the exercise of tight control over the scope and number of consumer devices and applications entering the organization, and “laissez-faire,” that is, the boundless tolerance for allowing consumer devices and applications into the workplace, they have acknowledged, voluntarily or not, that IT consumerization poses a new set of challenges that will need to be adjusted for. By broadening the scope of allowable devices and applications, for example, or by providing employees with IT allowances as job benefits, or by segmenting users based on consumerization profiles, companies around the globe have a variety of options to choose from.

Companies with workforces that are already embracing consumer IT and extracting measurable business value from it will be best positioned to capture those opportunities. Emerging markets might be one of them. Worldwide, emerging markets will serve as the primary engine of economic growth in the years to come, accounting for 70% of it.³ The increasing use of affordable consumer technologies in such economies will help drive that growth by opening untapped markets. While consumer IT might only be one factor in fueling this growth, it is certainly an important one. The Indian government, for example, has provided subsidized tablet PCs to some of its most rural areas,¹ while the southern state of Karnataka is issuing iPads to all its state ministers.⁴ The Four Seasons Hotel in India has adopted iPads in both its hotel lobbies and restaurants,² and an agricultural company in India has developed a mobile app that lets farmers use their personal cellphones to sell produce. Delivered mostly in a pictogram format, farmers can see what current market prices are for specific types of produce, how much of each type is needed and where it is needed. These are just a few examples of how consumer IT is applied and utilized in those countries.

New supply chain strategies are presenting additional opportunities to use consumer IT in emerging-

market workplaces. Already, Western firms are rethinking their traditional strategy of inventing at home and then exporting to developing markets. Of the Fortune 500, 98 companies now have R&D facilities located in China; 63 have them in India. In addition, vendors and manufacturers are increasingly concentrating on selling to emerging markets before entering mature markets.³ Consumer IT used in the workplace could help accelerate this ongoing trend. Many consumer tools are cheaper, easier to use, and quicker to implement than traditional enterprise IT, and they appeal more to emerging markets' entrepreneurial appetite.

Furthermore, consumer technology presents a first-of-its-kind opportunity to transform what has historically been a liability—lack of IT infrastructure—into an advantage. Such technologies, particularly when linked through cloud computing (such as software-as-a-service), can inexpensively compensate for the absence of enterprise systems that, in the past, limited the growth of organizations in developing economies.

For companies that have long boasted established systems, this development means they can no longer rely on their IT infrastructure to differentiate themselves from aspiring rivals. On the contrary, what used to function as a barrier of entry against upstart competitors might now become an obstacle for all players. Just as Japanese manufacturers in the 1980s learned from and avoided early adopters' mistakes, emerging-market companies might have the opportunity to leapfrog over an entire era of legacy systems by embracing consumer technologies. And while they may have smaller overall budgets, they can selectively deploy only those technologies—taken straight from the consumer market—that are most relevant to their needs.

While unknown, one might wonder what are the long-term implications of IT consumerization. A change in work culture seems inevitable, painting a world in which devices are easily substitutable, working hours surreptitiously extend into personal hours, office locations vanish or become irrelevant, and only the bandwidth to ac-



Employees in emerging markets see consumer IT as a tool that can drive innovation, increase personal productivity, and help companies attract and retain talent.



cess cloud-based applications counts. One can easily envision employees bringing their own technology tool-belts to a job interview, just like the apprentices had to do in the middle-aged guild model. Judged not only by experiences, but also by the ability to contribute to, and interlink with, existing members and work practices might be of new concern for employers that try to recruit in a post-PC era world. Admittedly, those are pure speculations, but business and IT executives alike will have to understand that the use of consumer technologies in the workplace—the “consumerization” of IT—is a global phenomenon. Throughout the world, they will have to address it, not with piecemeal measures, but with strategies that help them get the most out of these technologies while still maintaining adequate control. In emerging markets, employees see even greater possibilities for the use of consumer IT at work.

As the landscape becomes ever more competitive, companies operating in those markets cannot ignore these attitudes. To maintain high growth at the business or national level, leaders will have to harness their employees' enthusiasm for consumer IT as a tool for competitive advantage. **■**

References

1. BBC. India launches Aakash tablet computer priced at \$35. BBC South Asia (Oct. 2011); <http://www.bbc.co.uk/news/world-south-asia-15180831/>
2. Joshi, P. Use of Enterprise iPad on the rise in India. Business Standard (Feb. 2011); <http://www.business-standard.com/results/news/useenterprise-ipadthe-rise-in-india/424222/>
3. *The Economist*. The world turned upside down. (Apr. 15, 2010); <http://www.economist.com/node/15879369/>
4. *The Times of India*. On a tech high, Karnataka MLAs wish to long in to iPad3. (Mar. 26, 2012); http://articles.timesofindia.indiatimes.com/2012-03-26/bangalore/31239852_1_expensive-4g-model-ipad-64gb/
5. The World Bank. International Comparison Program (2011), <http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/ICPEXT/0,,contentMDK:20118237~menuPK:62002075~pagePK:60002244~piPK:62002388~theSitePK:270065,00.html/>
6. Zhen-Wei Qiang, C. Mobile telephony: A transformational tool for growth and development (2009); http://www.proparco.fr/jahia/webdav/site/proparco/shared/PORTAILS/Secteur_privé_developpement/PDF/SPD4_PDF/Christine-Zhen-Wei-Qiang-World-Bank-Mobile-Telephony-A-Transformational-Tool-for-Growth-and-Development.pdf/

Iris Junglas (ijunglas@cob.fsu.edu) is an assistant professor in Management Information Systems at Florida State University, Tallahassee, FL.

Jeanne Harris (jeanne.g.harris@accenture.com) is the Global Managing Director of IT Research at Accenture Institute for High Performance, Boston, MA.

research highlights

P. 92

Technical Perspective The Ray-Tracing Engine that Could

By Matt Pharr

P. 93

GPU Ray Tracing

By Steven G. Parker, Heiko Friedrich, David Luebke, Keith Morley, James Bigler, Jared Hoberock, David McAllister, Austin Robison, Andreas Dietrich, Greg Humphreys, Morgan McGuire, and Martin Stich

Technical Perspective

The Ray-Tracing Engine that Could

By Matt Pharr

AN EFFECTIVE APPROACH to building flexible software systems has been to make them extensible through embedded scripting languages. Languages like TCL, Python, and Lua have allowed programmers to orchestrate and customize the behavior of many software systems—examples include games, which are mostly written in C++ but often have AI for characters and other gameplay mechanics implemented in Lua, and high-performance computing applications written with Python code coordinating the execution of C++ or Fortran library code.

Most embedded scripting languages are interpreted, and thus are not suitable for the implementation of performance-critical inner loops. Furthermore, the runtime overhead of passing through the interface between the core system and the embedded language is too much to accept in many performance-focused domains, especially when frequent transitions are needed. Thus, most performance-oriented systems have not offered the option of programmability in the heart of their performance-critical parts.

However, in graphics, a programmable high-performance rasterization pipeline has been at the heart of interactive rendering for the last decade. Programmers of modern graphics processing units (GPUs) provide code for the pipeline's inner loops, writing “shaders” in C-based languages like HLSL or GLSL. The shader programming model is data-parallel; this model provides abundant parallelism, which maps well to the underlying SIMD hardware architecture. Although programmable GPUs have now been used in many domains beyond graphics for high-performance computation, it has been an open question whether it is possible to build GPU-targeted high-performance software systems that are themselves programmable.

The following paper by Parker et al. shows how to achieve both programmability and high performance in such a system. The domain of their system, OptiX, is interactive ray tracing for image synthesis. Ray tracing is a very flexible approach to rendering, and can simulate many important lighting effects more effectively than rasterization, but its use in interactive graphics has until recently been limited—prior to OptiX, users had the choice of high-performance ray-tracing systems that were insufficiently flexible, or highly flexible ray-tracing systems that had insufficient performance.

The authors have developed an elegant expression of the classic ray-tracing algorithm as a programmable data-flow graph assembled from user-supplied code at each stage. OptiX supplies highly optimized implementations of core geometric and parallel work scheduling algorithms that run between stages. Just like the GPU rasterization pipeline, the programmer has full control of the system's behavior at key points of programmability without needing to worry about the gritty details of high-performance GPU programming. OptiX uses the CUDA language for the user-supplied kernels; CUDA provides a data-parallel programming model that runs with high efficiency on GPUs.

