Innovative Technology for Computer Professionals

Computer

JANUARY 2011

http://www.computer.org

OUTLOOK

IEEE-CS PRESIDENT'S MESSAGE, P. 6

EIC'S MESSAGE, P. 9

INDUSTRY-ACADEMIA SYNERGY, P. 90

IEEE

IEEE computer society

Innovative Technology for Computer Professionals
Computer

Circulation: *Computer* (ISSN 0018-9162) is published monthly by the IEEE Computer Society. **IEEE Headquarters**, Three Park Avenue, 17th Floor, New York, NY 10016-5997; **IEEE Computer Society Publications Office**, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720-1314; voice +1 714 821 8380; fax +1 714 821 4010; **IEEE Computer Society Headquarters**, 2001 L Street NW, Suite 700, Washington, DC 20036. IEEE Computer Society membership includes $19 for a subscription to *Computer* magazine. Nonmember subscription rate available upon request. Single-copy prices: members $20.00; nonmembers $99.00.

Postmaster: Send undelivered copies and address changes to *Computer,* IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Canadian GST #125634188. Canada Post Corporation (Canadian distribution) publications mail agreement number 40013885. Return undeliverable Canadian addresses to PO Box 122, Niagara Falls, ON L2E 6S8 Canada. Printed in USA.

Editorial: Unless otherwise stated, bylined articles, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *Computer* does not necessarily constitute endorsement by the IEEE or the Computer Society. All submissions are subject to editing for style, clarity, and space.

Innovative Technology for Computer Professionals
Computer
http://computer.org/computer

CONTENTS

ABOUT THIS ISSUE

In this annual Outlook issue, we look ahead at research and technology that are likely to have an impact in the next few years. Topics addressed include the findings of a committee convened to identify key challenges to continued growth in computing performance and to outline a research agenda for meeting emerging 21st-century computing needs; nanostores, a new design approach that focuses on data-centric workloads and hardware-software codesign for upcoming technologies; the application of 3D silicon interposer technology to high-performance computational systems; and an apoptic computing project that has been working toward the long-term goal of programmed death by default for computer systems.

COMPUTING PRACTICES

COVER FEATURES

RESEARCH FEATURE

For more information on computing topics, visit the Computer Society Digital Library at www.computer.org/csdl.

IEEE Computer Society: http://computer.org
Computer: http://computer.org/computer
computer@computer.org
IEEE Computer Society Publications Office: +1 714 821 8380

Flagship Publication of the IEEE Computer Society

January 2011, Volume 44, Number 1

NEWS

MEMBERSHIP NEWS

COLUMNS

DEPARTMENTS

Computer Highlights Society Magazines

The IEEE Computer Society offers a lineup of 13 peer-reviewed technical magazines that cover cutting-edge topics in computing including scientific applications, design and test, security, Internet computing, machine intelligence, digital graphics, and computer history. Select articles from recent issues of Computer Society magazines are highlighted below.

IEEE Software

Small to medium enterprises increasingly participate in offshore software development. Key competitive SME abilities include detecting market niches and deploying highly flexible software development approaches. Therefore, learning how offshoring affects such capabilities, which are closely related to organizational learning, is crucial. "Operational and Strategic Learning in Global Software Development," in the November/December issue of *Software*, presents case studies from two German companies that engage in offshoring of software development.

Authors Alexander Boden, Bernhard Nett, and Volker Wulf of the University of Siegen highlight the different structures these companies have chosen for their development work and describe how they enact those structures.

IEEE Intelligent Systems

Retrieving useful Web news involves both filtering and keyword extraction. Because the layouts and styles of Web news pages differ from other webpages, it is especially important to accurately identify Web news for correct filtering. To do this, the authors of "News Filtering and Summarization on the Web" in the September/October issue of *IS* propose an automatic recognition method that uses classification rules for Web news based on a combination of URL, structure, and content attributes. After the automatic recognition and filtering, the system uses a new key-phrase extraction method from Web news content based on semantic relations.

IT Professional

TECHNOLOGY SOLUTIONS FOR THE ENTERPRISE

Can law enforcement agencies leverage open source to benefit the communities they serve? In some areas, recommending open source solutions is easy. In others, the arguments are less clear. Faced with limited funding and aging technology, some agencies have turned to using open source software as a cost-effective solution.

Authors Mun-Wai Hon, Greg Russell, and Michael Welch of Noblis look at four agencies' experiences with open source software in "Open Source Software Considerations for Law Enforcement" in the November/December issue of *IT Pro*.

IEEE Computer Graphics and Applications

In "Guest Editors' Introduction: Highlights from IEEE Pacific Visualization," which appears in the November/December issue of *CG&A*, Jarke van Wijk of Eindhoven University of Technology, Stephen North of AT&T, and Han-Wei Shen of Ohio State University introduce four articles based on papers from the 2010 IEEE Pacific Visualization Symposium. These articles, which strongly focus on visual design and applications, cover a range of applications in scientific visualization, information visualization, and graph visualization, demonstrating the field's breadth.

computing in SCIENCE & ENGINEERING

Creating the next generation of power-efficient parallel computers requires new mechanisms and methodologies for building parallel applications. Energy constraints have pushed technologies into a regime where parallelism will be ubiquitous, rather than limited to specialized high-end supercomputers. New execution models are required to span all scales, from desktop to supercomputer. Read "Advanced Architectures and Execution Models to Support Green Computing" and many other articles on green high-performance computing in the November/December issue of *CiSE*.

IEEE SECURITY & PRIVACY

Building Dependability, Reliability, and Trust

Cloud computing is quickly becoming the next wave of technological evolution in providing enterprise IT capabilities. Driving interest and investment in cloud computing are revolutionary changes to the economic model. This shift to on-demand and self-service IT functions creates new challenges, including regulating how internal business units purchase cloud services. But how does a business assess cloud providers' services for security, privacy, and service levels?

In "Cloud Provider Transparency: An Empirical Evaluation," in the November/December issue of *S&P*, author Wayne Pauley of EMC proposes an instrument for evaluating a cloud provider's transparency of security, privacy, and service-level competencies via its self-service Web portals and publications.

IEEE pervasive COMPUTING MOBILE AND UBIQUITOUS COMPUTING

Monitoring terrestrial high-arctic ecosystems is important because of their exposure to global warming. Conditions require a pervasive monitoring infrastructure that collects data automatically. Consequently, measurements that have traditionally been obtained manually must now be obtained with automatic systems.

In "Monitoring in a High-Arctic Environment: Some Lessons from MANA" in the October-December issue of *PvC*, authors Marcus Chang and Philippe Bonnet of IT University of Copenhagen relate details of the MANA project, which tackles these challenges with a sensor-network-based data acquisition system for year-round lake monitoring in Northeast Greenland. The article emphasizes issues that the researchers initially underestimated—the consequences of operating in a remote region, extreme weather's impact on system design and operator activities, and the demands resulting from the absence of a communication infrastructure.

IEEE Internet Computing

Most Web 2.0 applications let users associate textual information with multimedia content. Despite each application's lack of editorial control, these textual features are still the primary source of information for services such as search. Previous efforts to assess the quality of these features targeted single applications and focused on tags, neglecting the potential of other features. In "On the Quality of Information for Web 2.0 Services" in the November/December issue of *IC*, a team of authors from Brazil's Universidade Federal de Minas Gerais assesses and compares the quality of four textual features (title, tags, description, and comments) for supporting information services using data from YouTube, YahooVideo, and LastFM.

IEEE micro

In "Parallel Programming Models for Heterogeneous Multicore Architectures" in *Micro*'s September/October issue, a team of authors from Spain's Barcelona Supercomputing Center evaluates the scalability and productivity of six parallel programming models for heterogeneous architectures, and finds that task-based models using code and data annotations require minimum programming effort while sustaining nearly best performance. However, achieving this result requires both extensions of programming models to control locality and granularity and proper interoperability with platform-specific optimizations.

IEEE MultiMedia

Augmented reality expands the physical real-world environment to enhance users' perception of reality and experience by augmenting its elements with virtual computer-generated imagery. In the October-December issue of *MultiMedia*, Cristina Portalés, María J. Viñals, Pau Alonso-Monasterio, and Maryland Morant of the Universidad Politécnica de Valencia look at a state-of-the-art AR application in "AR-Immersive Cinema at the Aula Natura Visitors Center." The article showcases AR-Immersive Cinema, a novel AR application for museum exhibitions that is designed to merge past and present and immerse users in history.

IEEE Design&Test of Computers

D&T's November/December issue presents six articles that highlight challenges of, and approaches toward improving, design yield and reliability through postsilicon optimizations. The articles cover postsilicon adaptation and repair issues in a wide range of areas including analog circuits, embedded memories, and multicore systems. In "Guest Editors' Introduction: Managing Uncertainty through Postfabrication Calibration and Repair," Swarup Bhunia of Case Western Reserve University and Rahul Rao of IBM's T.J. Watson Research Center identify several highlights.

IEEE Annals of the History of Computing

What does the Ford mass-production story have to do with the history of computing? Everything. More than half a century before Toyota's just-in-time inventory captured hearts and minds around the industrial world, Ford was using state-of-the-art information technology to implement a networked manufacturing system. In "A Question of Scale: Networks, Systems, and Practice" in the October-December issue of *Annals*, Jonathan Aylen of the University of Manchester contends that historians of computing should continue to produce works about work, particularly work in the predigital age.

Editor: Bob Ward, *Computer*; bnward@computer.org

Planning for an Inevitable Future

Sorel Reisman
IEEE Computer Society 2011 President

As we celebrate its 65th birthday in 2011, the question is whether the IEEE Computer Society is prepared for the next 65, or 25, or even 5 years.

One morning about 30 years ago, I was about to head off to work (at IBM) when I noticed my eight-year-old son at the keyboard of my brand-new Apple II+ computer. My first reaction was to yell at him for playing with my new toy, but I was overtaken by an uncharacteristic moment of calm. For a few minutes I watched him and began to realize that something unique was going on. A child was using a computer console to key commands directly to an OS.

Okay, so you might argue that it was a really primitive console and a simple OS, but so what? After a couple of years, when the "computer revolution" really took hold with the introduction of the IBM PC, it became clear that a new social era was upon us—one that removed control of computing from the hands of engineers and scientists and turned it over to mere mortals.

Thirty years later, all of this is old news. Much has changed since the first days of personal computing, but one thing that hasn't is the need for everyone with a computer of any kind (Macintoshes included) to become a system administrator, a database administrator, a network administrator, and so on. We've all had to acquire some measure of the "IT skills" necessary to keep our computing environments operational and useful.

About three months ago, I had another aha! moment, this time while watching my 2½-year-old granddaughter navigate the menus of her dad's iPhone. It seems that others too have had this moment. In recent months, *The New York Times* and *USA Today* have presented reports describing toddlers' interest, obsession, and skills with the multitude of "iThings" that are starting to flood the consumer market. Journalists are likening this to the late 1940s, when television technology was first consumerized, eventually changing society worldwide.

ARE WE READY?

And so I began to think about what this could mean to us, as members of "the world's leading membership organization for computing professionals." While we celebrate the Computer Society's 65th birthday in 2011, considering the accelerating pace of the commoditization of computing, the question is whether the Society is prepared for the next 65, or 25, or even 5 years.

It's becoming increasingly clear to me that IT, however it's defined, is becoming part of society's physical infrastructure—along with roads and highways, the power grid, and water distribution. And it's also becoming clear that there's an increasingly large and growing audience of computing professionals who are not, strictly speaking, part of the IT culture the Computer Society has been serving since its inception. A stark reality is that with the advent of packaged IT products such as open source object libraries, outsourced IT cloud services, and even nanotechnology-embedded applications, the next generation of computer users will neither know nor care what lies under the hood of their computing appliances.

So the question that we must ask ourselves as Computer Society members and volunteers is, do we continue to focus our energies on a diminishing percentage of IT professionals who make all this happen, or should we be realigning our mission

Published by the IEEE Computer Society

0018-9162/11/$26.00 © 2011 IEEE

toward social changes that are clearly taking place, that are clearly upon us?

About four years ago, the CS, for a variety of reasons, began to transform our organizational structure from what it was then to what it is now—a slimmed-down, highly efficient, and cost-effective professional entity that serves about 75,000 members worldwide. That transformation, which has been very successful, has taken incredible dedication by executive director Angela Burgess, her staff, and the CS volunteers who have served in office since it began.

A NEW TRANSFORMATION

I believe it's time to consider a new transformation—one that addresses the societal changes I have described. To this end, I will be working with the new 2011 Board of Governors members to begin to plan for such a transformation. Working with me on this are the 2011 Executive Committee members: David Alan Grier, first vice president (and VP of publications); Jon Rokne, second vice president (and BoG secretary); Jim Moore, treasurer; Roger Fujii, VP of standards; Liz Burd, VP of educational activities; Paul Croll, VP of technical and conferences activities; Paul Joanneau, VP of professional activities; and Phil Laplante, Electronic Products and Services Committee chair.

Of course, we all owe a debt of gratitude to the 2010 Board under the leadership of Jim Isaak, without whose work and ideas these ambitions would never have become obvious. We will miss Stephen L. Diamond, Itaru Mimura, Christina M. Schober, Sattupathu V. Sankaran, Ann E.K. Sobel, and Jeffrey M. Voas, who are departing from the Board, but I'm sure that each of them will continue to work toward this transformation.

While we begin to plan for our inevitable future, we have to be practical and develop plans and activities that will continue to be the foundation of our current vision and goals. Although it's unlikely that the debate over whether we're academic or practitioner oriented will cease, another thing that won't cease will be our efforts to serve both kinds of members.

To this end, this year I'll be creating a new ad hoc Academic Advisory Board, similar to our Industry Advisory Board created two years ago, to help steer our academic foundation toward the future. We take great pride in both the quality of our publications and conferences and the work of all the volunteers who develop and organize them. But because of the technology-driven social changes that influence our professional lives—from the open access movement, to the popularity of e-learning, to the emergence and acceptance of webinars, podcasts, and so on—it's clear that we must constantly review and examine all the products and services we offer to our academic membership.

By the end of 2011, we'll have launched pilot Special Technical Communities in social networking, cloud computing, gaming, education, software engineering, and green computing.

At the same time, we won't decrease in any way the continued development of our practitioner-oriented products and services. For example, we're currently working with other IEEE societies to develop a new set of career retraining workshops to be launched later in 2011. Through this work, as part of a new member recruitment initiative, we'll soon be releasing a new, all-electronic bundle of software engineering products that include *IEEE Software* and *IT Professional* magazines, selected ReadyNotes, Essential Sets, and webinars related to software engineering, as well as discount coupons to enable members to more easily prepare for certification in our Certified Software Developer Associate/Professional programs.

Not forgotten in all of this are our students. To date we've been focusing our Build Your Career and Jobs Board on already-graduated CS members. This year, we'll be extending those websites to include career-building services for students that feature solicitation and publication of internship and co-op opportunities, as well as mentoring services. One of the new elements that we will introduce as part of this effort will be discussion forums to enable student members to interact with other students all over the world.

Key to this effort will be the extension of Instant Communities, the social network service conceived by 2010 President Jim Isaak, to a completely new service called Special Technical Communities. By the end of 2011, we'll have launched pilot STCs in social networking, cloud computing, gaming, education, software engineering, and green computing.

A foundational element of STCs are social networking technologies, without which it would be impossible to reach out to our geographically diverse members to enable them to work together in their shared fields of interests. One strength of both IEEE and the Computer Society is this geographic diversity. In 2011, in concert with IEEE's Globalization Initiative and eMembership option for emerging countries, I intend to reach out and visit chapters that sometimes don't get the attention they deserve, particularly in the BRIC (Brazil, Russia, India, and China) countries. While it's not always easy to visit everyone in person, I intend to organize a series of electronic meetings in which other senior volunteers and I will "meet" chapter members located outside North America.

AN OPEN INVITATION

Finally, I'd like to invite both members and nonmembers who want to communicate with me to participate in two social networking activities that I'll be using this year. The first is an extension to the President's blog that Kathy Land started in 2009 (www.computer.org/portal/web/the-presidents-discussion-corner).

We're relaunching that site with a slightly different message format. Each month, I'll be posting a very short video blog about an issue, announcement, or topic of interest, controversy, or importance that I invite you to visit and engage in via the threaded discussion there. The videos and their related discussions will be archived and available for revisiting throughout the year. The second activity that I'm launching is a new Twitter site: @ieeeCSPresident. I intend to use Twitter throughout the year to share my ideas, insights, and observations as I proudly represent all members of the IEEE Computer Society at conferences, chapter meetings, and other CS-related events.

Please let me know what you think. While we try very hard to meet our members' needs, we need your input to be sure we're doing a good job for you.

__Sorel Reisman,__ Managing Director of MERLOT, an international, higher education consortium, is a professor of information systems at California State University, Fullerton. Reisman received a PhD in Computer Applications from the University of Toronto. Contact him at sreisman@computer.org.

Selected CS articles and columns are available for free at http://ComputingNow.computer.org.

Print, Mobile, and Online

Ron Vetter
Editor in Chief

It will be important to optimize *Computer*'s content for mobile applications in the very near future so that readers can access information in a form that's convenient for them.

Welcome to the first issue of *Computer*'s 44th volume. Although I've been affiliated with *Computer* for many years, this is my first issue as editor in chief. I'm proud to follow Carl Chang, who effectively led the magazine during the past four years. Carl and previous EICs worked tirelessly to make *Computer* interesting, relevant, and of high quality. I will do my best to continue to meet these expectations as well as to position the magazine to take more advantage of the digital revolution.

EMBRACING THE FUTURE

All scientific publications, including *Computer*, face significant challenges. There's growing evidence that traditional publication models might not survive challenges by open access and Creative Commons licensing. In a digital environment, the traditional roles of both publisher and author are changing rapidly. As stated in the EIC search criteria, one of the primary tasks for *Computer*'s next EIC was to be able to move the magazine into new technology and science directions without compromising the rigor and high standards of its published print material.

One exciting development is in the mobile publishing arena. It's clear that information access via mobile devices has become increasingly commonplace for content publishers and readers alike. Many leading newspaper publishers have already started to convert their publications into fully electronic form as well as creating special versions for mobile phones and tablets. *USA Today* announced that it's reorganizing to focus on digital publishing, primarily on tablets. More recently, *The New York Times* declared that "at some point in the future" production of the printed newspaper will cease. One industry analyst states, "Clearly, digital publishing, primarily through tablets, is going to sweep through the entire publishing industry." It will be important to optimize *Computer*'s content for mobile applications in the very near future so that readers can access information in a form that's convenient for them.

Most people look for information using search engines. It's therefore essential to ensure that the Computer Society's intellectual property is highly ranked in search engines. Improved metadata, open and accessible indexing, and continued work on search engine optimization must be a high priority. In addition, during the next couple of years, the rapidly changing social media area will begin to mature.

As the Computer Society's online social media experimentation within Computing Now begins to show results, *Computer* needs to embrace those ideas that work. All kinds of interesting multimedia content, such as software simulations, audio podcasts, video interviews, PowerPoint lectures, and so on, can be part of this online presence. If you haven't already done so, I encourage you to visit the CN website (www.computer.org/computingnow) to check out some of the multimedia entries, CN Lab applications, and blogs.

EDITORIAL BOARD ADDITIONS

Publishing a high-quality magazine takes considerable time and effort. *Computer* is fortunate to have a high-caliber group of experts, including professional staff and volunteers, who are dedicated to bringing you highly relevant, leading-edge content. The masthead lists the people who help to make *Computer* a reality each month. In addition to the people who have served *Computer* in the past, I would like to introduce you to a few new additions for 2011.

0018-9162/11/$26.00 © 2011 IEEE

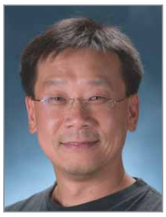
Charles R. Severance, Web editor, is a clinical associate professor in the School of Information at the University of Michigan. A founding faculty member of the Informatics concentration undergraduate degree program at the University of Michigan, Severance also works with the IMS Global Learning Consortium promoting and developing standards for teaching and learning technology. The author of three books and several refereed journal and conference papers, he has experience with serving as an expert on the Internet and technology as the cohost of several TV shows and a public radio call-in program. Severance received a PhD in computer science from Michigan State University.

Theresa-Marie Rhyne, advisory panel member, is a recognized expert in the field of computer-generated visualization. She is a consultant specializing in applying artistic color theories to visualization and digital media. In the 1990s, as a government contractor with Lockheed Martin Technical Services, she was the founding visualization leader of the US Environmental Protection Agency's Scientific Visualization Center. In the 2000s, she founded the Center for Visualization and Analytics and the Renaissance Computing Institute's Engagement Facility at North Carolina State University. Rhyne also serves as editor of the Visualization Viewpoints Department for *IEEE Computer Graphics & Applications* magazine. She received an MS in civil engineering from Stanford University and is a senior member of the IEEE Computer Society.

Rolf Oppliger, area editor for security and privacy, is an adjunct professor at the University of Zurich. Oppliger studied computer science, mathematics, and economics at the University of Berne, where he received an MSc and a PhD in computer science. Oppliger works in the area of information technology security. He has authored 11 books on this subject, frequently speaks at security-related conferences, and has published numerous papers and articles on IT security in scientific magazines and journals. He is the founder and owner of eSecurity Technologies, works for the Swiss federal administration, teaches at the University of Zurich, and serves as editor for the Artech House information security and privacy series. Oppliger is a senior member of the ACM and is a member of the IEEE Computer Society and the IACR.

Karl Ricanek, Identity Sciences column editor, is an associate professor in the Computer Science Department at the University of North Carolina, Wilmington. He is the founder and director of the Face Aging Group Research Lab (www.FaceAgingGroup.com) at UNCW, where he has been the primary project lead on more than $5 million in Department of Defense and intelligence funded research in biometrics. He has authored or coauthored more than 40 refereed articles and three book chapters in biometrics and pattern recognition, and has served as a program committee member for several biometric and related conferences. Ricanek received a PhD in electrical engineering from North Carolina A&T State University.

John Riedl, Social Computing column editor, is a professor in the Computer Science Department at the University of Minnesota. A founding editor in chief of *ACM Transactions on Interactive Intelligent Systems*, he has taught many courses in the areas of programming and systems at both the graduate and undergraduate levels, has authored numerous refereed journal and conference papers, has been the recipient of several teaching and best paper awards, and holds four US patents. Riedl received a PhD in computer sciences from Purdue University. He is a senior member of IEEE, a Fellow of the ACM, and a member of AAAI.



Kelvin Sung, editor of the Entertainment Computing column, is a professor of computing and software systems at the University of Washington, Bothell. His research focuses on studying the role of technology in supporting human communication, with recent work in the areas of serious games and topics related to teaching and learning computer graphics and foundational concepts in programming based on computer games. Sung played a key role in designing and implementing the Maya Renderer, an Academy Award-winning image generation system. He has authored numerous refereed journal and conference papers, has been the recipient of several grants, and holds two US patents. Sung received a PhD in computer science from the University of Illinois at Urbana-Champaign.

GET INVOLVED

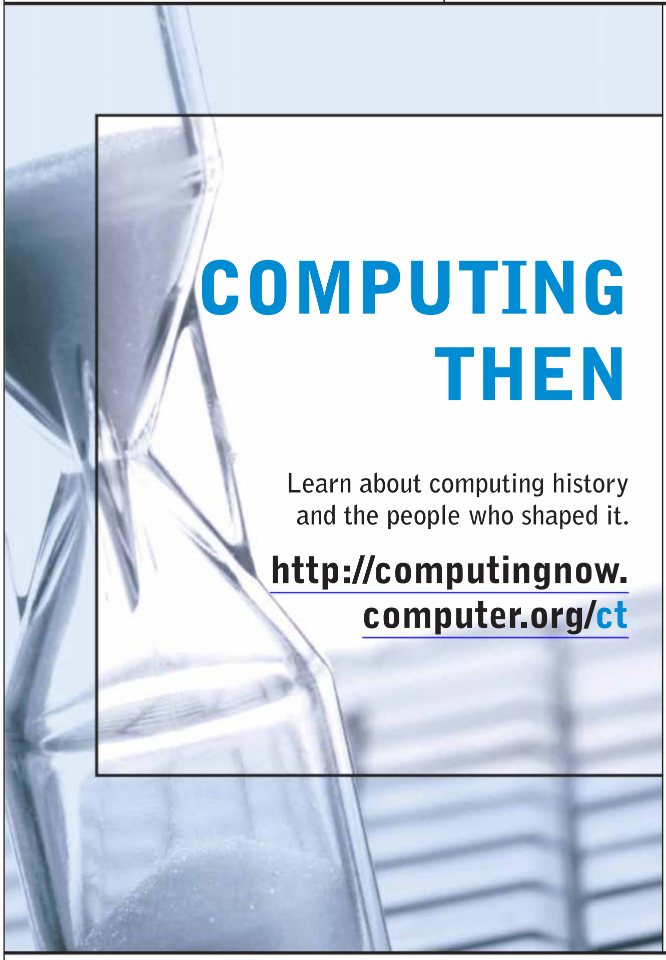
Because all members of the Computer Society receive *Computer* as a member benefit, it's important that the membership is well served. Working with editorial board members and staff, one of my top priorities is to get more members involved in *Computer*. Here are a few of the ways you can contribute:

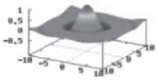
- *Submit an article.* ScholarOne Manuscript (https://mc.manuscriptcentral.com/com-cs), our totally electronic online service for processing manuscript submissions, provides complete author information and submission details.
- *Propose a special issue.* Contact the special issues editor (schilit@computer.org) to offer your suggestion or to receive information about submitting a special issue proposal.
- *Serve as a reviewer.* Indicate your interest in serving as a reviewer by sending an e-mail message containing your vita to computer-ma@computer.org.
- *Provide feedback.* We welcome your comments and encourage you to submit suggestions for topics to be covered in future issues.

I look forward to hearing from you and welcome your participation. C

***Ron Vetter** is a professor and past chair of the Department of Computer Science (www.uncw.edu/csc) at the University of North Carolina Wilmington. In 2007, he cofounded Mobile Education LLC (http://myMobEd.com), a technology company that specializes in developing interactive short message service applications. Vetter has served on numerous journal editorial boards and conference committees, including his current appointment as associate editor for Computing Now. Vetter received a PhD in computer science from the University of Minnesota, Minneapolis. Contact him at vetterr@uncw.edu.*

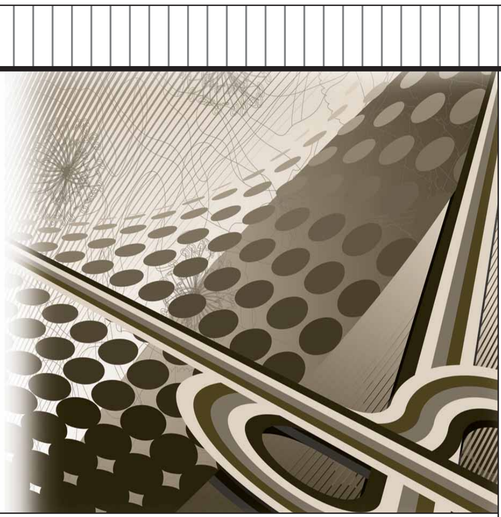
cn Selected CS articles and columns are available for free at http://ComputingNow.computer.org.





The Migration to the Middle

David Alan Grier, *George Washington University*



As we look to the future, we must not only anticipate a year of innovation and progress but also a migration of labor that will remake the field of digital technology.

The future caught me by the shoulder last month when I received an e-mail from Freddie.

Normally, I don't have particularly well-developed foresight. After all, I've spent a career in a field that has had a remarkably dull narrative of progress. Each year, hardware gets smaller, cheaper, faster. In the same period, software has become more comprehensive and easier to use. Occasionally, the story is punctured by deadlines missed or budget overruns. Even if we add to this mix the free-floating bits of malware and other problems, we still have a history that's remarkably triumphant and projects nothing but progress. Only when we look beyond our immediate circumstances do we get a hint of the forces that have shaped our industry and get the briefest glimpse of the future.

When I last saw Freddie, he was leaving town with his guitar in hand to begin a career in popular music. I predicted that he would likely make his name as a producer, arranger, or composer. He possessed the skills to handle any of these careers and also knew that he was unlikely to become prominent as a guitarist. Though guitars had once been the dominant instrument of popular music, they were no longer the center of most bands. A guitarist, no matter how skilled, could no longer expect the kind of attention given to the likes of Jimi Hendrix or Jimmy Page. No, the power in popular music was clearly in the hands of those who understood music theory and could make deals with capital.

Not long after his departure, Freddie surprised me with a photograph of his band playing on the stage of an influential music club in New York. It was a moment to be proud, as the club had been the original showcase for the punk and New Wave musicians of the early 1980s. It also evoked a bit of sadness, as Freddie was one of the last musicians to play in that building. The club had been purchased by a real estate developer and was about to be converted into a retail outlet for well-financed children with discriminating tastes.

EMPLOYMENT PATTERNS

While students can be a difficult group to track, they tend to follow a predictable pattern. I heard nothing from Freddie for three or four years. During that period, he fell in love with a young woman, married her, and fathered a child. When he finally resumed communications last summer, he gave me a quick summary of his life that included the fact that he was no longer working as a professional musician.

"What are you doing for a job?" I asked.

"Technology," was his quick reply. "I'm doing IT for a hedge fund. Perl. Java. Configuration management."

I then asked him how he had prepared for this career, as I recalled that he had spent more time in music studios than in computer laboratories. Freddie reminded me that he had originally taken a job managing a small network and had moved, step by step, to jobs with more sophisticated systems. "I did some studying on my own," he explained, "and learned from the work I did."

Freddie talked about his employer and the kinds of challenges he faced. He described situations that are all too familiar to anyone who has worked with software during the past 30 years: complicated interfaces, incompatible systems, unrealistic deadlines, inflated expectations.

As Freddie told his stories, I recognized that he might find some of these issues easier to address if he had additional training, access to best practices, or even more contact with peers who did the same kind of work. When I asked him who the senior technical staffer on his project was, Freddie paused.

Published by the IEEE Computer Society

"There really isn't one," he said.

"They all have the same kind of training as you?"

"Yes, but none of the others play guitar."

Depending on how his time as a professional but largely unpaid musician is viewed, Freddie has been in the workforce for a little more than a decade. According to the US Bureau of Labor Statistics, Freddie occupies one of the 855,000 jobs for computing professionals that has been created during that period. Those new jobs provide both an opportunity and challenge for the profession. As an opportunity, they demonstrate the world's need of technical skills. As a challenge, they demand that we recognize the forces that are changing the demography of our field, changes that are bringing new kinds of people into professional roles.

RAPID CHANGES

The demography of technological jobs isn't well understood, as the standard job categories for the field have changed rapidly over the past 30 years. The Standard Occupation Classifications for 1980, the categories that are used to track the labor force, included two job titles under the section Computer, Mathematical, and Operations Occupations. The first is "computer systems analyst." The second is "computer scientist, not elsewhere classified." The classifications didn't even include a category labeled "computer engineer." Such a job was assumed to be a position for specialized electrical engineers.

Of course, 30 years in the field of digital technology is a lifetime in any other industrial field. In 1980, the software industry had yet to be recognized by the business community. The term "software engineering" was just 12 years old and not widely used. Undergraduate programs in software engineering were few and far from standardized.

To understand why such categories are important, we only need to look at the demographic changes that occurred in agriculture during the 20th century. In 1900, fully 40 percent of the US workforce lived on farms and was engaged in agricultural production. A century later, that figured had dropped to 4 percent, and those workers provided food for a population that had increased by a factor of 3.5.

The farmers who left agriculture during the 20th century became part of the great movement of laborers who left the country and moved to the cities, where they became the new industrial workers of the era. It was known as the Prairie Brain Drain to the small cities of the west and as the Great Migration North to the sharecroppers of the south. "The major reason for the change," explained sociologist Daniel Bell, "was the huge increase in agricultural productivity during World War II and after, when the introduction of chemical fertilizers and pesticides raised agricultural productivity between 6 and 9 percent a year." Without these movements, the US would not have had enough labor to expand its industrial base during the 1920s and 1950s.

It isn't necessary to have a bachelor of science degree to be considered a software engineer.

Nothing in the world stands still. The US experienced a loss of manufacturing jobs after the 1960s. While some of this decline was caused by companies moving production outside the country, more was caused by the increasing productivity of manufacturing processes. The autoworker of 1970 could accomplish far more than the equivalent laborer of 1930. The workers released from employment in factories moved into the service industries, such as financial services, programming, and software engineering. They left Detroit and Akron and moved to Redmond, San Jose, and Waltham. Without this migration of workers, the US economy could not have expanded in the 1990s.

MIGRATING TOWARD THE MIDDLE

Having witnessed three major migrations of labor in the past century, we might wish to know if we are in the midst of a fourth. If we are seeing a new migration of technological workers, it is a migration to the middle, a decline in the number of people with either a great deal of training or almost no training, and an expansion of those who are modestly trained but still call themselves professionals

While growth in the computer occupations category during the past decade seems substantial, it isn't uniform across the different computing professions. The jobs that require both the most and the least training have declined during that time. The number of positions for computer engineers has declined by 5 percent, while the number of openings for computer operators has declined by 38 percent. This change isn't surprising. During the past decade, the tools and standards that support both kinds of jobs have increased, and some kinds of jobs, notably that of computer programmer, have moved to lower-cost labor markets.

The three kinds of positions that account for the 855,000 new jobs are squarely in the middle of the field: network system analysts, software engineers, and systems analysts. As defined by the Bureau of Labor Statistics, these fields all require substantial training beyond the basic skills of an operator but not the scientific education of a computer hardware engineer.

It isn't necessary to have a bachelor of science degree to be considered a software engineer. According to the Bureau of Labor Statistics, a software engineer is the leader of a programming or system development project, not necessarily a trained engineer. "Occupations are classified based on work performed," explains the Bureau of Labor Statistics. Only in some cases, notably the most restrictive professions, does it consider the "skills, education, and/or training needed to perform the work at a competent level."

Indeed, the number of students who complete a BS in software engineering has declined steadily during the past five years. It's now below the level of graduates in 2000. In many cases, employers have filled the software engineering jobs by importing trained engineers from other countries or by promoting individuals trained in different ways.

"Computer science as an academic discipline," writes the historian Nathan Ensmenger, "and computer programming as an occupation have struggled with various degrees of success to establish institutional boundaries" that identify them as professional occupations. They have been unable to establish common standards for admittance to the field. Those who can do the work, no matter how they may have been trained, can generally find work. Freddie, for example, doesn't have an engineering degree, yet the Bureau of Labor Statistics considers his job to be a software engineering position.

ALTERNATIVE TRAINING PROGRAMS

We shouldn't expect to fill our software engineering posts from the ranks of musicians. Yet we're seeing more and more individuals who are training for software development in programs that look less and less like computer science departments. Perhaps the most rapidly growing technical program is that of computer game design.

Game design programs have become popular at associate degree colleges, public and private, and are increasingly found in major educational institutions. They're filled with students who see themselves in the employ of Electronic Arts or some other game developer, living the rest of their natural lives in the fantasy land of *Donkey Kong* and *Halo*.

Fantasies are lovely things, of course, but they must do battle with the monsters of economic reality. More often than not, the fantasy falls before the law of supply and demand, and we discover that the wonderful world of Adam Smith allows no extra lives, no resurrections, and no secret commands to unlock special powers. Only a few programs can place all of their graduates with computer gaming companies. The rest go into more prosaic careers: Web design, network analysis, and, if they can do the work, software engineering.

If we are faced with a growing cohort of computing professionals who did not have a traditional professional education, then we must ultimately consider the question "Who do we consider to be a professional?" The answer is not straightforward. If we make the criteria too stringent, the Computer Society could evolve into a small organization of highly educated people who have little impact on the application of technology. If we make the definition too generous, we might find that we have a large body of members who don't have a common vocabulary that would allow them to discuss the issues relevant to the field. If we do nothing, we may watch the technical world march forward and leave us little to do.

Just one month ago, I saw plenty of evidence that the world was already looking elsewhere for technical information. While attending a technical conference in Silicon Valley, I discovered that most of attendees had more in common with Freddie than with me. During a coffee break, I had a lovely talk with a technology entrepreneur who had been the lead singer of a New York band. "Straight up rock 'n' roll," she explained, "a la Sheryl Crow or the Pretenders." As far as I could tell, I was the only member of IEEE or the ACM in attendance. Yet, as I listened to the talks and conversations, I could see that there was much the Computer Society could offer to these individuals.

"On a good night with the band," explained Freddie, "you were grateful to take home $50, if you could even get that. There's no future there." He sees enough future in software to cast his lot and the lot of his family with the field.

Part of that future seems clear to those of us who work in the computing profession. If past is prologue, software will become more complex and more inclusive. Its services will be available to us on small portable platforms, in giant infrastructures hidden in the cloud, and on everything in between. In all, the world of software will be faster, more interesting, more engaging.

Yet the future of software also includes more than a few people who earn their daily bread from technological work even though they once believed they could command the night as popular musicians. From our vantage point, it's difficult to discern who will create the next generation of software, how those workers will be trained, and what organizations will support their work. Yet we must learn to answer those questions if we are to stay relevant to the field, if we are to look ahead to anything more than the life of a struggling artist who can't earn more than $50 from an evening's gig. C

***David Alan Grier**, an associate professor of international science and technology policy at George Washington University, is the author of* Too Soon to Tell *(IEEE CS Press, 2009). Grier blogs at computer.org/theknownworld. The statistics for this column come from the Occupational Employment Survey of the Bureau of Labor Statistics and from the American Society for Engineering Education. Contact him at grier@computer.org.*

32 & 16 YEARS AGO

JANUARY 1979

PRESIDENT'S MESSAGE (p. 3) "It is estimated that the membership of the IEEE Computer Society will approach 40,000 by January 1979—an increase of about 25 percent over the past year. The Computer Society membership is now about one-fifth of that of the IEEE, by far the largest among the 30-odd societies/groups of the Institute. …"

SOFTWARE MANAGEMENT (p. 6) "SCM [Software Configuration Management] is intended to fill a void in the practice of managing software development projects. SCM does not differ substantially from the CM of hardware-oriented systems, which is generally well understood and effectively practiced. However, attempts to implement SCM have often failed because the particulars of SCM do not follow by direct analogy from the particulars of hardware CM and because SCM is an immature discipline that needs to be brought closer to maturity."

CHARGE COUPLED DEVICES (p. 16) "Although the CCD is a relatively new application of semiconductor technology, the CCD fabrication process is very similar to the standard silicon-gate process used to build *n*-channel dynamic MOS RAMs. New CCD structures also allow the cost-effective implementation of shift-register storage. CCDs are intended to fill the performance gap between high-speed RAMs and magnetic disks. … In addition to having shorter access times than disks, CCDs allow users and designers to take a modular approach to storage, where the price per bit is insensitive to capacity. …"

OPTICAL COMPUTING (p. 23) "Optical computing in the broadest sense is the acquisition and (or) manipulation of information by electromagnetic or acoustic waves or rays. It has been the subject of six conferences since 1972 and has spawned numerous special issues, survey papers, and books. The more important applications of optical computing are those in which information is manipulated or processed (rather than simply acquired) by an optical wavefront in which the input source is coherent laser light. We can achieve several basic operations in such an optical processor, and they have a number of applications in image and signal processing. Hybrid optical/digital processors are also of interest, since they are used in the final embodiment of any optical processor."

CONCURRENCY (p. 42) "Computing systems based on a data-flow organization of computation are presently under development in laboratories in the US and Europe. These systems are a fundamentally new style of concurrent (tightly coupled, distributed) computer, which could eventually supersede the conventional general-purpose (von Neumann) computer. Conventional computers and programs, even those capable of concurrent activity, are implicitly sequential and require any desired concurrency to be indicated explicitly. They are *control-flow* computers, and they can be contrasted with *data-flow* computers, which allow concurrent operations to be activated as soon as their input data are available."

HANDICAPPED EATING (p. 54) "Computer control of manipulators and machine tools is a highly developed technology; however, where space, cost, and power are limited, there is a continuing demand for simple but effective digital control systems capable of modest accuracy. We have developed a versatile system which permits severely handicapped individuals to eat independently, using a microcomputer-controlled manipulator. We … used a 'learning phase' to establish the point-to-point coordinates of the motion, regardless of the coordinate system in which the manipulator resides."

USER MANUALS (p. 72) "'Damn you! Damn you! Damn you! I spent $100 on your user manuals and I can't understand a thing they say!' So said an irate university student recently in a letter to a mainframe computer manufacturer. Such criticisms, while perhaps not often so vehement, are not at all uncommon. And yet, we can produce good manuals—if we carefully analyze the problems users are having. I propose three strategies: (1) reorganize the manual content and format; (2) create a new role for manual writers and the company organizations supporting them; and (3) speed up manual production and distribution. Each strategy solves a specific set of user problems."

FOOD PRODUCTION (p. 84) "Space-age technology offers a promising new way to monitor world food production, concluded scientists associated with the Large Area Crop Inventory Experiment at a symposium held at the Johnson Space Center last October. With modern computers and communication facilities, LACIE can gather timely weather data from all over the globe and use them to identify growing conditions in each area. More importantly, computers allow the large amount of data processing necessary to develop yield models that use these weather data to quantify the impact of climatic fluctuations on important food crops."

Editor: Neville Holmes; neville.holmes@utas.edu.au

0018-9162/11/$26.00 © 2011 IEEE

32 & 16 YEARS AGO

JANUARY 1995

PRESIDENT'S MESSAGE (p. 5) "Although our goal is to become more member- and customer-oriented, we must also consider several changes that are occurring in our field, our organization, and society at large. To address these changes, the plan has four central aims: acquire a global perspective in our activities, disseminate technical information with new electronic methods, keep pace with the evolution of our field, and better integrate our programs with member needs."