OptiX achieves programmability without sacrificing performance by eliminating the barrier between the core OptiX system code and the code provided by the user. It applies a specializing JIT compiler to both collections of code, allowing for not just inlining across the boundary between the two parts of the system but also for constant propagation and dead code elimination, thus generating a specialized version of the system. The core OptiX system can thus provide functionality that may not be needed in the end; the code for such functionality is removed when the system is compiled together.

This system implementation approach allows users of OptiX to implement, for example, custom representations of the 3D scene geometry as well as custom algorithms to simulate lighting and reflection—both key areas for customization in ray tracing. An indication of the authors' success in designing the right decomposition of the problem is the wide variety of applications of ray tracing—spanning not just rendering, but even audio simulation and collision detection—that have been implemented with OptiX. The resulting systems are close enough to peak efficiency that OptiX has quickly become the standard foundation for most GPU ray tracing.

One of the unexpected successes of the introduction of the programmable rasterization pipeline on GPUs has been the creativity programmers have shown in using the GPU rasterization pipeline in ways never imagined by its original designers. By putting both flexible and high-performance ray tracing in the toolbox of many more developers than before, it seems quite likely that OptiX will spark innovation in ways that are impossible to predict today.

This paper is a must-read for anyone who cares about writing extensible software systems that are also high-performance software systems. Although the target hardware architecture for this work is GPUs, the underlying ideas are equally applicable to high-performance software systems on CPUs. Today, with the availability of high-quality compiler toolkits like LLVM, the barrier to entry for implementing all sorts of systems in this manner is now quite low, while the potential advantages are significant. **□**

Matt Pharr (matt.pharr@gmail.com) is a software engineer in the Google [x] group at Google, Inc., Mountain View, CA.

© 2013 ACM 0001-0782/13/05

GPU Ray Tracing

By Steven G. Parker, Heiko Friedrich, David Luebke, Keith Morley, James Bigler, Jared Hoberock, David McAllister, Austin Robison, Andreas Dietrich, Greg Humphreys, Morgan McGuire, and Martin Stich

Abstract

The NVIDIA® OptiX™ ray tracing engine is a programmable system designed for NVIDIA GPUs and other highly parallel architectures. The OptiX engine builds on the key observation that most ray tracing algorithms can be implemented using a small set of programmable operations. Consequently, the core of OptiX is a domain-specific just-in-time compiler that generates custom ray tracing kernels by combining user-supplied programs for ray generation, material shading, object intersection, and scene traversal. This enables the implementation of a highly diverse set of ray tracing-based algorithms and applications, including interactive rendering, offline rendering, collision detection systems, artificial intelligence queries, and scientific simulations such as sound propagation. OptiX achieves high performance through a compact object model and application of several ray tracing-specific compiler optimizations. For ease of use it exposes a single-ray programming model with full support for recursion and a dynamic dispatch mechanism similar to virtual function calls.

1. INTRODUCTION

Many CS undergraduates have taken a computer graphics course where they wrote a simple ray tracer. With a few simple concepts on the physics of light transport, students can achieve high quality images with reflections, refraction, shadows, and camera effects such as depth of field—all of which present challenges on contemporary real-time graphics pipelines. Unfortunately, the computational burden of ray tracing makes it impractical in many settings, especially where interactivity is important. Researchers have invented many techniques for improving the performance of ray tracing,¹³ especially when mapped to high-performance architectural features such as explicit SIMD instructions¹² and Single-Instruction Multiple-Thread (SIMT)-based⁶ GPUs.¹ Unfortunately most such techniques muddy the simplicity and conceptual purity that make ray tracing attractive. Nor have industry standards emerged to hide these complexities, as Direct3D and OpenGL do for rasterization.

To address these problems, we introduce OptiX, a general purpose ray tracing engine. A general programming interface enables the implementation of a variety of ray tracing-based algorithms in graphics and non-graphics domains, such as rendering, sound propagation, collision detection, and artificial intelligence. This interface is conceptually simple yet enables high performance on modern GPU architectures and is competitive with hand-coded approaches.

In this paper, we discuss the design goals of the OptiX engine as well as an implementation for NVIDIA GPUs. In our implementation, we compose domain-specific compilation with a flexible set of controls over scene hierarchy,

acceleration structure creation and traversal, on-the-fly scene update, and a dynamically load-balanced GPU execution model. Although OptiX primarily targets highly parallel GPU architectures, it is applicable to a wide range of special- and general-purpose hardware, including modern CPUs.

1.1. Ray tracing, rasterization, and GPUs

Computer graphics algorithms for *rendering*, or image synthesis, take one of two complementary approaches. One family of algorithms loop over the pixels in the image, computing for each pixel, the first object visible at that pixel; this approach is called *ray tracing* because it solves the geometric problem of intersecting a ray from the pixel into the objects. A second family of algorithms loops over the objects in the scene, computing for each object the pixels covered by that object. Because the resulting per-object pixels (called *fragments*) are formatted for a raster display, this approach is called *rasterization*. The central data structure of ray tracing is a spatial index called an *acceleration structure*, used to avoid testing each ray against all objects. The central data structure of rasterization is the *depth buffer*, which stores the distance of the closest object seen at each pixel and discards fragments from invisible objects. While both approaches have been generalized and optimized greatly beyond this simplistic description, the basic distinction remains: ray tracing iterates over rays while rasterization iterates over objects. High-performance ray tracing and rasterization, both focus on rendering the simplest of objects: triangles.

Historically, ray tracing has been considered slow and rasterization fast. The simple, regular structure of depth-buffer rasterization lends itself to highly parallel hardware implementations: each object moves through several stages of computation (the so-called *graphics pipeline*), with each stage performing similar computations in data-parallel fashion on the many objects, fragments, and pixels in flight throughout the pipeline. As graphics hardware has grown more parallel it has also grown more general, evolving from specialized fixed-function circuitry implementing the various stages of the graphics pipeline into fully programmable processors that virtualize those stages onto hundreds or even thousands of small general-purpose cores. Today's graphics processing units, or GPUs, are massively parallel processors capable of performing trillions of floating-point math operations and rendering billions of triangles each second. The computational horsepower and power efficiency of modern GPUs has made them attractive for high-performance

The original version of this paper is entitled "OptiX: A General Purpose Ray Tracing Engine" and was published in *ACM Transactions on Graphics (TOG)—Proceedings of ACM SIGGRAPH*, July 2010, ACM

computing, from many of the fastest supercomputers in the world to science, math, and engineering codes on the desktop. All of which raises the question: can ray tracing be implemented efficiently and flexibly on GPUs?

1.2. Contributions and design goals

To create a high-performance system for a broad range of ray tracing tasks, several trade-offs and design decisions led to the following contributions:

- *A general, low level ray tracing engine.* OptiX is not a renderer. It focuses exclusively on the fundamental computations required for ray tracing and avoids embedding, rendering-specific constructs such as lights, shadows, and reflectance.
- *A programmable ray tracing pipeline.* OptiX shows that most ray tracing algorithms can be implemented using a small set of lightweight programmable operations. It defines an abstract ray tracing execution model as a sequence of user-specified programs, analogous to the traditional rasterization-based graphics pipeline.
- *A simple programming model.* OptiX avoids burdening the user with the machinery of high-performance ray tracing algorithms. It exposes a familiar recursive, single-ray programming model rather than ray packets or explicit vector constructs, and abstracts any batching or reordering of rays.
- *A domain-specific compiler.* The OptiX engine combines just-in-time compilation techniques with ray tracing-specific knowledge to implement its programming model efficiently. The engine abstraction permits the compiler to tune the execution model for available system hardware.

2. RELATED WORK

While numerous high-level ray tracing libraries, engines, and APIs have been proposed,¹³ efforts to date have been focused on specific applications or classes of rendering algorithms, making them difficult to adapt to other domains or architectures. On the other hand, several researchers have shown how to map ray tracing algorithms efficiently to GPUs and the NVIDIA® CUDA™ architecture,^{1,3,11} but these systems have focused on performance rather than flexibility.

Further discussion of related systems and research can be found in the original paper.¹⁰

3. A PROGRAMMABLE RAY TRACING PIPELINE

The core idea behind the OptiX engine is that most ray tracing algorithms can be implemented using combinations of a small set of programmable operations. This is directly analogous to the programmable rasterization pipelines employed by OpenGL and Direct3D. At a high level, these systems expose an abstract rasterizer containing lightweight call-backs for vertex shading, geometry processing, tessellation, and pixel shading operations. An ensemble of these program types, often used in multiple passes, can be used to implement a broad variety of rasterization-based algorithms.