EDITOR IN CHIEF'S MESSAGE (p. 6) "… We must help our readers acquire and maintain critical job skills—and learn about the products and tools—that help them compete in today's environment. Our content must address the industry's shift in emphasis from hardware to software engineering and computer communications. And in addition to the traditional peer-reviewed material that has always been our core content, we must pursue and develop articles that are shorter, more current, and application-related. …"

INTEL'S ERROR (p. 9) "The mistake in the Pentium's floating-point unit that caused computational errors in high-precision mathematics will affect only a few academic users, Intel insists. However, many in the industry were livid at the chipmaker for taking so long to admit the error, and many professionals, such as NASA scientists and aeronautical engineers, questioned whether high-precision work they'd done on the chip would need to be discarded due to the 'subtle flaw'."

SOFTWARE ENGINEERING (p. 11) "Many government agencies, led by the Department of Defense, are assertive in demanding better software development within their own organizations and from private industry. The most notable effort concerns the five-level Capability Maturity Model (CMM) developed for the government by the Software Engineering Institute. This model includes procedures for 'assessments' and the somewhat controversial 'evaluations.'"

MULTIPROCESSING (p. 40) "… Adaptive parallelism (AP) allows dynamic processor sets—the number of processors working on a computation may vary, depending on availability. For example, a program that begins executing on 64 processors (nodes) may finish on three, having executed on 65, 79 and 15 processors in the meantime. Even if a computation begins and ends on 64 processors, the 64 on which it ends are not necessarily the 64 on which it began. An AP program may be descheduled entirely during its run yet remain within the bounds of the AP model, because the number of available processors during a computation may go to zero."

VIDEOCONFERENCING (p. 77) "The Internet infrastructure is beginning to support videoconferencing applications in several ways. First, the emerging multicast backbone (or MBone) can efficiently send traffic from a single source over the network to multiple recipients. At the same time, many workstations attached to the Internet are being equipped with video capture and sound cards to send and receive video and audio data streams."

PRODUCT DESIGN (p. 81) "There is relentless pressure on designers to reduce the amount of time it takes to get new products to market. Gone are the days when hardware designers worked on their designs in isolation—and software designers went into high gear only after the hardware was built. Successful product design can now also involve more than one processor. If each processor requires a different development tool, designers can waste a lot of time learning how to use each one. …"

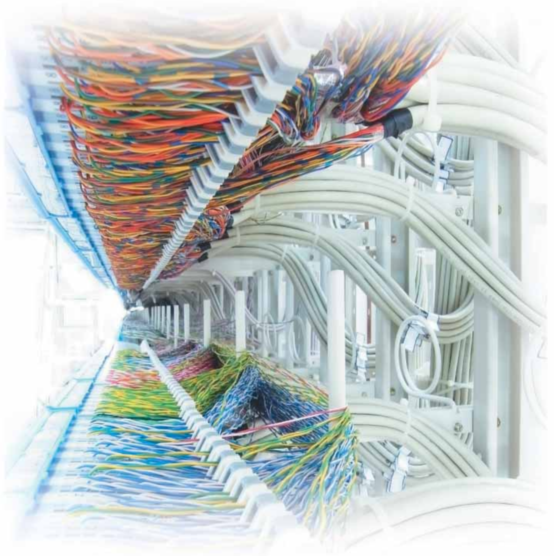
COMPUTING PATENTS (p. 93) "Reexamination is particularly important in the computer field, where more and more patents are granted and enforced as companies recognize their value. This growth in the number of applications has created an enormous challenge for the Patent Office to correctly determine which applications deserve a patent. The number of applications filed in Group 2300 (the Patent Office group that examines hardware and software computing technology) has more than doubled over the last six years, and the number of patents issued by that group has nearly doubled as well."

COMPUTER SCIENCE (p. 120) "Ever since the first computer science department was founded, debate has centered on whether computer science is in the same intellectual realm as biology or chemistry. From my perspective outside academia, CS is not a natural science, but more like engineering and mathematics.

"Computing has laws, but they're shallow, engineering-quality laws unworthy of an Important Science. … Strong and profound scientific laws are necessary if CS is to gain intellectual power to match computing's technological force. …"

PDFs of the articles and departments of the January issues of Computer *for 1979 and 1995 are available through the IEEE Computer Society's website: www.computer.org/computer.*

3D Displays without Glasses: Coming to a Screen near You



George Lawton

The most common 3D displays require users to wear special glasses, which has limited the technology's popularity. Now, researchers and vendors are working on glasses-free 3D displays.

A long-sought-after goal of the display industry is to give viewers the ability to watch still and moving images in 3D.

There are already 3D displays being sold with TVs and some laptops. However, the most cost-effective and widely available techniques rely on special glasses to create the 3D illusion, which is neither convenient nor optimal for viewers.

Many users don't like wearing the glasses or lose them easily, and they give some viewers headaches or eyestrain.

Now, though, researchers and vendors are making progress with autostereoscopic displays to generate 3D images without glasses.

Autostereoscopic technologies are just starting to show up in handheld cameras, camcorders, game devices, and phones by vendors such as Fuji and Nintendo.

The larger formats are being adapted for 3D billboards by companies including Alioscopy, Magnetic 3D, and Tridelity Display Solutions.

However, vendors are encountering challenges to making such displays technically and commercially viable for mass-market TVs.

"Although everyone would love to see a glasses-free solution, there are too many tradeoffs, primarily with image quality," said Chris Chinook, president of Insight Media, a market research firm.

"There is a growing body of evidence that 3D really has benefit," said Nick Holliman, senior lecturer at Durham University. His research found that people could do spatially related tasks—such as navigating a maze in a computer game—20 percent more accurately on a 3D screen.

However, 3D has a long way to go and numerous hurdles to clear before it can achieve commercial success.

BEGINNINGS

A person's eyes see the same scene from two slightly different angles. The human optical system combines the two images to yield an accurate 3D view of the scene. 3D displays work the same way.

The first autostereoscopic effect was incorporated into a painting by G.A. Bois-Clair in 1692 in France. He used the parallax-barrier technique, combining multiple images painted in alternating strips with a line of metal bars—arranged across the front of the painting—that blocked one set of strips from each eye.

Parallax-barrier techniques were applied to photographs in 1903 and movies in 1941.

In the 1920s, researchers began investigating an approach using a lenticular lens array—a group of lenses designed so that when viewed from slightly different angles by a person's eyes, they magnify different images—arranged across a display.

Work began on electronic autostereoscopic displays in the 1990s, said Durham University's Holliman.

The biggest-selling commercial 3D device has been the parallax-barrier-based Sharp SH 251iS mobile phone, which was introduced in 2000 and has sold about 2 million units, he estimated.

0018-9162/11/$26.00 © 2011 IEEE

AUTOSTEREOSCOPIC APPROACHES

There are two broad classes of autostereoscopic display: *multiview* and *light field*.

A multiview display uses optics to render two or more views of a scene, which the user's optical system fuses into a single 3D image.

Light-field displays, on the other hand, recreate the actual pattern of light—including its direction and angle of arrival—coming from all parts of a 3D object. This yields a better sense of depth.

Multiview displays

Multiview techniques render vertical slices of a single image on a 2D display.

Techniques. Parallax barriers, formed with either wires or a layer of material containing a series of carefully placed slits, block one set of views from each eye.

Lenticular lenses direct the light from one set of views to one eye and the other set to the other eye.

3D systems that use glasses rely on the multiview principle.

In one example of this approach, an image consists of two versions of the same scene, in different colors and spaced slightly apart. Glasses with colored lenses filter out one version of the scene to one eye and the other version to the other eye. The viewer's optical system then fuses them into a single 3D image.

The once-popular stereoscopes, invented in 1838, consisted of two lenses and cards with photos of the same image from slightly different angles. Viewers looking through the lenses at a card saw the image in 3D.

Multilight approaches use specially placed lights and prisms that alternately direct the light from multiple views of an object to each eye.

For this technique, Microsoft has developed an improved wedge lens and the ability to track a user's eyes so that the system can better target the light at one or more viewers.

Implementations. Sharp pioneered an LCD-based parallax-barrier filter that can be dynamically adjusted for different viewing angles.

When the filter is switched on, vertical elements turn opaque, occluding alternating lines of the screen from each eye.

A Hammacher Schlemmer 3D camera, the Nintendo 3DS game console, and Tridelity monitors employ this technology.

Alioscopy, LG, and Magnetic 3D are using lenticular technology in large advertising displays.

3M has produced the first commercial multilight display, used in Fuji cameras and in a new Texas Instruments cell-phone module.

Light-field techniques

Light-field displays generate pixel-like elements that transmit image-brightness data and information about the direction and angle that light rays would travel from an object to a viewer's eyes. This makes it easier for the optical system to interpret the image properly.

In contrast, multiview displays generate pixels with brightness but not angular information.

Light-field displays thus better enable the accurate convergence necessary for effective 3D viewing, said Insight Media's Chinook. They also reduce eyestrain.

In addition, the displays deliver more accurate depth information, which is important in medical- and scientific-visualization applications, said University of Valencia professor Manuel Martinez.

Integral imaging. The integral-imaging approach to autosteroscopy uses an array of lenses in front of a single camera or multiple cameras to capture the same scenes from slightly different perspectives. A special lens combines these images to generate a 3D image consisting of an array of viewing elements that transmit brightness and viewing-angle information.

This creates a 3D image in both vertical and horizontal planes, which makes it easier to maintain the effect for viewers of different heights or as a viewer stands up or sits down.

Multiview techniques create 3D images only in vertical planes.

Because integral-imaging views are blurry, Martinez explained, high-quality commercial integral-imaging displays will require much higher resolution than they currently have.

Toshiba spokesperson Kaori Hiroki said the company plans to introduce a new line of TVs this year based on integral-imaging technology.

Volumetric. Volumetric light-field displays—pioneered by vendors such as Actuality, which Optics for Hire acquired in 2009—create an image using a mirror or LED array that is spinning or vibrating within a glass cylinder or sphere.

Light transmitting an image comes from the spinning or vibrating element before reaching the user. As the element moves, it generates a separate version of the 3D image for each viewing angle. These views converge to create a 3D effect.

Sony has demonstrated a prototype volumetric display called Ray Modeler, which uses a rotating array of LED lights. This 3D display can produce color moving images viewable without glasses from any position around a cylinder.

A challenge for volumetric displays is the sheer amount of data required. Generating different images for each of the 360 degrees of viewing availability creates significant data-management requirements.

Guided light. Guided-light techniques use special holographic prisms that organize multiple high-end projectors' output into a single 3D image.

The Holografika HoloVizio, which Figure 1 shows, uses algorithms to determine how to recreate the light field from a 3D scene by directing each LCD projector to generate the appropriate pixel combination. The holographic grating on the front of the display combines the light from the separate projectors into a single 3D image for each viewer.

Holografika software developer Atilla Barsi said the display's projection technology and computer cluster are expensive. The company's display that measures 72 inches diagonally costs about $130,000, while the 32-inch display is about $13,000.

OBSTACLES

Because they entail complex projection and display technologies, autostereoscopic 3D systems cost substantially more than traditional 3D displays.

Outside of movie theaters, 3D displays are new and untested, while autostereoscopic technology is even more immature. And currently, resolution and performance aren't optimal.

To render multiple viewing angles for a group of users, lenticular and barrier techniques lose resolution.

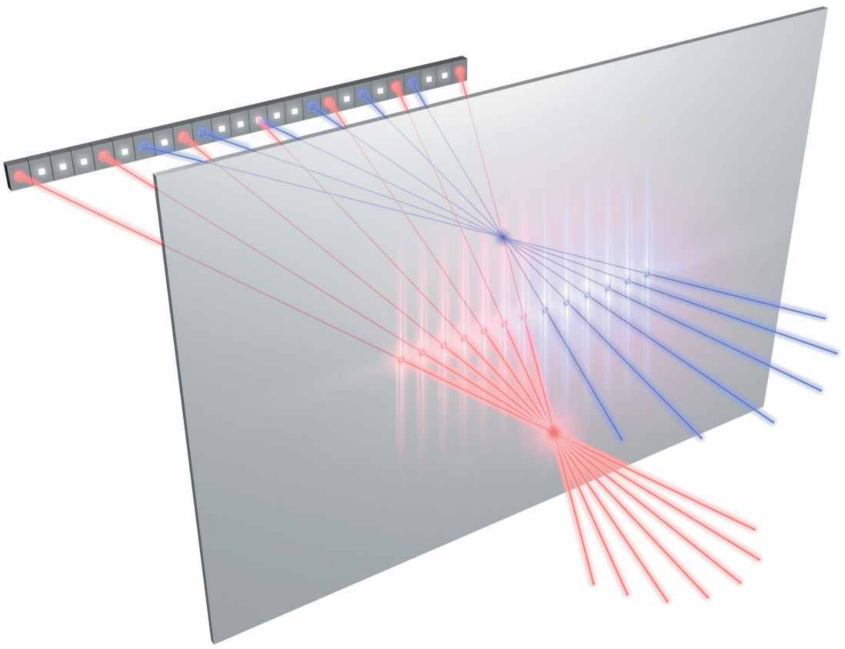
Moreover, they offer a limited number of viewing angles from which users can see 3D images, said Adrian Travis, senior scientist in Microsoft's Applied Sciences Group.

Andrew Jones, a research programmer at the University of Southern California's Institute for Creative Technologies, said the ICT is working on eye-tracking technology that could let viewers see 3D images from more vertical positions than are now possible.

Light-field techniques are considerably more expensive than multiview approaches because they require specialized optics and software tools, large amounts of data, and more computation.

Volumetric displays don't scale well because of the engineering challenges in building larger spinning elements that offer the necessary precision, said Jones.

The optics required cause guided-light displays to be large and expensive.



Source: Holografika

Figure 1. The Holografika HoloVizio display uses multiple projectors and other technologies to recreate the light patterns from a scene so that viewers see a 3D image.

Autostereoscopic 3D displays could prove helpful in teleconferencing, according to Jones. Expensive 3D displays could be particularly useful in high-end data-visualization and medical applications.

Durham University's Holliman said he doesn't expect autostereoscopic displays to replace glasses-based systems in movie theaters in the near future because they aren't cost-effective for large-screen venues.

But, he said, they could become popular in the consumer market—in phones, games, and TVs—in the next 5 to 10 years as manufacturers refine the approach.

The technology will take off in portable devices—such as cameras, photo displays, and game systems—because viewers can more easily adjust their positions to maintain the 3D effect, predicted Jennifer Colegrove, director of display technologies at market research firm DisplaySearch.

She also anticipated that 3D will be used increasingly in billboards and other advertising displays.

However, Colegrove said the majority of 3D displays will continue to require glasses because of autostereoscopic technology's higher costs and performance limitations.

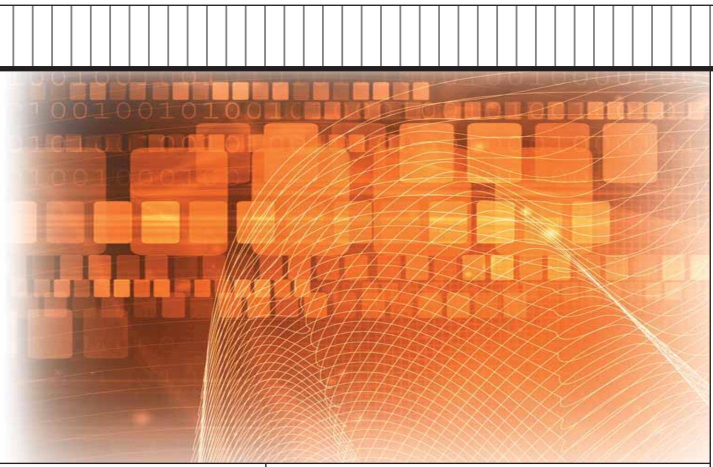
Between 2009 and 2018, DisplaySearch estimates, the overall worldwide 3D-display market will grow from $300 million to $22 billion, compared with an increase from $20 million to $2 billion for autostereoscopic displays. C

George Lawton is a freelance technology writer based in Guerneville, California. Contact him via his website at www.glawton.com.

Editor: Lee Garber, *Computer*; l.garber@computer.org

cn Selected CS articles and columns are available for free at http://ComputingNow.computer.org.

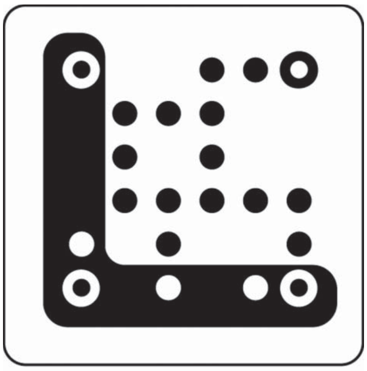
Scanning the Future with New Barcodes



Anyone looking at magazine advertisements, product packages, or even museum exhibits lately might notice that many of them have begun including a small square with an odd design inside.

These squares are 2D barcodes, which have symbols that represent data running horizontally and vertically. The 2D tags contain much more information and can perform many more functions than traditional one-dimensional universal product codes. UPCs consist of a series of thin and thick vertical lines and have been common on items such as product packaging since the mid-1970s.

2D barcodes typically store data such as a product's lot number and expiration date, and a URL for the manufacturer's website. They could also take a user to an advertiser's website. Someone looking at a movie poster could use a 2D barcode to go to an online trailer for the film.



Source: JAGTAG

Smartphone users take a photo of a JAGTAG, a type of 2D barcode, and send it via the handset to a JAGTAG server. Decoders translate the alphanumeric data the barcode represents. The system then consults a database to determine what the information means and sends it back to the user via multimedia messaging service.

Smartphone users photograph or scan a 2D barcode on an advertisement or product and text, e-mail, or send it via a format such as short message service. Users then connect either to a URL that takes them directly to a webpage with the desired information or to a clearinghouse that routes them to the barcode sponsor's server, which then sends them to the target webpage.

The 2D-barcode industry is still young with different levels of adoption by companies, demographic groups, and markets, said Microsoft Tag general manager Aaron Getz.

For example, the technology is widely embraced in Japan but still emerging in the US, noted Michael J. Liard, a research director with ABI Research, a market analysis firm.

The barcodes are most widely used for marketing, such as to enable customers to get coupons, obtain store or product information, or participate in promotional contests, according to Michael Becker, the Mobile Marketing Association's North America managing director.

A group of Harvard University business students conceptualized the first barcodes in 1932. The US Patent and Trademark Office awarded the first barcode patent in 1952.

A coalition of grocers and technology companies eventually standardized UPC technology, which began enabling product price scanning in 1974.

Denso Wave created 2D barcodes in 1994 for use in tracking vehicle parts through the manufacturing process. However, the tags have only recently begun widespread use in consumer settings.

2D barcodes use spaces, colors, and symbols—such as squares, dots, and triangles—to store information like letters, numbers, and punctuation, noted Ed Jordan, CEO of barcode vendor JAGTAG.

Direct-encoded symbologies provide the desired information, such as a URL, without users having to go to a server first. The barcode reader directly translates the symbols into the information.

Indirect symbologies store a short alphanumeric string that a reader decodes. The system then looks up the string's meaning in a database and returns the appropriate response.

There are about 20 types of 2D barcodes. Some have specialized applications, such as within the healthcare industry to identify stored blood types.

The tags generally contain symbols enclosed in a square or rectangle. The systems have various ways of making sure barcode readers can decode information in the correct order. For example, JAGTAGs do this via the four marked corners of its barcodes.

According to Mike Wehrs, CEO and president of barcode-system-provider Scanbuy, three nonproprietary 2D-barcode physical formats are most frequently used in the newly popular consumer applications. They differ in size, data capacity, and symbol type and configuration. Some have different capabilities such as error correction, which lets systems read the code even if the surface on which it appears has been damaged.

Quick Response encoding, a direct-encoded symbology that Denso Wave developed, is typically an inch square. A QR symbol can store a code with a maximum of 7,089 alphanumeric characters. The approach is commonly used in Japan, in applications such as manufacturing, logistics, and sales. There are several QR standards.

The direct-encoded *Data Matrix* barcodes vary in size from 2 millimeters to 14 inches square—the latter used on billboards—and can store up to 2,335 alphanumeric characters. Data Matrix barcodes are often used to track electronic and other types of components through the manufacturing process.

EZcode barcodes, an indirect approach that ETH Zurich developed, are a quarter-inch square, small enough to be attractive for use on product packaging and advertisements in publications. A scanner reads the barcode and uploads the code index to a server. The system consults a database and determines what the code means.

Numerous proprietary formats also exist including Microsoft Tags—released in May 2010—which use triangles and sometimes dots to represent data. The five-line color barcode is at least 0.75 inches square. The black-and-white versions are at least 0.875 inches square. Users can also create custom tags.

Users read the barcodes via a smartphone application. When decoded, the information points to a Microsoft server, which stores the represented information.

JAGTAGs are at least 0.75 inches square. Users photograph them with their smartphones and upload the image to a JAGTAG server. Decoders translate the tags, consult a database to determine what they mean, and send the desired information back to the user via multimedia messaging service.

Users typically utilize JAGTAGs to connect to websites to access many types of information such as weekly shopping coupons, sports highlights, and movie trailers.

According to Jordan, few phones have preinstalled barcode-scanning applications, which has slowed tag adoption. In the US, he noted, a minority of mobile-phone users even have smartphones. Moreover, the many types of 2D barcodes, the resulting market fragmentation, and the lack of widespread standardization are confusing to potential users, said Liard.

The lack of an established business model may also be holding back the market. Companies are still experimenting with various models, noted Jordan.

Nonetheless, said John Puterbaugh, founder and CEO of mobile-computing-services vendor Nellymoser, his company has studied the market and found that companies are interested in 2D barcodes.

As more consumers utilize smartphones and 2D barcodes, their use will increase, said Becker. And, Jordan added, companies will find new ways to employ them. C

Editor: Lee Garber, *Computer*; l.garber@computer.org

NEWS BRIEFS

Wireless Devices Provide Users with Mobile Wi-Fi Hotspots

The personal mobile hotspot—a relatively new approach that lets users obtain ubiquitous high-speed connectivity so that they can access the Web, e-mail, or other networked services wherever they are—is starting to become increasingly popular.

Unlike some other types of devices, the mobile hotspot lets users connect even if no traditional hotspot is available.

The devices let users of smartphones, laptops, iPods, tablets, gaming devices, or any other Wi-Fi-enabled mobile device access a cellular network to which they subscribe, noted Sprint spokesperson Caroline Semerdjian.

The mobile hotspots are basically Wi-Fi routers with wireless uplinks, according to Carl Howe, analyst in the Anywhere Consumer research group with the Yankee Group, a market research firm.

Traditional Wi-Fi hotspots, on the other hand, let users access services via a fixed router that links to a network with a wired connection. Personal mobile hotspots thus give users more flexibility by wirelessly connecting to a network.

Novatel Wireless introduced the technology via its credit-card-sized MiFi Intelligent Mobile Hotspot, according to company spokesperson Charlotte Rubin.

Other vendors, such as Cradlepoint, also make personal hotspots. And some newer Android phones contain embedded hotspots.

Phones connect via Wi-Fi to the mobile hotspots, which then link to the wireless network via cellular technology. Once connected, a phone can serve as an access point for other wireless devices. They could access the network simply by linking to the phone via Wi-Fi.

The mobile hotspots use various Wi-Fi technologies. For example, Sprint's devices use the IEEE 802.11b and 802.11g versions to communicate with the company's code-division multiple-access (CDMA) network, noted Semerdjian.

The battery-powered hotspots work via a Wi-Fi chip containing a transmitter, receiver, and antenna. They also have one or more antennas that let them connect to one or more types of cellular networks.

The devices provide communications-related services such as encryption, security, authentication, and GPS, Semerdjian noted.

They offer connections at the same speed as whatever Wi-Fi versions they work with. IEEE 802.11n is currently the fastest version, offering theoretical maximum data rates of 600 Mbps.

Being able to utilize one portable hotspot to provide any device with wireless connectivity can be more economical for a user than subscribing to separate data plans for each device.

But, according to Howe, carriers have mixed feelings about the devices. Some providers like them because they yield new income from a nonvoice service, he explained. However, he added, others prefer to focus their efforts on telephony, which is more lucrative.

As the hotspots get faster, they could replace wired broadband connections, particularly for consumers who want to consolidate their network-access services, he said.

Selling portable mobile hotspot services via monthly subscriptions may be difficult because users may not understand the technology's value or want yet another ongoing communications-related cost, according to Howe.

Also, if future smartphones and other machines include chips that provide mobile-hotspot capabilities, separate hotspot devices might become unnecessary.

Howe predicted the portable hotspots will be successful and will generate demand for connected radios, video players, game consoles, and other applications that would work via the devices.

*News Briefs written by **Linda Dailey Paulson**, a freelance technology writer based in Portland, Oregon. Contact her at ldpaulson@yahoo.com.*

ELECTRONIC PAINTBRUSH CAPTURES COLORS AND TEXTURES FROM OBJECTS FOR USE IN CREATING ART

Researchers have developed a high-tech brush that lets users create art by picking up images, video, audio, colors, and textures from objects and painting them onto a touch-screen canvas.

Experienced artists could use the device intuitively without having to work with typical computer-based interfaces such as mice, keyboards, or controllers, noted Stefan Marti, co-inventor of I/O Brush and a principal researcher and project leader at Samsung Research and Development.

I/O Brush looks like a typical paintbrush but contains a small camera that captures images, video, and audio. It also includes optical fibers among the bristles, a ring of LED lights for illumination, and four pressure sensors that measure the force the user exerts when painting. This determines the width and angle of the resulting brush stroke.

An embedded microelectromechanical inclinometer determines how the user rotates or angles the device. The system uses this information to create the types of strokes and textures a painter might utilize.

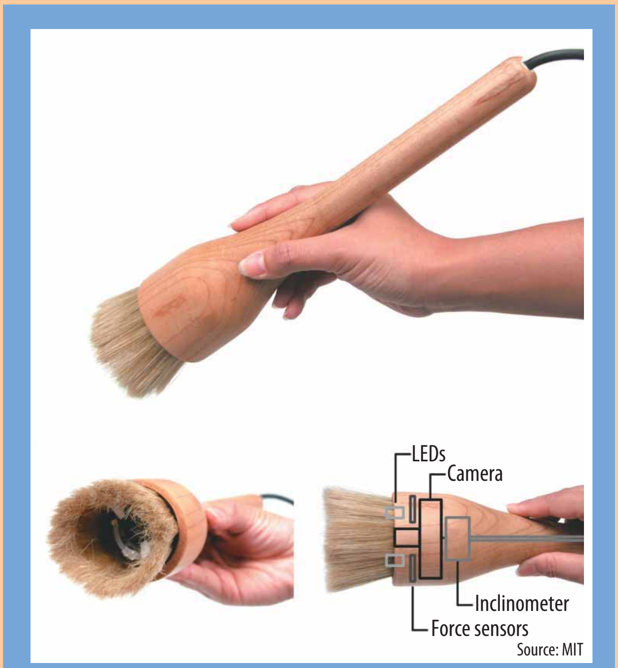
By applying the brush to the canvas in different ways, artists can distribute, sequence, rotate, and otherwise manipulate images, video, and audio as desired.

When touching an object, the brush captures and stores thumbnail images at about 16 frames per second. Upon touching the canvas—a large plasma, LCD, or rear-projection touch screen—the device lays down the thumbnails at 10 frames per second, overlapping them when necessary to create the desired image.

The I/O Brush can distinguish between touching an object to pick up image data and touching the canvas to apply the information. Algorithms identify when the brush is touching something but not moving, indicating that it's supposed to be acquiring images from an object; or when it is moving and thus should be placing images onto the canvas.

The resulting art pieces are touch-sensitive and interactive, and can play video thumbnails and audio related to the painting.

According to Marti, the researchers plan to create commercial consumer and professional, museum-quality I/O Brush versions. He said they have been in licensing and commercialization discussions with major software and hardware companies. C



Source: MIT

Researchers have built a high-tech brush that lets artists pick up images, video, audio, colors, and textures from objects and paint them onto a touch-screen canvas.

Automating a Building's Carbon Management

Geetha Thiagarajan, Venkatesh Sarangan, Ramasubramanian Suriyanarayanan, and Pragathichitra Sethuraman
TCS Innovation Laboratories, Chennai

Anand Sivasubramaniam
Pennsylvania State University

Avinash Yegyanarayanan
Tata Consultancy Services, Chennai

Buildings are the largest contributor to the world's carbon footprint, yet many building managers use only periodic audits to adjust resource consumption and carbon emission levels. The ECView framework leverages existing workflow systems to continually assess a building's carbon emissions in relation to daily weather, commuting and travel patterns, and changing government regulations.

The world is becoming increasingly more vigilant about energy use, and decision makers involved in resource allocation must consider environmentally beneficial, or *green*, solutions in managing systems. Energy consumption and carbon emissions are the two main concerns, and although the carbon footprint has several components, buildings appear to be the worst offenders in both categories. According to the US Green Building Council (www.usgbc.org), buildings account for nearly 72 percent of the US's electricity consumption and 39 percent of its carbon emissions.[1]

To determine a building's environmental impact and reduce its carbon footprint, managers must closely monitor the building's chief carbon contributors. At present, such monitoring consists of scheduling occasional audits to assess a building's resource consumption and emission levels and then basing any recommendations on the operational snapshot.

We believe this approach has serious deficiencies. A building's carbon footprint is the product of complex interplay among the building's structural and infrastructure characteristics, business processes and operational patterns, climate and weather dynamics, energy sources, workforce commute patterns, and government regulations. Because these disparate factors can change daily, any recommendation based on a snapshot will rapidly become invalid. A more effective approach is to continually track these influential factors and tune subsequent recommendations using a realistic portrait of energy use and carbon emissions.

When done manually, continual monitoring can be tedious, error-prone, and expensive, so it makes sense to use information technology (IT) for carbon management. Green solutions built around IT not only scale with building size, but they also keep pace with a building's operational dynamics. In addition, IT offers a way to encapsulate and repeat best practices in creating and applying green solutions, so even facilities with less experienced personnel can reap the benefits of expert carbon management. Finally, because IT has already permeated systems that contribute to an enterprise's carbon footprint, such as enterprise resource planning and workflow systems, those developing an IT system for carbon footprint management need not start from scratch. IT has a firm foundation in standardizing and securing distributed networked systems, which can be beneficial in building management.[2]

Recognizing the power of IT to facilitate carbon management, we developed ECView (Energy and Carbon View), an

Published by the IEEE Computer Society

0018-9162/11/$26.00 © 2011 IEEE

IT framework that assists managers in finding and maintaining solutions that reduce a building's carbon footprint. To test ECView's capabilities, we used it to continuously monitor and analyze the carbon footprint of a Tata Consultancy Services (TCS) office building in India over the course of a year. Using the insights our framework offered, we identified ways to reduce the TCS building's carbon footprint. Some of these strategies require zero capital expenditure.

FRAMEWORK FEATURES

ECView provides real-time carbon tracking, accounting, and asset management, and it supports a feature set that enables insights beyond what simple meter readings can provide.

Carbon tracking

ECView aims to transform building carbon management from periodic sampling to a real-time process that the facility manager can monitor and execute continuously. ECView starts by collecting data from sources such as building management and ERP systems and then applies analytic engines to process the data in role-based dashboards that facilitate a variety of insights valuable to decision makers. Each diverse functional unit—finance, sustainability, or facilities—has a different dashboard that contains real-time views of the data most relevant to that unit. ECView also generates curves that prioritize viable carbon abatement projects according to a metric the user chooses.

Asset management

To effectively manage a building's carbon emissions, an IT framework should track the health and performance of key infrastructure assets related to both the supply and demand sides of energy consumption. ECView performs all the required asset-keeping activities, such as benchmarking and tracking parameters related to the assets throughout their life cycle. In addition to these baseline functions, ECView can log the operational hours, outage durations, and maintenance history of key assets, as well as automatically raise alarms for scheduled preventive maintenance or expected forced outages and trigger appropriate workflows.

Multilevel monitoring support

Any IT solution should move beyond meters, integrating data from various sources and presenting it in a holistic fashion. The goal should be to provide managers with enough insight on carbon footprint contributors to make operational decisions that are more beneficial to the environment. The insights offered and analyses supported should not be determined solely by the amount of available instrumentation, and the system should be able to work around erroneous human inputs and faulty meter readings.

Two important characteristics of ECView differentiate it from traditional carbon management tools. First, the extent of analysis that traditional tools can support is closely tied to the available metering. In ECView, this dependency is minimal because the framework can support varying levels of facility metering through its internal analytical models. Second, ECView can analyze the carbon footprint at three different levels: resource, activity, and business process, leveraging the activity-based costing model[3] to apportion the resource-level carbon footprint to activities and business processes.

EMISSION MONITORING WITH METERS

The TCS building in our case study has a built-up area of 250,000 sq. ft. spread across five floors. It houses two departments, Infrastructure Services (IS) and Business Process Outsourcing (BPO), with roughly 3,000 cubicles, two datacenters (one for each department), and a cafeteria. The facility's peak energy demand is 3,000 kVA, and its average electricity consumption is roughly 1,012 MWh per month. An electric utility serves the facility; four in-house diesel generators, each rated at 1010 kVA, serve as backup.

> **To effectively manage a building's carbon emissions, an IT framework should track the health and performance of key infrastructure assets related to both the supply and demand sides of energy consumption.**

We monitored the three commonly accepted categories of carbon emissions, as specified in the *Greenhouse Gas Protocol on Corporate Accounting and Reporting Standards 2004,* (GHG protocol; www.ghgprotocol.org):

- *Scope 1.* Emissions from activities that a building controls directly, including emissions from in-house fuel combustion, refrigerant leakage, and fire extinguishers.
- *Scope 2.* Emissions from the utility company's generation of the electricity that a building consumes.
- *Scope 3.* Emissions from wastes, water, employee commuting, and business travel.

As these categories show, the carbon footprint has complex components that stem from both the building activities directly and from activities such as work-related travel, which contributes to carbon emissions indirectly from the vehicles used for commuting and business travel.

Our case study considered these three categories, focusing on emissions from in-house power plants, grid electricity consumption, and business travel and employee commuting—sources that cumulatively account for more than 95 percent of a service-sector office building's carbon footprint. In-house power plants (such as diesel generators) and grid electricity impact a building's energy bills, so they are interesting to track from a monetary cost perspective as well.

Scope 1 and 2 emissions

To track Scope 1 and Scope 2 emissions, ECView starts with readings from meters that individually track the electricity that the plant consumes from the utility company and in-house diesel generators. From this data and localized emission factors that are based on the utility's source mix, ECView arrives at the carbon footprint. A utility company can generate power from a mix of sources, including thermal, hydroelectric, nuclear, wind, and solar power. Each source emits different amounts of carbon during electricity generation; hence it is important to consider the utility company's source mix.

Figure 1 shows the TCS building's monthly consumption and aggregate Scope 1 and 2 emissions during the monitoring year.

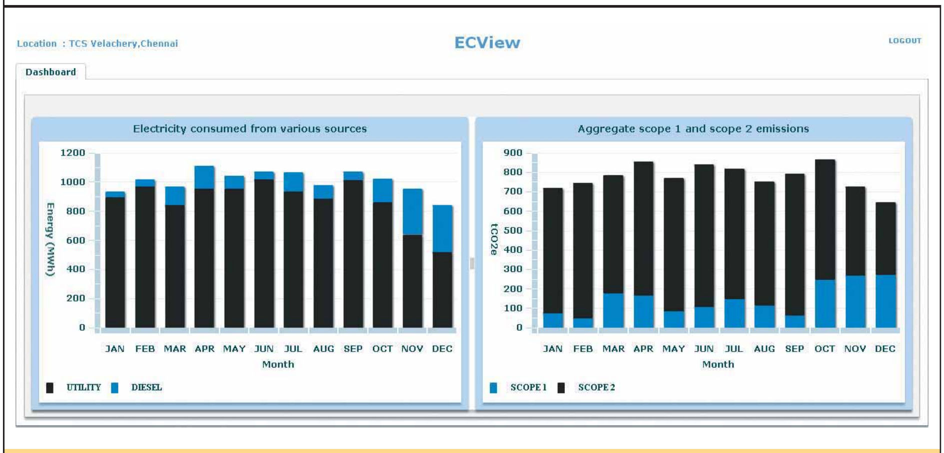

Location : TCS Velachery,Chennai
ECView
LOGOUT
Dashboard
Electricity consumed from various sources
Energy (MWh)
1200
1000
800
600
400
200
0
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
Month
■ UTILITY ■ DIESEL
Aggregate scope 1 and scope 2 emissions
tCO2e
900
800
700
600
500
400
300
200
100
0
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
Month
■ SCOPE 1 ■ SCOPE 2

Figure 1. Monthly electricity consumption and aggregate Scope 1 and 2 emissions for the TCS building. In the second graph, Scope 1 and 2 emissions are measured in tonnes of carbon-dioxide equivalent (TCO2e).

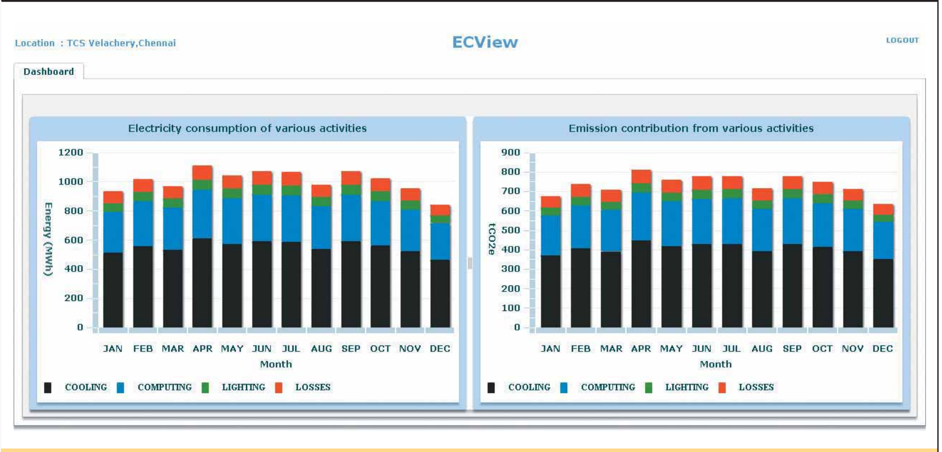

Location : TCS Velachery,Chennai
ECView
LOGOUT
Dashboard
Electricity consumption of various activities
Energy (MWh)
1200
1000
800
600
400
200
0
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
Month
■ COOLING ■ COMPUTING ■ LIGHTING ■ LOSSES
Emission contribution from various activities
tCO2e
900
800
700
600
500
400
300
200
100
0
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
Month
■ COOLING ■ COMPUTING ■ LIGHTING ■ LOSSES

Figure 2. Electricity consumption and carbon footprint by activity for the TCS building. The consumption breakdown is based on data from activity meters, but ECView can estimate consumption by activity even without such meters. Clearly, cooling is the most significant contributor, in large part because of India's climate. Losses refer to energy leaks and waste within the power distribution network.

ECView revealed that the facility sources around 87 percent of its electricity from the utility company and 13 percent from diesel generators. Of all the Scope 1 and 2 emissions, about 81 percent come from the utility company; 14 percent come from the diesel generators, and the remaining 5 percent come from liquified petroleum gas consumption (for cafeteria cooking) and refrigerant leakages (from chillers). Diesel consumption increases notably beginning in the tenth month of the study because, in that month, the utility company's regulations changed to prohibit industrial customers from drawing power between 6:00 pm and 10:00 pm. During such times, the facility met its electricity needs through the diesel generators, as evidenced by corresponding diesel consumption upswings.

Although knowing the facility's consumption at the resource level gave us a good idea about Scope 1 and 2 emissions, we still had several questions that resource-level monitoring could not answer: Which building activity consumes the most electricity? How much does each business unit contribute toward the building's carbon footprint? What measures can offer the highest footprint reduction for the investment?

To explore the answers to these questions, we used ECView to monitor the TCS building's consumption at a finer granularity. The building has meters that individually track the electricity that various activities consume, such as lighting, computing, and cooling. We fed readings from these activity-level meters into ECView, which then gave us an activity-oriented breakdown of the facility's electricity consumption and footprint.

As Figure 2 shows, cooling is the largest contributor, consuming 55 percent of the total electricity. This finding is understandable, given that the building is located in a tropical climate that is largely hot and humid. Computing and lighting consume 30 and 6.1 percent, respectively, and losses from equipment, distribution, and operations account for 8.7 percent.

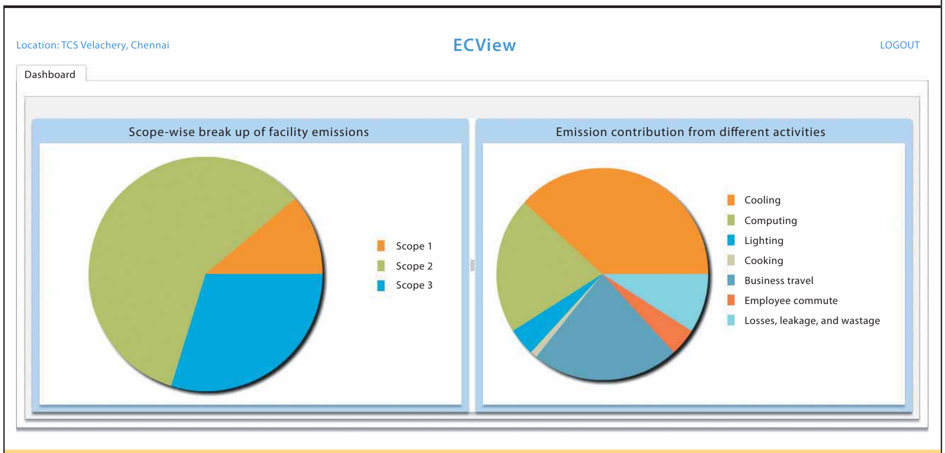

Location: TCS Velachery, Chennai
ECView
LOGOUT
Dashboard
Scope-wise break up of facility emissions
Scope 1
Scope 2
Scope 3
Emission contribution from different activities
Cooling
Computing
Lighting
Cooking
Business travel
Employee commute
Losses, leakage, and wastage

Figure 3. Carbon emission breakdown across emission type and activity. Travel was a significant contributor, although daily commuting was less influential, in part because employees tend to use public transportation and fuel-efficient motorcycles.