We have identified a corresponding programmable ray tracing execution model along with lightweight operations

that can be customized to implement a wide variety of ray tracing-based algorithms.⁹ These user-provided operations, which we simply call *programs*, can be combined with a user-defined data structure (*payload*) associated with each ray. The ensemble of programs together implement a particular client application's algorithm.

3.1 Programs

OptiX includes seven different types of these programs, each of which conceptually operates on a single ray at a time. In addition, a bounding box program operates on geometry to determine primitive bounds for acceleration structure construction. The combination of user programs and hardcoded OptiX kernel code forms the *ray tracing pipeline*, which is outlined in Figure 2. Unlike a feed-forward rasterization pipeline, it is more natural to think of the ray tracing pipeline as a call graph. The core operation, *rtTrace*, alternates between locating an intersection (*Traverse*) and responding to that intersection (*Shade*). By reading and writing data in user-defined ray payloads and in global device-memory arrays called *buffers*, these operations are combined to perform arbitrary computation during ray tracing.

Ray generation programs are the entry into the ray tracing pipeline. A single invocation of *rtContextLaunch* from the host will create many instantiations of these programs. A typical ray generation program will create a ray using a camera model for a single sample within a pixel, start a trace operation, and store the resulting color in an output buffer. But by distinguishing ray generation from pixels in an image, OptiX enables other operations such as creating photon maps, precomputing lighting texture maps (also known as baking), processing ray requests passed from OpenGL, shooting multiple rays for super-sampling, or implementing different camera models.

Intersection programs implement ray-geometry intersection tests. As the acceleration structures are traversed, the system will invoke intersection programs to perform geometric queries. The program determines if and where the ray touches the object and may compute normals, texture coordinates, or other attributes based on the hit position. An arbitrary number of attributes may be associated with each intersection. Intersection programs enable support for arbitrary surfaces beyond polygons and triangles, such as displacement maps, spheres, cylinders, high-order surfaces, or even fractal geometries like the Julia set in Figure 1. A programmable intersection operation is useful even in a triangle-only system because it facilitates direct access to native mesh formats.

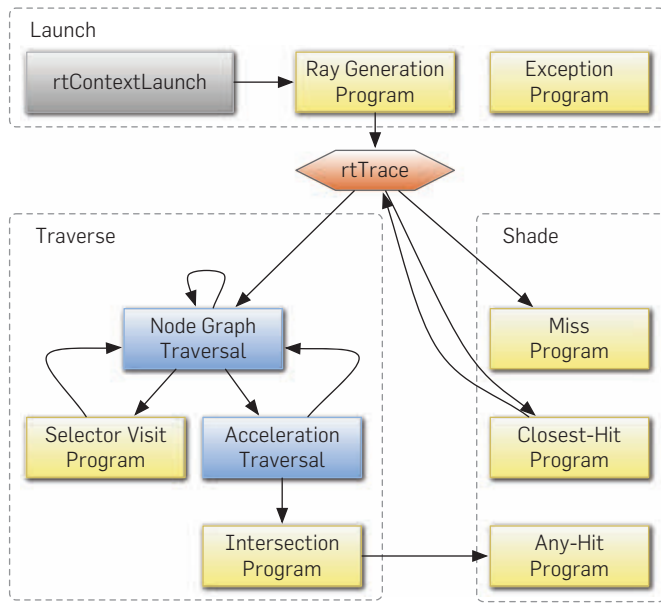
Closest-hit programs are invoked once traversal has found the nearest intersection of a ray with the scene geometry. This program type resembles surface shaders in classical rendering systems. Typically, a closest-hit program will perform computations like shading, potentially casting new rays in the process, and store resulting data in the ray payload.

Any-hit programs are called during traversal for every ray-object intersection that is found. The any-hit program allows the material to participate in object intersection decisions while keeping the shading operations separate from the geometry operations. It may optionally terminate the ray

Figure 1. Images from various applications built with OptiX. Top: Physically based light transport through path tracing. Bottom: Ray tracing of a procedural Julia set, photon mapping, large-scale line of sight and collision detection, Whitted-style ray tracing of dynamic geometry, and ray traced ambient occlusion. All applications are interactive.



Figure 2. A call graph showing the control flow through the ray tracing pipeline. The yellow boxes represent user-specified programs and the blue boxes are algorithms internal to OptiX. Execution is initiated by the API call `rtContextLaunch`. A built-in function, `rtTrace`, can be employed by the ray generation program to cast rays into the scene. This function may also be called recursively by the closest-hit program for shadow and secondary rays. The exception program is executed when the execution of a particular ray is terminated by an error such as excessive memory consumption.



using the built-in function `rtTerminateRay`, which will stop all traversal and unwind the call stack to the most recent invocation of `rtTrace`. This is a lightweight exception mechanism that can be used to implement early ray termination for shadow rays and ambient occlusion. Alternatively, the any-hit program may ignore the intersection using

`rtIgnoreIntersection`, allowing traversal to continue looking for other geometric objects. For instance, a program may choose to ignore an interaction based on a texture channel lookup to implement efficient alpha-mapped transparency without restarting traversal. Another use case for the any-hit program can be found in Section 6.1, where the application performs visibility attenuation for partial shadows cast by glass objects. Note that intersections may be presented out of order. The default any-hit program is a no-op, which is often the desired operation.

Miss programs are executed when the ray does not intersect any geometry in the interval provided. They can be used to implement a background color or environment map lookup.

Exception programs are executed when the system encounters an exceptional condition, for example, when the recursion stack exceeds the amount of memory available for each thread, or when a buffer access index is out of range. OptiX also supports user-defined exceptions that can be thrown from any program. The exception program can react, for example, by printing diagnostic messages or visualizing the condition by writing special color values to an output pixel buffer.

Selector visit programs expose programmability for coarse-level node graph traversal. For example, an application may choose to vary the level of geometric detail for parts of the scene on a per-ray basis.

3.2 Scene representation

An explicit goal of OptiX was to minimize the overhead of scene representation, rather than forcing a heavyweight scene graph onto users. The OptiX engine employs a simple but flexible structure for representing scene information and associated programmable operations, collected in a container object called the *context*. This representation is also the mechanism for binding programmable shaders to the object-specific data that they require.

Hierarchy nodes. A scene is represented as a lightweight graph that controls the traversal of rays through the scene. It can also be used to implement instancing two-level hierarchies for animations of rigid objects, or other common scene structures. To support instancing and sharing of common data, the nodes can have multiple parents.

Four main node types can be used to provide the scene representation using a directed graph. Any node can be used as the root of scene traversal. This allows, for example, different representations to be used for different ray types.

Group nodes contain zero or more (but usually two or more) children of any node type. A group node has an acceleration structure associated with it and can be used to provide the top level of a two-level traversal structure.

Geometry Group nodes are the leaves of the graph and contain the primitive and material objects described below. This node type also has an acceleration structure associated with it. Any non-empty scene will contain at least one geometry group.

Transform nodes have a single child of any node type, plus an associated 4×3 matrix that is used to perform an affine transformation of the underlying geometry.

Selector nodes have zero or more children of any node type, plus a single visit program that is executed to select among the available children.

Geometry and material objects. The bulk of the scene data is stored in the geometry nodes at the leaves of the graph. These contain objects that define geometry and shading operations. They may also have multiple parents, allowing material and geometry information to be shared at multiple points in the graph. As an example, consider Figure 3. The graph on the right shows a complete OptiX context for a simple scene with a pin-hole camera, two objects, and shadows. The ray generation program implements the camera, while a miss program implements a constant white background. A single geometry group contains two geometry instances with a single geometric index—in this case a bounding-volume hierarchy (BVH)—built over all underlying geometry in the triangle mesh and ground plane. Two types of geometry are implemented, a triangle mesh and a parallelogram, each with its own

set of intersection and bounding box programs. The two geometry instances share a single material that implements a diffuse lighting model and fully attenuates shadow rays via closest-hit and any-hit programs, respectively.

The diagram on the left of Figure 3 illustrates how these programs are invoked for 3 rays that traverse through the scene: 1. The ray generation program creates rays and traces them against the geometry group. This initiates the Traverse stage shown in Figure 2, executing parallelogram and triangle-mesh intersection until an intersection is found (2 and 3). If the ray intersects with geometry, the closest-hit program will be called whether the intersection was found on the ground plane or on the triangle mesh. The material will recursively generate shadow rays to determine if the light source is unobstructed. 4. When any intersection along the shadow ray is found, the any-hit program will terminate ray traversal and return to the calling program with shadow occlusion information. 5. If a ray does not intersect with any scene geometry, the miss program will be invoked.

Geometry Instance objects bind a *geometry object* to a set of *material objects*. This is a common structure used by scene graphs to keep geometric and shading information orthogonal.

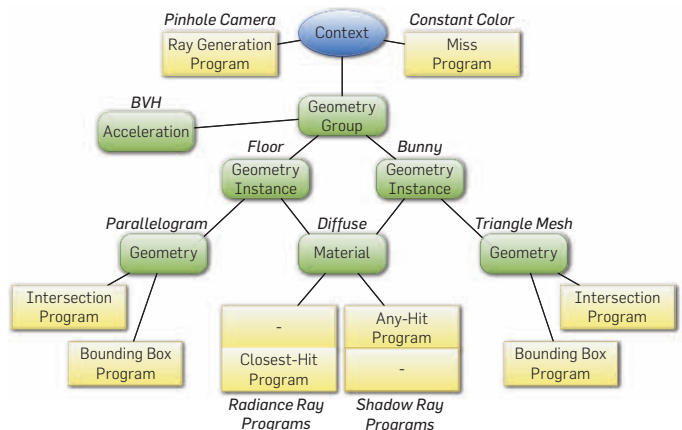
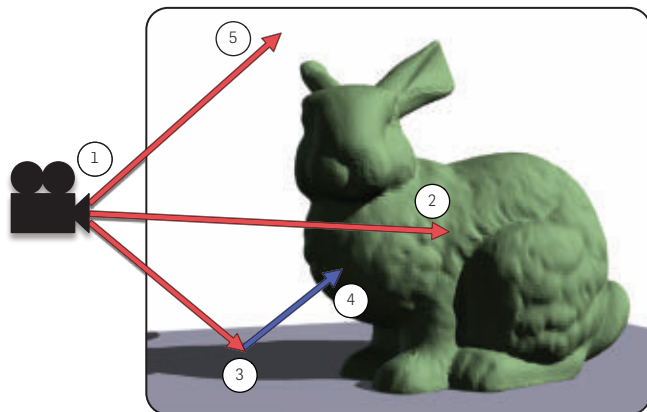
Geometry objects contain a list of geometric primitives. Each geometry object is associated with a bounding box program and an intersection program, both of which are shared among the geometry object's primitives.

Material objects hold information about shading operations, including the any-hit and closest-hit programs described in Section 3.1.

3.3. System overview

The OptiX engine consists of two distinct APIs. The *host API* is a set of C functions that the client application calls to create and configure a context, assemble a node graph, and launch ray tracing kernels. It also provides calls to manage GPU devices. The *program API* is the functionality exposed to user programs. This includes function calls for tracing rays, reporting intersections, and accessing data. In addition, several semantic variables encode state specific to ray tracing,

Figure 3. Example OptiX scene construction and execution.



for example, the current distance to the closest intersection. Printing and exception handling facilities are also available for debugging.

After using OptiX host, API functions to provide scene data such as geometry, materials, acceleration structures, hierarchical relationships, and programs, the application will then launch ray tracing with the `rtContextLaunch` API function that passes control to OptiX. If required, a new ray tracing kernel is compiled from the given user programs, acceleration structures are built (or updated) and data is synchronized between host and device memory, and finally, the ray tracing kernel is executed, invoking the various user programs as described above.

After execution of the ray tracing kernel has completed, its resulting data can be used by the application. Typically, this involves reading from output buffers filled by one of the user programs or displaying such a buffer directly, for example, via OpenGL. An interactive or multi-pass application then repeats the process starting at context setup, where arbitrary changes to the context can be made, and the kernel is launched again.

4. DOMAIN-SPECIFIC COMPILATION

The core of the OptiX host runtime is a just-in-time (JIT) compiler that serves several important functions. First, the JIT stage combines all of the user-provided shader programs into one or more kernels. Second, it analyzes the node graph to identify data-dependent optimizations. Finally, the resulting kernel is executed on the GPU using the CUDA driver API.

Generating and optimizing code for massively parallel architectures provide some challenges. One challenge is that code size and live state per computation must be minimized for maximum performance. Another challenge is structuring the code to reduce divergence. Our experience with OptiX highlights the interesting tensions between these sometimes conflicting requirements.

4.1. OptiX programs

The user-specified programs described in Section 3.1 are provided to the OptiX host API in the form of Parallel Thread Execution (PTX) functions.⁸ PTX is a virtual machine assembly language for NVIDIA's CUDA architecture, similar in many ways to the popular open source Low-Level Virtual Machine (LLVM) intermediate representation.⁵ Like LLVM, PTX defines a set of simple instructions that provide basic operations for arithmetic, control flow and memory access. PTX also provides several higher-level operations such as texture access and transcendental operations. Also similar to LLVM, PTX assumes an infinite register file and abstracts many real machine instructions. A JIT compiler in the CUDA runtime will perform register allocation, instruction scheduling, dead-code elimination, and numerous other late optimizations as it produces machine code targeting a particular GPU architecture.

PTX is written from the perspective of a single thread and thus does not require explicit lane mask manipulation operations. This makes it straightforward to lower PTX from a high-level shading language, while giving the OptiX runtime the ability to manipulate and optimize the resulting code.

NVIDIA's CUDA C/C++ compiler, `nvcc`, emits PTX and is currently the preferred mechanism for programming OptiX. Programs are compiled offline using `nvcc` and submitted to the OptiX API as a PTX string. By leveraging the CUDA C++ compiler, OptiX shader programs have a rich set of programming language constructs available, including pointers, templates, and overloading that come automatically by using C++ as the input language. A set of header files is provided that support the necessary variable annotations and pseudo-instructions for tracing rays and other OptiX operations. These operations are lowered to PTX in the form of a `call` instruction that gets further processed by the OptiX runtime.

4.2. PTX to PTX compilation

Given the set of PTX functions for a particular scene, the OptiX compiler rewrites the PTX using multiple PTX to PTX transformation passes, which are similar to the compiler passes that have proven successful in the LLVM infrastructure. In this manner, OptiX uses PTX as an intermediate representation rather than a traditional instruction set. This process implements a number of domain-specific operations including an ABI (calling sequence), link-time optimizations, and data-dependent optimizations. The fact that most data structures in a typical ray tracer are read-only, provides a substantial opportunity for optimizations that would not be considered safe in a more general environment.

One of the primary steps is transforming the set of mutually recursive programs into a non-recursive state machine. Although this was originally done to allow execution on a device that does not support recursion, we found benefits in scheduling coherent operations on the SIMT device and now employ this transformation even on newer devices that have direct support for recursion. The main step in the transformation is the introduction of a *continuation*, which is the minimal set of data necessary to resume a suspended function.

The set of PTX registers to be saved in the continuation is determined using a backward dataflow analysis pass that determines which registers are live when a recursive call (e.g., `rtTrace`) is encountered. A live register is one that is used as an argument for some subsequent instruction in the data-flow graph. We reserve slots on a per-thread stack array for each of these variables, store them on the stack before the call and restore them after the call. This is similar to a caller-save ABI that a traditional compiler would implement for a CPU-based programming language. In preparation for introducing continuations, we perform a loop-hoisting pass and a copy-propagation pass on each function to help minimize the state saved in each continuation.

Finally, the call is replaced with a branch to return execution to the state machine described below, and a label that can be used to eventually return control flow to this function. Further detail on this transformation can be found in the original paper.

4.3. Optimization

The OptiX compiler infrastructure provides a set of domain-specific and data-dependent optimizations

that would be challenging to implement in a statically compiled environment. These include:

- Elide transformation operations for node graphs that do not utilize a transformation node.
- Eliminate printing and exception related code if these options are not enabled in the current execution.
- Reduce continuation size by regenerating constants and intermediates after a restore. Since the OptiX execution model guarantees that object-specific variables are read-only, this local optimization does not require an interprocedural pass.
- Specialize traversal based on tree characteristics such as existence of degenerate leaves, degenerate trees, shared acceleration structure data, or mixed primitive types.
- Move small read-only data to constant memory or textures if there is available space.