Scope 3 emissions

According to the GHG protocol, the primary contributors for Scope 3 emissions in a service sector office building are business travel and daily commutes made by the employees. Typically, business travel requests are raised, approved, and reimbursed in an organization through enterprise workflow systems. As soon as the business travel workflow is completed, data pertaining to the travel is automatically extracted and sent to ECView. Employee commute data is collected through a Web-based questionnaire integrated with ECView.

Figure 3 shows the breakdown of the facility's overall annual carbon footprint across activities. We gained several insights from these results. One is, again, that cooling is the most significant footprint contributor. Another is that, despite the GHG protocol's recommendations, reporting Scope 3 emissions should *not* be optional, particularly for service sectors. As Figure 3 shows, business travel was a significant contributor. Although daily commuting did not contribute that much to the footprint, this finding might be specific to the TCS building, where employees mainly commute using public transportation and motorcycles with a 140-mpg fuel efficiency. The commuting scenario could be quite different in other locations.

TRACKING PER-ACTIVITY CONSUMPTION WITHOUT METERS

A typical service-sector office building consumes electricity for lighting, computing, and cooling, but not all buildings have meters that track electricity consumption for each activity individually. For these buildings, ECView estimates the energy consumed from various sources that relate to the activity. For lighting consumption, it uses the building design's watts-per-square-foot value, the design deviation factor, and the building's lighting operational pattern. For computing consumption, it uses desktop or server specifications, personnel count, utilization percentage, and operational patterns. Finally, for cooling consumption, it uses heat-gain equations based on the building's structural details, personnel count, internal load specifications, operational patterns, and local weather characteristics.

Comparing our actual case study results to ECView's estimations, we found an average error of 4.07 percent between actual and estimated values across all activities for the year, with minimum and maximum errors of 1.22 and 18.3 percent. We believe that these findings are close enough for ECView's practical use as a consumption estimator.

ESTIMATING DATACENTER CONSUMPTION

Because modern office buildings typically house datacenters, it is rapidly becoming essential for enterprises to monitor their datacenters' energy efficiency. As The Green Grid specifies (www.thegreengrid.org), the metric for datacenter efficiency is power usage effectiveness (PUE)—the ratio of total power entering a datacenter to the power required to support the IT infrastructure within the center. PUE is always greater than or equal to 1.0; the lower the value, the higher the efficiency.

A datacenter consumes power not only for the IT infrastructure but also to support equipment such as heating, ventilation, and air conditioning (HVAC) units; lighting; power distribution units; and uninterruptible power supplies (UPSs). Although a facility might be able to measure IT power directly from UPS panels, other devices might not lend themselves to such direct measuring. If the datacenter is inside an office building, for example (as in our case study), the chillers and air-handling units

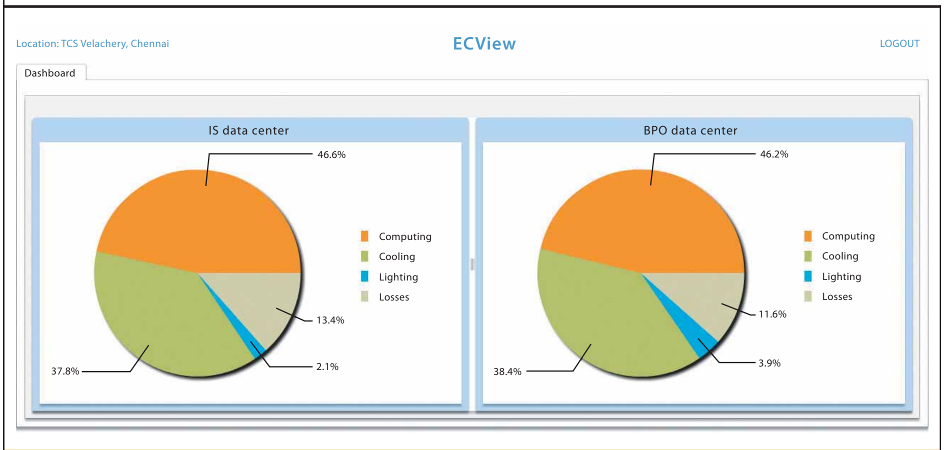


Figure 4. Breakdown of datacenter electricity consumption according to activity. ECView can apportion a datacenter's consumption even though a direct measure is for the building overall. In this view, the results reveal that the IS datacenter's power usage effectiveness is slightly better, in part because the IS datacenter has a lower proportion of lighting and cooling consumption.

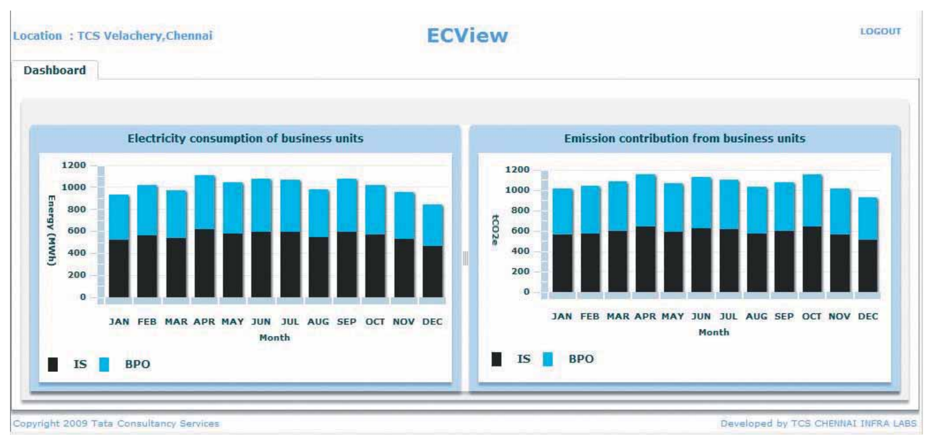

Location : TCS Velachery,Chennai
ECView
LOGOUT
Dashboard
Electricity consumption of business units
Energy (MWh)
1200
1000
800
600
400
200
0
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
Month
IS
BPO
Emission contribution from business units
tCO2e
1200
1000
800
600
400
200
0
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
Month
IS
BPO
Copyright 2009 Tata Consultancy Services
Developed by TCS CHENNAI INFRA LABS

Figure 5. Electricity consumption and overall carbon footprint in terms of business units in the TCS building. ECView uses models to apportion consumption to business units or processes, since most buildings do not have metering at that level.

cool the entire building, of which the datacenter is only one part. Consequently, the unit meters will not show the consumption from the datacenter alone. In this scenario, ECView estimates the energy consumed in cooling the datacenter.

Figure 4 shows the electricity consumption for the two datacenters in the TCS building. ECView obtained operational data on the cooling assets from the building management system (BMS)—a centralized controller for air conditioning, access, fire alarms, and so on, which most nonresidential buildings have. Using this data along with the datacenters' structural details and equipment characteristics, ECView estimated the cooling energy consumed through mathematical models and arrived at the PUE for each of the two TCS datacenters: the PUEs of 2.14 for the IS datacenter and 2.16 for the BPO datacenter show that the IS datacenter is marginally more efficient.

PROCESS-LEVEL APPORTIONING

In addition to providing perspectives at the resource and activity levels, ECView can show energy consumption at the process level or by business unit. Generally, a building will contain several enterprise processes, and knowing which process or unit contributes what percentage to the building's carbon footprint can be extremely valuable in channeling emission abatement efforts. Each process becomes aware of its own footprint, which can lead to individually customized abatement strategies. Such insights can also serve as input to chargeback models for shared facilities.

Apportioning footprint and consumption at the process level is not a straightforward task, however, since most buildings have no metering at this level. ECView uses activity-level data as well as process-specific parameters to apportion the building's carbon footprint across processes. For example, the apportionment of energy consumed for lighting is based on design watts per square foot, area occupied by the business unit, and personnel count. The apportionment of desktop computing energy is based on personnel count, and the apportionment of energy required for cooling is based on heat-gain models. Figure 5 shows the individual energy and carbon footprints for the two processes in the TCS building.

From this data, we inferred that the building's annual electricity consumption per person is 4.25 MWh—for IS and BPO, per-person consumption is 4.78 MWh and 3.73 MWh, respectively. Despite having more people and desktops and a 24/7 operation, BPO's per capita footprint is lower than that of IS. After further investigation, we found that the IS datacenter was consuming about 93 kW; the BPO datacenter, roughly 23 kW. A study of IS's datacenter revealed that some servers were energy guzzlers even though they had a low utilization rate. This finding suggested a need for virtualization and consolidation. When considered along with the results of earlier datacenter PUE studies, we concluded that a low PUE does not automatically translate to efficient operation. Because the PUE does not completely reflect efficient power use, we recommend adding metrics that tie the use of IT asset utilization to datacenter power consumption.

The annual per capita carbon footprint for the building is 4.49 tonnes of carbon-dioxide equivalent (TCO2e), a standard metric for measuring greenhouse gas emissions; for the IS and BPO departments, it is 5.05 TCO2e and 3.94 TCO2e, respectively. IS's per capita footprint is roughly 20 percent higher than BPO's because IS personnel travel much more often and IS's overall electricity consumption is higher than BPO's. This observation prompted a suggestion to change the travel policy to reduce the TCO2e attributable to travel.

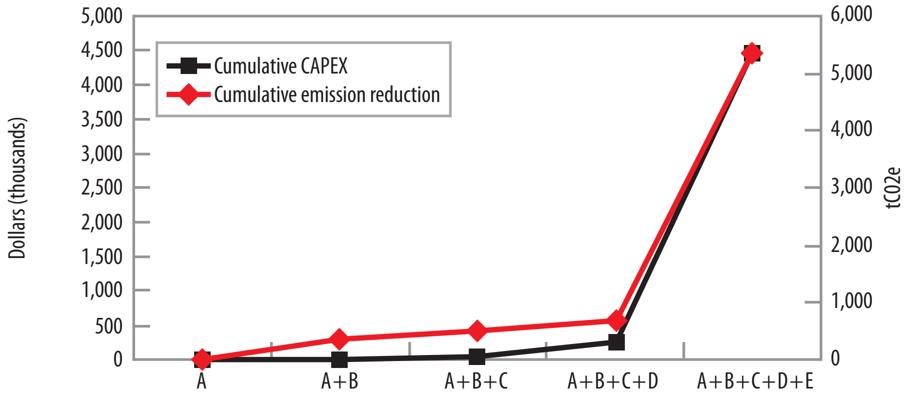


Figure 6. Carbon abatement measures using capital expenditure as the prioritizing index. Some of these measures, such as Option B, require no capital expenditure, yet they result in as much abatement as Option C, which requires a $25,000 capital expenditure. Option A represents the existing status (no abatement).

DECISION-MAKING SUPPORT

ECView has several features that aid decision making related to managing a building's carbon footprint, including the support for what-if carbon studies, an exhaustive database of carbon abatement measures, and the ability to chart optimal power purchases.

What-if carbon studies

ECView deepens the understanding of how business decisions and future actions affect a building's carbon footprint. Using the software's mathematical models to simulate hypothetical scenarios, decision makers can examine options related to the building's structural details and operational patterns. ECView does not restrict users to a menu of what-if options, which makes it possible to customize studies to clearly show the correlation between building operations and the carbon footprint.

For example, suppose the IS datacenter intends to expand its operations by adding 100 servers and that the servers' utilization percentage will not change. The manager might then want to explore how the additional servers would affect the building's annual electricity consumption and carbon footprint.

When we evaluated this what-if scenario with ECView, we found that improperly positioning the new servers would create hot spots that would greatly affect cooling and computing and cause the HVAC system to consume disproportionate energy amounts. ECView proposed a server placement strategy that would not create hot spots and showed that the building's annual energy consumption and carbon footprint would increase by 876 MWh (7.2 percent) and 641.5 TCO2e (5.01 percent), respectively.

Carbon abatement measures

From ECView's database of carbon abatement measures, users can customize an abatement strategy on the basis of the facility's assets, resource consumption, and operational pattern. Each measure has quantified values on the investment required, the carbon footprint reduction, and the payback period. Abatement curves prioritize measures according to a user-specified index.

Figure 6 shows curves that compare five abatement options in terms of capital expenditure. Option A is the as-is facility status with no abatement projects. Option B configures office desktops to hibernate during sustained inactive periods. Option C uses energy savers for lighting fixtures. Option D replaces the existing reciprocating chillers with absorption chillers. Finally, Option E purchases a dedicated windmill to provide the building with a green energy source. As the figure shows, a facility need not always spend a great deal for abatement. Some measures, such as Option B, are pure policy decisions.

Managing supply-side carbon

Most of the features we have described highlight managing the carbon footprint by optimizing the resource *demand*. ECView can also help managers more effectively manage the carbon footprint by optimizing the *supply.* Taking into account a building's location and energy demand profile, prevailing regulations, and available conventional and green energy sources, ECView can suggest the optimal green energy sourcing plan for that building.

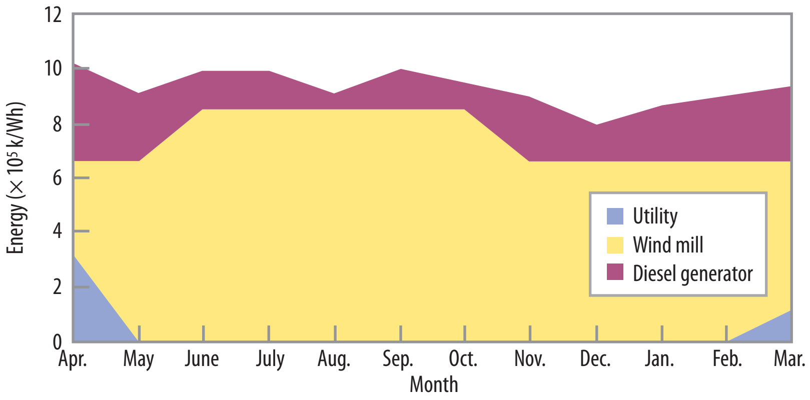


Figure 7. Optimal energy purchase plan for the TCS building in terms of cost. ECView ensures that any suggested plan satisfies availability and regulatory constraints. In this plan, a dedicated windmill provides most of the energy to the building.

Figure 7 shows the energy purchase plan suggested for the TCS building in our case study, which can reduce the facility's energy costs by 31 percent and its annual emissions by 59 percent.

Our case study of ECView shows that an IT framework can provide more insights into a building's carbon footprint than current intermittent energy audits. Even when a building does not have activity-level meters, ECView can estimate an activity's carbon use, proving that IT tools can fill the gaps and provide valuable insights. We have found, for example, that in a typical service-sector office building, HVAC is the largest individual carbon contributor followed by business travel.

ECView gives managers the freedom to explore carbon abatement measures and offers suggestions for optimizing power purchases. We have shown that in some cases carbon abatement strategies requiring zero investment could be as effective as those requiring a $25,000 investment. Armed with these insights and evidence that abatement options can be cost-effective, building managers have little reason to avoid green solutions for carbon management. C

References

1. US Department of Energy, "Buildings Energy Data Book 2009"; http://buildingsdatabook.eere.energy.gov.
2. D.F. Carr, "A New Place for IT," 16 Nov. 2009; www.informationweek.com.
3. H.T. Johnson and R.S. Kaplan, *Relevance Lost: The Rise and Fall of Management Accounting*, Harvard Business Press, 1991.

Geetha Thiagarajan *is a scientist at TCS Innovation Laboratories in Chennai. Her research interests include renewable and sustainable energy system analysis. Thiagarajan received a PhD in electrical engineering from the Indian Institute of Technology, Madras. Contact her at geetha1.t@tcs.com.*

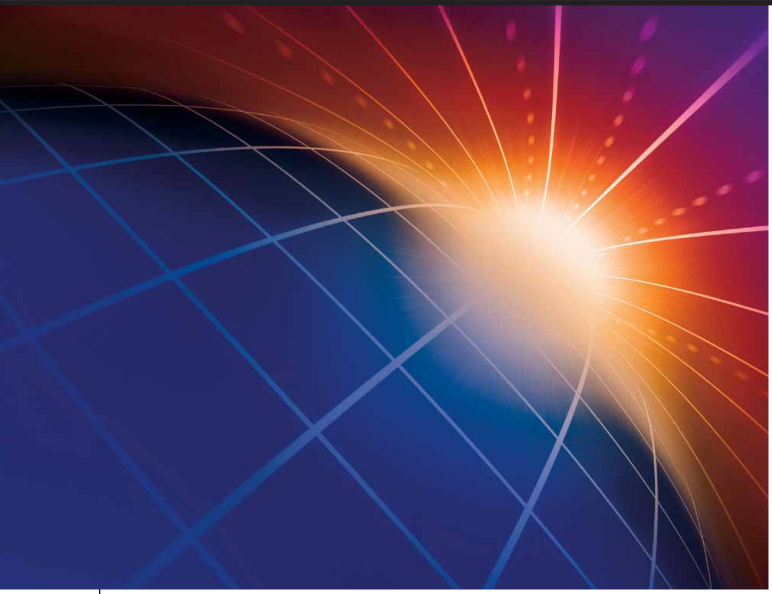
Venkatesh Sarangan *is a senior scientist at TCS Innovation Laboratories in Chennai. His research interests include computing and sustainability, radio-frequency ID systems, and wireless networking. Sarangan received a PhD in computer science and engineering from Pennsylvania State University. Contact him at venkatesh.sarangan@tcs.com.*

Ramasubramanian Suriyanarayanan *is a researcher at TCS Innovation Laboratories in Chennai. His research interests include developing IT-based solutions to help facilities, factories, and enterprises find economical green solutions to energy management. Suriyanarayanan received a BTech in computer science and engineering from SASTRA University, Tanjore. Contact him at ramasubramanian.suriyanarayanan@tcs.com.*

Pragathichitra Sethuraman *is a researcher at TCS Innovation Laboratories in Chennai. Her research interests include analyzing facilities' energy-consumption pattern, identifying the major carbon contributors, and determining appropriate methods to reduce the energy and carbon footprints. Sethuraman received a BTech in ceramic technology from Anna University. Contact her at pragathichitra.s@tcs.com.*

Anand Sivasubramaniam *is a professor of computer science and engineering at Pennsylvania State University and a consultant with Tata Consultancy Services. His research interests include computer architecture, operating systems and high-performance computing. Sivasubramaniam received a PhD in computer science from Georgia Tech. He is a senior member of IEEE and the ACM. Contact him at anand@cse.psu.edu.*

Avinash Yegyanarayanan *is an analyst at TCS. His research interests include datacenter power monitoring, building energy efficiency, and energy purchase planning applications. Previously, he was a researcher at TCS Innovation Laboratories in Chennai. Yegyanarayanan received a BTech in computer science from Vellore Institute of Technology. Contact him at avinash.y@tcs.com.*



Computing Performance: Game Over or Next Level?

Samuel H. Fuller, *Analog Devices Inc.*
Lynette I. Millett, *National Research Council*

The end of dramatic exponential growth in single-processor performance marks the end of the dominance of the single microproessor in computing. The era of sequential computing must give way to an era in which parallelism holds the forefront. Although important scientific and engineering challenges lie ahead, this is an opportune time for innovation in programming systems and computing architectures.

Last year, the Computer Science and Telecommunications Board (CSTB) of the US National Academy of Sciences released *The Future of Computing Performance: Game Over or Next Level?*[1] With sponsorship from the US National Science Foundation, the CSTB convened a committee of experts to identify key challenges to continued growth in computing performance and to outline a research agenda for meeting the emerging computing needs of the 21st century. These experts brought diverse perspectives in the fields of semiconductor technology, computer architecture, programming languages and methods, and applications to explore challenges to sustaining performance growth and meeting broad societal expectations for computing now and in the future.

The committee's observations, findings, and recommendations can be broadly summarized in two categories: energy and power constraints on growth in computing performance, and a proposed research agenda that emphasizes new approaches to software and parallelism to meet future expectations for performance growth.

COMPUTING GROWTH DEPENDENCE

Information technology has transformed how we work and live—and has the potential to continue doing so. IT helps bring distant people together, coordinate disaster response, enhance economic productivity, enable new medical diagnoses and treatments, add new efficiencies to our economy, improve weather prediction and climate modeling, broaden educational access, strengthen national defense, advance science, and produce and deliver content for education and entertainment.

These transformations have been made possible by sustained improvements in computer performance. We have been living in a world where information processing costs have been decreasing exponentially year after year. Moore's law—which originally referred to an empirical observation about the most economically favorable rate for industry to increase the number of transistors on a chip—has come to be associated with the expectation that microprocessors will become faster, communication bandwidth will increase, storage will become less expensive, and, more broadly, computers will become faster. Most notably, the performance of individual computer processors increased some 10,000 times over the past two decades, without substantial power consumption increases.

Although some might say they do not want or need a faster computer, users and the computer industry now depend on continuing this performance growth. Much IT

0018-9162/11/$26.00 © 2011 IEEE

innovation depends on taking advantage of computing performance's leading edge. The IT industry annually generates a trillion dollars and has even larger indirect effects throughout society.

This huge economic engine depends on a sustained demand for IT products and services, which in turn fuels demand for constantly improving performance. More broadly, virtually every sector of society—manufacturing, financial services, education, science, government, the military, and entertainment—now depends on this continued growth in computing performance to drive industrial productivity, increase efficiency, and enable innovation. The performance achievements have driven an implicit, pervasive expectation that future IT advances will occur as an inevitable continuation of the stunning advances IT has experienced in the past half-century.

Growth in single-processor performance has stalled—or at best is being increased only marginally over time.

Software developers themselves have come to depend on performance growth across several dimensions:

- adding visible features and ever more sophisticated interfaces to existing applications;
- increasing "hidden" (nonfunctional) value—such as improved security, reliability, and other trustworthiness features—without degrading the performance of existing functions;
- using higher-level abstractions, programming languages, and systems that require more computing power but reduce development time and improve software quality by making the development of correct programs and component integration easier; and
- anticipating performance improvements and creating innovative, computationally intensive applications even before the required performance is available at low cost.

Five decades of exponential performance growth have also made dominant the general-purpose microprocessor at the heart of every personal computer. This stems first from a cycle of economies of scale, wherein each computer generation has been both faster and less expensive than the previous one. Second, increased software portability lets current and forthcoming software applications run correctly and faster on new computers.

These economies have resulted from the application of Moore's law to transistor density, along with innovative approaches to effectively harness the new transistors that have become available. Software portability has been preserved by keeping instruction sets compatible over many generations of microprocessors, even as the underlying microprocessor technology underwent substantial enhancements, allowing investments in software to be amortized over long periods.

The success of this virtuous cycle dampened interest in the development of alternative computer and programming models. Alternative architectures might have been technically superior (for example, faster or more power-efficient) in specific domains, but, generally speaking, if they did not offer software compatibility, they could not easily compete in the marketplace and were overtaken by the ever-improving general-purpose processors available at relatively low cost.

SINGLE-PROCESSOR PERFORMANCE-GROWTH CONSTRAINTS

By the 2000s, however, it had become apparent that processor performance growth faced two major constraints.

First, the ability to increase clock speeds locked horns with power limits. The densest, highest-performance, and most power-efficient integrated circuits (ICs) are constructed from complementary metal-oxide semiconductor (CMOS) technology.

By 2004, the long-fruitful strategy of scaling down the size of CMOS circuits, reducing the supply voltage, and increasing the clock rate had become infeasible. Since a chip's power consumption is in proportion to the clock speed times the supply voltage squared, the inability to continue to lower the supply voltage halted developers' ability to increase the clock speed without increasing power dissipation.[2] The resulting power consumption exceeded the few hundred watts per chip level that can practically be dissipated in a mass-market computing system, as well as the practical limit of a few watts for mobile, battery-powered devices. The ultimate consequence has been that growth in single-processor performance has stalled—or at best is being increased only marginally over time.

Second, efforts to improve individual processors' internal architecture have netted diminishing returns. Many advances in the architecture of general-purpose sequential processors, such as deeper pipelines and speculative execution, have contributed to successful exploitation of increasing transistor densities. Today, however, there appears to be little opportunity to significantly increase performance by improving the internal structure of existing sequential processors.

The slowdown in processor performance growth, clock speed, and power since 2004 is evident in Figure 1, which also shows the continued, exponential growth in the number of transistors per chip. The original Moore's law projection of increasing transistors per chip remains unabated even as performance has stalled. The 2009 edition of the *International Technology Roadmap for Semiconductors* (www.itrs.net/Links/2009ITRS/Home2009.htm) predicts this growth continuing through the next decade, but we will probably be unable to continue increasing transistor density for CMOS circuits at the current pace for more than the next 10 years.

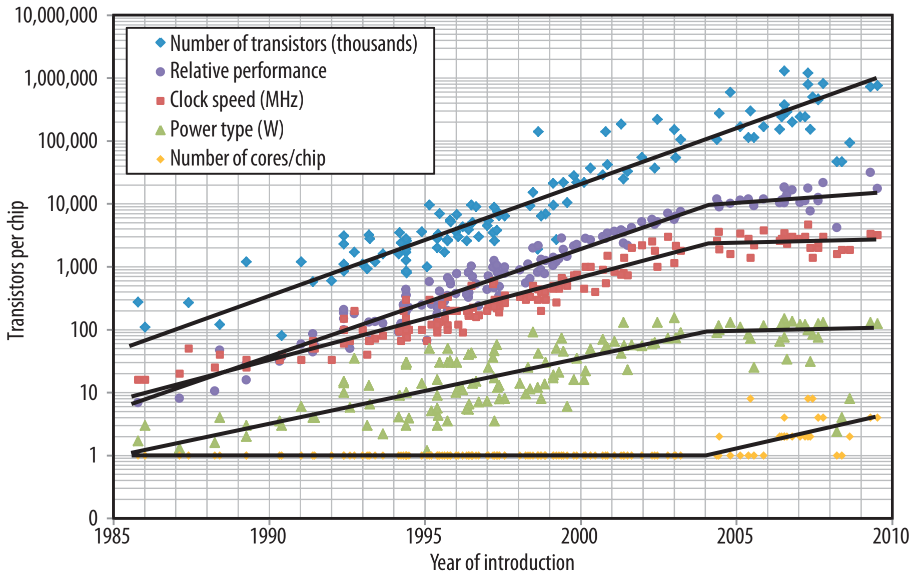


Figure 1. Transistors, frequency, power, performance, and processor cores over time. The original Moore's law projection of increasing transistors per chip remains unabated even as performance has stalled.

Figure 2 shows this expectation gap using a logarithmic vertical scale. In 2010, this gap for single-processor performance is approximately a factor of 10; by 2020, the gap will have grown to about a factor of 1,000. Most economic or societal sectors implicitly or explicitly expect computing to deliver steady, exponentially increasing performance, but these graphs show traditional single-processor computing systems will not match expectations.

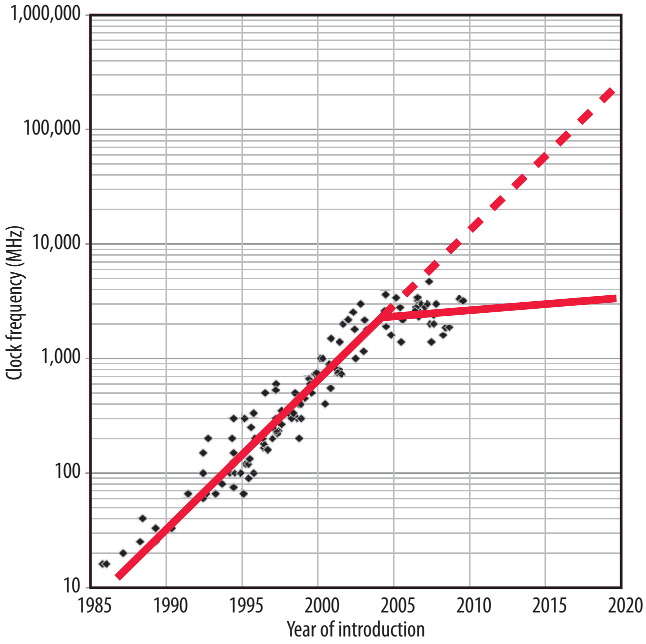


Figure 2. Historical growth in single-processor performance and a forecast of processor performance to 2020, based on the ITRS roadmap. A dashed line represents expectations if single-processor performance had continued its historical trend.

By 2020, we will see a large "expectation gap" for single processors. After many decades of dramatic exponential growth, single-processor performance is slowing and not expected to improve in the foreseeable future. Energy and power constraints play an important and growing role in computing performance. Computer systems require energy to operate, and, as with any device, the more energy needed, the more expensive the system is to operate and maintain. Moreover, the energy consumed by the system ends up as heat that must be removed. Even with new parallel models and solutions, the performance of most future computing systems will be limited by power or energy in ways the computer industry and researchers have yet to confront.

For example, the benefits of replacing a single, highly complex processor with increasing numbers of simpler processors will eventually reach a limit when further simplification costs more in performance than it saves in power. Power constraints are thus inevitable for systems ranging from handheld devices to the largest computing datacenters, even as the transition is made to parallel systems.

Total energy consumed by computing systems is already substantial and continues to grow rapidly in the US and elsewhere around the world. As is the case in other economic sectors, the total energy consumed by computing will come under increasing pressure.

Even if we succeed in sidestepping the limits on single-processor performance, total energy consumption will remain an important concern, and growth in performance will become limited by power consumption within a decade.

In short, the single processor and the sequential programming model that dominated computing since its birth in the 1940s will no longer be sufficient to deliver the continued growth in performance needed to facilitate future IT advances. Moreover, whether power and energy will be showstoppers or just significant constraints remains an open question. Although these issues pose major technical challenges, they will also drive considerable innovation in computing by forcing us to rethink the von Neumann model that has prevailed since the 1940s.

SOLVING WITH PARALLELISM

Future growth in computing performance must come from parallelism. Today, most software developers think and program using a sequential programming model to create software for single general-purpose microprocessors. The microprocessor industry has already begun to deliver parallel hardware in mainstream products with chip multiprocessors (CMPs), an approach that places new burdens on software developers to build applications that take advantage of multiple, distinct cores.

> **Whether power and energy will be showstoppers or just significant constraints remains an open question.**

Although developers have found reasonable ways to use two or even four cores effectively by running independent tasks on each one, they have not, for the most part, parallelized individual tasks to make full use of the available computational capacity. Moreover, if industry continues to follow the same trends, they will soon be delivering chips with hundreds of cores. Harnessing these will require new techniques for parallel computing, including breakthroughs in software models, languages, and tools. Developers of both hardware and software will need to focus more attention on overall system performance, likely at the expense of time to market and the efficiency of the virtuous cycle previously described.

The computer science and engineering communities have worked for decades on the hard problems associated with parallelism. For example, high-performance computing for science and engineering applications has depended on particular parallel-programming techniques such as implementing the message passing interface (MPI). In other cases, domain-specific languages and abstractions such as MapReduce[3] have provided interfaces with behind-the-scenes parallelism and well-chosen abstractions developed by experts—technologies that hide the complexity of parallel programming from application developers.

These efforts have typically involved a small cadre of programmers with highly specialized training in parallel programming who work on relatively narrow computing problems. None of this work has, however, come close to enabling widespread use of parallel programming for a wide array of computing problems.

A few research universities, including MIT, the University of Washington, and the University of California, Berkeley, have launched or revived research programs in parallelism. The topic has found a renewed focus in industry at companies such as Nvidia. However, these initial investments are not commensurate with the magnitude of the technical challenges or the stakes. Moreover, history shows that technology advances of this sort often take a decade or more.[4] The results of such research are needed today to sustain historical trends in computing performance, which already puts us a decade behind. Even with concerted investment, there is no guarantee that widely applicable solutions will be found. If they cannot be, we need to know this quickly so that we can explore other avenues.

MEETING THE CHALLENGES

Current technological challenges affect not only computing but also the many sectors of society that now depend on advances in IT and computation. These suggest national and global economic repercussions. At the same time, the crisis in computing performance has pointed to new opportunities for innovation in diverse hardware and software infrastructures that excel in metrics other than single-processor performance, such as low-power consumption and aggregate delivery of throughput cycles. There are opportunities for major changes in system architectures. Further, we need extensive investment in whole-system research to lay the foundation for next-generation computing environments.

The CSTB committee's recommendations are broadly aimed at federal research agencies, the computing and information technology industry, and educators, and they fall into two categories. The first is research. The best science and engineering minds must be brought to bear on our most daunting challenges. The second category is practice and education. Better practice in developing computer hardware and software today will provide a foundation for future performance gains. Education will empower the emerging generation of technical experts to understand different and in some cases not-yet-developed parallel models for thinking about IT, computation, and software.

RESEARCH RECOMMENDATIONS

The committee urges investment in several crosscutting areas of research, including algorithms, broadly usable parallel programming methods, rethinking the canonical computing stack, parallel architectures, and power efficiency.

Researchers must invest in and develop algorithms that can exploit parallel processing. Today, relatively little software is explicitly parallel. To obtain the desired performance, many—if not most—software designers must grapple with parallelism. For some applications, they might still be able to write sequential programs, leaving it to compilers and other software tools to extract the parallelism in the underlying algorithms. For more complex applications, it might be necessary for programmers to write explicitly parallel programs. Parallel approaches are already used in some applications when no viable alternative is available. The committee believes that careful attention to parallelism will become the rule rather than the exception.

Further, it will be important to invest in research on and development of programming methods that will enable efficient use of parallel systems by typical programmers as well as by experts in parallel systems. Many of today's programming models, languages, compilers, hypervisors, and operating systems are targeted primarily at single-processor hardware. In the future, these layers will need to target, optimize programs for, and be optimized themselves for explicitly parallel hardware.

Rethinking programming models

The intellectual keystone of this endeavor is rethinking programming models so that programmers can express application parallelism naturally. This will let parallel software be developed for diverse systems rather than specific configurations, and let system software deal with balancing computation and minimizing communication among multiple computational units.

This situation is reminiscent of the late 1970s, when programming models and tools were inadequate for building substantially more complex software. Better programming models—such as structured programming in the 1970s, object orientation in the 1980s, and managed programming languages in the 1990s—have made it possible to produce much more sophisticated software. Analogous advances in the form of better tools and additional training will be needed to increase programmer productivity for parallel systems.

The ability to express application parallelism so that an application runs faster as more cores are added would provide a key breakthrough. The most prevalent parallel-programming languages do not provide this performance portability. A related question is what to do with the enormous body of legacy sequential code, which will only contribute to substantial performance improvements if it can be parallelized.

Experience has shown that parallelizing sequential code or highly sequential algorithms effectively is exceedingly difficult in general. Writing software that expresses the type of parallelism required to exploit chip multiprocessor hardware requires new software engineering processes and tools, including new programming languages that ease the expression of parallelism, and a new software stack that can exploit and map the parallelism to diverse and evolving hardware. It will also require training programmers to solve their problems with parallel computational thinking.

The models themselves might or might not be explicitly parallel; whether or when most programmers should be exposed to explicit parallelism remains an open question. A single, universal programming model might or might not exist, so multiple models—including some domain-specific ones—should be explored.

We need additional research in the development of new libraries and programming languages, with appropriate compilation and runtime support that embodies the new programming models. It seems reasonable to expect that some programming models, libraries, and languages will be suited for a broad base of skilled but not superstar programmers. They could even appear on the surface to be sequential or declarative. Others, however, will target efficiency, contributing to the highest performance for critical subsystems that are to be extensively reused, and thus be intended for a smaller set of expert programmers.

Advances in the form of better tools and additional training will be needed to increase programmer productivity for parallel systems.

System software

Research should also focus on system software for highly parallel systems. Although today's operating systems can handle some modest parallelism, future systems will include many more processors whose allocation, load balancing, and data communication and synchronization interactions will be difficult to handle well. Solving those problems will require a rethinking of how computation and communication resources are viewed, much as demands for increased memory led to the introduction of virtual memory a half-century ago.

Long-term efforts should focus on rethinking the canonical computing "stack"—applications, programming language, compiler, runtime, virtual machine, operating system, hypervisor, and architecture—in light of parallelism and resource-management challenges. Computer scientists and engineers typically manage complexity by separating the interface from its implementation. In conventional computer systems, developers do this recursively to form a computing stack of applications, programming language, compiler, runtime, virtual machine, operating system, hypervisor, and architecture components.

Whether today's conventional stack provides the right framework to support parallelism and manage resources remains unclear. The structure and elements of the stack itself should be a focus of long-term research exploration.

Rethinking architecture

We must invest in research on and development of parallel architectures driven by applications, including enhancements to chip multiprocessor systems and conventional data-parallel architectures, cost-effective designs for application-specific architectures, and support for radically different approaches. In addition, advances in architecture and hardware will play an important role. One path forward continues to refine CMPs and associated architectural approaches. We must determine if today's CMP approaches are suitable for designing most computers.

We must invest in research on and development of parallel architectures driven by applications.

The current CMP architecture, the heart of this architectural franchise, keeps companies investing heavily. But CMP architectures bring their own challenges. We must determine if large numbers of cores work in most computer deployments, such as desktops and even mobile phones. We must then see how the cores can be harnessed temporarily, in an automated or semiautomated fashion, to overcome sequential bottlenecks. Leveraging the mechanisms and policies to best exploit locality, we should keep data stored close to other data that might be needed at the same time or for particular computations, while avoiding communications bottlenecks. Finally, we must address how to handle synchronization and scheduling, how to address the challenges associated with power and energy, and what the new architectures mean for such system-level features as reliability and security.

Using homogeneous processors in CMP architectures provides one approach, but computer architectures that include multiple or heterogeneous cores, some of which might be more capable than others, or even use different instruction-set architectures, could prove most effective. Special-purpose processors have long exploited parallelism, notably graphics processing units (GPUs) and digital signal processor (DSP) hardware. These have been successfully deployed in important segments of the market. Other niches like these could be filled by GPUs and DSPs, or computing cores that use more graphics for GPU support of general-purpose programs might cause differences between the two approaches to blur.

Perhaps some entirely new architectural approach will prove more successful. If systems with CMP architectures cannot be effectively programmed, an alternative will be needed. Work in this general area could sidestep conventional cores and view the chip as a tabula rasa with billions of transistors, translating into hundreds of functional units. The effective organization of these units into a programmable architecture is an open question. Exploratory computing systems, based on programmable gate arrays, offer a step in this direction, but we need continued innovation to develop programming systems that can harness the field-programmable gate array's potential parallelism.

Another place where fundamentally different approaches might be needed could involve alternatives to CMOS. There are many advantages to sticking with today's silicon-based CMOS technology, which has proven remarkably scalable over many generations of microprocessors, and around which an enormous industrial and experience base has been established. However, it will also be essential to invest in new computation substrates whose underlying power efficiency promises to be fundamentally better than that of silicon-based CMOSs. Computing has benefited in the past from order-of-magnitude performance improvements in power consumption in the progression from vacuum tubes, to discrete bipolar transistors, to ICs first based on bipolar transistors, to N-type metal-oxide-semiconductor (NMOS) logic, to CMOS. No alternative has approached commercial availability yet, although some show potential.

In the best case, investment will yield devices and manufacturing methods as yet unforeseen that will dramatically surpass the CMOS IC. Worst case, no new technology will emerge to help solve current problems. This uncertainty argues for investment in multiple approaches as soon as possible, and computer system designers would be well advised not to expect one of the new devices to appear in time to obviate the development of new, parallel architectures built on proven CMOS technology.

We need better performance immediately. Society cannot wait the decade or two it would take to identify, refine, and apply a new technology that might not materialize. Moreover, even if researchers do discover a groundbreaking technology, the investment in parallelism would not be wasted because its advances would probably exploit the new technology as well.

Power efficiency

Because energy consumption and power dissipation increasingly limit computing systems, developing efficient power sources is critical. We must invest in research to make computer systems more power-efficient at all system levels, including software, application-specific approaches, and alternative devices. R&D efforts should address ways in which software and system architectures can improve power efficiency, such as exploiting locality and the use of domain-specific execution units. R&D should also be aimed at making logic gates more power-efficient. Such efforts should address alternative physical devices beyond incremental improvements in today's CMOS circuits.

Exploiting parallelism alone cannot ensure continued growth in computer performance. Developers have many potential avenues for investigating better power efficiency, some of which require sustained attention to known engineering issues and others that require further research. These include the following approaches:

- Redesign the delivery of power to and removal of heat from computing systems for increased efficiency.
- Design and deploy systems in which we use the absolute maximum fraction of power to do the computing, and less for routing power to the system and removing heat from it. New standards—including ones that set ever more aggressive targets—might provide useful incentives for the development of better techniques.
- Develop alternatives to the general-purpose processor that exploit locality.
- Develop domain-specific or application-specific processors analogous to GPUs and DSPs that provide better performance and power consumption characteristics than general-purpose processors for other specific application domains.
- Investigate possible new, lower-power device technology beyond CMOS.

Additional research should focus on system designs and software configurations that reduce power consumption, such as when resources are idle, or reducing power-consumption mapping applications to domain-specific and heterogeneous hardware units, limiting the amount of communication among disparate hardware units.

Although the shift toward CMPs will let industry continue to scale the performance of CMPs based on general-purpose processors for some time, general-purpose CMPs will eventually reach their own limits. CMP designers can trade off single-thread performance of individual processors against lower energy dissipation per instruction, thus allowing more instructions by multiple processors while holding the chip's power dissipation constant. However, that is possible only within a limited range of energy performance.

Beyond some limit, however, lowering energy per instruction by processor simplification can lead to degradation in overall CMP performance. This happens when processor performance starts to decrease faster than energy per instruction, which then requires new approaches to create more energy-efficient computers.

It might be that general-purpose CMPs will prove to be an inefficient solution in the long run, and we will need to create more application-optimized processing units. Tuning hardware and software toward a specific type of application provides a much more energy-efficient solution.

However, the current design trend is away from building customized solutions, because increasing design complexity has caused nonrecurring engineering costs for designing chips to grow rapidly. High costs limit the range of potential market segments to the few that have volume high enough to justify the initial engineering investment. A shift to more application-optimized computing systems, if necessary, demands a new design approach that would let application-specific chips be created at reasonable cost.