Furthermore, the rewrite passes are allowed to introduce substantial modifications to the code, which can be cleaned up by additional standard optimization passes such as dead-code elimination, constant propagation, loop-hoisting, and copy-propagation.

5. EXECUTION MODEL

Fundamentally, ray tracing is a highly parallel MIMD operation. In any interesting rendering algorithm, rays will rapidly diverge even if they begin together in the camera model. At first blush, this is a challenge for GPUs that rely on SIMT execution for efficiency. However, it should be observed that execution divergence is only temporary; a ray that hits a glass material temporarily diverges from one that hits a painted surface, yet they both quickly return to the core operation of tracing rays - a refraction or reflection in the former case and a shadow ray in the latter.

Consequently, the state machine described in Section 4 provides an opportunity to reconverge after temporary divergence. To accomplish this, we link all of the transformed programs into a monolithic kernel, or *megakernel*, an approach that has proven successful on modern GPUs.¹ This approach minimizes kernel launch overhead but potentially reduces processor utilization as register requirements grow to the maximum across constituent kernels. OptiX implements a megakernel by linking together a set of individual user programs and traversing the state machine induced by execution flow between them at runtime.

5.1. Megakernel execution

A straightforward approach to megakernel execution is simple iteration over a switch-case construct. Inside each case, a user program is executed and the result of this computation is the case, or state, to select on the next iteration. Within such a state machine mechanism, OptiX may implement function calls, recursion, and exceptions.

Figure 4 illustrates a simple state machine. The program states are simply inserted into the body of the switch statement. The state index, which we call a *virtual program counter (VPC)*, selects the program snippet that will be executed next. Function calls are implemented by setting the VPC directly,

virtual function calls are implemented by setting it from a table, and function returns simply restore the state to the continuation associated with a previously active function (the virtual return address). Furthermore, special control flows such as exceptions manipulate the VPC directly, creating the desired state transition in a manner similar to a lightweight version of the *setjmp/longjmp* functionality provided by C.

5.2. Fine-grained scheduling

While the straightforward approach to megakernel execution is functionally correct, it suffers serialization penalties when the state diverges within a single SIMT unit.⁶ To mitigate the effects of execution divergence, the OptiX runtime uses a fine-grained scheduling scheme to reclaim divergent threads that would otherwise lay dormant. Instead of allowing the SIMT hardware to automatically serialize a divergent switch's execution, OptiX explicitly selects a single state for an entire SIMT unit to execute using a scheduling heuristic. Threads within the SIMT unit that do not require the state simply idle that iteration. The mechanism is outlined in Figure 5.

We have experimented with a variety of fine-grained scheduling heuristics. One simple scheme that works well determines a schedule by assigning a static prioritization over states. By scheduling threads with like states during execution, OptiX reduces the number of total state transitions made by a SIMT unit, which can substantially decrease execution time over the automatic schedule induced by the serialization hardware. Figure 6 shows an example of such a reduction.

As GPUs evolve, different execution models may become practical. For example, a streaming execution model² may be useful on some architectures. Other architectures may provide hardware support for acceleration structure traversal or other common operations. Since the OptiX engine does not

Figure 4. Pseudo-code for a simple state machine approach to megakernel execution. The state to be selected next is chosen by a switch statement. The switch is executed repeatedly until the state variable contains a special value that indicates termination.

```
state = initialState;
while( state != DONE )
  switch(state) {
    case 1:      state = program1();  break;
    case 2:      state = program2();  break;
    ...
    case N:      state = programN();  break;
  }
```

Figure 5. Pseudo-code for megakernel execution through a state machine with fine-grained scheduling.

```
state = initialState;
while( state != DONE ) {
  next_state = scheduler();
  if(state == next_state)
    switch(state) {
      // Insert cases here as before
    }
}
```

prescribe an execution order between the roots of the ray trees, these alternatives could be targeted with a rewrite pass similar to the one we presently use to generate a megakernel.

6. APPLICATION CASE STUDIES

This section presents some example use cases of OptiX by discussing the basic ideas behind a number of different applications. More examples can be found in Parker et al.¹⁰

6.1. Whitted-style ray tracing

The OptiX SDK contains several example ray tracing applications. One of these is an updated re-creation of Whitted's original sphere scene (Figure 7).¹⁴ This scene is simple, yet demonstrates important features of the OptiX engine.

The sample's ray generation program implements a basic pinhole camera model. The camera position, orientation, and viewing frustum are specified by a set of program variables that can be modified interactively. The ray generation program begins the shading process by shooting a single ray per pixel or, optionally, performing adaptive antialiasing via supersampling. The material *closest-hit* programs are then responsible for recursively casting rays and computing a shaded sample color. After returning from the recursion,

Figure 6. The benefit of fine-grained scheduling with prioritization, as achieved when rendering 7. Bars represent the number of state executions per pixel. A substantial reduction can be seen by scheduling the state transitions with a fixed priority, as described in Section 5.2.

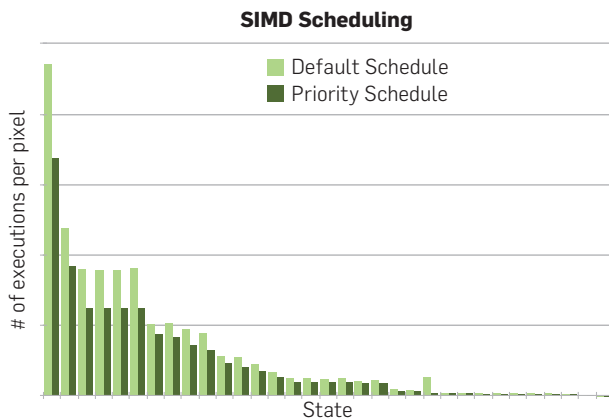
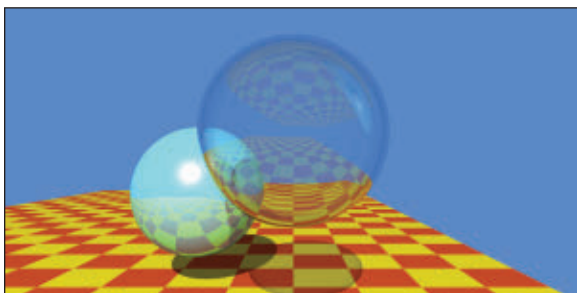


Figure 7. Re-creation of Whitted's sphere scene with user-specified programs: sphere and rectangle intersection; glass, procedural checker, and metal hit programs; sky miss program; and pinhole camera with adaptive anti-aliasing ray generation. Runs at over 100 fps on a GeForce GTX680 at 1k by 1k resolution.



the ray generation program accumulates the sample color, stored in the ray payload, into an output buffer.

The application defines three separate pairs of intersection and bounding box programs, each implementing a different geometric primitive: a parallelogram for the floor, a sphere for the metal ball, and a thin-shell sphere for the hollow glass ball. The glass ball could have been modeled with two instances of the plain sphere primitive, but the flexibility of the OptiX program model gives us the freedom to implement a more efficient specialized version for this case. Each intersection program sets several attribute variables: a geometric normal, a shading normal, and, if appropriate, a texture coordinate. The attributes are utilized by material programs to perform shading computations.

The ray type mechanism is employed to differentiate radiance from shadow rays. The application attaches to the materials' *any-hit* slots for shadow rays, a trivial program that immediately terminates a ray. This early ray termination yields high efficiency for mutual visibility tests between a shading point and the light source. The glass material is an exception, however: here, the any-hit program is used to attenuate a visibility factor stored in the ray payload. As a result, the glass sphere casts a subtler shadow than the metal sphere.

6.2. NVIDIA design garage

NVIDIA Design Garage is a sophisticated interactive rendering demonstration intended for public distribution. The top image of Figure 2 was rendered using this software. The core of Design Garage is a physically-based Monte Carlo path tracing system⁴ that continuously samples light paths and refines an image estimate by integrating new samples over time. The user may interactively view and edit a scene as an initial noisy image converges to the final solution.

To control stack utilization, Design Garage implements path tracing using iteration within the ray generation program rather than recursively invoking *rtTrace*. The pseudo-code of Figure 8 summarizes.