Additional research should focus on system designs and software configurations that reduce power consumption.

PRACTICE AND EDUCATION RECOMMENDATIONS

Implementing the proposed research agenda, although crucial for progress, will take time. Meanwhile, society has an immediate and pressing need to use current and emerging CMP systems effectively. Efforts in current development and engineering practices and education are also important. The CSTB committee encourages development of open interface standards for parallel programming to promote cooperation and innovation by sharing rather than proliferating proprietary programming environments.

Private-sector firms are often incentivized to create proprietary interfaces and implementations to establish a competitive advantage. However, a lack of standardization can impede progress because the presence of so many incompatible approaches deprives most from achieving the benefits of wide adoption and reuse—a major reason industry participates in standards efforts. The committee encourages the development of programming-interface standards that can facilitate wide adoption of parallel programming even as they foster competition in other areas.

We must develop tools and methods for transforming legacy applications to parallel systems. Whatever long-term success we achieve in the effective use of parallel systems from rethinking algorithms and developing new programming methods will probably come at the expense of the backward- and cross-platform compatibility that has been an IT economic cornerstone for decades. To salvage value from the nation's current, substantial IT investment, we must seek ways to bring sequential programs into the parallel world.

The committee urges industry and academia to develop "power tools" that will help experts migrate legacy code to tomorrow's parallel computers. In addition, emphasis should be placed on tools and strategies to enhance code creation, maintenance, verification, and the adaptation of parallel programs.

Computer science education must increase the emphasis on parallelism by using a variety of methods and approaches to prepare students for the shape of the computing resources they will work with through their careers. We must encourage those who will develop the future's parallel software. To sustain IT innovation, we will need a workforce that is adept in writing parallel applications that run well on parallel hardware, in creating parallel software systems, and in designing parallel hardware.

Both undergraduate and graduate students in computer science, as well as those in other fields that intensively use computing, will need to be educated in parallel programming. The engineering, science, and computer science curricula at both the undergraduate and graduate levels should begin to incorporate an emphasis on parallel computational thinking, parallel algorithms, and parallel programming.

With respect to the computer-science curriculum, given that no general-purpose paradigm has emerged, universities should teach diverse parallel-programming languages, abstractions, and approaches until effective ways of teaching and programming emerge. The necessary shape of the needed changes will not be clear until reasonably general parallel-programming methods have been devised and shown to be promising.

In relation to this goal, we must improve the programming workforce's ability to cope with parallelism's new challenges. On the one hand, this will involve retraining today's programmers. On the other, it will demand developing new models and abstractions to make parallel programming more accessible to typically skilled programmers.

The end of dramatic exponential growth in single-processor performance also ends the single microprocessor's dominance. The era of sequential computing must give way to a new era in which parallelism holds the forefront. There is no guarantee we can make parallel computing as common and easy to use as yesterday's sequential single-processor computer systems, but unless we aggressively pursue the efforts suggested by the CSTB committee's recommendations, it will be *game over* for growth in computing performance. This has larger implications. If parallel programming and software efforts fail to become widespread, the development of exciting new applications that drive the computer industry will slow and affect many other parts of the economy.

Although important scientific and engineering challenges lie ahead, this is an opportune time for innovation in programming systems and computing architectures. We have already begun to see diversity in computer designs to optimize for such considerations as power and throughput. The next generation of discoveries will likely require advances at all levels of the computing systems' hardware and software to achieve the *next level* of benefits to society. C

Acknowledgments

We thank the entire membership and staff of the Committee on Sustaining Growth in Computing Performance for their contributions to this article: Samuel H. Fuller (chair), Luiz Andre Barroso, Robert P. Colwell, William J. Dally, Dan Dobberpuhl, Pradeep Dubey, Mark D. Hill, Mark Horowitz, David Kirk, Monica Lam, Kathryn S. McKinley, Charles Moore, Katherine Yelick, Lynette I. Millett, and Jon Eisenberg. Committee member Mark Horowitz curated the data for the graphs. Support for this project was provided by the National Science Foundation under award CNS-0630358.

References

1. National Research Council, *The Future of Computing Performance: Game Over or Next Level?* Nat'l Academies Press, 2010.
2. R.H. Dennard et al., "Design of Ion-Implanted MOSFETS with Very Small Physical Dimensions," *IEEE J. Solid State Circuits*, vol. 9, no. 5, 1974, pp. 256-268.
3. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Comm. ACM*, vol. 51, no. 1, 2008, pp. 107-113.
4. National Research Council, *Evolving the High-Performance Computing and Communications Initiative to Support the Nation's Information Infrastructure*, Nat'l Academies Press, 1995.

Samuel H. Fuller *is the CTO and vice president of research and development at Analog Devices Inc. He received a PhD in electrical engineering from Stanford University. He is an IEEE Fellow. Contact him at sam.fuller@analog.com.*

Lynette I. Millett *is senior staff officer at the Computer Science and Telecommunications Board, National Research Council of the National Academy of Sciences. She received an MSc in computer science from Cornell University. Contact her at lmillett@nas.edu.*

The Computer Science and Telecommunications Board of the National Academies will host a symposium planned for 22 February 2011 in Washington, D.C., to explore future directions in sustaining computing performance improvements. See http://cstb.org for more information.

COVER FEATURE

From Microprocessors to Nanostores: Rethinking Data-Centric Systems

Parthasarathy Ranganathan, *HP Labs*

The confluence of emerging technologies and new data-centric workloads offers a unique opportunity to rethink traditional system architectures and memory hierarchies in future designs.

What will future computing systems look like?

We are entering an exciting era for systems design. Historically, the first computer to achieve terascale computing (10^{12}, or one trillion operations per second) was demonstrated in the late 1990s. In the 2000s, the first petascale computer was demonstrated with a thousand-times better performance. Extrapolating these trends, we can expect the first exascale computer (with one million trillion operations per second) to appear around the end of this next decade.

In addition to continued advances in performance, we are also seeing tremendous improvements in power, sustainability, manageability, reliability, and scalability. Power management, in particular, is now a first-class design consideration. Recently, system designs have gone beyond optimizing operational energy consumption to examining the total life-cycle energy consumption of systems for improved environmental sustainability. Similarly, in addition to introducing an exciting new model for delivering computing, the emergence of cloud computing has enabled significant advances in scalability as well as innovations in the software stack.

Looking further out, emerging technologies such as photonics, nonvolatile memory, 3D stacking, and new data-centric workloads offer compelling new opportunities. The confluence of these trends motivates a rethinking of the basic systems' building blocks of the future and a likely new design approach called *nanostores* that focus on data-centric workloads and hardware-software codesign for upcoming technologies.

THE DATA EXPLOSION

The amount of data being created is exploding, growing significantly faster than Moore's law. For example, the amount of online data indexed by Google is estimated to have increased from 5 exabytes (one exabyte = 1 million trillion bytes) in 2002 to 280 exabytes in 2009[1]—a 56-fold increase in seven years. In contrast, an equivalent Moore's law growth in computing for the corresponding time would deliver only a 16-fold increase.

This data growth is not limited to the Internet alone, but is pervasive across all markets. In the enterprise space, the size of the largest data warehouse has been increasing at a cumulative annual growth rate of 173 percent[2]—again, significantly more than Moore's law.

New kinds of data

Some common categories for data growth include those pertaining to bringing traditionally offline data online and

0018-9162/11/$26.00 © 2011 IEEE

to new digital media creation, including webpages, personal images, scanned records, audio files, government databases, digitized movies, personal videos, satellite images, scientific databases, census data, and scanned books. A recent estimate indicates that 24 hours of video are uploaded on YouTube every minute. At HD rates of 2-5 Mbps, that is close to 45-75 terabytes of data per day. Given that only about 5 percent of the world's data is currently digitized,[3] growth in this data category is likely to continue for several more years.

More recently, large-scale sensor deployment has contributed to the explosion in data growth. Developments in nanoscale sensors have enabled tracking multiple dimensions—including vibration, tilt, rotation, airflow, light, temperature, chemical signals, humidity, pressure, and location—to collect real-time data sampled at very fine granularities. These advances have motivated researchers to discuss the notion of developing a "central nervous system for the earth (CeNSE)"[4] with intriguing sample applications of rich sensor networks in areas including retail sales, defense, traffic, seismic and oil explorations, weather and climate modeling, and wildlife tracking. This vision will lead to data creation and analysis significantly beyond anything we have seen so far.

Looking ahead, it's clear that we're only at the beginning of an even more fundamental shift in what we do with data.

The pervasive use of mobile devices by a large part of the world's population, and the ability to gather and disseminate information through these devices, contributes to additional real-time rich data creation. For example, at the time of Michael Jackson's death in June 2009, Twitter estimated about 5,000 tweets per minute, and AT&T estimated about 65,000 texts per second. Currently, over a 90-day period, 20 percent of Internet search queries are typically "new data."[1]

Significantly, this large-scale growth in data is happening in combination with a rich diversity in the type of data being created. In addition to the diversity in media types—text, audio, video, images, and so on—there is also significant diversity in how the data is organized: structured (accessible through databases), unstructured (accessed as a collection of files), or semistructured (for example, XML or e-mail).

New kinds of data processing

This growth in data is leading to a corresponding growth in data-centric applications that operate in diverse ways: capturing, classifying, analyzing, processing, archiving, and so on. Examples include Web search, recommendation systems, decision support, online gaming, sorting, compression, sensor networks, ad hoc queries, cubing, media transcoding and streaming, photo processing, social network analysis, personalization, summarization, index building, song recognition, aggregation, Web mashups, data mining, and encryption. Figure 1 presents a taxonomy of data-centric workloads that summarizes this space.

Compared to traditional enterprise workloads such as online transaction processing and Web services, emerging data-centric workloads change many assumptions about system design. These workloads typically operate at larger scale (hundreds of thousands of servers) and on more diverse data (structured, unstructured, rich media) with I/O-intensive, often random, data access patterns and limited locality. In addition, these workloads are characterized by innovations in the software stack targeted at increased scalability and commodity hardware such as Google's MapReduce and BigTable.

Looking ahead, it's clear that we're only at the beginning of an even more fundamental shift in what we do with data. As an illustrative example, consider what happens when we search for an address on the Web.

In the past, this request would be sent to a back-end webserver that would respond with the image of a map showing the address's location. However, in recent years, more sophisticated data analysis has been added to the response to this query. For example, along with just accessing the map database, the query could potentially access *multiple data sources*—for example, satellite imagery, prior images from other users, webpages associated with location information, overlays of transit maps, and so on. Beyond just static images, *dynamic* data sources can be brought into play—such as providing live traffic or real-time weather information, current Twitter feeds, or live news or video. *Synthetic* data such as images from user-provided 3D models of buildings or outputs from trend analyzers and visualizers also can be superimposed on the map.

Adding personalization and contextual responses to the mix introduces another layer of data processing complexity. For example, different data can be presented to the user based on the last two searches prior to this search, or on the user's prior behavior when doing the same search, or on locational information (for example, if the current location matches the location where the user searched previously).

Social networks and recommendation systems add yet another layer of data processing complexity. Examples include on-map visualization of individuals' locations drawn from social networks, inferred preferences, and prescriptive recommendations based on social trends. Advertisements and, more generally, business monetization of search, adds another layer of data processing in terms of accessing more data sources and more sophisticated algorithms for user preference and content relevance.

Response time	Real-time	Real-time or interactive responses required
	Background	Response time is not critical for user needs
Access pattern	Random	Unpredictable access to regions of datastore
	Sequential	Sequential access of data chunks
	Permutation	Data is redistributed across the system
Working set	All	The entire dataset is accessed
	Partial	Only a subset of data is accessed
Data type	Structured	Metadata/schema/type are used for data records
	Unstructured	No explicit data structure, for example, text/binary files
	Rich media	Audio/video and image data with inherent structures and specific processing algorithms
Read vs. write	Read heavy	Data reads are significant for processing
	Write heavy	Data writes are significant for processing
Processing complexity	High	Complex processing of data is required per data item; examples: video transcoding, classification, prediction
	Medium	Simpler processing is required per data item; examples: pattern matching, search, encryption
	Low	Dominated by data access with low compute ratio; examples: sort, upload, download, filtering, and aggregation

Figure 1. Data-centric workload taxonomy.

In many cases, all this data processing comes with fairly small latency requirements for response, even requiring real-time responses in some scenarios.

This scenario shows how, from simple Web search and content serving, online data processing is evolving to allow more complex meaning extraction across multiple data repositories, and more sophisticated cross-correlations, including more complex I/O movement. In a continuum of data processing operations including upload/ingress; download/egress; search (tree traversal); read, modify, write; pattern matching; aggregation; correlation/join; index building; cubing; classification; prediction; and social network analysis, recent trends show a strong movement toward operations with more complex data movement patterns.[5]

Similar trends can be seen in enterprise data management across the information→insight→outcome life cycle. There is an increasing emphasis on real-time feeds of business information, often across multiple formal or ad hoc data repositories, reduced latencies between events and decisions, and sophisticated combinations of parallel analytics, business intelligence, and search and extraction operations. Jim Gray alluded to similar trends in scientific computing when discussing a new era in which scientific phenomena are understood through large-scale data analysis.[6] Such trends can also be seen in other important workloads of the future, with applications like computational journalism, urban planning, natural-language processing, smart grids, crowdsourcing, and defense applications. The common traits in all these future workloads are an emphasis on complex cross-correlations across multiple data repositories and new data analysis/compute assumptions.

Together, this growing complexity and dynamism in extraction of meaning from data, combined with the large-scale diversity in the amount of data generated, represent an interesting inflection point in the future data-centric era. The "Implications of Data-Centric Workloads for System Architectures" sidebar provides additional information about this trend for system designs.

IT'S A NEW WORLD—AN INFLECTION POINT IN TECHNOLOGY

Concurrently, recent trends point to several potential technology disruptions on the horizon.

On the compute side, recent microprocessors have favored multicore designs emphasizing multiple simpler cores for greater throughput. This is well matched with the large-scale distributed parallelism in data-centric workloads. Operating cores at near-threshold voltage has been shown to significantly improve energy efficiency.[7] Similarly, recent advances in networking show a strong growth in bandwidth for communication between different compute elements at various system design levels.

However, the most important technology changes pertinent to data-centric computing relate to the advances in and adoption of nonvolatile memory. Flash memories have been widely adopted in popular consumer systems—for example, Apple's iPhone—and are gaining adoption in the enterprise market—for example, Fusion-io.

Figure 2 shows the trends in costs for these technologies relative to traditional hard disks and DRAM memories. Emerging nonvolatile memories have been demonstrated to have properties superior to flash memories, most notably phase-change memory (PCM)[8] and, more recently, memristors.[9] Trends suggest that future nonvolatile

IMPLICATIONS OF DATA-CENTRIC WORKLOADS FOR SYSTEM ARCHITECTURES

An important trend in the emergence of data-centric workloads has been the introduction of complex analysis at immense scale, closely coupled with the growth of large-scale Internet Web services. Traditional data-centric workloads like Web serving and online transaction processing are being superseded by workloads like real-time multimedia streaming and conversion; history-based recommendation systems; searches of text, images, and even videos; and deep analysis of unstructured data—for example, Google Squared.

From a system architecture viewpoint, a common characteristic of these workloads is their general implementation on highly distributed systems, and that they adopt approaches that scale by partitioning data across individual nodes. Both the total amount of data involved in a single task and the number of distributed compute nodes required to process the data reflect their large scale. Additionally, these workloads are I/O intensive, often with random access patterns to small-size objects over large datasets.

Many of these applications operate on larger fractions of data in memory. According to a recent report, the amount of DRAM used in Facebook for nonimage data is approximately 75 percent of the total data size.[1] While this trend partly reflects the low latency requirements and the limited locality due to complex linkages between data for the Facebook workload, similar trends for larger memory capacities can be seen for memcached servers and TPC-H benchmark winners over the past decade. Similarly, search algorithms such as the one from Google have evolved to store their search indices entirely in DRAM. These trends motivate a rethinking of the balance between memory and disk-based storage in traditional designs.

Interestingly, datasets and the need to operate on larger fractions of the data in-memory continue to increase, there will likely be an inflection point at which conventional system architectures based on faster and more powerful processors and ever deeper memory hierarchies are not likely to work from an energy perspective (Figure A). Indeed, a recent exascale report identifies the amount of energy consumed in transporting data across different levels as a key limiting factor.[2] Complex power-hungry processors also are sometimes a mismatch with data-intensive workloads, leading to further energy inefficiencies.

Recent data-centric workloads have been characterized by numerous commercially deployed innovations in the software stack—for example, Google's BigTable and MapReduce, Amazon's Dynamo, Yahoo's PNUTS, Microsoft's Dryad, Facebook's Memcached, and LinkedIn's Voldemort. Indeed, according to a recent presentation, the software stack behind the very successful Google search engine was significantly rearchitected four times in the past seven years to achieve better performance at increased scale.[3]

The growing importance of this class of workloads, their focus on large-scale distributed systems with ever-increasing memory use, the potential inadequacy of existing architectural approaches, and the relative openness to software-level innovations in the emerging workloads offer an opportunity for a corresponding clean-slate architecture design targeted at data-centric computing.

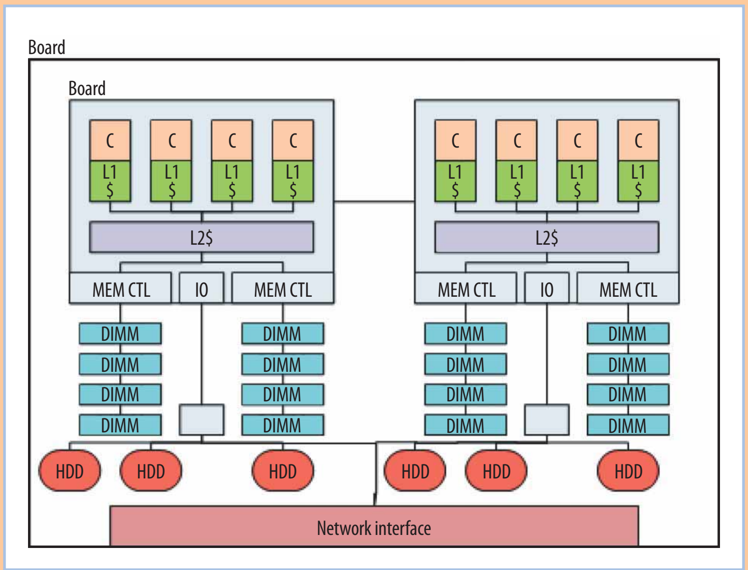


Figure A. Changing workload trends motivate a rethinking of the traditional designs with deep hierarchies.

References

1. J. Ousterhout et al., "The Case for RAM Clouds: Scalable High-Performance Storage Entirely in DRAM," *ACM SIGOPS Operating Systems Rev.*, vol. 43, no. 4, 2009, pp 92-105.
2. P. Kogge ed., "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," 2008; http://www.science.doe.gov/ascr/Research/CS/DARPA%20exascale%20-%20hardware%20(2008).pdf.
3. J. Dean, "Challenges in Building Large-Scale Information Retrieval Systems," keynote talk, *Proc. 2nd Ann. ACM Conf. Web Search and Data Mining* (WSDM 09), ACM Press, 2009; http://wsdm2009.org/proceedings.php.

memories can be viable DRAM replacements, achieving competitive speeds at lower power consumption, with nonvolatility properties similar to disks but without the power overhead. Additionally, recent studies have identified a slowing of DRAM growth due to scaling challenges for charge-based memories.[10,11] The adoption of NVRAM as a DRAM replacement can potentially be accelerated due to such limitations in scaling DRAM.

Density and endurance have been traditional limitations of NVRAM technologies, but recent trends suggest that these limitations can be addressed. Multilevel designs can achieve increased density, potentially allowing multiple layers per die.[12] At a single chip level, 3D die stacking using through-silicon vias for interdie communication can further increase density. Such 3D stacking also has the additional advantage of closely integrating the processor

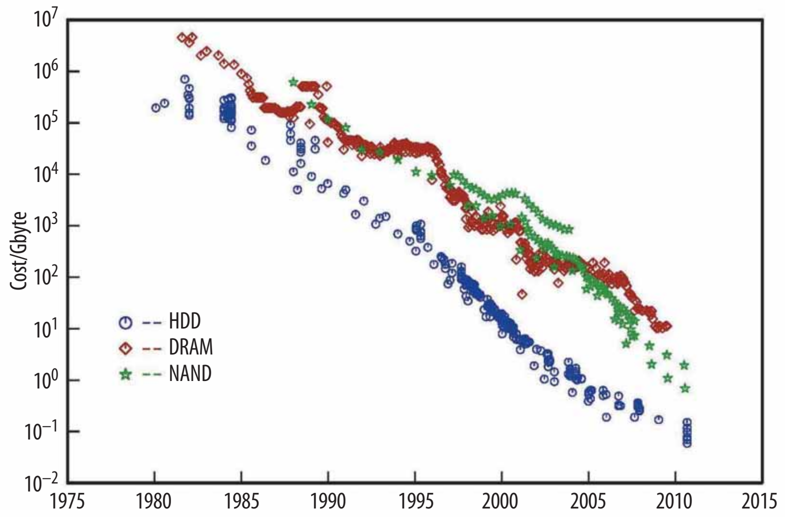


Figure 2. Nonvolatile memory cost trends. These trends suggest that future nonvolatile memories can be viable DRAM replacements.

and memory for higher bandwidth and lower power (due to short-length low-capacitance wires). Structures like wire bonding in system-in-package or package-on-package 3D stacking are already integrated into products currently on the market, such as mobile systems, while more sophisticated 3D-stacking solutions have been demonstrated in the lab.

In terms of endurance, compared to flash memories, PCMs and memristors offer significantly better functionality—10^7-10^8 writes per cell compared to the 10^5 writes per cell for flash. Optimizations at the technology, circuit, and systems levels have been shown to further address endurance issues, and more improvements are likely as the technologies mature and gain widespread adoption.[11,13]

More details about emerging nonvolatile memories can be found in several recent overviews and tutorials [14,15]—for example, HotChips 2010 (www.hotchips.org).

These trends suggest that technologies like PCM and memristors, especially when viewed in the context of advances like 3D die stacking, multicores, and improved networking, can induce more fundamental architectural change for data-intensive computing than traditional approaches that use them as solid-state disks or as another intermediate level in the memory hierarchy.

NANOSTORES: A NEW SYSTEM ARCHITECTURE BUILDING BLOCK?

The confluence of these various trends—future large-scale distributed data-centric workloads with I/O-intensive behavior, innovations in the software stack, and the emergence of new nonvolatile memories potentially timed with the end of scaling for DRAM—offers a unique opportunity to rethink traditional system architectures and memory hierarchies in future designs.

Nanostores offer one such intuitive, and potentially advantageous, way to leverage this confluence of application and technology trends. We coined the term nanostores as a duality of microprocessors to reflect the evolution to nanotechnology and the emphasis on data instead of compute. The key property of nanostores is the colocation of processors with nonvolatile storage, eliminating many intervening levels of the storage hierarchy. All data is stored in a single-level nonvolatile memory datastore that replaces traditional disk and DRAM layers—disk use is relegated to archival backups.

For example, a single nanostore chip consists of multiple 3D-stacked layers of dense silicon nonvolatile memories such as PCMs or memristors, with a top layer of power-efficient compute cores. Through-silicon vias are used to provide wide, low-energy datapaths between the processors and the datastores. Each nanostore can act as a full-fledged system with a network interface. Individual such nanostores are networked through onboard connectors to form a large-scale distributed system or cluster akin to current large-scale clusters for data-centric computing. The system can support different network topologies, including traditional fat trees or recent proposals like HyperX.[16]

In terms of physical organization, multiple nanostore chips are organized into small daughter boards (microblades) that, in turn, plug into traditional blade server boards. Given the heat dissipation characteristics of the design, we also can envision newer packaging technologies for the broader solution. Figure 3 illustrates an example dematerialized datacenter design[17] in which the individual blade servers connect to an optical backplane “spine” with optimized airflow and packaging density.

Power and thermal issues are important concerns with 3D stacking and limit the amount of compute that a nanostore can include. Figure 4 illustrates how additional, more powerful, compute elements can be added to create a “hierarchy of computes” that back up the on-chip computation in the nanostore. This also enables repurposing the design so that nanostores act more like current systems—with powerful compute elements and deep hierarchies to data—if needed for applications such as legacy workloads.

There is a wide range of possible implementations for this high-level organization. There are numerous design choices in terms of the provisioning, organization, and balance of the compute, storage, and network per nanostore

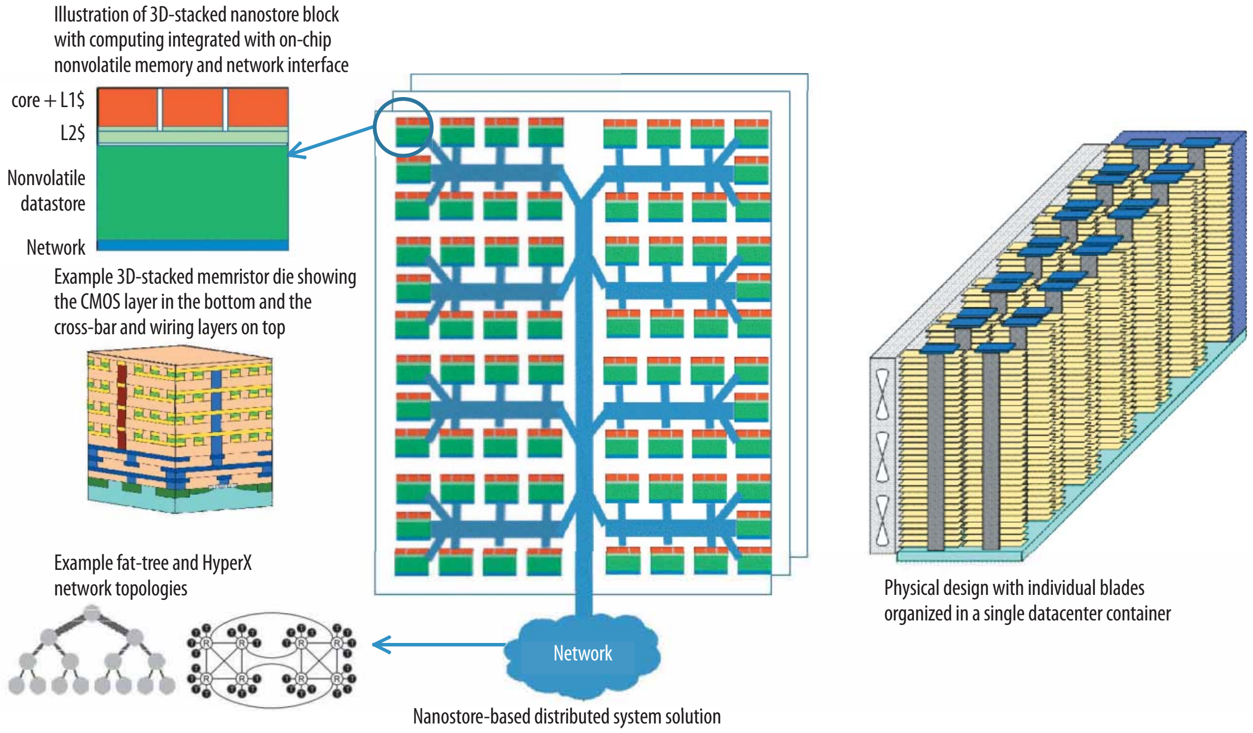


Illustration of 3D-stacked nanostore block with computing integrated with on-chip nonvolatile memory and network interface
core + L1$
L2$
Nonvolatile datastore
Network
Example 3D-stacked memristor die showing the CMOS layer in the bottom and the cross-bar and wiring layers on top
Example fat-tree and HyperX network topologies
Network
Nanostore-based distributed system solution
Physical design with individual blades organized in a single datacenter container

Figure 3. Nanostores colocate processors and nonvolatile memory on the same chip and connect to one another to form a larger cluster for data-centric workloads.

as well as the sharing model across the individual nodes and the network topology, including potential differences between the on-chip, onboard, and cross-cluster networks. Constraints on the design choices include technology- and circuit-level parameters such as the size of the die and the yield, the number of 3D-stacked or intradie (random) layers, as well as packaging constraints such as the power/thermal budget per chip or board.

Similarly, the software models vary depending on the specific architecture built with nanostores. A pure nanostore-to-nanostore architecture is well matched with a large-scale distributed shared-nothing system abstraction, similar to current data-centric workloads. Each nanostore can be viewed as a complete independent system executing the software stack needed to implement a data-parallel execution environment like MapReduce. Alternate designs that include multiple levels of compute will require more sophisticated software models. For example, we could consider software models similar to those under consideration for future desktop systems for coordination across general-purpose and graphics-processing units.

BENEFITS AND CHALLENGES

While similar to some of the principles used in designs such as ActiveStorage,[18] IRAM,[19] and RAMCloud,[20] the nanostore design is unique in its colocation of power-efficient computing with a single-level nonvolatile datastore and a large-scale distributed design well matched with data-centric workloads. This combination provides several benefits.

The single-level datastore enables improved performance due to faster data access in terms of latency and bandwidth. Flattening the memory hierarchy and the increased energy efficiency of NVRAM over disk and DRAM also improve energy efficiency. The large-scale distributed design facilitates higher performance from increased parallelism and higher overall data/network bandwidth. This design also improves energy efficiency by partitioning the system into smaller elements that can leverage more power-efficient components such as simpler cores. Beyond operational energy, this design also has the potential to reduce the embedded energy[17] in the system, which can lead to more sustainable designs.

An illustrative implementation provides a better estimate of these benefits. In this example, the nanostore die size is 100 mm^2, similar to the cost-effective design point for memories.

Let's assume cores in the compute layer are based on low-voltage power-efficient microarchitectures with simple SRAM cache hierarchies. Different organizations are possible for the compute layer—in the number of cores (1 to 128), clock frequency (100 MHz to 2 GHz), issue width and pipeline depth (2-way simple to 4-way deep), and L2 cache size (512 Kbytes or 1 Mbyte per core); the limiting factor will be the power density at the socket (currently 32 watts/cm^2). For our projected timeframe, we expect 3D stacking to provide significant bandwidth (up to 32 Gbytes per second in the PicoServer design) between the processor and stacked memory, and 80 Gbps (2 x 40-Gbps NICs) networking bandwidth per system (in an equivalent traditional architecture). Assuming 25-nm technology, 8 layers of 3D, and intra-die stacking, a single node that groups nine nanostores to provide 8 + 1 redundancy can provide one-half to one terabyte of nonvolatile memory (depending on assumptions around single-level or multilevel cells) with teraops of local compute colocated with the storage (assuming simple low-power processors) and about 256 Gbytes of aggregate datastore bandwidth.

The latencies to access the data are expected to be competitive with DRAM (within about a factor of two), but at much lower energy consumption (current estimates vary from 2 to 10 picojoules/bit compared to 24 picojoules/bit for DRAM).[9] Compared to traditional disks or existing flash memories, this configuration provides several orders-of-magnitude better latencies and energy efficiency.

These numbers demonstrate the potential for better performance at improved energy efficiency with these designs. This improvement stems from more energy-efficient memory technologies, compute colocation leading to lower energy consumption in the hierarchy, and more energy-efficient balanced designs. While these are peak numbers, we have also experimented with simulation numbers for common data-centric kernels that address the key dimensions of the taxonomy discussed in Figure 1. Our results indicate significant improvements in performance and energy efficiency. For I/O-bound workloads, this can often be a few orders of magnitude higher performance at close to an order of magnitude better energy efficiency.[21] At the same time, achieving this potential presents numerous challenges.

Scalability

Given the smaller capacities of per-socket storage, the number of individual elements in the system increases dramatically. This can potentially increase the stress on the networking subsystem in terms of bandwidth contention (particularly for all-to-all communication), topological complexity and port count, and power.

Endurance

Based on current estimations of expected densities and endurance, in theory, storage wearout can occur in two years for PCMs or 11 years for memristors. However, in practice not all applications sustain rates at that level, and the average across the application is much lower, leading to much longer lifetimes across the array. Wear-leveling schemes must still be used to spread writes across the entire memory to prevent early failure of hot data blocks. Assuming a previously proposed approach—start-gap wear leveling—at an efficiency of 90 percent of optimal wear-leveling (shown to be realistic for OLTP/database workloads)[13] and using the memory write bandwidths from our simulations, we estimate per-socket lifetimes of 7-18 years for our benchmarks on the PCM-based design. Nevertheless, techniques that carefully manage wearout warrant further study.

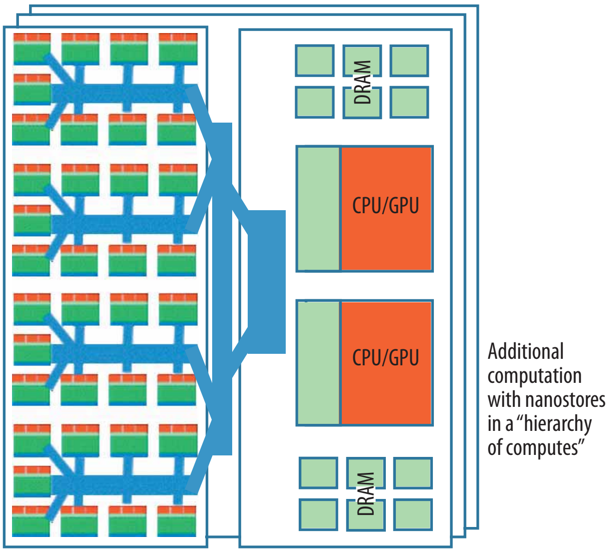

DRAM
CPU/GPU
CPU/GPU
DRAM
Additional computation with nanostores in a "hierarchy of computes"

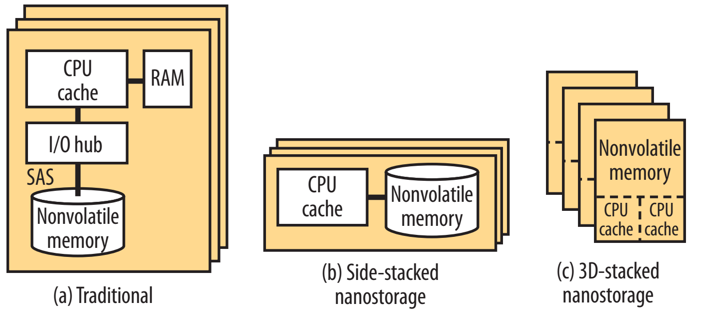
Figure 4. Adding more powerful compute elements can create a hierarchy of computes that backs up on-chip computation in the nanostore.

Cost

As Figure 2 shows, current flash memories are about an order of magnitude more costly on a dollar-per-byte basis compared to disk. NVRAM has the potential to lower these costs by allowing more aggressive stacking and simpler fabrication processes. The improved energy efficiency of nanostores can also further lower total costs of ownership. Based on these observations, we expect the nanostore design to be competitive in costs compared to traditional designs, but this needs to be validated with further study.

Design choices

Several interesting design questions remain to be answered. How well do nanostore designs perform compared to aggressive extrapolations of existing approaches? Are the expected benefits significant enough to warrant the change? How do the benefits change across the range of data-centric workloads? How do the benefits break down? Do we need to rethink the balance of compute, data, and network for this new architecture? What are the implications of specific design choices and technol-



(a) Traditional

(b) Side-stacked nanostorage

(c) 3D-stacked nanostorage

Figure 5. Three different designs offer tradeoffs for data-centric workloads: (a) traditional design, (b) nanostorage side-stacked design, and (c) nanostorage 3D-stacked design.

ogy extrapolations? In particular, what is the sensitivity to the network bandwidth assumptions and packaging limitations?

There is significant potential, and need for, additional architectural innovation in this space.

Software and systems design

Software scalability is an important issue. From 1998 to 2009, Google's infrastructure is reported to have scaled performance (queries processed per day) by 1,000 times while scaling the infrastructure by 1,000 times.[22] While large-scale deployments of data-centric workloads have been demonstrated, the sizing of the system will have to carefully consider latency requirements—for example, response time for a search request. Similarly, current software stacks include decades of development with assumptions of seek limitations of traditional magnetic disks. While some recent studies have revisited such assumptions—for example, byte-persistent file systems[23]—there is a potential, and need, for additional software innovation around datastores for nonvolatile memory-based storage systems.

Modeling and benchmarking

Any new architecture redesign comes with associated challenges in modeling and benchmarking.[24] Our focus on the combination of multiple future technologies for emerging workloads poses several challenges in the choice of benchmarks, technology parameters, and baseline systems appropriate to this longer timeframe.

To evaluate alternate designs and their tradeoffs, we need to study large-scale clusters running distributed workloads operating on large volumes of data. We also need to examine tradeoffs at the full system level, including computing, networking, memory, and storage layers. Conventional architecture simulators not only lack the ability to cope with this level of system scale, but also the modeling means for storage and network subsystems at a distributed systems level. There is also a combinatorial explosion in the design space from various assumptions at the fine-grained and coarse-grained architectural levels, as well as the choice of technology and workload parameters.

An appropriate evaluation methodology is required to systematically reason about this large design space. Similarly, a key need is the availability of a representative set of the emerging distributed workloads that drive these data-centric markets. New approaches are needed in this space as well.[5]

MATCHING NANOSTORES TO DATA-CENTRIC WORKLOADS

The benefits of colocating compute close to nonvolatile datastores can be achieved with different designs, each with different tradeoffs for specific data-centric workloads. As Figure 5 shows, these tradeoffs are best illustrated by comparing three designs—a traditional design with DRAM memory and solid-state disks, a nanostore 3D-stacked design (similar to our discussions so far), and an alternate nanostore side-stacked design that colocates the compute with the nonvolatile datastore, but separately off the memory bus (with the nonvolatile store replacing traditional disks as before).

From a data-centric workload point of view, a good way to reason across these designs is to consider the amount of raw compute processing capacity that can be applied per unit data, at both global and local levels, and the bottlenecks in the hardware and software that limit the ability to use this compute capacity.

The traditional design is likely to work well for compute-heavy workloads with small data bandwidth per cycle—for example, video transcoding—or workloads in which the hot and cold working set sizes are orders of magnitude apart—for example, image archiving. Workloads that require additional bandwidth for the underlying data and can leverage data-partitioned parallelism—for example, MapReduce workloads, sorts, clickstreams, and log analysis—can benefit from the nanostore side-stacked and nanostore 3D-stacked designs.

Rewriting the software to leverage the improved datastore latencies can provide additional benefits—until the network becomes a bottleneck. For parallel workloads that can be rewritten to use fine granularity with limited cross-cluster communication (filtering, aggregation, textual search, and so on), the nanostore 3D-stacked design is likely to work best—until the compute becomes a bottleneck. More work is needed in software for effective parallelization, but the cost and energy advantages may prove these measures to be worthwhile.

The trends toward growing pressure for improved bandwidth and latency in data-centric workloads, ongoing progress in parallelizing software, and improvements in local interconnection networks support using a nanostore design for future systems, but hybrid designs also may emerge.

GREEN CLOUDS AND BLACK SWANS

It has been said that the essence of science is cumulative.[25] The emergence of new software and computing models, including cloud computing and the increased emphasis on data and data-centric applications, combined with exciting advances in technology—such as 3D-stacked nonvolatile memories, optics, and more power-efficient computation—provide one such opportunity for cumulative benefits. Such an opportunity represents a potential *black swan event*—a high-impact, infrequent event that in hindsight is very predictable—that presages future system architecture designs.[26]

One trend that is both logical and disruptive is colocating computing closer to the data, which will in turn lead to new building blocks for future systems. As stand-alone building blocks, a large number of individual nanostores can communicate over emerging optical interconnects and support large-scale distributed data-centric workloads. The key aspects of this approach are large-scale distributed parallelism and balanced energy-efficient compute in close proximity to the data. Together, these features allow nanostores to potentially achieve significantly higher performance at lower energy.

While such designs are promising, they are by no means inevitable, and several important design questions and challenges still remain to be addressed. Nanostores enable a rich architectural space that includes heterogeneous designs and integrated optics. There are also interesting opportunities for software optimizations including new interfaces and management of persistent datastores.

The improvements in performance, energy efficiency, and density in future system architectures will likely enable new applications across multiple larger, diverse data sources; the corresponding hardware-software codesign also provides rich opportunities for future research.

This article is intended to fuel the discussion that is needed in the broader community to start examining new, more disruptive, alternate architectures for future data-centric systems. C

Acknowledgments

This article greatly benefited from the input and support of several people, notably John Sontag, Jichuan Chang, Mehul Shah, David Roberts, Trevor Mudge, Niraj Tolia, Steven Pelley, Kevin Lim, Greg Astfalk, and Norm Jouppi.