In Design Garage, each material employs a closest-hit program to determine the next ray to be traced, and passes that back up using a specific field in the ray payload. The closest-hit program also calculates the throughput of the current light bounce, which is used by the ray generation to maintain the cumulative product of throughput over the complete light path. Multiplying the color of the light source hit by the last ray in the path yields the final sample contribution.

OptiX's support for C++ in ray programs allow materials to share a generic closest-hit implementation that implements

Figure 8. Pseudo-code for iterative path tracing in Design Garage.

```
float3 throughput = make_float3( 1, 1, 1 );
payload.nextRay = camera.getPrimaryRay();
payload.shootNextRay = true;

while( payload.shootNextRay == true ) {
    rtTrace( payload.nextRay, payload );
    throughput *= payload.throughput;
}
sampleContribution = payload.lightColor * throughput;
```

light-loops and other core lighting operations, while a specific Bidirectional Scattering Distribution Function (BSDF) model implements importance sampling and probability density evaluation. Design Garage implements a number of different physically-based materials, including metal and automotive paint. Some of these shaders support normal and specular maps.

While OptiX implements all ray tracing functionality of Design Garage, an OpenGL pipeline implements final image reconstruction and display. This pipeline performs various post processing stages such as tone mapping, glare, and filtering using standard rasterization-based techniques.

6.3. Image space photon mapping

Image space photon mapping (ISPM)⁷ is a real-time rendering algorithm that combines ray tracing and rasterization strategies (Figure 9). We ported the published implementation to the OptiX engine. That process gives insight into the differences between a traditional vectorized serial ray tracer and OptiX.

The ISPM algorithm computes the first segment of photon paths from the light by rasterizing a “bounce map” from the light’s reference frame. It then propagates photons by recursively ray tracing until the last scattering event before the eye. At each scattering event, the photon is deposited into an array that is the “photon map.” Indirect illumination is then gathered in image space by rasterizing a small volume around each photon from the eye’s viewpoint. Direct illumination is computed by shadow maps and rasterization.

Consider the structure of a CPU-ISPM photon tracer. It launches one persistent thread per core. These threads process photon paths from a global atomic work queue. ISPM photon mapping generates incoherent rays, so traditional packet strategies for vectorizing ray traversal do not help with this process. For each path, the processing thread enters a while-loop, depositing one photon in a global photon array per iteration. The loop terminates upon photon absorption.

Trace performance increases with the success of fine-grain scheduling of programs into coherent units and decreases with the size of state communicated between programs. Mimicking a traditional CPU-style of software

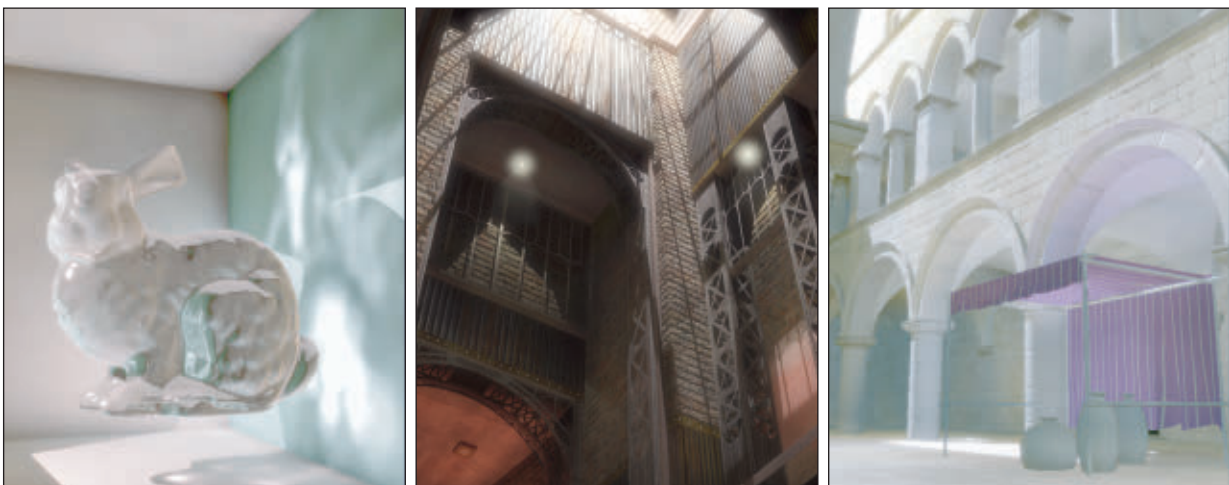
architecture would be inefficient under OptiX because it would require passing all material parameters between the ray generation and hit programs and a variable iteration while-loop in the closest-hit program. OptiX-ISPM therefore follows an alternative design that treats all propagation iterations as co-routines. It contains a single ray generation program with one thread per photon path. A recursive closest-hit program implements the propagate-and-deposit iterations. This allows threads to yield between iterations so that the fine-grained scheduler can regroup them.

7. SUMMARY AND FUTURE WORK

The OptiX system provides a general-purpose and high performance ray tracing API. OptiX balances ease of use with performance by presenting a simple programming model, based on a programmable ray tracing pipeline for single-ray user programs that can be compiled into an efficient self-scheduling megakernel. Thus the heart of OptiX is a JIT compiler that processes *programs*, snippets of user-specified code in the PTX language. OptiX associates these programs with nodes in a graph that defines the geometric configuration and acceleration data structures against which rays are traced. Our contributions include a low-level ray tracing API and associated programming model, the concept of a programmable ray tracing pipeline and the associated set of program types, a domain-specific JIT compiler that performs the megakernel transformations and implements several domain-specific optimizations, and a lightweight scene representation that lends itself to high-performance ray tracing and supports, but does not restrict, the structure of the application scene graph. The OptiX ray tracing engine is a shipping product and already supports a wide range of applications. We illustrate the broad applicability of OptiX with multiple examples ranging from simplistic to fairly complex.

While OptiX already contains a rich set of features and is suitable for many use cases, it will continue to be refined and improved. For example, we (or a third party developer) can add support for further high level input languages, i.e. languages that produce PTX code to be consumed by OptiX. In

Figure 9. ISPM real-time global illumination. A recursive closest-hit program in OptiX implements the photon trace.




addition to language frontends, we are planning support for further backends. Because PTX serves only as an intermediate representation, it is possible to translate and execute compiled megakernels on machines other than NVIDIA GPUs. OptiX has a CPU fallback path that employs this approach.

One downside of OptiX, like any compiler, is that performance of the compiled kernel does not always match a hand-tuned kernel for a specific use-case. We continue to explore optimization techniques to close that gap. The original paper discusses performance in more detail.

We have also discovered tradeoffs in the compile-time specialization of kernels that achieve high performance, but result in small delays when assumptions are violated and a kernel must be regenerated. In the future, the system may choose to fall back to a generalized kernel to maintain slightly degraded interactivity while a new specialized kernel is compiled.

Acknowledgments

The car, frog, and engine model in Figure 1 are courtesy of TurboSquid. The bunny model in Figures 3 and 9 is courtesy of the Stanford University Graphics Lab. Phil Miller was instrumental in keeping the effort on track. The authors benefited greatly from groundwork and numerous conversations on ray tracing with members of NVIDIA Research and the SceniX team. 

References

1. Aila, T., Laine, S. Understanding the efficiency of ray traversal on GPUs. In *Proceedings of High-Performance Graphics 2009* (2009), 145–149.
2. Gribble, C.P., Ramani, K. Coherent ray tracing via stream filtering. In *Proceedings of the 2006 IEEE Symposium on Interactive Ray Tracing* (2008), 59–66.
3. Horn, D.R., Sugerma, J., Houston, M.,

4. Hanrahan, P. Interactive k-d tree gpu raytracing. In *ISD '07: Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games* (2007), ACM, New York, NY, USA, 167–174.
5. Kajiya, J.T. The rendering equation. In *Computer Graphics (Proceedings of ACM SIGGRAPH)* (1986), 143–150.
6. Lattner, C., Adve, V. LLVM: A compilation framework for lifelong program analysis & transformation. In *CGO '04: Proceedings of the 2004 International Symposium on Code Generation and Optimization* (2004).
7. Lindholm, E., Nickolls, J., Oberman, S., Montrym, J. NVIDIA Tesla: A unified graphics and computing architecture. *IEEE Micro* 28 (2008), 39–55.
8. McGuire, M., Luebke, D. Hardware-accelerated global illumination by image space photon mapping. In *Proceedings of the 2009 ACM SIGGRAPH/EuroGraphics conference on High Performance Graphics* (2009).
9. NVIDIA. PTX: Parallel Thread Execution ISA Version 2.3 (2011). http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/ptx_isa_2.3.pdf.
10. NVIDIA. NVIDIA OptiX Ray Tracing Engine Programming Guide Version 2.5 (2012). <http://www.nvidia.com/object/optix.html>.
11. Parker, S.G., Bigler, J., Dietrich, A., Friedrich, H., Hoberock, J., Luebke, D., McAllister, D., McGuire, M., Morley, K., Robison, A., Stich, M. OptiX: A general purpose ray tracing engine. In *ACM Transactions on Graphics (TOG) – Proceedings of ACM SIGGRAPH* (2010).
12. Popov, S., Günther, J., Seidel, H.P., Slusallek, P. Stackless kd-tree traversal for high performance gpu ray tracing. In *Computer Graphics Forum (Proceedings of Eurographics)*, vol. 26, no. 3 (Sept. 2007), 415–424.
13. Wald, I., Benthin, C., Wagner, M., Slusallek, P. Interactive rendering with coherent ray tracing. In *Computer Graphics Forum (Proceedings of Eurographics 2001)*, vol. 20, (2001).
14. Wald, I., Mark, W.R., Günther, J., Boulos, S., Ize, T., Hunt, W., Parker, S.G., Shirley, P. State of the art in ray tracing animated scenes. In *STAR Proceedings of Eurographics 2007* (2007), 89–116.
15. Whitted, T. An improved illumination model for shaded display. *Commun. ACM* 23, 6 (1980), 343–349.

Steven G. Parker, Heiko Friedrich, David Luebke, Keith Morley, James Bigler, Jared Hoberock, David McAllister, Austin Robison, Andreas Dietrich, Greg Humphreys, and Martin Stich (sparker, hfriedrich, dluebke, kmorley, jbigler, jhoberock, davemc, arobison, adietrich, ghumphreys, mstich}@nvidia.com), NVIDIA, Santa Clara, CA.

Morgan McGuire (morgan@cs.williams.edu), NVIDIA and Williams College.

© 2013 ACM 0001-0782/13/05



Association for
Computing Machinery

Advancing Computing as a Science & Profession



You've come a long way.
Share what you've learned.



ACM has partnered with MentorNet, the award-winning nonprofit e-mentoring network in engineering, science and mathematics. MentorNet's award-winning **One-on-One Mentoring Programs** pair ACM student members with mentors from industry, government, higher education, and other sectors.

- Communicate by email about career goals, course work, and many other topics.
- Spend just **20 minutes a week** - and make a huge difference in a student's life.
- Take part in a lively online community of professionals and students all over the world.



Make a difference to a student in your field.
Sign up today at: www.mentornet.net
Find out more at: www.acm.org/mentornet

MentorNet's sponsors include 3M Foundation, ACM, Alcoa Foundation, Agilent Technologies, Amylin Pharmaceuticals, Bechtel Group Foundation, Cisco Systems, Hewlett-Packard Company, IBM Corporation, Intel Foundation, Lockheed Martin Space Systems, National Science Foundation, Naval Research Laboratory, NVIDIA, Sandia National Laboratories, Schlumberger, S.D. Bechtel, Jr. Foundation, Texas Instruments, and The Henry Luce Foundation.

CAREERS

Maharishi University of Management Computer Science Dept Assistant/Associate Professor of Computer Science

The Computer Science Department at Maharishi University of Management invites applications for a full-time faculty position beginning Fall 2013. Qualifications include Ph.D. in Computer Science (or closely related area), or M.S. and seven years of professional software development experience. Candidates will be considered for Assistant, Associate, or full Professor depending on experience and qualifications.

The primary responsibility is teaching computer science courses at the MS level. Applications will be reviewed as they are received until the position is filled. To apply, email curriculum vitae (pdf file) to cssearch2011@mum.edu.

For further information, see <http://www.mum.edu/> and <http://mscs.mum.edu/>. MUM is located in Fairfield, Iowa, and is an equal opportunity employer.

State University of New York at Binghamton Department of Computer Science Assistant Professor Positions

Applications are invited for a tenure-track Assistant Professor Positions beginning Fall 2013 with specialization in the information security or systems-level security area. The Department has about 800 majors, including 63 full-time PhD students. Junior

faculty have a significantly reduced teaching load for at least the first three years. This position is part of a campus initiative in the cybersecurity area. Apply online at: <http://binghamton.interviewexchange.com>

First consideration given to applications received by **April 25, 2013**.

We are an EE/AA employer.

University of Cape Town Department of Computer Science Lecturer or Senior Lecturer

The Department of Computer Science seeks to appoint two academics at Lecturer or Senior Lecturer level.

The successful candidate must have a PhD at the time of appointment. Applicants should provide clear information about their research and teaching experience as this can influence the level of appointment. They will be expected to develop and teach Computer Science courses, to carry out research, to contribute to departmental administration and to supervise postgraduate students.

Our BSc Honours degrees are accredited by the British Computer Society and we have a large cohort of MSc and PhD students. The Department hosts the UCT Centre in ICT for Development, and also specialises in telecommunications, visual computing, high-performance computing, digital libraries, artificial intelligence and security. A specialization in one of these areas will be an advantage.

The annual remuneration packages for 2013, including benefits but excluding a 10% annual scarce skills allowance, are:

- ▶ Lecturer: R 457 223
- ▶ Senior Lecturer: R 562 173

To apply please e-mail the application form at <http://web.uct.ac.za/depts/sapweb/forms/hr201.doc> and all other relevant documentation as indicated on the form, to Edith Graham (Ref: SR465/13), Staff Recruitment and Selection, UCT, by 31 May 2013.

E-mail: edith.graham@uct.ac.za;
Telephone: +27 21 650 5405;
Department website: www.cs.uct.ac.za

UCT is committed to the pursuit of excellence, diversity and redress. Our Employment Equity Policy is available at

<http://hr.uct.ac.za/policies/ee.php>.

University of Central Florida (UCF) Center for Research in Computer Vision (CRCV) Faculty Positions

The University of Central Florida (UCF) has recently established a University level Center for Research in Computer Vision (CRCV). The common



Florida Institute of Technology
High Tech with a Human Touch™

Florida Institute of Technology offers a master's program and Ph.D. program in Human-Centered Design (HCD).



- Candidates with backgrounds and degrees in engineering, science and human factors, as well as arts and architecture are encouraged to apply.
- A graduate degree in HCD supports independent scholarly work, opportunities in academia or pursuit of advanced research and leadership in government, industry and business.
- Current research is in: cognitive engineering, life-critical systems, complexity analysis for HCD, human-centered organization design and management, modeling and simulation, advanced interaction media, creativity and design thinking, functional analysis, industrial design, and usability engineering.
- Internationally connected with best research and professional institutions in HCD.

For more information:
(321) 309-4960 • dcaballe@fit.edu
Visit: <http://research.fit.edu/hcdi>
150 W. University Blvd., Melbourne, FL 32901

EN-109-213

ACM Inroads
The magazine for
computing educators worldwide
Paving the way toward excellence in computing education

<http://inroads.acm.org>

ACM Association for Computing Machinery
Advancing Computing as a Science & Profession

goal of the center is to strongly promote basic research in computer vision and its applications.

CRCV is looking for multiple exceptional tenured or tenure-track faculty members, at all levels in the Computer Vision area. Of particular interest are mid-career and senior candidates with a strong track record of publications and research funding. CRCV will offer competitive salaries and start-up packages, and UCF provides generous benefits. Faculty hired at CRCV will be tenured in the Electrical Engineering & Computer Science department and will be required to teach a maximum of two courses per academic year and be expected to bring in substantial external research funding. In addition, Center faculty are expected to have a vigorous program of graduate student mentoring and are encouraged to involve undergraduates in their research.

Applicants must have a Ph.D. in an area appropriate to Computer Vision by the start of the appointment and a strong commitment to academic activities, including teaching, scholarly publications and sponsored research. Prefer applicants with an exceptional record of scholarly research and, at the senior levels, be highly recognized for

their technical contributions and leadership in their areas of expertise. In addition, successful candidates must be strongly effective teachers.

To submit an application, please go to: <http://www.jobswithucf.com/postings/34681>

Applicants must submit all required documents at the time of application which includes the following: Research Statement; Teaching Statement; Curriculum Vitae; and a list of at least three references with address, phone numbers and email address.