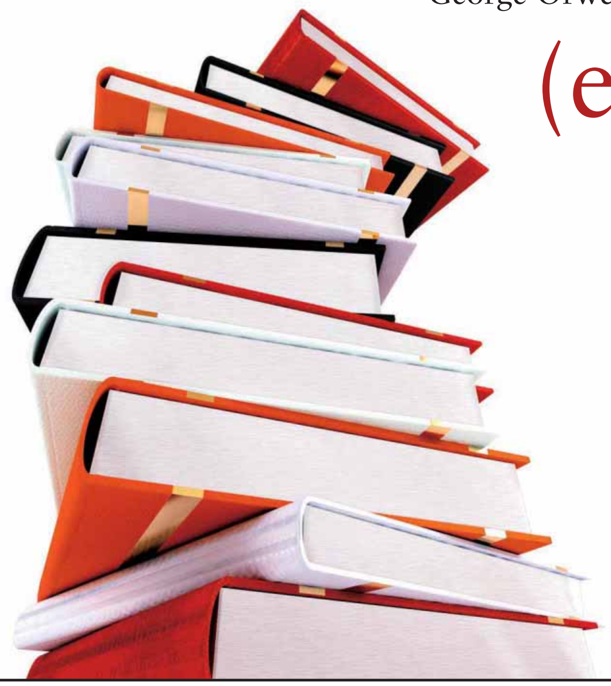
References

1. M. Mayer, “The Physics of Data,” Xerox PARC Forum Distinguished Lecture, 2009; www.parc.com/event/936/innovation-at-google.html.
2. R. Winter, “Why Are Data Warehouses Growing So Fast?” 2008; www.b-eye-network.com/view/7188.
3. P. Lyman and H.R. Varian, “How Much Information?” 2003; www2.sims.berkeley.edu/research/projects/how-much-info-2003.
4. R.S. Williams, “A Central Nervous System for the Earth,” *Harvard Business Rev.*, vol. 87, no. 2, 2009, p. 39.
5. M. Shah et al., “Data Dwarfs: Motivating a Coverage Set for Future Large Data Center Workloads,” *Proc. Workshop Architectural Concerns in Large Datacenters* (ACLD 10), 2010; sites.google.com/site/acldisca2010.
6. J. Gray, “E-Science: The Next Decade Will Be Exciting,” 2006; http://research.microsoft.com/en-us/um/people/gray/talks/ETH_E_Science.ppt.
7. B. Zhai et al., “Energy-Efficient Near-Threshold Chip Multi-Processing,” *Proc. Int’l Symp. Low Power Electronics and Design* (ISLPED 07), IEEE Press, 2007, pp. 32-37.
8. C. Lam, “Cell Design Considerations for Phase Change Memory as a Universal Memory,” *Proc. Int’l Symp. VLSI Technology, Systems, and Applications* (VLSI-TSA 08), IEEE Press, 2008, pp. 132-133.
9. D.B. Stukov et al., “The Missing Memristor Found,” *Nature*, vol. 453, 2008, pp. 80-83.
10. ITRS Roadmap, www.itrs.net/links/2009ITRS/Home2009.htm.
11. B.C. Lee et al., “Architecting Phase Change Memory as a Scalable DRAM Alternative,” *Proc. 36th Ann. Int’l Symp. Computer Architecture* (ISCA 09), ACM Press, 2009, pp. 2-13.
12. D. Lewis and H-H. Lee, “Architectural Evaluation of 3D Stacked RRAM Caches,” *Proc. IEEE 3D System Integration Conf.*, IEEE Press, 2009, pp. 1-4.
13. K. Moinuddin et al., “Scalable High-Performance Main Memory System Using Phase-Change Memory Technology,” *Proc. 36th Ann. Int’l Symp. Computer Architecture* (ISCA 09), ACM Press, 2009, pp. 24-33.
14. R. Bez, U. Russo, and A. Redaelli, “Nonvolatile Memory Technologies: An Overview,” *Proc. Workshop Technology Architecture Interaction: Emerging Technologies and Their Impact on Computer Architecture*, 2010, pp.44-65.
15. N. Jouppi and Y. Xie, tutorial, “Emerging Technologies and Their Impact on System Design,” *Proc. 15th Int’l Conf. Architectural Support for Programming Languages and Operating Systems* (ASPLOS 10), 2010; www.cse.psu.edu/~yuanxie/ASPLOS10-tutorial.html.
16. J. Ahn et al., “HyperX: Topology, Routing, and Packaging of Efficient Large-Scale Networks,” *Proc. Conf. High-Performance Computing Networking, Storage and Analysis* (SC 09), ACM Press, 2009, pp. 1-11.
17. J. Meza et al., “Lifecycle-Based Data Center Design,” *Proc. ASME Int’l Mechanical Eng. Congress & Exposition* (IMECE 10), Am. Soc. Mechanical Engineers, 2010; www.asmeconferences.org/congress2010.
18. E. Riedel et al., “Active Disks for Large-Scale Data Processing,” *Computer*, June 2001, pp. 68-74.
19. D. Patterson et al., “A Case for Intelligent DRAM: IRAM,” *IEEE Micro*, vol. 17, no. 2, 1997, pp. 33-44.

20. J. Ousterhout et al., "The Case for RAMClouds: Scalable High-Performance Storage Entirely in DRAM," *ACM SIGOPS Operating Systems Rev.*, vol. 43, no. 4, 2009, pp. 92-105.
21. J. Chang et al., "Is Storage Hierarchy Dead? Co-located Compute-Storage NVRAM-Based Architectures for Data-Centric Workloads," tech. report HPL-2010-114, HP Labs, 2010.
22. J. Dean, "Challenges in Building Large-Scale Information Retrieval Systems," keynote talk, *Proc. 2nd Ann. ACM Conf. Web Search and Data Mining* (WSDM 09), ACM Press, 2009; http://wsdm2009.org/proceedings.php.
23. J. Condit et al., "Better I/O through Byte-Addressable, Persistent Memory," *Proc. ACM SIGOPS 22nd Ann. Symp. Operating Systems Principles* (SOSP 09), ACM Press, 2009, pp. 133-146.
24. B. Dally, "Moving the Needle: Effective Computer Architecture Research in Academe and Industry," keynote talk, *Proc. 37th Int'l Symp. Computer Architecture* (ISCA 10), ACM Press, 2010; www.hipeac.net/node/2903.
25. R. Hamming, "You and Your Research," 1986; www.cs.virginia.edu/~robins/YouAndYourResearch.pdf.
26. P. Ranganathan, "Green Clouds, Red Walls, and Black Swans," keynote presentation, *Proc. 7th IEEE Int'l Conf. Autonomic Computing* (ICAC 10), IEEE Press, 2010; www.cis.fiu.edu/conferences/icac2010.

***Parthasarathy Ranganathan** is a distinguished technologist at HP Labs, where he is the principal investigator for the exascale datacenter project. His research interests are in systems architecture and energy efficiency. Partha received a PhD in electrical and computer engineering from Rice University. Contact him at partha.ranganathan@hp.com.*

cn Selected CS articles and columns are available for free at http://ComputingNow.computer.org.



COVER FEATURE

Bridging the Interconnection Density Gap for Exascale Computation

Daniel S. Stevenson and Robert O. Conn, *RTI International*

Three-dimensional silicon interposers are the basis for an architectural approach that bridges the interconnection density between FR-4 printed circuit boards and silicon ICs. The promise is increased performance and less complexity relative to monolithic silicon-based designs.

The next generations of high-performance computational hardware pose daunting challenges for system designers. Power must come down while processor-to-memory and processor-to-processor bandwidth goes up, and the distance between components must shrink substantially. Although several architectural approaches are on the horizon, designs featuring the use of large-area, 3D silicon interposer (3DSI) technologies, traditional bare die, and the newest 3D integrated circuits (ICs) of stacked bare die appear to be the most promising. Indeed, in our view, the assembly of bare die on various silicon substrates appears to be the most efficient way to meet the design challenges driven by exascale computing and other high-performance applications that are likely to appear within the next decade.

For the past five years, we have been researching the application of 3DSI technology to high-performance computational systems, exploring such requirements as fault tolerance for exascale system designs. Initially, we focused on reconfigurable computer designs, but more recently, we have been examining traditional high-performance computational systems. From our work, we have developed considerable insight into power and signal distribution, thermal management, vertical module integration, and module interconnection. Building on that insight, we have evolved the idea of a multistrata 3DSI, in which each stratum is optimized for a unique function.

The idea of attaching unpackaged silicon chips directly to silicon wafers, and other substrates, to create miniaturized circuit boards is not new. However, until recently, assemblies have been limited to 1 sq. in. circuit boards primarily because of two major manufacturing problems: the delamination of metal traces from the silicon surface during accelerated aging and the yield loss in systems comprising partially tested die.

Several developments are lifting that size restriction. Improvements in the fabrication process and the industry's shift from aluminum to copper have resulted in substantially better metal trace adhesion. The development of failure-tolerant structures in the patterned metal is now enabling the production of reliable large-area (up to 100 × 125 mm) circuit board equivalents from silicon wafers. Finally, bare die testing has come a long way in recent years, perhaps because of the demand for bare die in 3D applications. It is now possible to get fully tested bare memory die chips and to use nondestructive techniques to test unpackaged die.

0018-9162/11/$26.00 © 2011 IEEE

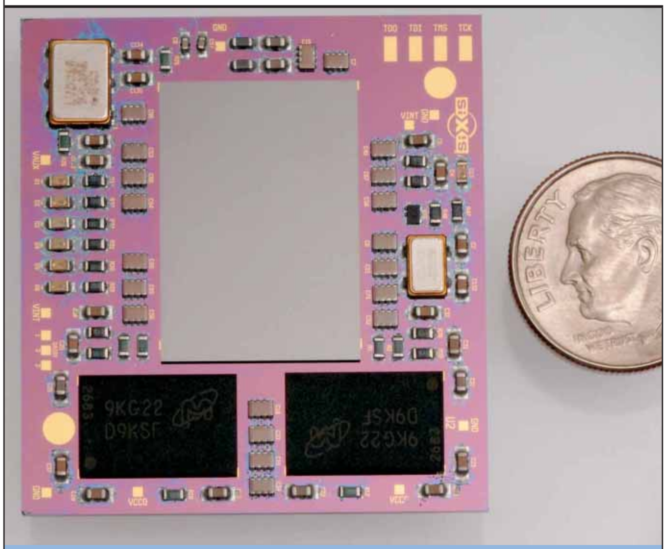


Figure 1. RTI-fabricated 3DSI with five metal layers. The design features one bare die field-programmable gate array, two packaged DDR2 memories, oscillators and passive devices, two topside signal and power distribution layers each, a single backside redistribution layer, and 80-μ TSVs. The overall size is 37 × 40 mm.

THE TECHNOLOGY

3DSI technology involves the use of standard complementary metal oxide semiconductor (CMOS) fabrication methods, primarily multilayer metal and through-silicon vias (TSVs), to manufacture silicon substrates for the interconnection of both packaged and unpackaged semiconductor devices. 3DSI technology is suitable for interconnection of the current generation of ICs as well as the 3D ICs that will appear in next-generation devices.

Current 3DSI designs typically feature power and signal distribution traces on both the front and back of the interposer and embedded passive devices, such as capacitors, resistors, and inductors. TSVs allow the vertical interconnection of front and backside metal layers. Future designs will incorporate active devices in the form of optoelectronics and power regulation and will be fabricated as multiple fusion-bonded silicon strata optimized for specific functions.

According to Philip E. Garrou of Microelectronic Consultants of North Carolina, only a few 3DSI designs larger than 25 × 25 mm have reached fabrication, and we are aware of only one announced commercial product using 3DSIs approaching this size (www.i-micronews.com/news/Xilinx-brings-3D-TSV-interconnects-commercialization-phase,5693.html).

Figure 1 shows an RTI-fabricated 3DSI for a third-party design.[1] RTI has also fabricated test wafers with large metal areas (100 × 120 mm) that have successfully undergone temperature-cycle accelerated-aging tests for 10,000 cycles.

HIGH-PERFORMANCE DEMANDS

3DSI technology and design involve a complex set of engineering tradeoffs to realize the driving applications' conflicting requirements, including

- *power distribution*—how to manage resistive losses and voltage fluctuations when passing large currents through thin films;
- *distribution of general-purpose signals less than 1 GHz*—how to manage signal integrity on the interposer surface and through the TSVs;
- *high-performance memory bus design*—how to manage signal integrity and propagation delay across multiple bus traces to meet the stringent requirements of double data rate, standard 3 (DDR3) memories;
- *electromagnetic radiation and interference*;
- *clock distribution*;
- *thermal management*—how to manage placement and cooling techniques to keep components within their operating temperature range;
- *parts partitioning*—how to partition parts between the silicon interposer and the package substrate, considering the fabrication process and material selection; and
- *system-reliability factors,* including how to test unassembled components and package assembled devices.

Decisions in any of these domains can have significant implications for performance in the others. For example, TSVs that are optimal for power perform poorly for high-speed signals and vice versa, which means that TSV size has many implications for power and signal distribution. Complicating this design tradeoff decision is the immaturity of the process for fabricating wafers with multiple via sizes. Another challenge is the very-high-current power distribution (as much as tens of amps for some parts), which is impossible to support over 1 cm of copper even 4 μ thick. TSVs connecting 3DSI power planes to FR-4 substrate power planes must be balanced with the local chips' power needs.

REQUIREMENTS FOR A NEAR-FUTURE SYSTEM

To explore the advantages of 3DSI, we considered a high-performance system that might be available in 5 to 10 years, adopting many of the system assumptions from the aggressive silicon-system strawman architecture described in DARPA's exascale computing study.[2]

As Figure 2 shows, the DARPA architecture defines a processing node and a processing group. The processing node consists of an 800-core processor chip connected to 16 DRAM chips (possibly in 4-die stacks). Each processing node interfaces to 12 router nodes, which provide a dragonfly interconnection topology.[3] A processing group comprises 12 processing nodes.

We can make four assumptions on the basis of the ITRS 2009 (www.itrs.net; system drivers and assembly) and the JEDEC DDR4 (www.simmtester.com/page/news/shownews.asp?num=13216) roadmaps:

- multicore processing die with approximately 800 cores;
- commercial off-the-shelf stacked DDR4 memory die with 640 data bus pins, supporting more than 8 gigatransfers per second and greater than 4-GHz clock rates;
- off-chip serializer-deserializer circuit operation at 40 Gbps; and
- TSV aspect ratios of 10:1.

Reducing signal distance

Performance was one of the original motivations for shrinking the packaging. As the original Cray-1 proved, reducing the distance that signals must travel increases clock rates, improves performance, and reduces power. Cray-1 designers bent the system chassis into a distinctive cylindrical shape and placed speed-dependent portions of the system on the axis-facing-end of PCBs to exploit shorter wire lengths. ICs were so close that no wire needed to be more than four feet long (www.cray.com/About/History.aspx). Building on this basic idea, designers can use 3D packaging techniques to reduce the distance and associated delay between critical components. 3D interposers, for example, can reduce the maximum distance between memory and CPUs to less than 10 mm, substantially improving memory performance.

Reduction of signal propagation delays is arguably the most significant area in which 3DSI technology can have a positive impact. Properly applied, the technology can make systems faster by making them more compact, thus reducing the distances that signals must travel between components. Both vertical integration and the elimination of chip packaging contribute to making 3D ICs more compact. Also, for typical ball-grid-array devices, getting rid of the packages eliminates solder balls and other sources of signal discontinuities.

Intramodule interconnection

The interconnection should aim to minimize the number of electrical connections into and out of the computational module while maximizing bandwidth and providing adequate power and ground connections.

To meet that goal, we would use several design approaches in concert: integrated power and passives, interposer specification and design partitioning that maximizes component interconnection within the interposer, and high-speed serializer-deserializer connections rather than parallel buses (to the extent possible). We would also use fiber optics for module-to-module interconnection and connection to high-speed peripherals, such as a redundant array of independent disks with Infiniband interfaces.

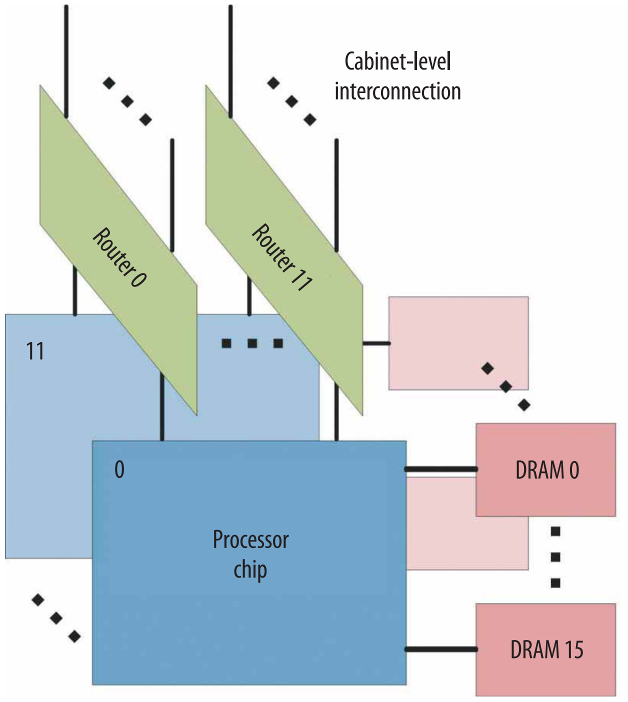


Figure 2. Architecture of a high-performance system likely to appear in the next decade. Each processing node has an 800-core chip connected to 16 DRAM chips. Each processing group consists of 12 processing nodes.

To reach the goals for the exascale processing node, local memory should be as close as possible to the associated computational die. The memory layout bus should minimize lateral signal routing and use TSVs optimized to support the expected 4-GHz clock rates.

In the exascale architecture, processing nodes would communicate with each other through serialization-deserialization signals carried over fiber-optic cables. To minimize signal integrity loss, the required optoelectrical conversion devices, such as laser modulators, drivers, receiver arrays, and trans-impedance amplifiers, need to be less than 10 mm from the computational die. The routing of signals between the processor die and optoelectronic converters could take place on the topside metal layers in the interposer and would not require TSVs.

A notional floorplan

We have worked out a series of methods for interfacing multimode fiber to 3DSIs using directly modulated vertical-cavity surface-emitting lasers (VCSELs). For the sample exascale architecture, bandwidth between processing nodes argues for the use of modulators and

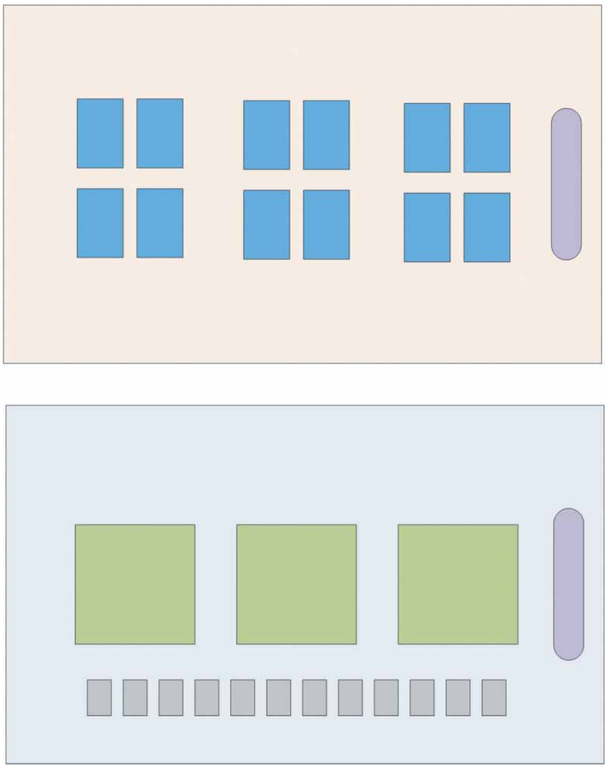


Figure 3. A notional floorplan for 3DSIs supporting three processing nodes and memory. The top side depicts the processor die, along with discrete fiber-optic interfaces to the router nodes. The backside depicts three groups of stacked memory die directly opposite the processor die, which aims to minimize memory bus length. The oval areas represent arrays of contact pads for general-purpose I/O and power input.

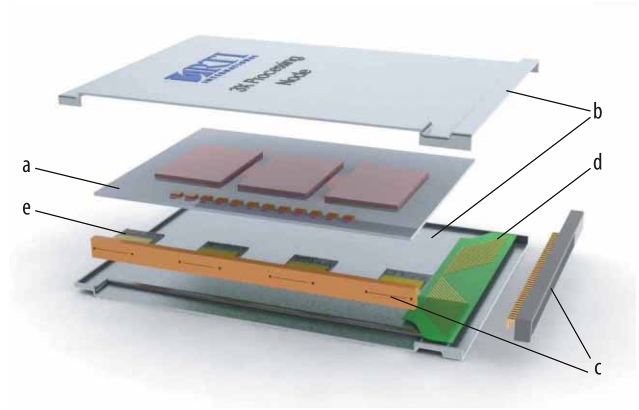


Figure 4. Exploded view of the processing node 3DSI as a packaged device. The view shows (a) the interposer, (b) clamshell lids, and (c) optical and electrical connectors. The 3DSI's edge and the electrical connector attach to (d) a small FR-4 board. The optical connector uses a fiber ribbon terminated in (e) a V-groove chip to provide coupling to optical transceivers on the 3DSI.

single-mode guided-wave techniques for on-interposer optical signal distribution and single-mode fiber for connections between processing and router nodes. To minimize fiber counts and connector size, we anticipate the use of coarse wavelength-division multiplexing, which would require planar waveguide structures on the 3DSI for multiplexing and demultiplexing multiple wavelengths onto individual fibers.

Figure 3 shows a notional floorplan for the 3D interposer (topside and backside) that illustrates these ideas. The floorplan assumes the use of multiple optical devices, but it might be possible to use a single silicon photonic IC instead.

Figure 4 is an exploded diagram depicting a potential package design for the processing node 3DSI, in which all high-speed (greater than 1 GHz) I/O is carried over fiber interfaces. An electrical connector provides connections for input power and low-speed electrical I/O, such as test and boundary scan interfaces. Although the diagram does not show heat sinks, the design assumes air cooling, which is a major design objective of DARPA for its exascale efforts. We have also looked at packaging and mechanical designs for 3DSI applications that assume cold plates for heat removal.

We favor a power distribution that is based on embedding power converters into the 3DSI. This approach implies the fabrication of active devices in the 3DSI, but enables a power input of 48 Vdc. We distribute this low current supply laterally to multiple point-of-load power converters, which then step down the supply to the low voltages that the various die attached to the 3DSI require. Our approach thus maximizes the vertical distribution of high current power while minimizing resistive losses.

Figure 5 shows both the 2D and 3D interconnection of a 12-node processing group and a 12-node routing group. Because the dragonfly architecture requires the full interconnection of processing and routing nodes, each processor is connected to each node. Each group is instantiated as four packaged devices edge-mounted to a PCB that provides for low-speed signal interconnection and regulated power input. Optical fiber to the router system carries all high-speed signals for processor-node to router-node communication. We estimate a volume of about 1,000 cc for the combined router and processing groups.

INTERCONNECTION FOR A PROCESSING NODE ARRAY

For architectures with directly interconnected processing nodes, as in a hypercube, the design can further reduce external interconnection rather than communicating through a router. In one proposed architecture,[4] a macrochip integrates a 9 × 9 processing node array (vertically integrated multicore CPU and memory) with a silicon photonic network for local message routing.

In this direct interconnection arrangement, interconnections fall into three domains: on-chip, on-module and system, and on-interposer.

On-chip interconnection

On-chip interconnection is at the lithographic limits of today's IC technology, with copper line widths below 100 nm. However, to connect to current packaging substrates, I/O pads must be 100 µ on a side—about a thousand times larger than the signal traces. Chip vendors planning to use 3DSIs can build next-generation ICs with pads matched to the much smaller 3DSI dimension (as low as a 10-µ pad pitch) and increase I/O density, as well as improve performance through reduced I/O parasitics. Packaging limitations in current large computational chips restrict I/O pins to about 1,000 to 2,000. To maximize bandwidth and minimize power, we expect exascale processing chips to have on the order of 5,000-10,000 I/O pins.

On-module and system interconnections

Transmission-line copper will remain the predominant medium for on-module chip interconnection and will likely continue to be the medium of choice for traces on a PCB and for signal and power connectors on packaged parts. Line width should be no more than 100 µ but is more often 150 µ. Signal wires are large enough that wire resistance—at nearly zero ohms—is not usually a consideration. Thus, high-frequency transmission effects on signals are a concern, and the PCB design must pay special attention to transmission lines, termination resistors, differential signals, and reference voltages.

To deal with intermodule signals, we recommend using traditional copper signaling and coping with attendant transmission line issues. Because most intermodule signals will first go through a TSV on a 3DSI, signal speed will be somewhat limited but will suffice for below 1 GHz, such as for system housekeeping and monitoring. High-speed (greater than 1 GHz) data lines can also use copper with more severe 3DSI design constraints. Next-generation TSVs will need to support full-speed (greater than 12 GHz) signals.

In a high-density system using 3DSIs, size is a primary concern, so the highest-speed interconnection should be optical. First, copper connectors are too large for systems with 3DSIs. Second, relative to copper connectors, designers can pack much more bandwidth into an optical cable of a given dimension. In general, 3DSIs can route all high-speed intermodule and system signals into optical cables, which can then be connected as desired to other system modules or parts. Designers should locate optoelectronic components that support optical-to-electrical-to-optical (OEO) conversions adjacent to serialization-deserialization I/O on the processing devices.

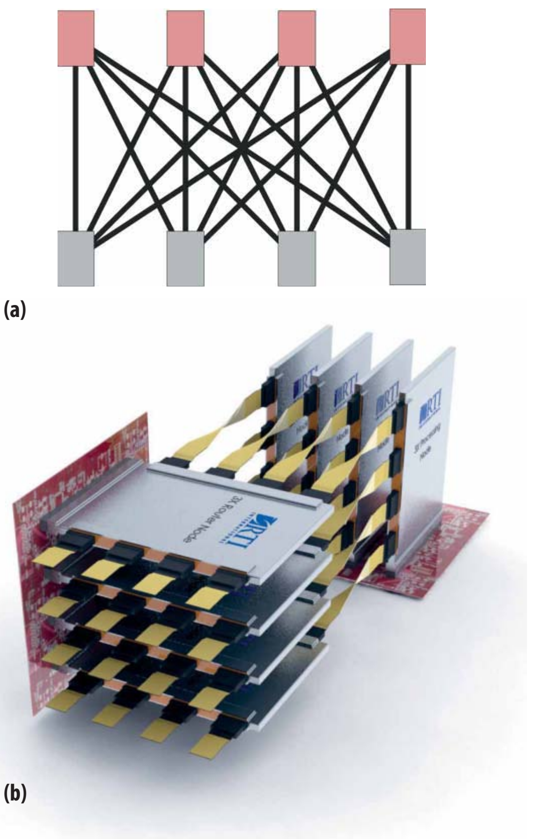


Figure 5. 2D versus 3D interconnection of a 12-node processing group and 12-node router group. Relative to (a) a traditional 2D approach, (b) the 3D mechanical design simplifies the implementation of interconnection topology, eliminating many problematic wire crossovers. In (b), fiber ribbon stubs represent connections to cabinet-level routing.

On-interposer interconnection

On-interposer layout is a new design domain without well-established design rules. Because signal trace dimensions are between those of an IC and an FR-4 PC board, trace widths are roughly from 5 µ to 10 µ, and line thicknesses are 1 µ to 2 µ, very high interconnect density is possible. In a 3DSI, designers can route 1,000 signals off the bottom of an FPGA die in two layers, as opposed to six layers for a packaged part on an FR-4 PC board.

The 3DSI signal traces are small—typically 8 µ × 2 µ —making them somewhat resistive even over several millimeters. From this view alone, 3DSI signal speed would appear to be slower per centimeter of wire than the FR-4 signal speed. However, because the distances on the 3DSI are substantially shorter than those on the FR-4, actual 3DSI propagation delays between devices can be less than those for the FR-4.

By modeling 3DSI signal lines in the resistance-capacitance domain, designers can avoid or minimize most problems associated with transmission lines, such as

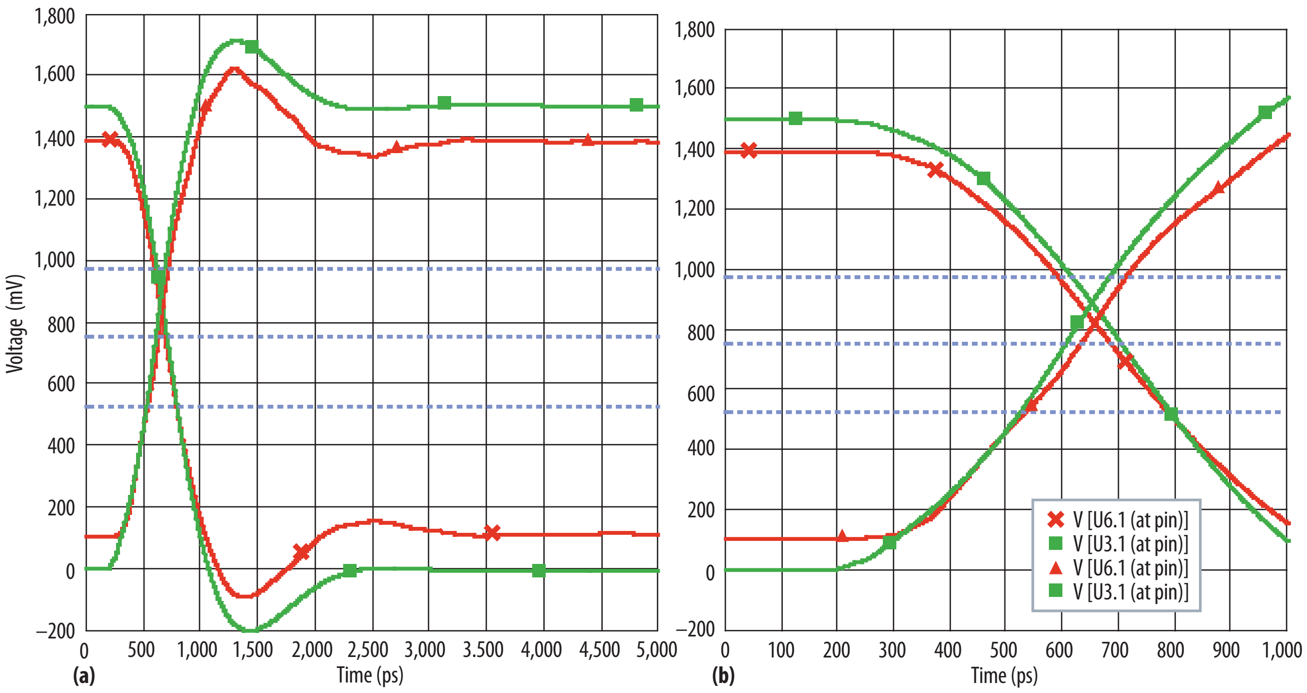


Figure 6. Comparison of 3DSI and FR-4 PCB (50 ohms) signal propagation characteristics. Results are for (a) overshoot over 3,500 ps and (b) timing over 900 ps. I/O voltage is 1.5 V. The 2-cm 3DSI wire (green) is 8 μ × 2 μ over 4 μ of silicon dioxide over a ground plane. The FR-4 wire (red) is 4 cm long by 100 μ × 10 μ.The 3DSI is actually 65 ps faster than the graphs indicate, since the model used to run the comparison includes approximately 65 ps of package transit time.

termination resistors, extra I/O power, cross-talk, and length matching. By adjusting the width of a copper signal line, they can control the time delay down the line. Thus, long address lines would be wider than short address lines rather than the awkward PCB technique of zigzagging a wire to add length.

On the basis of past efforts[5] to model for the signal integrity of 1-GHz signals traversing TSV, diameters are likely to be between 10 μ and 20 μ. A maximum TSV aspect ratio of 10:1 would give a 100-μ silicon thickness for signals, and such thinned wafers would require special handling. As an alternative, we have developed a concept for high-integrity 12-GHz signals traversing a 60-μ TSV on a 500-μ-thick 3DSI that we plan to evaluate through modeling and experimental validation.

In the future, glass might replace silicon in fabricating 3D interposers. Glass is advantageous because of its electronic properties, such as being a pure insulator rather than a semiconductor, which would improve high-speed signal distribution. Glass also has a potentially lower cost because panel processing methods already exist from LCD monitor fabrication. However, at present, silicon substrates are likely to prevail because of their larger manufacturing infrastructure base and overall process maturity.[6] We envision a time when glass displaces silicon interposers in cost-sensitive consumer electronics applications. For high-performance systems, the need to incorporate active components in the interposer for power conversion, and regulation and for optoelectronic conversion, plus the presence of high-speed signals, argue for using both glass and silicon in a stratified interposer.

SIGNAL AND POWER INTEGRITY ISSUES

Figure 6 shows comparative signal propagation characteristics for a 3DSI and an FR-4 PCB wire. (Although our simulation used a 2-cm 3DSI wire and 4-cm FR-4 PCB wire, a 3DSI wire is typically much less than half the length of an equivalent FR-4 wire.) We optimized the FR-4 signals as data sheet transmission lines, and the 3DSI signals as an unterminated resistance-capacitance wire. The IBIS output driver model we used for the comparative simulation is from the Xilinx Virtex 5 LVCMOS15_F_16.

The DARPA exascale report[2] discusses the need for orders-of-magnitude power reduction to meet the demands of high-performance computation. Boosting system power efficiency by a factor of 10 will require significant work on chip architecture, particularly on memory devices. According to private communication with Paul Franzon, an Exascale Study Group member, one way to reduce memory access power is to dramatically increase the number of I/O lines between memory and CPU.

Packaging the architecture using 3DSI technology can result in some savings. With shorter signal traces, less power is required for signals greater than 1 GHz. Essentially, because less of the energy launched down a short trace is lost as electromagnetic radiation, less launch power is required to reach the destination.

In our experience, layout for 3D ICs entails design considerations that fall between those used on PCB design, where 50- and 100-ohm design rules are common, and ASICs, where trace impedance is rarely a design concern. 3DSIs, with their resistive wires, provide more options in designing for signal integrity, and often make it possible to avoid termination resistors, resulting in even more power savings. In fact, in a case study in which we compared a design with 3DSIs with one using FR-4 transmission lines, we estimated a 30 percent reduction in module power.

MULTISTRATA DESIGN

In geology, a stratum is a layer of rock or soil with internally consistent characteristics that distinguish it from contiguous layers. We have applied this concept to an approach for designing and fabricating 3DSIs that will meet the performance requirements for exascale computation. Our approach uses five strata within the interposer, with each stratum optimized for a unique functional role—power distribution, power regulation, optical interconnection, passive devices, and signal distribution. It is possible to fabricate each stratum individually using processes and materials appropriate to its function, and to test each stratum before it is fusion-bonded into a single 3DSI structure.

Power distribution

Even with vertical power integration, the metal traces in this layer must be as thick as possible, which with existing 3DSI technology is between 2 µ and 4 µ. Future generations will allow a thickness of 10 µ or more. The insulator layer between power/ground planes in this stratum must be thin with a high dielectric constant to maximize capacitance.

This built-in decoupling capacitance in the interposer minimizes the need to use discrete capacitors in the module design to provide clean power. A dielectric that is 0.5 µ thick ($k = 4$) can provide 70 pF of capacitance per square mm. Power TSVs must be designed for low resistance, which implies TSVs with a large diameter relative to that of signal TSVs. In contrast, current power TSVs are one-size-fits-all. A possible compromise is to use small-diameter TSVs for signal routing and, for power routing, to use same-size multiple vias instead of a single TSV with a larger diameter.

Power regulation

The high current requirements of components, such as the current generation of CPUs, GPUs, and FPGAs, require minimizing lateral power distribution to avoid excess resistive (I2R) losses. Future chip designs with even lower voltage logic levels make this problem more acute. An effective way to deal with it is to integrate power regulation into an interposer stratum that enables options for vertical power. To minimize system I2R drops, our preferred approach is to provide input power to the regulator strata as a high-voltage (12-, 24-, or 48-V) source, which we can also use to provide any required lateral power distribution.

This stratum contains multiple integrated switching or linear regulator devices near each point of load, providing localized regulation that in turn offers opportunities to manage system power—from lowering voltages when things are slow to turning off unused sections. Additionally, bringing power into the device at, say, 48 Vdc reduces the power pin count from hundreds or thousands to tens.

Because less of the energy launched down a short trace is lost as electromagnetic radiation, less launch power is required to reach the destination.

Passive devices

Integrating resistors, capacitors, and inductors into the 3DSI can have several benefits. In addition to reducing the number of discrete parts required during assembly, integration into a dedicated layer makes it possible to reduce the distance among passives and the active parts to which they connect. We can place bypass capacitors in the vertical stack between the chip drawing the power and the power regulators. The bypass capacitors can be large-area plate capacitors or barrel capacitors in TSVs. This decreased distance substantially lowers the inductance typically seen in power circuits, resulting in a much smaller capacitance requirement. Such integrated bypass capacitors can be as close as 1 mm from an actual I/O drive transistor.

Optical

The use of serialization-deserialization signals for external interconnection (network interfaces, hard drives, and module interconnection) will increase clock rates for these signals. Signal reach in the electrical domain (without regeneration) becomes an increasingly challenging design task for 3DSIs. Optical interconnection using both single- and multimode optics is a viable option because the available bandwidth-distance product for fiber is three to six orders of magnitude greater than in copper, according to ISO/IEC 11801. We prefer to place OEO devices as close as is practical to serialization-deserialization sources. This preference, along with the emergence of high-speed CMOS photonic transceivers,[7] suggests the use of active silicon photonic devices in this stratum.

Signal distribution

For clock rates of 1 GHz or higher, low capacitance and controlled impedance are major design considerations for stackup (material selection and thickness) and layout. In these cases, signal energy tends to be confined to the outer 0.5 µ or less of a trace (skin effect), so trace metal can be thinner than what is required for power distribution. Minimizing capacitance between traces is also important, so the stratum's signal layers should have a thick, low-*k* dielectric that separates the power and signal traces.

INTERFACING TO TRADITIONAL AND ORGANIC CIRCUIT BOARDS

We have successfully experimented with several options for interfacing 3DSIs to traditional FR-4 PCBs and organic circuit boards, including the use of anisotropic conductive materials as a high-density connector and 3DSIs packaged as a traditional ball-grid-array device. The major design consideration for this interface is the mechanical stresses that arise from the coefficient-of-thermal-expansion (CTE) mismatch between silicon and the FR-4 PCB.

For ball-grid-array devices, the CTE mismatches between the silicon-organic interface impose a limit on package size. Over multiple power cycles, metal fatigue leads to solder ball fracture and part failure. Underfill materials can mitigate this problem, but they can also complicate the manufacturing process. Our solution is to use a high-density connector that enables easy removal of an interposer for rework, repair, and testing. Commercially available anisotropic conductive films come close to meeting this need, but are less suitable for rework than we prefer and are not suitable for the preassembly testing of bare die and 3DSIs.

The ideal material is an anisotropic conductive film that would permanently adhere to the I/O pads on a 3DSI and would have conductor densities to support a 10-µ pad pitch. To provide a system connector function, the circuit-board side of the ideal film would not be adhesive, but would maintain contact with I/O through low-pressure clamping. For testing, the film would adhere to a fixture designed to test bare die or to preassembled interposer strata. To our knowledge, this material does not yet exist, but RTI's work on a patterned grid of nanomaterial pads shows promise.

COOLING OPTIONS

Clearly, getting power into a vertically integrated system poses a challenge. Removing the resulting heat is another design issue with many tradeoffs and possible approaches. Our sample processing module assumes air cooling, and we have studied packaging architectures based on the use of cold plates. These designs have vertically stacked processing nodes, with cold plates filling the space between adjacent nodes. The result is higher density for the electronics and the potential for lower signal transit latency between nodes. We see examples of the cold-plate approach in test heads of large semiconductor chip testers.

Designs based on microfluidic cooling are also interesting. Fluid channels etched into the bulk silicon of vertically integrated die,[8] a stratum of a 3DSI, or a combination of these are all possible approaches.

Multistrata 3DSIs that integrate both active and passive components provide signal and power distribution that addresses many of the design challenges inherent in exascale computational systems. Relative to PCB and organic substrates, 3DSI-based designs provide higher density routing and I/O as well as improved reliability and repairability.

In light of these advantages, the technology is gaining interest. Silicon photonics and 3D glass interposers are active research topics, and we expect to see them become viable and practical technologies incorporated into 3DSIs in the near future. Considering the potential performance advantages of 3DSI-based designs relative to other alternatives, we expect to see many more practical designs using this complex but promising technology. C

Acknowledgments

We thank Bill Batchelor, Dorota Temple, Dean Malta, Alan Huffman, and Paul Franzon for their critical technical feedback on the concepts described in this article.

References

1. E. Vick, S. Goodwin, and D. Temple, "Electrical Demonstration of TSV Interconnects and Multilevel Metalization for 3D Si Interposer Applications," *Proc. 43rd Int'l Symp. Microelectronics* (IMAPS 10), Int'l Microelectronics and Packaging Soc., 2010, pp. 7-14.
2. P. Kogge et al., "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," Sept. 2008; http://users.ece.gatech.edu/mrichard/ExascaleComputingStudyReports/exascale_final_report_100208.pdf.
3. J. Kim et al., "Technology-Driven, Highly-Scalable Dragonfly Topology," *Proc. 35th Int'l Symp. Computer Architecture* (ISCA08), IEEE CS Press, 2008, pp. 77-88.
4. A. Krishnamoorthy et al., "Computer Systems Based on Silicon Photonic Interconnects," *Proc. IEEE*, July 2009, pp. 1337-1361.
5. F. Liu et al., "Electrical Characterization for 3D Through-Silicon-Vias," *Proc. 60th Electrical Components and Technology Conf.*, IEEE Press, 2010, pp. 1100-1105.

6. Yole Development, “3D Glass and Silicon Interposers: Technologies, Applications, and Markets, ” Sept. 2010; www.i-micronews.com/reports/3D-Silicon-Glass-Interposers/156.
7. T. Pinguet et al., “40-Gbps Monolithically Integrated Transceivers in CMOS Photonics,” *Proc. Soc. Photo-Optical Eng.* (SPIE 08), vol. 6898, SPIE, 2008, pp. 689805-689805-14.
8. C. Hidrovo and K.E. Goodson, *Active Microfluidic Cooling of Integrated Circuits, in Electrical, Optical and Thermal Interconnections for 3D Integrated Systems*, Artech, 2008, pp. 293-330.

***Daniel S. Stevenson** is a senior research engineer at RTI International’s Center for Materials and Electronic Technologies. His research interests include application of 3D silicon interposers, high-performance computational and communications systems, and information assurance. Stevenson received an MS in physics from the University of North Carolina at Chapel Hill. He is a senior member of IEEE. Contact him at danstevenson@rti.org.*

***Robert O. Conn** is a senior research engineer at RTI International’s Center for Materials and Electronic Technologies. His research interests include the application of 3D silicon interposers, large-scale reconfigurable computing systems, and sustainable energy systems. Conn received a BSEE in electrical engineering from the University of California, Berkeley. He is a member of IEEE. Contact him at rconn@rti.org.*

cn Selected CS articles and columns are available for free at http://ComputingNow.computer.org.

IEEE computer society

PURPOSE: The IEEE Computer Society is the world's largest association of computing professionals and is the leading provider of technical information in the field.
MEMBERSHIP: Members receive the monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.
COMPUTER SOCIETY WEBSITE: www.computer.org
OMBUDSMAN: To check membership status or report a change of address, call the IEEE Member Services toll-free number, +1 800 678 4333 (US) or +1 732 981 0060 (international). Direct all other Computer Society-related questions—magazine delivery or unresolved complaints—to help@computer.org.
CHAPTERS: Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.
AVAILABLE INFORMATION: To obtain more information on any of the following, contact Customer Service at +1 714 821 8380 or +1 800 272 6657:

- Membership applications
- Publications catalog
- Draft standards and order forms
- Technical committee list
- Technical committee application
- Chapter start-up procedures
- Student scholarship information
- Volunteer leaders/staff directory
- IEEE senior member grade application (requires 10 years
- practice and significant performance in five of those 10)

PUBLICATIONS AND ACTIVITIES

***Computer*:** The flagship publication of the IEEE Computer Society, *Computer*, publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.
Periodicals: The society publishes 13 magazines, 18 transactions, and one letters. Refer to membership application or request information as noted above.
Conference Proceedings & Books: Conference Publishing Services publishes more than 175 titles every year. CS Press publishes books in partnership with John Wiley & Sons.
Standards Working Groups: More than 150 groups produce IEEE standards used throughout the world.
Technical Committees: TCs provide professional interaction in more than 45 technical areas and directly influence computer engineering conferences and publications.
Conferences/Education: The society holds about 200 conferences each year and sponsors many educational activities, including computing science accreditation.
Certifications: The society offers two software developer credentials. For more information, visit www.computer.org/certification.

NEXT BOARD MEETING
2–4 Feb. 2011, Long Beach, Calif., USA

IEEE

revised 2 Dec. 2010

EXECUTIVE COMMITTEE

President: Sorel Reisman*
President-Elect: John W. Walz*
Past President: James D. Isaak*
VP, Standards Activities: David Alan Grier (1st VP)*
Secretary: Jon Rokne (2nd VP)*
VP, Educational Activities: Elizabeth L. Burd*
VP, Member & Geographic Activities: Sattupathu V. Sankaran†
VP, Publications: David Alan Grier*
VP, Professional Activities: James W. Moore*
VP, Technical & Conference Activities: John W. Walz*
Treasurer: Frank E. Ferrante*
2011–2012 IEEE Division VIII Director: Susan K. (Kathy) Land, CSDP†
2010–2011 IEEE Division V Director: Michael R. Williams†
***Computer* Editor in Chief:** Carl K. Chang†
**voting member of the Board of Governors* *†nonvoting member of the Board of Governors*

BOARD OF GOVERNORS

Term Expiring 2011: Elisa Bertino, Jose Castillo-Velázquez, George V. Cybenko, Ann DeMarle, David S. Ebert, Hironori Kasahara, Steven L. Tanimoto
Term Expiring 2012: Elizabeth L. Burd, Thomas M. Conte, Frank E. Ferrante, Jean-Luc Gaudiot, Paul K. Joannou, Luis Kun, James W. Moore
Term Expiring 2013: Pierre Bourque, Dennis J. Frailey, Atsuhiro Goto, André Ivanov, Dejan S. Milojicic, Jane Chu Prey, Charlene (Chuck) Walrad

EXECUTIVE STAFF

Executive Director: Angela R. Burgess
Associate Executive Director; Director, Governance: Anne Marie Kelly
Director, Finance & Accounting: John Miller
Director, Information Technology & Services: Ray Kahn
Director, Membership Development: Violet S. Doan
Director, Products & Services: Evan Butterfield
Director, Sales & Marketing: Dick Price

COMPUTER SOCIETY OFFICES

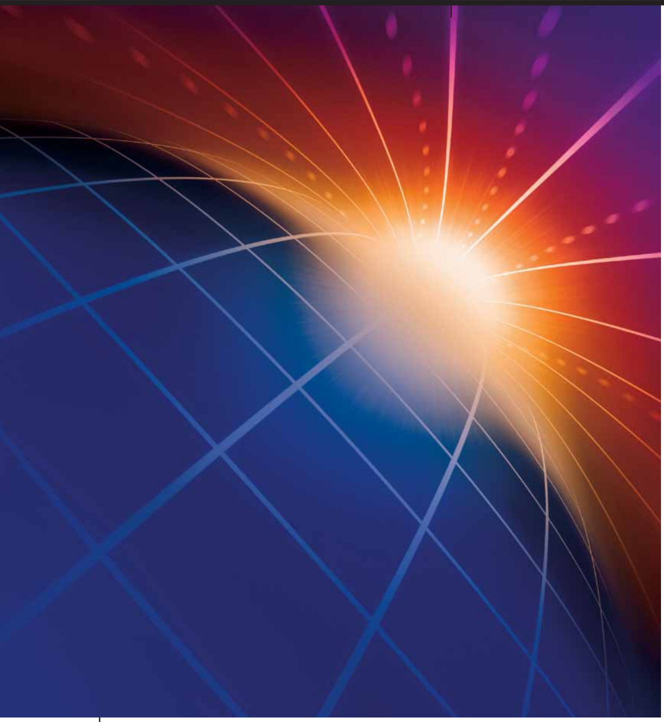
Washington, D.C.: 2001 L St., Ste. 700, Washington, D.C. 20036
Phone: +1 202 371 0101 • **Fax:** +1 202 728 9614
Email: hq.ofc@computer.org
Los Alamitos: 10662 Los Vaqueros Circle, Los Alamitos, CA 90720-1314
Phone: +1 714 821 8380
Email: help@computer.org

MEMBERSHIP & PUBLICATION ORDERS

Phone: +1 800 272 6657 • **Fax:** +1 714 821 4641
Email: help@computer.org
Asia/Pacific: Watanabe Building, 1-4-2 Minami-Aoyama, Minato-ku, Tokyo 107-0062, Japan
Phone: +81 3 3408 3118 • **Fax:** +81 3 3408 3553
Email: tokyo.ofc@computer.org

IEEE OFFICERS

President: Moshe Kam
President-Elect: Gordon W. Day
Past President: Pedro A. Ray
Secretary: Roger D. Pollard
Treasurer: Harold L. Flescher
President, Standards Association Board of Governors: Steven M. Mills
VP, Educational Activities: Tariq S. Durrani
VP, Membership & Geographic Activities: Howard E. Michel
VP, Publication Services & Products: David A. Hodges
VP, Technical Activities: Donna L. Hudson
IEEE Division V Director: Michael R. Williams
IEEE Division VIII Director: Susan K. (Kathy) Land
President, IEEE-USA: Ronald G. Jensen



Apoptotic Computing: Programmed Death by Default for Computer-Based Systems

Roy Sterritt, *University of Ulster*

Inspired by the cellular self-destruct mechanisms in biological apoptosis, apoptotic computing offers a promising means to develop self-managing computer-based systems.

At the 2009 International Joint Conference on Artificial Intelligence, researchers warned that the nightmare scenarios depicted in sci-fi films such as *2001: A Space Odyssey*, the *Terminator* and *Matrix* series, *Minority Report*, and *I, Robot* could come true. “Scientists fear a revolt by killer robots” proclaimed the UK’s *Sunday Times*,[1] which highlighted alarming findings at the conference that mankind might lose control of computer-based systems that carry out a growing share of society’s workload, from chatting on the phone to waging war, and have already reached a level of indestructibility comparable with the cockroach. For instance, unmanned predator drones, which can seek out and kill human targets, have already moved out of the movie theatre and into the theatre of war in Afghanistan and Iraq. While presently controlled by human operators, these drones are moving toward more autonomous control. Similar devices may also soon appear above city streets to carry out domestic surveillance. Samsung, the South Korean electronics giant, has developed autonomous sentry robots with “shoot to kill” capability to serve as armed border guards.[1]

To provide for this future, the Apoptotic Computing project has been working since 2002 toward the long-term goal of *programmed death by default* for computer-based systems.[2-6] Motivated by the apoptosis mechanisms in multicellular organisms, apoptotic computing can be considered a subarea of bio-inspired computing, natural computing, or autonomic systems. Two example applications are autonomic agent-based environments and swarm space exploration systems.

BIOLOGICAL APOPTOSIS

Developing a self-managing computer system is the vision of autonomic computing.[7-9] As the “Autonomic System Properties” sidebar explains, an autonomic computing system is analogous to the biological nervous system, which automatically maintains homeostasis (metabolic equilibrium) and controls responsiveness to external stimuli. For example, most of the time you are not consciously aware of your breathing rate or how fast your heart is beating, while touching a sharp knife with your finger results in a reflex reaction to move the finger out of danger.[10]

If you cut yourself and start bleeding, you treat the wound and carry on without thinking about it, although pain receptors will induce self-protection and self-configuration to use the other hand. Yet, often the cut will have caused skin cells to be displaced down into muscle tissue.[11] If the cells survive and divide, they have the potential to grow into a tumor. The body’s solution to this situation is cell self-destruction (with mounting evidence

0018-9162/11/$26.00 © 2011 IEEE

that some forms of cancer are the result of cells not dying fast enough, rather than multiplying out of control, as previously thought).

Biologists believe that cells are programmed to commit suicide through a controlled process known as *apoptosis*.[12] The term is derived from the Greek word for "to fall off," in reference to dead leaves falling from trees in autumn; likewise, cells "fall off" the living organism and die. As Figure 1a shows, a cell's constant receipt of "stay alive" signals turns off the self-destruct sequence.[3] When these signals cease, the cell starts to shrink, internal structures decompose, and all internal proteins degrade; thereafter, the cell breaks into small, membrane-wrapped fragments to be engulfed by phagocytic cells for recycling. Figure 1b contrasts apoptosis, also known as "death by default,"[11] with *necrosis*, the unprogrammed death of a cell due to injury—inflammation and the accumulation of toxic substances.[13]

Recent research indicates that cells receive orders to kill themselves when they divide.[12] The reason appears to be self-protection. An organism relies on cell division for maintenance and growth, but the process is also dangerous: if just one of the billions of cells in a human body locks into division, the result is a tumor. The suicide and reprieve controls can be likened to the dual keys of a nuclear missile: the suicide signal (first key) turns on cell growth but at the same time activates a sequence that leads to self-destruction, while the reprieve signal (second key) overrides the self-destruct sequence.[14]

AUTONOMIC AGENTS

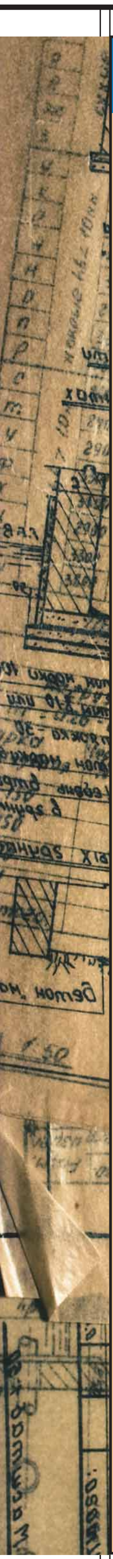
Autonomic computing depends on many disciplines for its success; not least of these is research in agent technologies. There are no assumptions that an autonomic architecture must use agents, but agent properties—adaptability, autonomy, cooperation, and so on—complement the paradigm's objectives. In addition, there are arguments for designing complex systems with multiple agents,[15] providing such systems with inbuilt redundancy and greater robustness,[16] and for retrofitting legacy systems with autonomic capabilities that may benefit from an agent-based approach.[17]

Figure 2 shows a basic *autonomic element* (AE), which consists of a *managed component* (MC) and an *autonomic manager* (AM).[18] The AM can be a stationary agent—for instance, a self-managing cell[19] that contains functionality for measurement and event correlation and provides support for policy-based control. AMs communicate via means such as self-* event messages.

Mobile agents can also play a role in autonomic systems. Their ability to reduce network load, overcome network latency, encapsulate protocols, execute asynchronously and autonomously, adapt dynamically, reflect natural heterogeneity, and maintain robustness and fault tolerance can make it easier for AMs within different systems to cooperate.

APOPTOSIS IN AGENT-BASED AUTONOMIC ENVIRONMENTS

Michael S. Greenberg and colleagues first proposed agent destruction to facilitate security in mobile-agent



AUTONOMIC SYSTEM PROPERTIES

The general properties of an autonomic, or self-managing, computing system consist of four objectives that represent broad system requirements, and four attributes that identify basic implementation mechanisms.[1,2]

An autonomic system has the following objectives:

- *Self-configuration*. The system must be able to readjust itself automatically, either to support a change in circumstances or to assist in meeting other system objectives.
- *Self-healing*. In reactive mode, the system must effectively recover when a fault occurs, identify the fault, and, when possible, repair it. In proactive mode, the system monitors vital signs to predict and avoid health problems, or to prevent their reaching undesirable levels.
- *Self-optimization*. The system can measure its current performance against the known optimum and has defined policies for attempting improvements. It can also react to the user's policy changes within the system.
- *Self-protection*. The system must defend itself from accidental or malicious external attacks, which requires an awareness of potential threats and the means to manage them.

To achieve these self-managing objectives, a system must be

- *self-aware*—aware of its internal state;
- *self-situated*—aware of current external operating conditions and context;
- *self-monitoring*—able to detect changing circumstances; and
- *self-adjusting*—able to adapt accordingly.

Thus, to be autonomic a system must be aware of its available resources and components, their ideal performance characteristics, and current status. It must also be aware of interconnection with other systems, as well as rules and policies for adjusting as required. Operating in a heterogeneous environment requires relying on open standards to communicate with other systems.

These mechanisms do not exist independently. For example, to successfully survive an attack, the system must exhibit self-healing abilities, with a mixture of self-configuration and self-optimization. This not only ensures the system's dependability and continued operation but also increases self-protection from similar future attacks. Self-managing mechanisms must also ensure minimal disruption to users.

References

1. R. Sterritt, "Towards Autonomic Computing: Effective Event Management," *Proc. 27th Ann. IEEE/NASA Software Eng. Workshop* (SEW 02), IEEE CS Press, 2002, pp. 40-47.
2. R. Sterritt and D. Bustard, "Autonomic Computing—A Means of Achieving Dependability?" *Proc. 10th IEEE Int'l Conf. and Workshop Eng. of Computer-Based Systems* (ECBS 03), IEEE CS Press, 2003, pp. 247-251.

systems.[20] They described an event in which network operators decommissioned a computer named omega.univ.edu and moved its work to other machines. A few years later, the operators assigned a new computer the old name and, to everyone's surprise, e-mail arrived, much of it three years old; the mail had survived "pending" on Internet relays waiting for omega.univ.edu to come back up. Greenberg's team considered a similar scenario in which mobile agents—not rogue agents but ones carrying proper authenticated credentials—carried out work that was out of context rather than the result of abnormal procedures or system failure. In this circumstance, the mobile agents could cause substantial damage—for example, deliver an archaic upgrade to part of the network operating system, bringing down the entire network.

Misuse involving mobile agents can occur in several forms. Agents can accidentally or unintentionally misuse hosts due to, say, race conditions or unexpected emergent behavior in those agents. In addition, external bodies acting upon agents, either deliberately or accidentally, can lead to their misuse by hosts or other agents—for example, due to damage, breaches of privacy, harassment, social engineering, event-triggered attacks, or compound attacks.

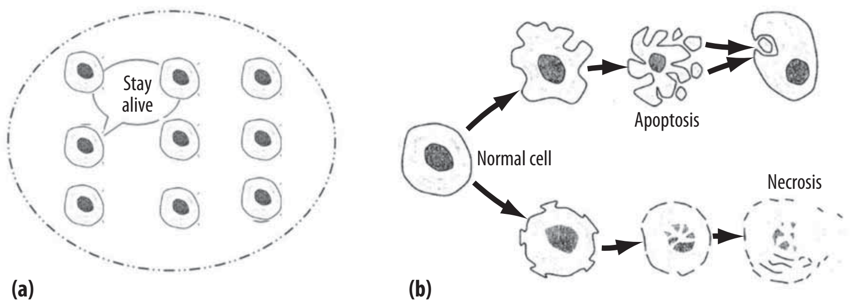


Figure 1. Biological apoptosis. (a) A cell's constant receipt of "stay alive" signals turns off its programmed self-destruct sequence. (b) Apoptosis versus necrosis due to injury.

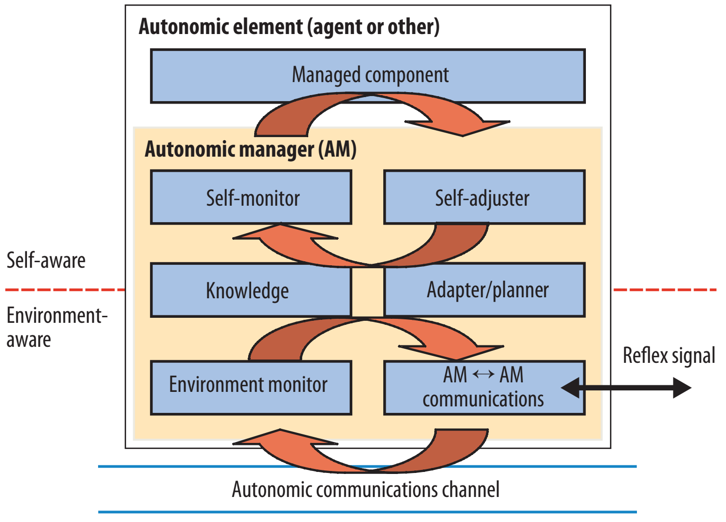


Figure 2. An autonomic element consists of a managed component and an autonomic manager, which can be a stationary agent. The AM ↔ AM communications module includes heartbeat monitoring and pulse monitoring. AMs communicate through an autonomic channel via such means as self-* event messages.

Encryption can prevent situations in which portions of an agent's binary image—monetary certificates, keys, information, and so on—could be copied when visiting a host. However, agent execution requires decryption, which provides a window of vulnerability.[20] This situation is analogous to the body's vulnerability during cell division.[3]

Figure 3 shows a high-level view of a simple autonomic environment with three AEs (a typical system has hundreds, thousands, or even millions of AEs). Each AE is an abstract view of Figure 1, and in this case the MCs represent self-managing computer systems. These AEs can have many other lower-level AEs—for example, an autonomic manager for the disk drive—while at the same time residing within the scope of a higher-level AM such as a system-wide local area network domain's AE.

Within each AM, heartbeat monitors (HBMs) send "I am alive" signals to ensure the continued operation of vital processes in the MC and to immediately indicate if any fail. The AM has a control loop that continually monitors and adjusts, if necessary, metrics within the MC, yet vital processes in the MC can also be safeguarded by an HBM that emits a heartbeat signal as opposed to its being polled by the AM, avoiding lost time (time to next poll) by the AM to notice a failure (note in Figure 3 that the left-hand AE has an HBM between the AM and a process on the MC).

Because each AM is aware of its MC's health via the continuous control loop, it can share this information by sending a pulse signal ("I am un/healthy") to another AM—in Figure 3, for example, from the left-hand AE to the middle AE. This not only allows self-managing options if the machines are, say, sharing workload as a cluster but protects the AM itself as the pulse signal also acts as

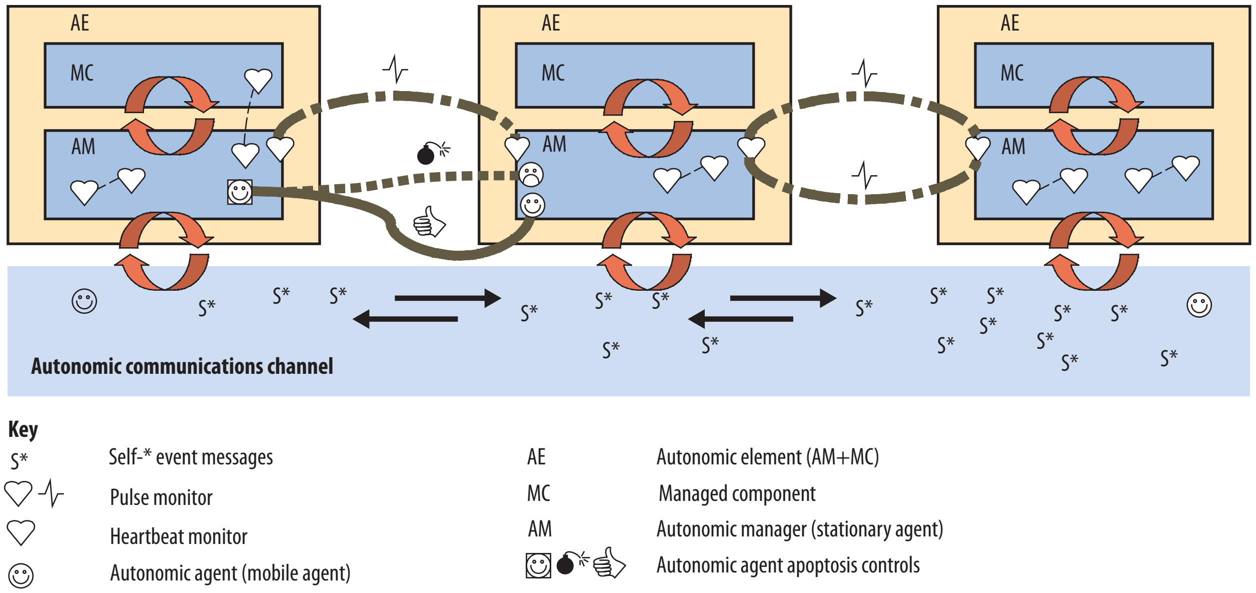

AE
MC
AM
AE
MC
AM
AE
MC
AM
S*
Autonomic communications channel
Key
S* Self-* event messages
Pulse monitor
Heartbeat monitor
Autonomic agent (mobile agent)
AE Autonomic element (AM+MC)
MC Managed component
AM Autonomic manager (stationary agent)
Autonomic agent apoptosis controls

Figure 3. Simple autonomic environment consisting of AEs with autonomic agents (stationary and mobile), heartbeat monitors ("I am alive"), pulse monitors ("I am un/healthy"), and apoptosis controls ("stay alive/self-destruct").

an HBM signal from one AM to another. Thus, if an AE's vital process fails, the neighboring AM will immediately become aware of it and, for example, try to restart the failed AE or initiate a failover to another AM. This pulse signal can also act as a reflex signal between AMs warning of an immediate incident—a more direct solution than the AM's processing numerous event messages to eventually determine an urgent situation.

Because AMs also monitor the external environment (the second control loop), they have a view of their local environment's health. They can encode such information into the pulse signal along with self-health data (just as our hearts have a double beat). The double-pulse signals between the right-hand and center AEs in Figure 3 represent this situation.

AMs can dispatch mobile agents to work on their behalf—for example, to update a set of policies. To help provide self-protection in these situations, AMs can send apoptosis signals ("stay alive/self-destruct") to such agents by either authorizing continued operation or by withdrawing such authorization—for example, if the policies become out of date. Figure 3 depicts both scenarios.

We refer to the absence of a "stay alive" signal resulting in agent self-destruction as *strong* apoptotic computing, or programmed death by default, while *weak* apoptotic computing involves an explicit self-destruct signal—similar in principle to the garbage collection method first used by Lisp and by many languages since or the destructor method in object orientation. The differences in these approaches are subtle but important. Only a built-in default death can guarantee true system safety. For example, you would never rely on a self-destruct signal getting through to an agent containing system password updates in a hostile environment. Likewise, a robot with adaptive capabilities could learn to ignore such a signal. That said, clearly not all circumstances require a death-by-default mechanism. However, we believe that many researchers using *programmed death* under the apoptosis descriptor should be using programmed death by default.

There is a concern that denial-of-service attacks could prevent "stay alive" signals from reaching their target and thereby induce unintentional agent self-destruction. DoS attacks could likewise interrupt terminate signals, resulting in potentially dangerous scenarios. DoS-immune architectures are thus a critical part of next-generation self-managing systems.

SWARM SPACE EXPLORATION SYSTEMS

Space exploration missions by necessity have become increasingly autonomous and adaptable. To develop more self-sustainable exploration systems, NASA is investigating the use of biologically inspired swarm technologies.[3] As Figure 4 shows, the idea is that swarms of small spacecraft offer greater redundancy (and, consequently, greater protection of assets), lower costs and risks, and the ability to explore more remote regions of space than a single large craft.

The Autonomous NanoTechnology Swarm mission (http://ants.gsfc.nasa.gov), a collaboration between NASA's Goddard Space Flight Center and its Langley Research Center, exploits swarm technologies and AI techniques to develop revolutionary architectures for both spacecraft and surface-based rovers. ANTS consists of several submissions:

- The Saturn Autonomous Ring Array consists of a swarm of 1,000 pico-class spacecraft, organized as 10 subswarms with specialized instruments, to perform in situ exploration of Saturn's rings to better understand their constitution and how they were formed. SARA will require self-configuring structures for nuclear propulsion and control as well as autonomous operation for both maneuvering around Saturn's rings and collision avoidance.
- The Prospecting Asteroid Mission (PAM) also involves 1,000 pico-class spacecraft but with the aim of exploring the asteroid belt and collecting data on particular asteroids of interest for potential future mining operations.
- The Lander Amorphous Rover Antenna (LARA) will implement new NASA-developed technologies in the field of miniaturized robotics, which could form the basis of remote lunar landers launched from remote sites, as well as offering innovative techniques to allow rovers to move in an amoeboid fashion over the moon's uneven terrain.

The ANTS architecture emulates the successful division of labor exhibited by low-level social-insect colonies. In such colonies, with sufficiently efficient social interaction and coordination, a group of specialists usually outperforms a group of generalists. To accomplish their specific mission goals, ANTS systems likewise rely on large numbers of small, autonomous, reconfigurable, and redundant worker craft that act as independent or collective agents.[21] The architecture is self-similar in that ANTS system elements and subelements can be structured recursively,[22] and it is self-managing, with at least one ruler (AM) per ANTS craft.

NASA missions such as ANTS provide a trusted private environment, eliminating many agent security issues and enabling system designers to focus on ensuring that agents are operating in the correct context and exhibiting emergent behavior within acceptable parameters.

In considering the role of the self-destruct property inspired by apoptosis, suppose one of the worker craft in the ANTS mission was operating incorrectly and, when coexisting with other workers, was causing undesirable emergent behavior and failing to self-heal correctly. That emergent behavior could put the mission in danger, and ultimately the ruler would withdraw the "stay alive" signal.[3] Likewise, if a worker or its instrument was damaged, either by colliding with another worker or (more likely) an asteroid, or during a solar storm, the ruler would withdraw the "stay alive" signal and request a replacement worker. Another worker would then self-configure to take on the role of the lost worker to ensure optimal balanced coverage of tasks to meet the scientific goals. If a ruler or messenger was similarly damaged, its ruler would withdraw the "stay alive" signal and promote a worker to play its role.

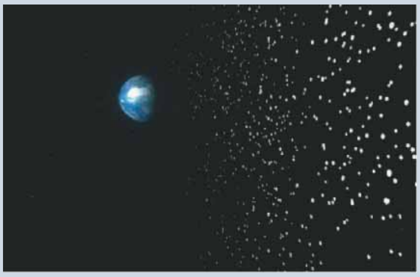


Figure 4. NASA's new space exploration paradigm calls for missions involving thousands of small spacecraft rather than a single large craft. Image courtesy of NASA.

THE EVOLVING STATE OF THE ART

Several researchers have investigated the apoptotic computing concept and its potential applications.

Christian Tschudin initially suggested using apoptosis in highly distributed systems.[23]

James Riordan and Dominique Alessandri proposed apoptosis as a means to automatically counter the increasing number of security vulnerabilities that hackers publish and exploit before systems administrators can close them.[24] They described an apoptosis service provider that, should a system vulnerability be found, could release a message into the environment to trigger various preconfigured responses to shut down the system or warn a responsible party.

Leszek Lilien and Bharat Bhargava argued for apoptosis as a means to secure atomic bundles of private data, in which the process is activated when detectors determine a credible threat exists to the bundle by any host, including the bundle's destination.[25]

In drawing parallels between biology and computing, Steve Burbeck proposed four interconnected principles for managing evolving systems, one of which is apoptosis.[26] As an example of this principle, he cited the Blue Screen of Death, a programmed response to an unrecoverable error. Burbeck argued that a computer, like a metazoan cell, should be able to sense its own rogue behavior, such as downloading uncertified code, and disconnect itself from the network.

M.M. Olsen, N. Siegelmann-Danieli, and H.T. Siegelmann developed a multiagent system called HADES that can protect itself via "life" protocols—which control the replication, repair, movement, and self-induced death of each agent—and a "rescue" protocol.[27]

Madihah Mohd Saudi and colleagues researched apoptosis with respect to security systems, focusing on network problems, and later applied it specifically to worm attacks.[28,29]

Finally, David Jones implemented apoptotic self-destruct and "stay alive" signaling while investigating memory requirements in inheritance versus an abstract-oriented approach.[30]

The majority of these applications fall into the weak apoptotic computing (programmed death) category, and would likely benefit from, instead, utilizing a strong (programmed death by default) approach. They also highlight the strong need for standards and trust requirements, with the immediate challenge of developing a DoS-resistant architecture.

The human body regulates vital functions such as heartbeat, blood flow, and cell growth and death, all without conscious effort. We must develop computer-based systems that can perform similar operations on themselves without constant human intervention.

Promising apoptotic computing applications have been developed for data objects, highly distributed systems, services, agent systems, and swarm systems. However, more applied work is needed in other areas, and researchers must address the challenges around trust—until then, users are not likely to embrace a system with self-destruct capabilities.

The case has been made that all computer-based systems should be autonomic.[31] Likewise there is a compelling argument that all such systems should be apoptotic, especially as computing becomes increasingly pervasive and ubiquitous. Apoptotic controls should cover all levels of human-computer interaction from data, to services, to agents, to robotics. With recent headline incidents of credit card and personal data losses by organizations and governments, and scenarios once relegated to science fiction becoming increasingly possible, programmed death by default is a necessity.

We are rapidly approaching the time when new autonomous computer-based systems and robots should undergo tests, similar to ethical and clinical trials for new drugs, before they can be introduced. Emerging research from apoptotic computing could guide the safe deployment of such systems. C

Acknowledgments

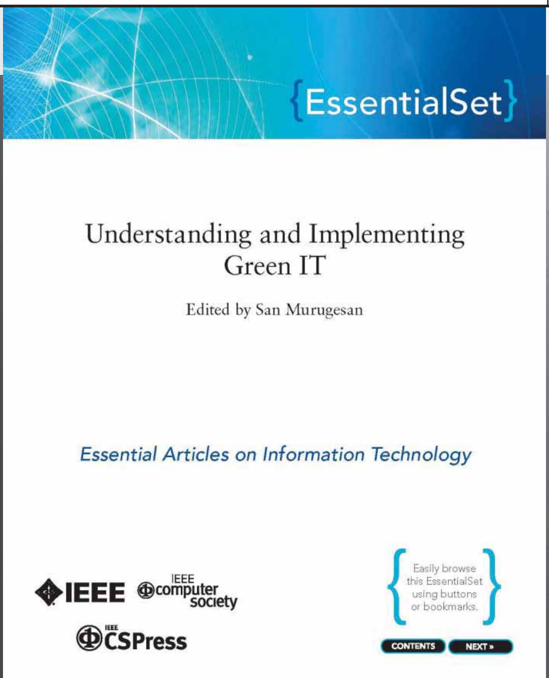
The author is supported by the University of Ulster's Computer Science Research Institute and School of Computing and Mathematics. Some of the research described in this article is patented and patent-pending by Roy Sterritt and Mike Hinchey (Lero—the Irish Software Engineering Research Centre) through NASA and assigned to the US government. In memory of our dear friend and colleague Scott Hamilton.

References

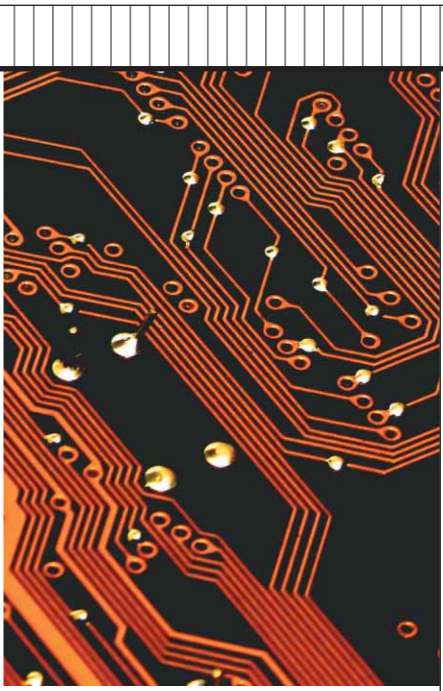
1. J. Arlidge, "Scientists Fear a Revolt by Killer Robots," *The Sunday Times*, 2 Aug. 2009; http://technology.timesonline.co.uk/tol/news/tech_and_web/article6736130.ece.
2. R. Sterritt and M. Hinchey, "SPAACE IV: Self-Properties for an Autonomous & Autonomic Computing Environment—Part IV: A Newish Hope," *Proc. 7th IEEE Int'l Conf. and Workshops Eng. of Autonomic and Autonomous Systems* (EASe 10), IEEE CS Press, 2010, pp. 119-125.
3. R. Sterritt and M. Hinchey, "Apoptosis and Self-Destruct: A Contribution to Autonomic Agents?" *Proc. 3rd Int'l Workshop Formal Approaches to Agent-Based Systems* (FAABS 04), LNCS 3228, Springer, 2004, pp. 262-270.
4. R. Sterritt and M. Hinchey, "Engineering Ultimate Self-Protection in Autonomic Agents for Space Exploration Missions," *Proc. 12th IEEE Int'l Conf. and Workshops Eng. of Computer-Based Systems* (ECBS 05), IEEE CS Press, 2005, pp. 506-511.
5. R. Sterritt and M. Hinchey, "From Here to Autonomicity: Self-Managing Agents and the Biological Metaphors That Inspire Them," *Proc. 8th Int'l Conf. Integrated Design and Process Technology* (IDPT 05), Soc. for Design and Process Science, 2005, pp. 143-150.
6. R. Sterritt and M. Hinchey, "Biologically-Inspired Concepts for Autonomic Self-Protection in Multiagent Systems," M. Barley et al., eds., *Safety and Security in Multi-Agent Systems: Research Results from 2004-2006*, LNCS 4324, Springer, 2009, pp. 330-341.
7. J.O. Kephart and D.M. Chess, "The Vision of Autonomic Computing," *Computer*, Jan. 2003, pp. 41-52.
8. M.G. Hinchey and R. Sterritt, "Self-Managing Software," *Computer*, Feb. 2006, pp. 107-109.
9. S. Dobson et al., "Fulfilling the Vision of Autonomic Computing," *Computer*, Jan. 2010, pp. 35-41.
10. R.A. Lockshin and Z. Zakeri, "Programmed Cell Death and Apoptosis: Origins of the Theory," *Nature Reviews Molecular Cell Biology*, July 2001, pp. 542-550.
11. Y. Ishizaki et al., "Programmed Cell Death by Default in Embryonic Cells, Fibroblasts, and Cancer Cells," *Molecular Biology of the Cell*, Nov. 1995, pp. 1443-1458.
12. J. Klefstrom, E.W. Verschuren, and G. Evan, "c-Myc Augments the Apoptotic Activity of Cytosolic Death Receptor Signaling Proteins by Engaging the Mitochondrial Apoptotic Pathway," *J. Biological Chemistry*, 8 Nov. 2002, pp. 43224-43232.
13. M. Sluyser, ed., *Apoptosis in Normal Development and Cancer*, Taylor & Francis, 1996.
14. J. Newell, "Dying to Live: Why Our Cells Self-Destruct," *Focus*, Dec. 1994; http://members.fortunecity.com/templarseries/Yahoo/Omegaman/apoptosi.html.
15. N.R. Jennings and M. Wooldridge, "Agent-Oriented Software Engineering," J. Bradshaw, ed., *Handbook of Agent Technology*, AAAI/MIT Press, 2000.
16. M.N. Huhns, V.T. Holderfield, and R.L.Z. Gutierrez, "Robust Software via Agent-Based Redundancy," *Proc. 2nd Int'l Joint Conf. Autonomous Agents and Multiagent Systems* (AAMAS 03), ACM Press, 2003, pp. 1018-1019.
17. G. Kaiser et al., "Kinesthetics eXtreme: An External Infrastructure for Monitoring Distributed Legacy Systems," *Proc. Autonomic Computing Workshop—5th Ann. Int'l Workshop Active Middleware Services* (AMS 03), IEEE Press, 2003, pp. 22-30.

18. R. Sterritt and D. Bustard, “Towards an Autonomic Computing Environment,” *Proc. 14th Int’l Workshop Database and Expert Systems Applications* (DEXA 03), IEEE CS Press, 2003, pp. 694-698.
19. E. Lupu et al., “AMUSE: Autonomic Management of Ubiquitous Systems for e-Health,” *Concurrency and Computation: Practice and Experience*, Mar. 2003, pp. 277-295.
20. M.S. Greenberg et al., “Mobile Agents and Security,” *IEEE Comm. Magazine*, July 1998, pp. 76-85.
21. P.E. Clark et al., “ANTS: A New Concept for Very Remote Exploration with Intelligent Software Agents,” *Eos, Trans., American Geophysical Union*, vol. 82, no. 47, 2001.
22. S. Curtis et al., “ANTS (Autonomous Nano Technology Swarm): An Artificial Intelligence Approach to Asteroid Belt Resource Exploration,” *Proc. 51st Int’l Astronautical Congress, Int’l Astronautical Federation*, 2000; http://ants.gsfc.nasa.gov/documents.d/iaf2000-ants.pdf.
23. C. Tschudin, “Apoptosis—The Programmed Death of Distributed Services,” J. Vitek and C. Jensen, eds., *Secure Internet Programming: Security Issues for Mobile and Distributed Objects*, LNCS 1603, Springer, 1999, pp. 253-260.
24. J. Riordan and D. Alessandri, “Target Naming and Service Apoptosis,” *Proc. 3rd Ann. Workshop Recent Advances in Intrusion Detection* (RAID 00), LNCS 1907, Springer, 2000, pp. 217-225.
25. L. Lilien and B. Bhargava, “A Scheme for Privacy-Preserving Data Dissemination,” *IEEE Trans. Systems, Man, and Cybernetics, Part A: Systems and Humans*, May 2006, pp. 503-506.
26. S. Burbeck, “Complexity and the Evolution of Computing: Biological Principles for Managing Evolving Systems,” v2.2, white paper, 9 Apr. 2007; http://evolutionofcomputing.org/Complexity%20and%20Evolution%20of%20Computing%20v2.pdf.
27. M.M. Olsen, N. Siegelmann-Danieli, and H.T. Siegelmann, “Robust Artificial Life via Artificial Programmed Death,” *Artificial Intelligence*, Apr. 2008, pp. 884-898.
28. M.M. Saudi et al., “An Overview of Apoptosis for Computer Security,” *Proc. Int’l Symp. Information Technology* (ITSim 08), IEEE Press, 2008, pp. 2534-2539.
29. M.M. Saudi et al., “An Overview of STAKCERT Framework in Confronting Worms Attack,” *Proc. 2nd IEEE Int’l Conf. Computer Science and Information Technology* (ICCSIT 09), IEEE Press, 2009, pp. 104-108.
30. D. Jones, “Implementing Biologically-Inspired Apoptotic Behaviour in Digital Objects: An Aspect-Oriented Approach,” MSc dissertation, Open University, UK, 2010; http://www.apoptotic-computing.org/media/Apoptotic-Computing-MSc-Dissertation.pdf.
31. R. Sterritt and M. Hinchey, “Why Computer-Based Systems Should Be Autonomic,” *Proc. 12th IEEE Int’l Conf. Workshops Eng. Computer-Based Systems* (ECBS 05), IEEE CS Press, 2005, pp. 406-414.

***Roy Sterritt** is a faculty member in the School of Computing and Mathematics and a researcher in the Computer Science Research Institute at the University of Ulster, Northern Ireland. His research focuses on the engineering of computer-based systems, in particular self-managing/autonomic systems. Sterritt received a BSc in computing and information systems and an MA in business strategy from the University of Ulster. He is a member of IEEE and the IEEE Computer Society. Contact him at r.sterritt@ulster.ac.uk.*



Using Modeling and Simulation to Evaluate Enterprises' Risk Exposure to Social Networks



Anna Squicciarini and Sathya Dev Rajasekaran, *Pennsylvania State University*
Marco Casassa Mont, *HP Labs*

An analytic methodology involving modeling and simulation could help decision makers determine how their employees' use of social networks impacts their organization, identify how to mitigate potential risks, and evaluate the financial and organizational implications of doing so.

As social networking becomes more pervasive within organizations and business environments, employers must be aware of the potential security and business risks. Enterprises face not only productivity loss due to employees' spending time on social networking activities but also the threat of information leakage caused by incautious posts or explicit references to private business information.[1] Employees might post information that could negatively impact the company's reputation, write complaints about internal organizational issues, or even directly defame the organization they work (or worked) for.

While social networking makes business information readily available to a broad audience, including customers, business competitors, and partners, hackers also can access such information, potentially gaining a competitive advantage and causing the targeted enterprise financial losses. The risk of attackers' exploiting social networking data warehouses is also on the rise as more tools are available to them, such as data aggregator and data mining tools.

Decision makers—including chief technology officers, chief information officers, and chief information security officers—are now "getting serious about social networks,"[2] and many companies are proactively studying this phenomenon to understand its possible benefits and risks. Companies can use social networks as a resource, for example, to do extensive background checking about potential employees or to promote their business. However, they're also exploring mitigating approaches to the potential risks of using these networks.