Applicants for this position will also be considered for position numbers 38406 and 37361.

University of Chicago
Department of Computer Science
Associate Professor - Req # 01629

The Department of Computer Science at the University of Chicago invites applications from qualified candidates for faculty positions at the rank of Associate Professor in the area of machine learning. Outstanding researchers working in both the theory of machine learning and applications to areas such as natural language processing, computer vision, and computer systems are encouraged to apply.

Candidates must have a doctoral degree in computer science or a related field and be several years beyond the Ph.D. Candidates are expected to have established an outstanding independent research program and will be expected to contribute to the Department's undergraduate and graduate teaching programs.

The University of Chicago has the highest standards for scholarship and faculty quality, is dedicated to fundamental research, and encourages collaboration across disciplines.

The Department of Computer Science (cs.uchicago.edu) is the hub of a large, diverse computing community of two hundred researchers focused on advancing foundations of computing and driving its most advanced applications. Long distinguished in theoretical computer science and artificial intelligence, the Department is now building strong systems and machine learning groups. The larger community in these areas at the University of Chicago includes the Computation Institute, the Toyota Technological Institute, the Department of Statistics, and Argonne's Mathematics and Computer Science Division.

The Chicago metropolitan area provides a diverse and exciting environment. The local economy is vigorous, with international stature in banking, trade, commerce, manufacturing, and transportation, while the cultural scene includes multiple cultures, vibrant theater, world-renowned symphony, opera, jazz, and blues. The University is located in Hyde Park, a Chicago neighborhood on the Lake Michigan shore just a few minutes from downtown on an electric commuter train.

All applicants must apply through the University's Academic Career Opportunities website <http://tinyurl.com/at80lkr> and must upload a curriculum vitae with a list of publications, a succinct outline of research plans, a one-page teaching statement and a reference contact list consisting of three people. Review of completed applications will continue until the position is filled.

The University of Chicago is an Affirmative Action / Equal Opportunity Employer.



ADVERTISING IN CAREER OPPORTUNITIES

How to Submit a Classified Line Ad: Send an e-mail to acmm mediasales@acm.org. Please include text, and indicate the issue/or issues where the ad will appear, and a contact name and number.

Estimates: An insertion order will then be e-mailed back to you. The ad will be typeset according to CACM guidelines. NO PROOFS can be sent. Classified line ads are NOT commissionable.

Rates: \$325.00 for six lines of text, 40 characters per line. \$32.50 for each additional line after the first six. The MINIMUM is six lines.

Deadlines: 20th of the month/2 months prior to issue date. For latest deadline info, please contact: acmm mediasales@acm.org

Career Opportunities Online: Classified and recruitment display ads receive a free duplicate listing on our website at: <http://jobs.acm.org>

Ads are listed for a period of 30 days.

For More Information Contact:

**ACM Media Sales,
at 212-626-0686 or
acmm mediasales@acm.org**

ACM Transactions on Reconfigurable Technology and Systems

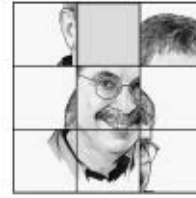


This quarterly publication is a peer-reviewed and archival journal that covers reconfigurable technology, systems, and applications on reconfigurable computers. Topics include all levels of reconfigurable system abstractions and all aspects of reconfigurable technology including platforms, programming environments and application successes.

www.acm.org/trets
www.acm.org/subscribe



Association for
Computing Machinery



DOI:10.1145/2447976.2447998

Peter Winkler

Puzzled

Ant Alice's Adventures

These three puzzles involve my favorite ant, Ant Alice. Like all ants on this page, Alice moves at exactly one centimeter per second in whichever direction she happens to be facing; if she meets another ant head on, both immediately reverse direction and walk away from each other, each still at speed 1 cm/sec. Figuring out how Alice and her friends behave is surprisingly easy if viewed the right way; here's a tip: Certain physical principles may play a role in your reasoning.

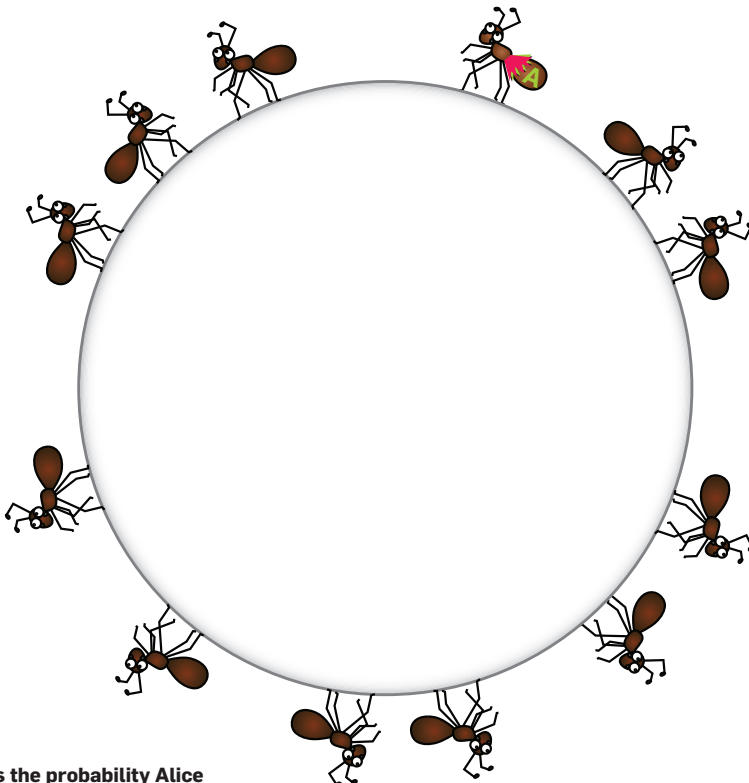
1. Ant Alice is the middle ant of 25 ants on a meter-long stick, some facing east, some facing west. (We may assume ants are tiny compared to the distances between them, so they can be thought of as moving points.)

At a signal, all begin to march in whichever direction they are currently facing, bouncing and reversing direction whenever two collide. Those reaching the end of the stick fall off and float gently to the ground (no ants were harmed

in the creation of these puzzles). How long must we wait before we are sure Alice has fallen off the stick?

2. Suppose the ants' initial positions, and the directions they face, are uniformly random. What is the probability that when Alice falls off the end she was initially facing?

3. Suppose Alice is one of only 12 ants, each initially placed uniformly at random on a circle of length (circumference) one meter (see the figure here). Each ant initially faces clockwise or counterclockwise with equal probability. At a signal, they begin marching (and bouncing off one another) according to the usual rules. What is the probability that 100 seconds later Alice will find herself exactly where she began?



What is the probability Alice ends up where she started?

Readers are encouraged to submit prospective puzzles for future columns to puzzled@cacm.acm.org.

Peter Winkler (puzzled@cacm.acm.org) is William Morrill Professor of Mathematics and Computer Science at Dartmouth College, Hanover, NH.

ILLUSTRATION BY PETER WINKLER

Computing Reviews

The Best Reviews and
Notable Books & Articles
of 2012

Online and in print

computingreviews.com

A daily snapshot of what is new and hot in computing.



ACM ANNOUNCES NEW CHANGES TO EXPAND ACCESS TO PUBLICATIONS

ACM is pleased to announce important changes to our publications policy aimed at increasing exposure to and the impact of our publications on the global computing community. These changes further empower ACM authors to more widely distribute and make available their work, while at the same time taking advantage of the numerous benefits and prestige of publishing with ACM.

Some highlights of the new policy:

- ACM authors now have the option of retaining copyright and other important ownership rights by selecting one of ACM's "Copyright Transfer", "Exclusive License to Publish", or "Non-Exclusive Permission to Publish" Agreements
- ACM authors now have the option of making their work perpetually free to the world via the ACM DL platform by participating in the new ACM Open Access publication program.
- ACM Special Interest Groups (SIGs) now have the option of making their ACM conference proceedings freely available to the world on a limited basis via the ACM Digital Library, SIG, or conference websites.

For detailed information about these new options, as well as information about ACM's complete publications policy, please visit <http://authors.acm.org>.

acm Author Rights

- CHOOSE your rights option
- POST on your own websites
- DISTRIBUTE via ACM Author-Izer
- REUSE your own work
- CREATE derivative works
- RETAIN perpetual rights