Initial mitigation measures include blocking the incriminated sites,[3] updating security policies to address admissible social network use, and introducing new rules and guidelines. However, these reactive approaches[4] often fail if decision makers don't understand the causes of the risks and the impact of choices made in attempting to mitigate them. For example, blocking social network sites from office machines helps reduce the amount of time employees spend on the sites, but it has no effect on employees working from home or using personal devices.

Current operational security guidelines for safe online behaviors often simply suggest using common sense and following generic processes. Clearly, such advice doesn't help prevent or mitigate possible attacks. Traditional risk-

Published by the IEEE Computer Society

assessment methodologies such as ISO 2700x[5] provide practical advice and suggest security-driven assessment processes. However, their recommendations and guidelines must be refined and contextualized to the social network problem.

We propose an alternative approach that helps decision makers reason about social network use and possible risk exposure and aids in exploring investment options to mitigate risks. Our approach uses modeling languages to represent the involved processes, users, and systems in the enterprise; the threat environments; and relevant security control points. It also simulates the implications of possible threats and the impact of adopting different type of controls.[6]

THREATS AND RISKS

In addition to using social engineering and phishing attacks,[7] attackers might indirectly obtain sensitive information from a single user's profile in a social network or by combining different pieces of information, obtained by cross-correlating multiple profiles from users belonging to one or more social networks.

The data that social network sites store can be attacked in several ways. Two major approaches are *vertical* attacks, which focus on a specific social network site, and *horizontal* attacks, which occur across multiple social networks. For each of these approaches, the external observer can focus on a specific person or a group of people.

Vertical attacks

A vertical attack targets the profiles of one or more people participating in a social network. For example, a user might post sensitive personal information in a profile, along with job information and data that breaches the enterprise's security and confidentiality policy—for example, salary or tax information. A profile often also includes other potentially sensitive information, such as full contact information, relationship status, and other social traits. In addition, attackers can infer interesting business information when employees simultaneously update their profiles. For example, consider a group of users working in the technology field and living in the Philadelphia area. Suppose that after a certain date, many of them start looking for job opportunities elsewhere. Such action might imply an upcoming crisis in their specific field.

Data aggregation can increase a vertical attack's effectiveness. There are at least two ways to aggregate data from different profiles. First, an attacker can build a comprehensive user profile by collecting attributes disclosed in the profiles of different users who share a subset of attributes. Attackers can easily create numerous ad hoc accounts and lure the victim to befriend them, giving them access to the victim's profiles. For example, consider a Facebook group, "Employees at the UAHO Company in Lewiston." If any two users reveal the office address and phone number, the work contact information is automatically known for all individuals in that group. As recent studies have disclosed, this is a powerful, yet underestimated, attack vector in social networking.[7]

Alternatively, an analysis of comments, hypertext, or public message boards can reveal important content. For example, by text-mining certain occurrences of words (such as a project name), attackers might partially infer information about the project's evolution, status, and outcome. Social applications and widgets can represent powerful tools to collect additional users' profile data.

A vertical attack targets the profiles of one or more people participating in a social network.

Horizontal attacks

Our analysis of horizontal attacks is complementary to the vertical analysis. While previous observations hold true, now the sources of information involve multiple social network sites. For example, with regard to aggregated data, an attacker can cross-correlate and complement the attributes of a user's profile by getting information from that user's profiles on other sites such as LinkedIn and Facebook, as long as the link between the profiles is obvious—for example, the same registered name.

In addition to actual hacking tools, social network sites provide tools that hackers can use to gather data. For example, Orkut (www.orkut.com) provides Polls, an application that lets an observer collect data on certain topics. A hacker could pose questions aimed at discovering a company's ongoing projects or its employee satisfaction level. Another potentially effective tool, Buzz (http://twitterbuzz.com), reveals a company's Twitter activity.

Companies or employees should own and hold their own profiles. However, at present, there is no effective control over the veracity of profile information. An attacker could easily create a bogus account and pretend to be an employee of a certain company. Social attacks could serve to gain additional information from targeted employees.

LinkedIn is also exploitable. For a small price, LinkedIn lets users access other users' profiles using fine-grained internal searching tools. For less than $25 per month, hackers can access user profiles—for example, by using a LinkedIn Premium account, even if the users aren't part of the hacker's network of friends and colleagues; conduct detailed searches and retrieve related information, references, and résumés; and read descriptions of work activities.

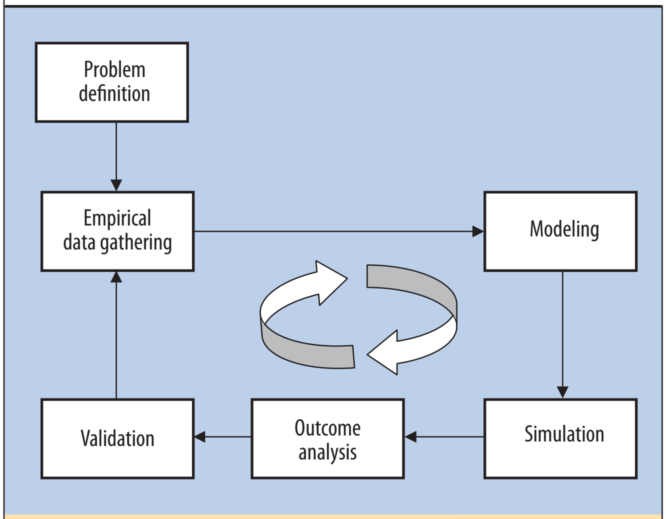


Figure 1. Overview of proposed methodology. The various steps in the methodology include defining the problem, gathering empirical data, modeling and simulation, and outcome analysis and validation. Multiple iterations might be required.

DECISION SUPPORT FOR RISK ASSESSMENT

Best security practices, risk-assessment guidelines, and methodologies described in documents such as ISO 2700x help decision makers explore and understand risks and how to act on them based on common sense, security, and risk-management criteria. However, these documents typically offer only general guidelines and must be interpreted and applied to a specific context. Decision makers might have to make assumptions about existing risks or threats—people's behaviors, organizational culture, attackers' motivations, and so on—types of data that could leak, and possible control points. Then they must analyze the implications of these possibilities, sometimes with no empirical data. Because of environmental variability, evaluating the impact of certain decisions—for example, investments in control points—might not be trivial.

Our methodology for assessing risks in enterprises uses the scientific method shown in Figure 1 and is based on initial R&D work in security and identity analytics.[6,8] In addition to empirical data gathering, modeling, simulation, outcome analysis, and verification and validation, our methodology might require multiple refinement steps.

Modeling and simulation techniques represent the involved system processes and human behaviors along with the relevant cause-and-effect relationships. This lets a decision maker explore the impact of different investment choices and controls on the outcomes in the domain of interest, such as data leakage, security risks, and costs. Changing assumptions and parameters can generate different predictions. In this context, modeling and simulation provide strategic decision support capabilities as they let decision makers identify the most suitable investment options. Enterprises can adopt this methodology to obtain support for risk assessment in the context of social networking.

A decision maker might be interested in the answers to the following key questions:

- What is the current level of risk exposure due to employees using social networks? What does this mean in terms of potential data leakage?
- What are the consequences, in terms of risks and costs, of making certain investment or policy decisions? What are the best investments? What impact would they have on data leakages?

In general, decision makers can act on different investment *levers* (that is, means of achieving desired outcomes) to change employees' behaviors, based on available resources and investments. Common levers used to enforce internal security policies include

- *Education*. Organizations can invest in educating employees about threats when using social networks, as well as correct behaviors.
- *Monitoring, awareness, and punishment*. Investments can target monitoring user behaviors (using auditing or logging), creating awareness of unacceptable behaviors, and punishing such behaviors based on collected evidence.
- *Technical control points*. An organization can invest in controls, such as software and hardware solutions to monitor, intercept, and block data leakages—for example, by using e-mail interceptors, digital rights management solutions, and black-listing and blocking of websites.

Implementing any of these levers isn't just a matter of technical or management issues. Decision makers must consider their economic impact as well as the consequences on employee productivity and morale.

In this context, a decision maker must understand, at least qualitatively, the impact employees have on data leakage, given a specific threat environment. A decision maker must also assess the impact of the available controls. Restrictive controls, such as strong forms of access control[9] or auditing, might be costly, and they might not even be effective in certain domains because people can find creative ways to bypass them. The involved population of employees might also negatively react to some controls, or certain controls might require an adaptation phase, with incremental risk mitigation.

Acting on different investment levers can have different outcomes, depending on the specific targeted threats, employee behavior (for example, compliant, noncompliant, or disruptive), and the types of accessed social networks. For a highly skilled, professional workforce, some additional education courses might provide the right level of awareness

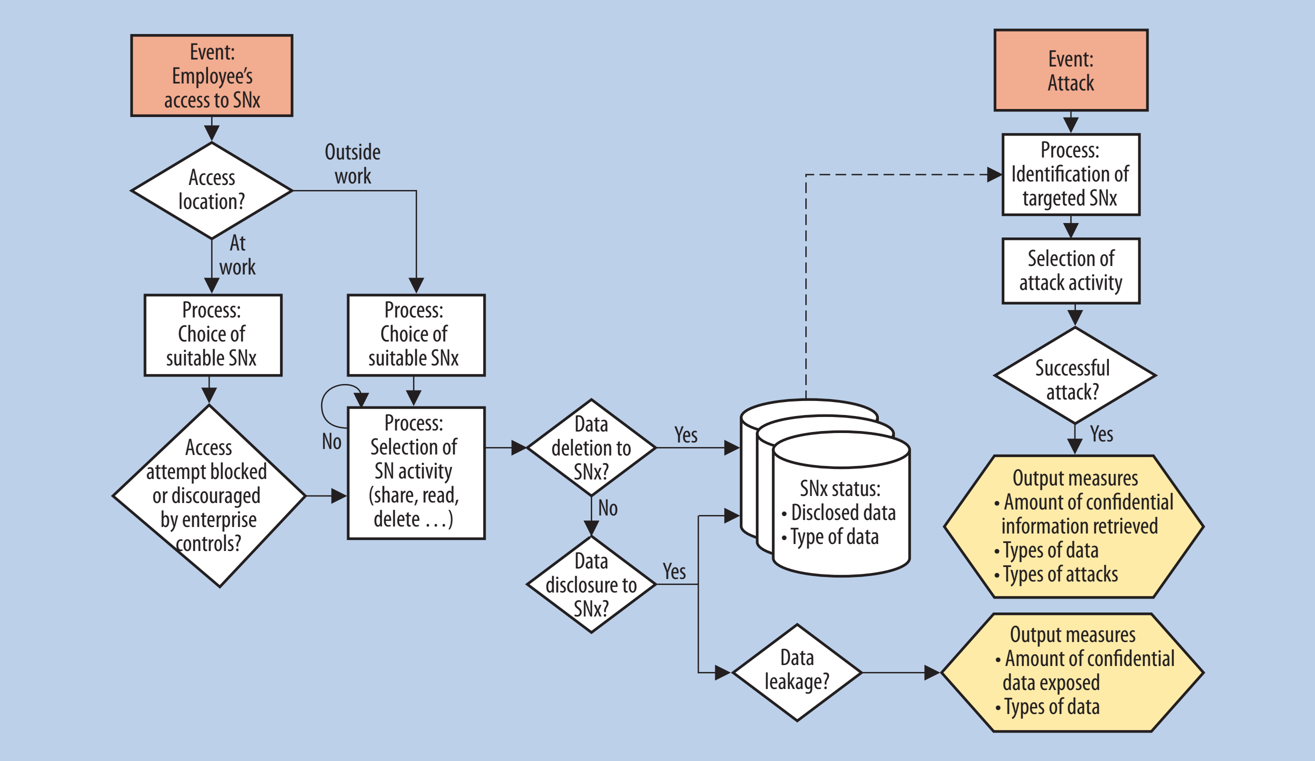

Event: Employee's access to SNx
Access location?
Outside work
At work
Process: Choice of suitable SNx
Process: Choice of suitable SNx
Access attempt blocked or discouraged by enterprise controls?
No
Process: Selection of SN activity (share, read, delete . . .)
Data deletion to SNx?
Yes
No
Data disclosure to SNx?
Yes
SNx status: • Disclosed data • Type of data
Data leakage?
Event: Attack
Process: Identification of targeted SNx
Selection of attack activity
Successful attack?
Yes
Output measures • Amount of confidential information retrieved • Types of data • Types of attacks
Output measures • Amount of confidential data exposed • Types of data

Figure 2. Conceptual model for enterprise data leakage due to employee participation in social network sites. The model represents employees' home and work activities involving interactions with social networks; potential disclosures of sensitive data and mitigations of security control points (controls); and attacker activities aimed at accessing and correlating sensitive information.

about how to deal with risks involved in posting to external social network sites. For some other employees, however, monitoring and punishment controls or control points blocking access to social network sites might be necessary.

Figure 2 shows a high-level model of the processes and activities involved in employees' use of social networks, within and outside the organization, along with related attacks. This model focuses on the entities and elements of interest—that is, employees' interactions with social networks, data leakage incidents, mitigation factors introduced by adopting different control points, and attackers' activities. For simplicity, we omitted some other elements, such as the users' change of attitude in the face of successful attacks or punishments.

The rationale underpinning this conceptual model is based on observations of a "day in the life" of employees. As such, the model factors in employees' activities involving interactions with multiple social networks, carried out on a daily basis, both at work and at home. (SNx indicates a social network site in the range of available social network sites.) These activities include the possibility that personal and business data is posted to external social network sites. The model explicitly accounts for how the adopted controls mitigate the risks involved in these activities and employees' behaviors by considering the controls' effectiveness (for example, access-blocking controls have no effect at home). It also considers the fact that disclosed data can be exploited by skilled attack agents (that is, external observers, such as hackers and other attackers), whose activities are driven by different motivations.

An analysis of the implications of involved risks must consider several key elements.

First, a population of users (employees) displays different behaviors and attitudes to social networks. In general, users are more or less likely to use social networks and to disclose information based on skills, education level, and awareness of potential punishment.

We represent user behavior through a probabilistic model that builds on a set of assumptions related to users' activities and their shared habits. Specifically, we model each user as an autonomous agent, triggered by an event (represented as a negative exponential probability distribution) involving an attempt to access a social network. The model explicitly represents different processes and decision points, including the following:

- selection of a social network (SNx) with which users will interact in different contexts, such as at work or at home;
- attempt to access SNx (users' attitudes toward social networks depend on their education, awareness, and where they operate);

- impact of the available control points if the user is within the organizational boundaries; and
- in case of successful access, activities involving adding, reading, and deleting information. We use binomial probability distributions to model the likelihood that accidental or deliberate data leakages will occur when information is added to SNx.

The model explicitly represents these activities and measures the number of exposed confidential data items, their types, where they were exposed (types of social networks), and so on. Clearly, when implemented in real-world settings, these activities and related assumptions must be refined by grounding the model in a specific context and organizational reality.

The model measures the amount, type, and values of data that attackers access based on their success rate.

The social network sites are the second critical element. We model a social network as a data storage system and set of services that let users add, share, delete, or read information. A social network is characterized by the number of its subscribers and volume of stored data, modeled as a fixed set of input parameters, obtained empirically. External entities might also access and read this information.

A social network enforces a defined level of security and privacy controls—for example, access control, subscriber authentication, and privacy preferences. We model all of these aspects explicitly as a fixed set of input parameters, obtained by empirically analyzing the social network. Depending on the implemented privacy controls, the attacker might face different challenges in accessing the site's data.

The third element is the attack agent. An attack agent's goal is to access confidential data stored at one or more social network sites. Our model makes assumptions about the population of attackers that operates to harness and exploit data. Each attack agent is triggered by an *attack event* (modeled with negative exponential probability distributions that qualify attack frequency) and characterized by

- a profile that includes the attackers' motivations and skills in performing their actions; and
- the likelihood that attackers could successfully gain access to one or more social networks, based on their skills and the available tools and types of attacks.

The model measures the amount, type, and values of data that attackers access based on their success rate. To effectively calculate the risk level and the actual impact of data leakage, we must make assumptions about the threats—that is, the number of attackers, their skills, and their motivations. For simplicity, our model considers three threat types (low, medium, and high), each characterized by a different instantiation of these parameters (number of attackers, skill level, and motivations).

Levers are the fourth key element. Levers that enterprise decision makers might act on include control points (CP_L), education (ED_L), and monitoring and punishment (MP_L). The model represents their adoption levels within a (0, 3) range, where 0 = none, 1 = low, 2 = medium, and 3 = high. The model represents the impact of these levers using probability distributions that describe the likelihood of certain data leakage events not happening as mitigated by related investments. We omit the details of the probability distributions because of space limitations.

IMPLEMENTED MODEL AND EXPERIMENTAL EVALUATION

To provide decision support, we refine and instantiate the conceptual model for a specific enterprise context. We must then tune and validate the implemented model to ensure that it reflects the current enterprise situation. We can then use the model for "what if" analysis and to predict relevance to decision makers, even in the absence of empirical data, by exploring the space of variables of interest.

We've fully implemented an instance of the model for illustration purposes using the AnyLogic modeling and simulation framework (www.xjtek.com). For the sake of this discussion, we deliberately kept the implemented model simple. We could factor in other aspects of interest, based on available information.

In general, a model-based approach lets decision makers explore in advance various assumptions and choices (for example, about suitable investment levers) and provide related explanations and predictions of their impacts in terms of costs and risk exposure. We use Monte Carlo simulations to obtain statistically significant outcomes from the model over a predefined simulation period. A set of input parameters characterizes the implemented model. For users, we include level of education, number of social networks used, and types of information they're likely to post. For attackers, we look at skill level, motivation, and types of attacks they're likely to carry out. Finally, for controls, we use education, monitoring, and technical control points, each characterized by a likelihood of preventing data leakage.

Simulation details

We considered a population of 15,000 employees using social networks. We ran simulations over a period of three years, tracking the impact of the employees' social network use on a set of output measures, including:

- amount of data leaked;
- types (value) of data leaked;
- amount of data successfully accessed by attackers (exploited data leakages), based on the various categories of attacks;
- types (value) of data accessed by attackers; and
- the enterprise's monetary investment in each implemented lever.

We can compare these output measures across multiple types of simulations (based on different assumptions, such as investment choices made on different levers) to determine which investments produce outcomes that are more relevant to decision makers. We could further combine these measures to produce metrics that are more compelling to decision makers, such as overall investment cost and overall security risk for the enterprise. We derive these two metrics from lower-level measures, such as amount and type of data. We approximate overall cost using the costs of investing in various controls, and approximate overall risk using the potential risk of data leakages and the amount of actually leaked information.

Here, we model the overall investment cost for the enterprise as the weighted sum of fixed/cost initial investment and variable cost/maintenance cost. We model the fixed/cost initial investment as a linear equation with respect to the levers. Making specific investments is an initial/one-off cost for the organization. The organization can use control points, education, and monitoring/punishment levers.

The variable cost/maintenance cost considers the period in which the investments in levers are made and the number of employees in the enterprise. We model the variable cost as

$$\text{Variable\_Cost}(t) = d(t) * \text{CP\_L} * t + e(t) * \text{ED\_L} * t + f(t) * \text{MP\_L},$$

where CP_L, ED_L, and MP_L vary in the (0, 3) range and $d(t)$, $e(t)$, and $f(t)$ are variable weights reflecting different impacts and costs of levers at different points. The cost of levers might change over time because of factors such as software updates and license renewal, training/education sessions, and monitoring system upgrades. To function properly, organizations might need to update or renew software used to restrict user information annually. They might need to conduct regular training/education sessions to update employees on current risks, or the employees could find a way to bypass the enterprise's monitoring system. Hence a monitoring system upgrade would be mandatory.

In the example under analysis, we represent the overall risk for the enterprise deriving from an attacker as

$$\text{Overall Risk Exposure} = \text{function(skill\_level, motivation)} * \text{Info\_Actually\_Disclosed(value)},$$

where the attacker's skill_level value ranges from 0 to 2 (that is, none, basic, and high). Similarly, the attacker's motivation, which determines attack frequency, can have values in the 0 to 2 range. The Info_Actually_Disclosed is information that is actually disclosed on the social network site and accounts for the information's value to the organization (that is, type of information).

Given more details about a specific organizational reality, we could have included other metrics, such as productivity and impact on morale. How we model productivity depends on the organization being considered, including factors such as the organization's mission and the employees' work activities. To assess the impact on morale, the organization can periodically collect employees' opinions using questionnaires or surveys and use the results as a metric.

The cost of levers might change over time because of factors such as software updates and license renewal, training/education sessions, and monitoring system upgrades.

We assume that when users first join a social network, they expose little or no information in their profiles, but they perform updates (involving additional data disclosures) on a regular basis. We use a random distribution to check if the user is performing an update. The update frequency can increase based on the user's profile. We assume that a highly educated employee is less likely to disclose personal and confidential information on the social network.[10]

Our model considers a horizontal attack based on a discrete bivariate probability distribution driven by the attacker's skills and motivations. We assume that enterprise decision makers can act on three different levers in a (0, 3) range. The higher the lever's value, the more effective it is. Of course, the higher the value, the higher the investment cost as well.

Technical control points include interception and prevention solutions and filtering systems. Our model assumes an initial installation cost (for example, $100,000), taking fixed and variable costs into account.

We modeled the fact that trained users are more aware of the implications of using social networks and tend to expose less information than they would have without training. We modeled a user's impact in terms of reducing the leaked information using a triangular distribution. The minimum, maximum, and mode values increase as the training lever increases. The impact of education is transitory, decreasing over time unless new investments are made. It also requires times to be effective due to training

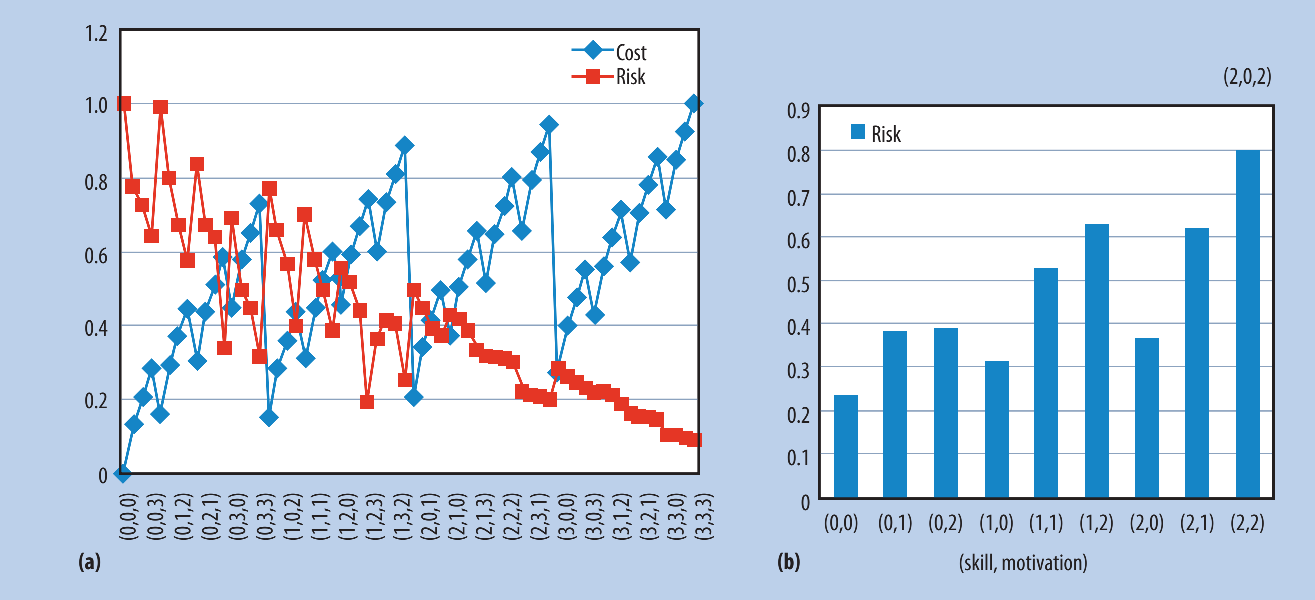


Figure 3. Experimental results: (a) cost and risk simulation outcomes for each combination of the three levers, and (b) variation of risk based on the attacker's motivation and skill.

or retraining of employees.

Monitoring and punishment involve logging, auditing, and subsequent punishment of users who violate the rules. This control's impact is multifaceted. If the organization detects illegal user actions, it can restrict the user's access to the social network from the workplace. However, the user can still access the social network from outside the enterprise. Strong punishment can be a further deterrent, but, in general, its impact is hard to model. So, we simply modeled the fact that the higher the degree of monitoring within an organization, the higher the degree of data leakage detected and people punished.

Our model considers this factor to be public knowledge, and, as such, it discourages other users from engaging in similar activities. Specifically, we model this effect by changing the probability that a person will use social networks after colleagues are punished. The involved costs increase over time because of possible upgrades.

Experiments

Once the model is fully instantiated and validated, we use it to perform experiments by making assumptions about the existing threat environment. In our example, this would let decision makers explore the impact of different investments for the three lever types.

We present a few outcomes that could be deduced from experiments performed using our model and Monte Carlo simulations.

The first set of experiments focused on the impact, in terms of data leakage risks and costs, derived from the activities of just one attacker. We arbitrarily fixed the attacker skill level at 0 and the motivation level at 2. We covered all possible combinations of the three investment levers a decision maker can act on (4 × 4 × 4 combinations of investment values), totaling 64 what-if experiments, each involving 100 simulations of the correspondent configured model. Figure 3a plots the normalized values of risk and cost against the different combinations of levers (*x*-axis), expressed as triples—control, training, monitoring. Of course, models based on different options would produce different outcomes.

We observed some interesting tradeoffs between risks and costs. In the real world, decision makers must balance risk exposure with costs, due to limited resources. Thus, although it's the safest combination, they would be unlikely to choose the (3, 3, 3) combination of the three levers.

A decision maker analyzing these outcomes might identify a few combinations of interest, providing a reasonable tradeoff between normalized values of risk exposure and costs. Based on the assumptions made in the implemented model, (3, 0, 0) is the most effective combination of investment options for the given threat level. The (2, 0, 2) combination is the second-best option.

To illustrate different uses of our approach, we also performed a second set of simulations by making different assumptions about the attacker's characteristics, motivation, and skill to determine how these elements would impact the overall risk. In these simulations, the decision maker uses the (2, 0, 2) combination of investment levers to minimize data leakage at a reasonable cost. An attacker's skill and motivation both vary in the (0, 2) range. We performed a related simulation to calculate the risk for all nine combinations of the attacker's skill and motivation.

Figure 3b illustrates the simulation's outcomes. As we might expect, the risk associated with an enterprise is at its highest when the attacker is both highly skilled and highly motivated, corresponding to the (2, 2) combination. Based on our model's assumptions, an enterprise's risk depends more on the motivation level than the attacker's skill. For instance, an attacker with skill level 0 and motivation level 2 would cause more risk to the enterprise than an attacker with skill level 1 and motivation level 0. The higher the attacker's motivation level, the more frequent the attacks. Multiple attempts will increase the attacker's understanding of the types of information available in the targeted social network, and, through successive attempts, the attacker would eventually obtain the relevant information. This would not be the case for a highly skilled attacker with low motivation. Here, the attacker would try to get the information in only a few attempts or give up in case of failure.

Findings

In a real case study, we could validate the model and its outcomes against empirical data. However, it's often difficult to gather this information. In its absence, models based on educated guesses and empirical observations of cause-and-effects relationships can provide qualitative analysis and relative indications of the impacts of choices and decisions. In our example, an analysis of the experimental results leads to a few conclusions.

First, given the amount of data exposed in social networks and the lack of built-in security mechanisms protecting enterprises from data leakages, an attack could succeed as long as the attacker is well-motivated, determined, and sufficiently skilled to bypass the authentication and privacy controls.

In addition, users' education and awareness play a key role. The more the users are aware of the threats associated with data disclosure, the more reluctant they will be to reveal private information and the less likely to fall victims to attacks.

Finally, using a combination of levers can mitigate the loss as much as possible, although at a cost for the enterprise. Our predictive modeling and simulation approach can help determine the most suitable combination for a given context.

Our future research activities include grounding and further validating this approach in a specific context, based on empirical data collected in a social study, by targeting a set of selected organizations.

Acknowledgments

US National Science Foundation grant NSF-CNS-0831360 partially funded Anna Squicciarini's work. Part of this work was sponsored and carried out in the context of the Cloud & Security Lab (CSL) Security and Identity Analytics project at HP Labs, Bristol.

References

1. C. Saran, "Policies Needed to Limit Social Networking Risks, Says KPMG," *Computer Weekly*, 10 Jan. 2008; www.computerweekly.com/Articles/2008/01/10/228852/policies-needed-to-limit-social-networking-risk-says.htm.
2. O. Ross, "CIOs Getting Serious about Social Networking," *ZdNet*, 25 Feb. 2009; www.zdnet.com/news/cios-getting-serious-about-social-networking/272809.
3. D. Raywood, "Companies Encouraged to Restrict Social Networking Access," *SC Magazine*, 4 Mar. 2009; www.scmagazineuk.com/companies-encouraged-to-restrict-social-networking-access/article/128244.
4. "IT Managers Lack Knowledge of Web 2.0 Use on Their Networks," *SC Magazine*, 2 Mar. 2010; www.scmagazineuk.com/it-managers-lack-knowledge-of-web-20-use-on-their-networks/article/164872.
5. Int'l Organization for Standardization, ISO 27001, *Information Security Management*, 2005; www.iso.org/iso/catalogue_detail?csnumber=42103.
6. B. Monahan, "Gnosis: HP Labs Modeling and Simulation Framework," *Systems Security Lab*, 2009; www.hpl.hp.com/research/systems_security/gnosis.html.
7. E. Zhelevam and L. Getoor, "To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles," *Proc. World Wide Web Conf.*, ACM Press, 2009, pp. 531-540.
8. HP Labs, "Identity Analytics Project," 2009; www.hpl.hp.com/personal/Marco_Casassa_Mont/Projects/IdentityAnalytics/IdentityAnalytics.htm.
9. B. Carminati and E. Ferrari, "Privacy-Aware Collaborative Access Control in Web-Based Social Networks," *Proc. IFIP Data and Application Security Workshop*, ACM Press, 2008, pp. 81-96.
10. A. Acquisti and R. Gross, "Imagined Communities: Awareness, Information Sharing and Privacy on Facebook," *Proc. Privacy-Enhancing Technologies (PET) Workshop*, LNCS 4258, Springer, 2006, pp. 36-58.

***Anna Squicciarini** is an assistant professor in the College of Information Science and Technology at Pennsylvania State University. Her research interests include access control for distributed systems, privacy, security for Web 2.0 technologies, and grid computing. Squicciarini received a PhD in computer science from the University of Milan, Italy. She is a member of IEEE. Contact her at acs20@psu.edu.*

***Sathya Dev Rajasekaran** is a software engineer at Cisco Systems, where he works on datacenter switches. His research interests include security for peer-to-peer systems and security and privacy in social networks. He received an MS in electrical engineering from Pennsylvania State University. Contact him at sxr338@psu.edu.*

***Marco Casassa Mont** is a senior research scientist at the HP Labs in Bristol, UK–Secure Systems Lab. His research interests include strategic aspects of security, privacy, and identity and access management. He is also a lead architect and technologist in the UK collaborative Ensuring Consent and Revocation (Encore) project, focusing on privacy management. Casassa Mont received an MSc in computer science from the University of Turin, Italy. He is a senior member of IEEE and a member of the UK Institute of Information Security Professionals. Contact him at marco.casassa-mont@hp.com.*

CAREER OPPORTUNITIES

CISCO

Cisco Systems, Inc. is accepting resumes for the following position:

San Jose/Milpitas/Santa Clara, CA

Software/QA Engineer

(Ref#: SJ11)

Debug software products through the use of systematic tests to develop, apply, and maintain quality standards for company products.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

Nokia Inc.

has the following open positions in

Redwood City, CA

Consulting Service Engineer

Exp. in the IT/telecom industry to involve working with Object oriented analysis & design using C++; Client server application architecture & multi-threaded application development & design patterns under multi-platform like Windows & Linux; & other duties/skills required. [Job ID: NOK-10RC-CSE]

Sunnyvale, CA

Senior Software Engineer

Exp. in programming in JAVA, develop software applications in J2EE framework; work with database Technologies to include Oracle, MySql, Postgres; applications Servers: Tomcat, Weblogic, JBoss; Web servers: Apache; & other duties/skills required. [Job ID: NOK-SV10-SSE]

Mail resume to: Nokia Recruiter,
3575 Lone Star Circle, Ste. 434
Ft. Worth, TX 76177
& note specific Job ID#.

PURDUE UNIVERSITY, Department of Computer Science, Assistant Professor Position. The Department of Computer Science at Purdue University invites applications for tenure-track positions at the assistant professor level beginning August 2011. Outstanding candidates in all areas of Computer Science will be considered. Specific needs that have been identified include theory and software engineering. Candidates with a multi-disciplinary focus are encouraged to apply. The Department of Computer Science offers a stimulating and nurturing academic environment. Forty-four faculty members direct research programs in analysis of algorithms, bioinformatics, databases, distributed and parallel computing, graphics and visualization, information security, machine learning, networking, programming languages and compilers, scientific computing, and software engineering. Information about the department and a detailed description of the open position are available at http://www.cs.purdue.edu. All applicants should hold a PhD in Computer Science, or a closely related discipline, be committed to excellence in teaching, and have demonstrated potential for excellence in research. The successful candidate will be expected to teach courses in computer science, conduct research in field of expertise and participate in other department and university activities. Salary and benefits are highly competitive. Applicants are strongly encouraged to apply online at https://hiring.science.purdue.edu. Hard copy applications can be sent to: Faculty Search Chair, Department of Computer Science, 305 N. University Street, Purdue University, West Lafayette, IN 47907. Review of applications will begin on November 10, 2010, and will continue until the positions are filled. Purdue University is an Equal Opportunity/Equal Access/Affirmative Action employer fully committed to achieving a diverse workforce.

BUSINESS OPERATIONS ANALYST, Miami, FL: Coordinate business operations for warehousing management, operating policies, plans in line with organizational level short term/long term objectives. Work with Sales Director to manage supply chain of all products. Reply to: FEM Mitchell Group USA, LLC, 1 South East 3rd Ave, #1860 Miami, FL 33131.

DATA ARCH (Enterprise Sys), Bridgeport, CT: Deve, execute proj plans for data admin (deve, ops, control, backup, recovery, security, etc.), integration (completeness, consolidation/updating, organization) of all databases; set database standards, governance crite-

CISCO

Cisco Systems, Inc. is accepting resumes for the following position:

Bellevue, WA

Network Consulting Engineer

(Ref#: BEL1)

Responsible for the support and delivery of Advanced Services to company's major accounts.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

Nokia Siemens Networks US, LLC (NSN)

has the following open positions in

Redmond , WA

Converged Core Engineer

Knowledge of network technologies to include TCP/IP, routing/switching, network security, network management, DNS, Radius, & LDAP; & mobile telephony core networks, home location register, & usage of network analyzers & other duties/skills required. [Job ID# NSN-10WA-CCE]

AT&T OSS Technical Manager

Work with & understand telecom network operations, NetAct platform knowledge as well as GSM/UMTS network technology & network architecture, & IP principles & networking; & other duties/skills required. [Job ID: NSN-WA10-AOTM]

Mail resume to: NSN Recruiter, MS 4C-1-1580
6000 Connection Dr.
Irving , TX 75039
& note specific Job ID#.

ria. Perform data mgmt of enterprise database sys. Utilize Oracle, MS SQL, .NET, MS Windows Server Systems, SAN Storage, BI, Data Warehousing concepts, Data Integration, XML, SDLC, Apache, PHP, Oracle, Java apps in multiplat env. Test, troubleshoot, maint exis sys. Reply to: Human Resources Mgr, Janet Finch, City of Bridgeport, 45 Lyon Terrace, Human Resources, Room 223, Bridgeport, CT 06604. Ref job class code G437.

NETWORK ADMINISTRATOR sought by Computing Concepts, Inc. of E. Rutherford, NJ, to maintain n/work h/ware & s/ware at client's site in Florham Park, NJ. Monitor n/work to ensure n/work availability to all system users & perform necessary maintenance to support n/work availability. Plan, coord & implmt n/work security measures. Install, configure & support an organizations' local area network, wide area network, & internet system or a segment of a n/work system. Bachelor's deg in Engg. Resumes to: Computing Concepts, Inc., Attention: Mario Giacone, 185 E. Union Ave, E. Rutherford, NJ 07073.

NOKIA INC. has the following exp./degreed position in Burlington, MA: Senior Software Engineer: Define the architecture of a software product; work with Core Java Programming, JSP/Servlets, J2EE & related web technologies; XML, HTML, Javascript, CSS & XSLT; database modeling & other web storage technologies & other duties/skills required. [Job ID: NOK-10MA-SRSWE]. Senior Engineer: Work with mobile device software like Palm OS, Windows CE, BREW; embedded device software in development, error fixing, debugging, configuring software builds, trouble shooting, build & integration; use Version control systems such as SVN, CVS, or MS Visual Sourcesafe, & C, C++; & other duties/skills required. [Job ID: NOK-MA10-SRENG]. Send resume to: Nokia Recruiter, 3575 Lone Star Circle, Ste. 434, Ft. Worth, TX 76177 & note Job ID.

SR. CONSULTANTS, Austin, TX, Ascendant Technology. Analyze business reqs., develop/deploy apps for e-Commerce. Req. MA (or foreign equiv) Comp. Sci. Resume only to C. Jones, HR Mgr., Ref. # 081215, 16817 167th Ave. NE, Woodinville, WA 98072.

SOFTWARE ENGINEER: (San Diego, CA) Ability to develop, maintain & upgrade software & firmware; analyze user needs, develop software solutions, implement support software development test routines for verification of prototypes & production; evaluate & recommend automated test tools & strategies for software & hardware continu-

UF | UNIVERSITY of FLORIDA

The Foundation for The Gator Nation

Assistant/Associate/Full Professor

Department of Computer and Information Science and Engineering

The Department of Computer and Information Science and Engineering at the University of Florida invites applications for tenure track positions at any rank with appointment starting August 2011 or thereafter. While, exceptionally strong candidates in any area of computer science or computer engineering are encouraged to apply, we are especially interested in the following areas: bioinformatics, computational science and engineering, cyber-physical autonomous systems, database systems, energy-aware/green computing, exascale computing, human-centered computing, human-computer interaction, medical informatics, modeling, physics-based animation. security, and virtual reality.

All applicants should hold a Ph.D. in Computer Science, Computer Engineering, or a closely related discipline, and should be committed to excellence in teaching and research. Salary and support are competitive and depend on background and experience. The CISE Department currently has 32 tenure track faculty and a student body of 548 graduate and 551 undergraduate students. The Department encompasses a wide range of research areas including computer graphics, simulation, and art; computer systems; computer vision and intelligent systems; database and information systems; high performance computing and algorithms; and networks and security.

Applicants should submit online, a current curriculum vitae and a statement of career objectives to **http://www.cise.ufl.edu/employment/faculty**. Please make reference to position **00004236**. Applicants are requested to have three letters of recommendation submitted electronically as explained in the aforementioned website. Applications will not be reviewed until all recommendation letters have been received. The search committee will accept applications until the positions are filled.

Please complete our Application Data Card located at: **http://www.hr.ufl.edu/job/datacard.htm**. Use position number **00004236**. This information is returned directly to the Faculty Development Office. Submission of this information is entirely voluntary.

For more information about the Department, please visit our website: University of Florida :: Department of Computer and Information Science and Engineering (CISE) **http://www.cise.ufl.edu/**.

The University of Florida is an equal opportunity institution dedicated to building a broadly diverse and inclusive faculty and staff. Minorities, women and those from other underserved groups are encouraged to apply.

University of Illinois at Urbana-Champaign

The Department of Electrical and Computer Engineering invites applications for faculty positions in all areas of electrical and computer engineering, particularly in the areas of energy and power systems, microelectronics, photonics, and computing. Applications are encouraged from candidates whose research programs are in traditional, nontraditional, and interdisciplinary areas of electrical and computer engineering. Our department is engaged in exciting new and expanding programs for research, education, and economic development, with strong ties to industry.

Applicants for positions at the assistant professor level must have an earned Ph.D., excellent academic credentials, and an outstanding ability to teach effectively at both the graduate and undergraduate levels. Successful candidates will be expected to initiate and carry out independent research and to perform academic duties associated with our B.S., M.S., and Ph.D. programs. Senior level appointments with tenure are available for persons of international stature.

Our faculty carry out research in a broad spectrum of areas and are supported by world-class facilities and programs for international work, including the Coordinated Science Laboratory, the Information Trust Institute, the Micro and Nanotechnology Laboratory, the Beckman Institute for Advanced Science and Technology, and several industrial centers. The department has one of the leading programs in the United States granting approximately 300 B.S. degrees, 100 M.S. degrees, and 60 Ph.D. degrees annually.

In order to ensure full consideration by the Search Committee, applications must be received by January 15, 2011. Salary will be commensurate with qualifications. Preferred starting date is August 16, 2011, but is negotiable. Please go to http://jobs.illinois.edu and upload a cover letter, CV, research statement, teaching statement and three references to submit your application. For inquiry, please call 217-333-2301 or email ece-recruiting@illinois.edu.

Illinois is an Affirmative Action /Equal Opportunity Employer and welcomes individuals with diverse backgrounds, experiences, and ideas who embrace and value diversity and inclusivity (www.inclusiveillinois.illinois.edu).

CALIFORNIA STATE UNIVERSITY, SACRAMENTO

Department of Computer Science

Tenure-Track Assistant Professor position to begin in August 2011 in Information Assurance and Computer Security. Ph.D. in Computer Science, Computer Engineering, or closely related field required by the time appointment is made. For detailed position information, including application procedure, please see http://www.csus.edu/hr/faculty/vacancies.htm or http://www.ecs.csus.edu/csc. Screening will begin March 1, 2011, and continue until position is filled. To apply, send cover letter, vita including a list of publications, statement of research and teaching interests, transcripts of all college work including undergraduate work (unofficial copies acceptable for application process), names and phone numbers of at least three references familiar with teaching and research potential to: Search Committee, Department of Computer Science, California State University, Sacramento, 6000 J Street, Sacramento, CA 95819-6021. Incomplete applications will not be considered. AA/EO employer. Clery Act statistics available.

ous improvement of released products. Bachelor's degree in Computer Science or related field. Foreign equivalency OK. Resume to California MedTech, LLC, HR Dept. 15870 Bernardo Center Drive, San Diego, CA 92127.

PROGRAMMER ANALYST - Dsgn, dvlp, test & implmt applic s/w utilizing knowl of & exp w/ Oracle 11i/11.0.3/10.7, Forms 4.5/6.0/6i, Reports 2.5/4.5/6.0/6i, XML Publisher, Toad, SQL Navigator, Discoverer 3i/4i, SQL*Loader, Work Flow Builder 2.6, Oracle, SQL Server, Java, J2EE, Unix & Windows 00/NT; Req MS Comp Sci, Eng or rel. Mail resumes to Nanda Infotech Services Inc. dba EDP Inc. 1332 Street Road Bensalem, PA 19020.

NOKIA SIEMENS NETWORKS US, LLC (NSN) an exp./degreed position in Atlanta, GA. Radio Access Network Engineer: Exp. in project, technical support & equipment network implementation for radio network controller elements (2G BSC/BTS/3G RNC/NodeB/LTE) in wireless communications networks; understand Circuit Switch & Packet Switch Core network elements; & other duties/skills required. [Job ID: NSN-GA10-RANE]. Mail resume to NSN Recruiter, MS: 4C-1-1580, NSN 6000 Connection Dr., Irving, TX 75039 & note Job ID#.

NETWORK ADMINISTRATOR: Dsgn, dvlp & test comp n/works utilizing knowl of & exp w/TCP IP n/works w/OSI model using Cisco, Juniper, Firewall, Load Balancers. Installation & admin of active directory, DNS, DHCP, WINS, File Print servers, VMS. Exp w/VMware Vsphere, fiber channel n/work, SAN is reqd. HTTP, HTTPS, SSL, SSH, SMTP, POP3, DNS, FTP preferred. Req MS Comp Sci, Engg, Bus. or rel. Mail resumes to Saphire Solutions Inc. 523 Green St., 2nd Fl, Iselin NJ 08830.

PROGRAMMER ANALYST: Dsgn, dvlp, test & implmt, Oracle 9i/10g/11g Apps, Oracle Spatial, Oracle Intermedia RAC, JAVA/J2EE, AWK. Reqs MS Comp Sci, Engg or rel. Mail resumes to Sunmerge Systems Inc., 15 Corporate Place South, Ste. 430, Piscataway, NJ 08854.

HEWLETT-PACKARD COMPANY has an opportunity for the following position in Palo Alto, CA and at various unanticipated sites throughout the U.S. Principal Architect: Reqs: Exp w/impltntn of BAC, CMDB, DDM & RC; Knwldg of IT infrastructure & Open Archtctr incl multi-vendor HW & base sw archtctrs; knwldg of ntwrk & commnctns concepts & dv-

SYRACUSE UNIVERSITY

Department of Electrical Engineering and Computer Science (EECS)

Tenure-Track Assistant Professor

The department invites applications for a faculty member at the tenure-track Assistant Professor level in the discipline of Computer Engineering. The ideal candidate would be conducting research in the areas of Virtualization & Cloud Computing, Sustainable & Green Computing, or Grid Computing. The new faculty member should be capable of collaborating with existing EECS faculty members who have active research interests in these areas.

Preference will be given to candidates who can develop successful inter-disciplinary research proposals. We are particularly interested in candidates with prior participation in successful proposal-writing.

A doctorate in computer engineering or computer science is required at the time of employment.

Start-up funds commensurate with the needs of the individual will be provided for the successful candidate.

Review of received applications and interview decisions will begin in December 2010 and continue until the position is filled. The starting date is August 22, 2011. For more information or to apply, please visit the following web site: **www.sujobopps.com** (Job #026988). For detailed information about the Department of EECS, please see the following web site: **http://www.eecs.syr.edu**.

The University and surrounding areas offer a vibrant intellectual and cultural atmosphere, a diverse ethnic community, great public education systems, affordable homes, and many other assets that make it a great place to live and work.

SYRACUSE UNIVERSITY · FOUNDED A·D· 1870 · SUOS CULTORES SCIENTIA CORONAT

Syracuse University is an Affirmative Action/Equal Opportunity Employer.

Nokia Siemens Networks US, LLC (NSN)

has the following open positions in

Irving, TX

Technical Support Engineer

Integrate & troubleshoot SGSN network elements; troubleshooting customer network involving UMTS, GPRS, EDGE technologies & protocols & analyzing protocols; work with telecommunications open standards; & other duties/skills required. [Job ID# NSN-TX10-TSE]

Application Engineer

Deploy, customize, integrate, & test end-to-end MMSC network elements & Bash script, Java, Python, & C/C++; & other duties/skills required. [Job ID: NSN-TX10-APPL]

RNC Technical Support Engineer

Exp. involving RAN network elements; provide customer support exp. to involve analyzing GSM, UMTS, & GPRS network protocols & work with all GSM, UMTS, & GPRS network elements; 2G/3G RAN planning & optimization; & other duties/skills required. [Job ID: NSN-F10-RNC]

Converged Core Engineer

Provide customer support exp. to involve analyzing GSM GPRS network protocols & working with all GSM GPRS network elements; & other duties/skills required. [Job ID: NSN-F10-CCE]

Mail resume to: NSN Recruiter, MS 4C-1-1580
6000 Connection Dr.
Irving , TX 75039
& note specific Job ID#.

lpmts (TCP/IP, X.25, SNA; Exp & knwldg of client/srvr tchnlgy, ntwrkng systems & sltns, middleware concepts, IT security, web tchnlgs, Storage/SAN mgmt, data warehsg, internet-Intranet portal tchnlgs; knwldg of BMC Atrium CMDB, BMC Confgrtn; Mgmt./Marimba IMB CC-MDB & MOM, EMC Applctn Dscvry Mgr, CA CMDB; Prjct Mgmt skills. Also requires Bachelors in EE, Engrng, CS or rel & 7 yrs. exp in job offered or rel. Wk is to be performed at Palo Alto, CA & at various unanticipated sites throughout the U.S. List full name, address & email address on resume. Send resume & refer to Job# PALARE2. Please send resumes with job number to Hewlett-Packard Company, 19483 Pruneridge Ave., MS 4206, Cupertino, CA 95014. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

SAP NETWEAVER ADMINISTRATOR. Responsible for administration, monitoring and maintaining the technical and performance infrastructure of advanced SAP NetWeaver environments to support complex SAP distributed client/server applications. Participate in small- and large-scale projects including analysis and monitoring of advanced SAP NetWeaver platform technologies supported by Advanced Business Application Programming (ABAP) and JAVA programming languages, change management and testing techniques. Send applications to: M.Gagne, Mailstation F110, Massachusetts Mutual Life Insurance Company, 1295 State Street, Springfield, MA 01111; Please Reference Job ID: NC50123132.

BUSINESS ANALYST GLOBAL SYSTEMS. Plymouth, MN (mult pos). Apply & mng Oracle e-bus suite apps usage to deliver process effi towards ops, clinical, mgmt, resourcing & fin bus processes. Job can be performed anywhere U.S. but reports to co headqrts in Raleigh NC. Telecommuting permissible. Req MS in CS, Engg, CIS, related or foreign equiv & 3 yrs exp job offered or comp-rel occp. Exp incld identifying design & execute key RICE components; design Oracle e-bus suite module config & setup, incld Oracle Proj Mgmt, Oracle Proj Res Mgmt, Oracle Proj Costing, Oracle Proj Billing & Oracle Rec; & testing bus scenarios. Cov Let & res to INC Research, LLC, Attn SWSS, 3321 Bee Cave Road, Austin, TX 78746. No calls.

SOFTWARE CONSULTING ENGINEER, exp. required, to work in Sugar Land, TX.

Nokia Inc.

has the following open positions in

Mountain View, CA

Senior Software Engineer: Exp. in software development; exp. developing software for mobile phones with C++ (mobile development exp.), Version control software, & Email & Personal Information Manager (PIM) application development; exp. working with Symbian. [Job ID# NOK2-10MV-SRS]

Principal Architect: Exp. in software engineering (architecture & development); design & development on Symbian OS device; design & analysis with unified modeling language; development exp. in embedded systems, C/C++, Visual C++ & Java; & software development tools & other duties/skills required. [Job ID NOK-MV10-PAR]

Principal Engineer: Exp. in solving complex application design on mobile devices to include software development working with C++; Object Orient Design, E-mail protocols & Activesync; Symbian, S60, & Agile software development methodologies using Scrum; & other duties/skills required. [Job ID: NOK-MV10-PE]

Principal UX Designer: Exp. involving the design of user interfaces for mobile devices in the following areas: Browser chrome (not content), Java applications, web runtime behaviors for mobile operating system involving operating system handling & behaviors, not application design & text input for web technologies coordinated with text input for native applications on mobile operating systems & other duties/skills required: [Job ID: NOK-MV10-PUXD]

Senior Software Engineer: Design & develop user interface components & modules for mobile devices; develop software for mobile phones for major mobile software developers & develop software using C/C++ for mobile devices, & other duties/skills required. [Job ID: NOK-MV10-SRSE]

Mail resume to Nokia Recruiter, 3575 Lone Star Circle, Ste. 434, Ft. Worth, TX 76177 & note specific Job ID#.

BAYLOR UNIVERSITY

Computer Science Department Chair

Chartered in 1845 by the Republic of Texas, Baylor University is the oldest university in Texas and the world's largest Baptist University. Baylor's mission is to educate men and women for worldwide leadership and service by integrating academic excellence and Christian commitment within a caring community. Baylor is actively recruiting new faculty with a strong commitment to the classroom and an equally strong commitment to discovering new knowledge as Baylor aspires to become a top tier research university while reaffirming and strengthening its distinctive Christian mission as described in Baylor 2012 (www.baylor.edu/vision/).

The Department of Computer Science seeks qualified candidates for the faculty chair position. Viable candidates possess high academic standards and Christian convictions matched by an active participation in a community of faith. Additionally, s/he must have a Ph.D. in Computer Science or related area, leadership experience, a commitment to undergraduate and graduate education, a strong research record that includes significant external funding, effective communication and organization skills.

The Department: The Department offers a CSAB-accredited B.S. in Computer Science degree, a B.A. degree with a major in Computer Science, a B.S. in Informatics with a major in Bioinformatics, and a M.S. degree in Computer Science. The Department has 15 full-time faculty, over 370 undergraduate majors and 30 master's students. The Department's greatest strength is its dedication to the success of the students and each other. Interested candidates may contact any faculty member to ask questions and/or visit the web site of the School of Engineering and Computer Science at http://www.ecs.baylor.edu.

The University: Baylor University, situated on a 500-acre campus next to the Brazos River. It annually enrolls more than 14,000 students in over 150 baccalaureate and 80 graduate programs through: the College of Arts and Sciences; the Schools of Business, Education, Engineering and Computer Science, Music, Nursing, Law, Social Work, and Graduate Studies; plus Truett Seminary and the Honors College. For more information see http://www.baylor.edu.

Application Procedure: Applications, including detailed curriculum vitae, a statement demonstrating an active Christian faith, and contact information for three references should be sent to: Chair Search Committee, Department of Computer Science, Baylor University, One Bear Place #97356, Waco, TX 76798-7356.

Appointment Date: Fall 2011. For full consideration, applications should be received by January 1, 2011. However, applications will be accepted until the position is filled.

Baylor is a Baptist university affiliated with the Baptist General Convention of Texas. As an Affirmative Action/Equal Employment Opportunity employer, Baylor encourages minorities, women, veterans, and persons with disabilities to apply.

BAYLOR UNIVERSITY

Assistant, Associate or Full Professor of Computer Science

Chartered in 1845 by the Republic of Texas, Baylor University is the oldest university in Texas and the world's largest Baptist University. Baylor's mission is to educate men and women for worldwide leadership and service by integrating academic excellence and Christian commitment within a caring community. Baylor is actively recruiting new faculty with a strong commitment to the classroom and an equally strong commitment to discovering new knowledge as Baylor aspires to become a top tier research university while reaffirming and strengthening its distinctive Christian mission as described in Baylor 2012 (www.baylor.edu/vision/).

The Department of Computer Science seeks a productive scholar and dedicated teacher for a tenure-track position beginning August, 2011. All specializations will be considered with particular interest in game/simulated environments and mobile computing. The ideal candidate will hold a terminal degree in Computer Science or a closely related field, demonstrate scholarly capability in his or her area of specialization, and exhibit a passion for teaching and mentoring at the graduate and undergraduate level. For position details and application information please visit: http://www.ecs.baylor.edu

The Department: The Department offers a CSAB-accredited B.S. in Computer Science degree, a B.A. degree with a major in Computer Science, a B.S. in Informatics with a major in Bioinformatics, and a M.S. degree in Computer Science. The Department has 15 full-time faculty, over 370 undergraduate majors and 30 master's students. The Department's greatest strength is its dedication to the success of the students and each other. Interested candidates may contact any faculty member to ask questions and/or visit the web site of the School of Engineering and Computer Science at http://www.ecs.baylor.edu.

The University: Baylor University, situated on a 500-acre campus next to the Brazos River. It annually enrolls more than 14,000 students in over 150 baccalaureate and 80 graduate programs through: the College of Arts and Sciences; the Schools of Business, Education, Engineering and Computer Science, Music, Nursing, Law, Social Work, and Graduate Studies; plus Truett Seminary and the Honors College. For more information see http://www.baylor.edu.

Application Procedure: Please submit a letter of application, current curriculum vitae, and transcripts. Include names, addresses, and phone numbers of three individuals from whom you have requested letters of recommendation to: Jeff Donahoo, Ph.D., Search Committee Chair, Baylor University, One Bear Place #97356, Waco, Texas 76798-7356, Materials may be submitted to: Jeff_Donahoo@baylor.edu

Appointment Date: Fall 2011. For full consideration, applications should be received by January 1, 2011. However, applications will be accepted until the position is filled.

Baylor is a Baptist university affiliated with the Baptist General Convention of Texas. As an Affirmative Action/Equal Employment Opportunity employer, Baylor encourages minorities, women, veterans, and persons with disabilities to apply.

Submit resume to ABB Inc., www.abb.com, Careers. Must reference job code NB50692394.

NOKIA SIEMENS NETWORKS US, LLC (NSN) has an exp./degreed position in Irving, TX. Solution Architect: Develop customer solution architecture in mobile packet Core, mobile backhaul & end to end performance area of competence; work with 2G/3G (GPRS/UMTS) packet core design, dimensioning & optimization; & other duties/skills required. [Job ID: NSN-10F-SA]. Mail resume to NSN Recruiter, MS: 4C-1-1580, 6000 Connection Dr., Irving, TX 75039 & note specific Job ID#.

NOKIA INC. has the following exp/degreed position in San Diego, CA: Project Manager, Manufacturing SW & Tools: Creates project plans and handles technology development and deployment of CDMA and TDS-CDMA ODM projects; also works with production test development and other duties/skills required. [Job ID: 11CA-PM]. Mail resume to: Attn: Nokia Recruiter, 3575 Lone Star Circle, Ste. 434, Ft. Worth, TX 76177 & note Job ID.

PROGRAMMER ANALYST: Dsgn, implmt, customize, upgrade SharePoint 2007 Portal sites in Medium or Large Server farm envrmts using C#, ASP.NET, SQL Server 2005/2008, Visual Studio 2005/2008, SharePoint Designer, XML, Web Based Applications, .Net framework 3.0/3.5, WCF, Windows Server 2003/2008, IIS 5.1/6.0/7.0.; SharePoint Search Configuration; InfoPath Forms Development; Req MS in Comp Sci, Eng or rel. Mail resumes to Infologitech Inc., 50 Cragwood Rd, Ste 209, South Plainfield, NJ 07080.

NOKIA SIEMENS NETWORKS US, LLC (NSN) a position in Atlanta, GA. Solution Consultant (IP & Telecoms): Drive & manage consulting-led sales & support system-led sales; deliver consulting to customers & IP Network operations & planning in service provider environments; knowledge of Packet Core planning & optimization & telecom planning exp. involving GSM, GPRS, & UMTS elements; knowledge of product portfolio to include Packet Core network elements; & other duties/skills required. CCIE cert. required [Job ID: NSN-GA10-SC]. Mail resume to NSN Recruiter, MS: 4C-1-1580, NSN 6000 Connection Dr., Irving, TX 75039 & note Job ID#.

PROJECT MANAGER, SENIOR SOLUTIONS ARCHITECT (NY, NY) Utilize knowl of ETL Tools, Composite, Middleware Tech MQ, TIBCO, Perl & Java Tech to provide IT solutions for Prime Brokerage & Fund Services. Apply Business Rules

CISCO

Cisco Systems, Inc. is accepting resumes for the following position:

Bentonville, AR

Software Engineer

(Ref#:BEN1)

Responsible for the definition, design, development, test, debugging, release, enhancement and maintenance of networking software.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

CISCO

Cisco Systems, Inc. is accepting resumes for the following position:

Boxborough, MA

Program Manager

(Ref#:BOX8)

Manage and develop large engineering programs from concept to delivery.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

CISCO

Cisco Systems, Inc. is accepting resumes for the following position:

Lawrenceville, GA

Customer Support Engineer

(Ref#:LV6)

Responsible for providing technical support regarding the company's proprietary systems and software.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

Nokia Siemens Networks US, LLC (NSN)

has the following open positions in

Redmond , WA

Converged Core Engineer

Working with network technologies to include TCP/IP, routing/switching, network security, network management, DNS, radius, LDAP, & Unix administration or undertaken Unix administration tasks on Unix & other duties/skills required. [Job ID# NSN-WA10F-CCEN]

OSS Engineer

Provide technical support & consultancy of OSS throughout USA/Canada; install, commission, troubleshoot, integrate OSS platforms with Oracle in Linux/Unix, HP Unix/RedHat Linux & Windows platform; engage in system acceptance testing of OSS regional cluster systems & perform bull backups of OSS system & other duties/skills required. [Job ID: NSN-WA10-OSSE]

Mail resume to: NSN Recruiter, MS 4C-1-1580
6000 Connection Dr.
Irving , TX 75039
& note specific Job ID#.

using JRules, Savvion BPM to define workflows for Security Management Apps. BS/MS or foreign equiv in rel. fld + rel. exp. Resumes: IT& E Business Consulting Services INC., 469 7th Avenue, 4th Fl., Ste. 17, NY, NY 10018.

DYNAMIC CONSULTING PARTNERS INC., Computer Systems Analyst. Job Duties: Design and implement technology solutions and advise best practices using Microsoft's Great Plains financial systems Software. Analyze client's business, find functional and technical areas of improvements, and advise clients to best achieve their objectives using knowledge and progressive experience in the area of expertise. Requirements: Bachelor's degree in Management Information System. 5 years of progressive experience in Business Consulting work. Contact and send resume: mali@dcpglobal.biz.

NOKIA INC. has the following exp./degreed position in Irving, TX. Senior Application Specialist: Work with the business community to review system requirements & propose, prototype & develop system solutions in SAP & i2; supply chain management & related logistics & planning processes & other duties/skills required. [Job ID: NOK-TX10-SAS]. Mail resume to: Nokia Recruiter, 3575 Lone Star Circle, Ste. 434, Ft. Worth, TX 76177 & note JobID.

NOKIA INC. has an exp/degreed position in Mountain View, CA: Software Engineer: Design & implement next generation mobile messaging/email applications to involve software application development in C++; & other duties/skills required. [Job ID: NOK-MV10F-SE]. Mail resume to: Nokia Recruiter, 3575 Lone Star Circle, Ste. 434, Ft. Worth, TX 76177 & note Job ID#.

HEWLETT-PACKARD COMPANY has an opportunity for the following position in Rio Rancho, New Mexico. Competency Centre Engineer: Reqs: 6yrs exp in srvc & sprt on more than 1 Enterprise Oper System (linux, HPUX, Windows, AIX, Tru64Unix & Solaris) encompassing clustering sw applctns. 6 yrs exp. in srvc & sprt on Netware Oper systems. Must have knwldg of Netware in SAN envmt., Securepath or equiv, MPIO & its underlying archtctr. 6 yrs exp of high level Enterprise Virtual Array (EVA) exp., incl. detailed log & analysis. Performance Analysis or Complex sltns & ability to trblshoot critical situations where data or production is at risk. 6yrs exp of Array, SAN & NAS exp. incl. Brocade & Cisco FC switches, MSA & HSG arrays & NAS Srvrs. Expert knwldg of HP Storage Works 9100 Extreme Data Systems (ExDS9000), incl. cnfgrtn & trblshootg critical issues. Ex-

CISCO

Cisco Systems, Inc. is accepting resumes for the following position:

Miami, FL

Systems Engineer

(Ref#:MIA2)

Provide business-level guidance to the account team or operation on technology trends and competitive threats, both at a technical and business level.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

Nokia Inc.

has the following open positions in

Irving, TX

IT Manager

Exp. in the IT industry to involve project/process management methodologies to include understanding of telecom products & IT system integrations, customer care business processes & business models in telecom industry, project management key practices (involving Agile/Waterfall & iterative model); & other duties/skills required. [Job ID: NOK-TX10-ITM]

Senior Software Engineer

Exp. involving analysis & system design; design, specification, implementation, unit testing, debugging & delivery of system components, work with embedded software development, C++, Object Oriented Analysis & design, Symbian & Series 60 operating system; & other duties/skills required. [Job ID: NOK-TX10-SSE]

Mail resume to: Nokia Recruiter,
3575 Lone Star Circle, Ste. 434
Ft. Worth, TX 76177
& note specific Job ID#.

Cisco Systems, Inc. is accepting resumes for the following position:

San Bruno/San Francisco, CA

Technical Lead

(Ref#: SBR7)

Leads engineering groups on projects to design, develop or test hardware or software products.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

Cisco Systems, Inc. is accepting resumes for the following position:

Richardson, TX

Technical Marketing Engineer

(Ref#:RIC13)

Responsible for enlarging company's market and increasing revenue by marketing, supporting, and promoting company's technology to customers.

Please mail resumes with reference number to Cisco Systems, Inc., Attn: J51W, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

pert knwldg of HP Entemeta MPIO strg sltn, incl. archtctr interoperability b/w Oper system & HP EVA. Detailed knwldg of AIX Oper system. Adv. Acct mgmt knwldg & exp. Customer rltnshp & problem solvng exp. Adv. Tchncl knwldg & exp in at least 6 of the follwg areas: HP EVA & MSA iSCSI connectivity; EVA Bus. Copy/RSM & Cont. Access; Exp. w/installing/configuring/supporting HP Blade tchnlgs; EVA Cluster extensions for Linux or Windows; Strg Wrks Strg Mirror (or NSI Doubletake); Microsoft of Linux Clustering sw; Strg Wrks Secure Path & MPIO on various oper systems; Strg Wrks Virtual Replicator & Strg Wrks Volume Grwth; Prof in areas incl Strg area ntwrks, HBA Drivers, LUNs & fiber channel switches, strg cnsldtn, scalable & highly avail Ntwrk-Attchd Strg; Optimal cnfgrtn of MS Exchng & MS SQL Srvr Applctn in EVA envmts. Certificates Req: Master ASE in High Avlblty & Clustering; Master ASE in strg area. Also reqs: Bachelors in CS, CE or rel. & 6 yrs. exp. in job offered or rel. If no degree or unrelated degree, employer will accept one yr of exp for each yr not completed toward req'd degree. List full name, address & email address on resume. Send resume & refer to Job# RIOCFO2. Please send resumes with job number to Hewlett-Packard Company, H1-6E-28, 5400 Legacy Drive, Plano, TX 75024. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

NOKIA SIEMENS NETWORKS US, LLC (NSN) has the following exp./degreed open position in Redmond, WA. Technical Lead, SCA Mobile Network: Commission, integrate, & troubleshoot GSM elements involving NSS networks elements to include tech. support into customer networks; knowledge of GSM/GPRS/UMTS networks (MSC, MGW, MSS, & SGSN) & work with remote & local migration/re-home support to include Data fill checks, MOP creation, & CIQs; & other duties/skills required. [Job ID# NSN-NSN-WA10F-TL]. Send resume to: NSN Recruiter, MS 4C-1-1580, 6000 Connection Dr., Irving, TX 75039 & note Job ID#.

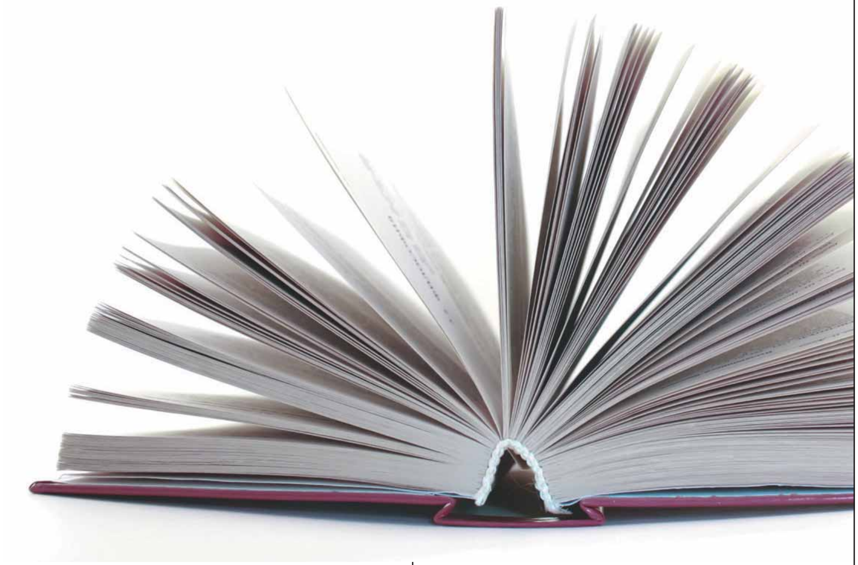
BUSINESS OBJECTS ADMINISTRATOR - Dsgn, install, configure, maintain, performance tune & support Business Objects, Enterprise XI R3.1/R2 Ste. of Products. Expertise in BOBJ XI R3/R2 Servers/Users/Groups/Folders Mgmt, LCM, CCM, SIA, Crystal Reports/ Xcelsius 2008,QAAWS,MS Windows 2003/2000,Apache/Tomcat, Oracle10g/9i, SQL Server reqd. Req MS in Comp Sci, Eng, Math or rel. Mail resumes to Professionals LLC, 1000 N West St, Suite # 1283, Wilmington, DE 19801.

NOKIA SIEMENS NETWORKS US, LLC (NSN) has an exp./degreed position in Irving, TX. Technical Order Configuration Expert: Exp. involving radio & telecommunications working with telecom solution providers/suppliers to include products, such as SDH, Broadband Access, IPC, MW, CIC, solutions, DWDM – business development, network planning, solution architecture, commercial mgmt. (pricing/cost), focusing on Metro & Long Haul DWDM products & Layer 1 through 3 exp. & other duties/skills required. [Job ID: NSN-TX10-TOCE]. Mail resume to NSN Recruiter, MS: 4C-1-1580, 6000 Connection Dr., Irving, TX 75039 & note specific Job ID#.

SYSTEMS ANALYST, Fairless Hills, PA and other client locations: Dsgn, dev, impl, & customize applns & systems using Oracle, SQL Server, Informatica, Siebel, ABAP/4, Toad, Teradata, Java, Visual Basic, Shell Scripts, Perl, Python, Epiphany, Erwin, QTP, Quality Center, Clear Quest, Solaris, Linux, & Win 2000/XP/NT. BS in CS, MIS, CIS, or Eng (any) W/5yrs exp. Salary DOE. Email hr@virgo-systemsinc.com or mail Virgo, Inc, 333 N. Oxford Valley Road, Suite # 405, Fairless Hills, PA 19030 & ref KA1010.

NOKIA SIEMENS NETWORKS US, LLC (NSN) has the following exp./degreed position in Atlanta, GA. Customer Service Engineer (Senior RF Engineer): Plan & perform complex technical activities requiring knowledge in multiple areas of radio communication to service the customer involving cellular network technologies to include working with GSM/GPRS/EDGE & radio network design functions as well as planning and optimization; & other duties/skills required. [Job ID# NSN-10FGA-SRFE]. Send resume to: NSN Recruiter, MS 4C-1-1580, 6000 Connection Dr., Irving, TX 75039 & note Job ID#.

SOFTWARE ENGINEER: Dsgn, dvlp Universes & complex Adhoc reports. Integrate ESRI maps w/Webl reports. Expertise & exp in BO XI R3, Designer, Web Intelligence XI R3, Dashboard Manager, Crystal Reports 2008, Xcelsius, ArcGIS 9.3. Windows XP/2000/NT, UNIX, Teradata, Oracle 10g, SQL Server, PL/SQL, HTML, JSP, Perl VB, Shell Script & UNIX. SAS EG w/ ESRI exp a plus. Req MS in Comp Sci, Eng or rel. Res: Professionals LLC, 1000 N West St, Ste. # 1283, Wilmington, DE 19801.



The Game Maker's Apprentice: Game Development for Beginners, Jacob Habgood and Mark Overmars. This book and its accompanying toolset show how to create nine different games in a range of genres, including action, adventure, and puzzles games—complete with professional-quality sound effects and visuals. It discusses game design theory and features practical examples of how it can be applied to make games more fun to play.

Game Maker lets users create games using a simple drag-and-drop interface, so no prior coding is necessary. It also includes an optional programming language for adding advanced features, with more information available at http://gamemaker.nl. The book includes a CD containing Game Maker software and all the game projects created for the book—plus a host of professional-quality graphics and sound effects for use in the reader's own games.

Apress, ISBN-978-1-59059-615-9, 336 pp.

Silicon Earth: Introduction to the Microelectronics and Nanotechnology Revolution, John D. Cressler. We are in the center of the most life-changing technological revolution Earth has known. In just 60 years, a single invention—the silicon transistor—has produced the most sweeping and pervasive set of changes ever, reshaping the core of human existence on a global scale.

Using nonintimidating language, with an intuitive approach and minimal math, the author addresses the scientific and engineering underpinnings of microelectronics and nanotechnology, as well as how this technology transforms many interdisciplinary fields. Special "widget deconstruction" chapters address the inner workings of ubiquitous micro- and nano-enabled technologies, such as cellphones, flash drives, GPS devices, DVDs, and digital cameras.

Cambridge University Press; www.cambridge.org; 978-0-521-70505-9; 508 pp.

Leading the Virtual Workforce: How Great Leaders Transform Organizations in the 21st Century, Karen Sobel Lojeski. With today's world markets unsteady, unemployment on the rise, housing foreclosures up, and asset values down, the political landscape is shifting. Under such conditions, people often look to leaders to soothe battered nerves. But in today's environment, few leaders can be relied upon.

Interviews with representatives of companies such as IBM, Merck, Western Union, Alcatel-Lucent, HP, and AT&T provide detailed case studies that address what's different about leadership today and how to become a great leader in the Digital Age. Key topics include dispelling common myths and reshaping old leadership models.

Wiley; 978-0-470-42280-9; 155 pp.

Game Usability: Advancing the Player Experience, Katherine Isbister and Noah Schaffer. Computers were once the domain of geeks who enjoyed dealing with a difficult interface. But making the interface really intuitive and useful took computers far beyond the geek crowd. Suddenly, a new factor became crucial to software's success: the user experience.

Today, developers apply and extend these ideas, while game companies take their creations beyond the hardcore gamer market. People who love challenges are happy to master complicated or highly genre-constrained interfaces. Yet as the market expands in step with the growth of interest in casual games, game companies realize that usability matters, particularly to mainstream audiences. If it's not seamless, easy to use, and engaging, players will lose interest.

This book gives game designers a better understanding of how player characteristics impact usability strategy, offering specific methods and measures to employ in game usability practice. It also includes practical advice on how to include usability in already tight development timelines, and how to advocate for usability and communicate results to higher-ups effectively.

Morgan Kaufmann; www.elsevierdirect.com; 978-0-123-74461-6; 512 pp.

Send book announcements to newbooks@computer.org.

Computer Society Winner Nets $50,000 Prize

For his multiple processor-based project, "Automated Parallelization through Dynamic Analysis," Kevin Michael Ellis, 18, of Portland, Oregon, recently received a $1,000 IEEE Computer Society prize at the 2010 Intel International Science and Engineering Fair (ISEF) in San Jose, California. Ellis also received the $50,000 Intel Foundation Young Scientist Award, one of the top prizes at ISEF.

Founded by the nonprofit educational organization Society for Science & the Public in 1950, Intel ISEF is the world's largest precollege science fair.

This year, the competition included 1,611 young scientists from 59 countries, regions, and territories. The Intel Foundation awarded $8,000 to each of 19 "Best of Category" winners and also provided $1,000 grants to the winners' schools and the affiliated science fairs they represent. Dozens of other corporate, academic, government, and science-focused sponsors provided more than 600 additional awards and prizes.

INTEL ISEF COMPETITION

ISEF-affiliated science fairs throughout the world bring together competitors from the 9th through 12th grades. Students must develop a hypothesis, a procedure for testing the hypothesis, and a detailed research plan. Once the project is approved, contestants begin experimenting, observing and collecting data in a project journal, and subsequently interpret the data and observations before drawing conclusions in a final report for presentation.

Complete details on past, current, and future ISEF competitions, as well as resources for locating affiliated regional science fairs, are available at www.societyforscience.org.

COMPUTER SOCIETY AWARDS

Seven competitors at Intel ISEF 2010 received cash awards from the IEEE Computer Society. Based on scores compiled by a team of expert volunteer judges fielded by the Computer Society, winners of IEEE Computer Society awards at ISEF 2010 were

First Award—$1,000: **Kevin Michael Ellis**, 18, Portland, Oregon

"Automated Parallelization through Dynamic Analysis"

Second Award—$500: **Dylan Freedman**, 16, Carmel, California

"A Novel Approach to Text Compression Using *N*-Grams"

Third Award—$350: **Vighnesh Leonard Shiv**, 16, Portland, Oregon

"BeatHoven: Identifying and Inventing Solutions to Obstacles Hindering Automatic Transcription of Polyphonic Music of a Single Instrument"

Team First Award—$500 for each team member: **Spencer August Berglund**, 14, and **David Alexandre Joseph Campeau**, 15, Rochester, Minnesota

"Autonomous Robotic Rubik's Cube Solver"

Team Second Award—$400 for each team member: **Akash Krishnan**, 15, and **Mathew Fernandez**, 16, Portland, Oregon

"The Classification and Recognition of Emotions in Prerecorded Speech"

Award winners also receive a gift certificate for any Computer Society publication and a one-year subscription to a Society magazine of their choice.

OTHER HONORS

Said lead IEEE Computer Society judge Lowell Johnson, "Each year the winners of Computer Society awards usually win several other honors, usually comparable to ours monetarily. However, this year our first-place winner (Ellis) won approximately a dozen other awards, including the Intel Foundation Young Scientist Award of $50,000, which is the second-highest of the Intel Grand Awards. I do not remember this ever happening before. Kevin's work was thoughtful, well-researched in terms of the literature, and he was rigorous in testing his approach."

Published by the IEEE Computer Society

Ellis received, among other honors, the $3,000 Intel ISEF Best of Category Award for computer science. Computer Society honoree Vighnesh Leonard Shiv earned a fourth award in the same category, as well as a $5,000 scholarship to Oregon Institute of Technology and $2,000 in United Technologies corporate stock. Computer Society award winner Dylan Freedman earned an $8,000 scholarship subsidy from the US Office of Naval Research. Society honorees Akash Krishnan and Mathew Fernandez won an all-expense paid trip to attend the 2011 European Union Contest for Young Scientists, "Genius Scholarships" to Sierra Nevada College, and a $5,000 Intel ISEF Best of Category Award for team projects.

The overall Intel ISEF winner for 2010 was Amy Chyao, 16, of Richardson, Texas, who received the new $75,000 Gordon E. Moore Award for developing a photosensitizer for photodynamic therapy, an emerging cancer treatment that uses light energy to activate a drug that kills cancer cells.

The Intel International Science and Engineering Fair moves to Los Angeles, California in 2011. Further information about ISEF is available at www.societyforscience.org.

Society Publication Seeks Volunteer EIC

The IEEE Computer Society seeks applicants for the position of editor in chief of *Annals of the History of Computing*, serving a two-year term, starting 1 January 2013. Prospective candidates are asked to provide (as PDF files) by **1 March 2011** a complete curriculum vitae, a brief plan for the publication's future, and a letter of support from their institution or employer. For more information on the search process and to submit application materials, please contact Robin Baldwin, rbaldwin@computer.org.

QUALIFICATIONS AND REQUIREMENTS

Candidates for any Computer Society editor-in-chief position should possess a good understanding of industry, academic, and government aspects of the specific publication's field. In addition, candidates must demonstrate the managerial skills necessary to process manuscripts through the editorial cycle in a timely fashion. An editor in chief must be able to attract respected experts to the publication's editorial board. Major responsibilities include

- actively soliciting high-quality manuscripts from potential authors and, with support from publication staff, helping these authors get their manuscripts published;
- identifying and appointing editorial board members, with the concurrence of the Publications Board;
- selecting competent manuscript reviewers, with the help of editorial board members, and managing timely reviews of manuscripts;
- directing editorial board members to seek special-issue proposals and manuscripts in specific areas;
- providing a clear, broad focus through promotion of personal vision and guidance where appropriate; and
- resolving conflicts or problems as necessary.

Applicants should possess recognized expertise in the computer science and engineering community and must have clear employer support.

REAPPOINTMENTS

Other IEEE Computer Society publications have editors in chief who are currently standing for reappointment to a second two-year term. The IEEE Computer Society Publications Board invites comments on the tenures of the individual editors. Editors in chief standing for reappointment to terms in 2012-2013 are:

- Gabriel Taubin, *IEEE Computer Graphics & Applications*
- Simon Liu, *IT Professional*
- John R. Smith, *IEEE Multimedia*
- Nigel Davies, *IEEE Pervasive Computing*
- Ravi Sandhu, *IEEE Transactions on Dependable & Secure Computing*
- Ivan Stojmenovic, *IEEE Transactions on Parallel & Distributed Systems*
- Bashar Nuseibeh, *IEEE Transactions on Software Engineering*
- Kevin Skadron, *Computer Architecture Letters*

Send comments to:

Ed Zintel, ezintel@computer.org for *IEEE Multimedia*

Bob Ward, bnward@computer.org for *IT Professional*

Kathy Santa Maria, ksantama@computer.org for all *Transactions* and *Letters*

Jennifer Stout, jstout@computer.org for *IEEE Computer Graphics & Applications* and *IEEE Pervasive Computing*.

Board of Governors Votes to Amend Bylaws

At a recent meeting, the IEEE Computer Society Board of Governors voted to amend two articles of the Society's bylaws. Article IV (http://bit.ly/gnVY0F) was changed to reflect changes to the makeup of the Executive Committee. Article XII (http://bit.ly/gbVHQX) was edited to clarify duties and membership qualifications for the Awards Committee.

View the revised bylaws online. Changes to Article IV are available via http://bit.ly/gnVY0F, while changes to Article XII are available via http://bit.ly/gbVHQX. Deletions are marked in strikeout text, and insertions are underlined.

CALLS FOR ARTICLES FOR IEEE CS PUBLICATIONS

IEEE Intelligent Systems seeks submissions for a September/October 2011 issue on brain informatics.

BI is an emerging multidisciplinary research field that focuses on the mechanisms underlying the human information-processing system. It investigates the essential functions of the brain, ranging from perception to thinking and encompassing such areas as multiperception, attention, memory, language, computation, heuristic search, reasoning, planning, decision making, problem solving, learning, discovery, and creativity.

The guest editors are soliciting papers developing brain data grids and brain research support portals; cognitive architectures and their relations to functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG); and brain data modeling and formal conceptual models of human brain data, among other topics.

Articles are due by **24 February**. Visit www.computer.org/portal/web/computingnow/iscfp5 to view the complete call for papers.

CALENDAR

FEBRUARY

9-10 Feb: ESSoS 2011, Int'l Symp. on Eng. Secure Software and Systems, Madrid, Spain.; http://distrinet.cs.kuleuven.be/events/essos2011

12-16 Feb: HPCA 2011, IEEE Int'l Symp. on High-Performance Computer Architecture, San Antonio, Texas; http://hpca17.ac.upc.edu

18-20 Feb: EAIT 2011, 2nd Int'l Conf. on Emerging Applications of Information Technology, Kolkata, India; https://sites.google.com/site/csieait2011

MARCH

13-15 Mar: ISORC 2011, IEEE 14th Int'l Symp. on Object/Component/Service-Oriented Real-Time Distributed Computing, Newport Beach, California; http://dream.eng.uci.edu/isorc2011

20-25 Mar: ACSEAC 2011, African Conference on Software Engineering and Applied Computing, Cape Town, South Africa; www.acseac.org

21-25 Mar: SIMUTools 2011, 4th Int'l ICST Conf. on Simulation Tools and Techniques, Barcelona, Spain; www.simutools.org

APRIL

8-9 Apr: ICCP 2011, IEEE Int'l Conf. on Computational Photography, Pittsburgh; www.cs.cmu.edu/~ICCP2011/index.html

10-15 Apr: InfoCom 2011, 30th IEEE Int'l Conf. on Computer Communications, Shanghai; www.ieee-infocom.org

11-16 Apr: ICDE 2011, 30th IEEE Int'l Conf. on Data Engineering, Hanover, Germany; www.icde2011.org

MAY

1-5 May: VTS 2011, 29th VLSI Test Symp., Dana Point, California; www.tttc-vts.org/public_html/new/2011/index.php

16-18 May: ISSST 2011, IEEE Int'l Symp. on Sustainable Systems and Technology, Chicago; www.ieee-issst.org

16-20 May: IPDPS 2011, 25th IEEE Int'l Parallel & Distributed Processing Symp., Anchorage, Alaska; www.ipdps.org

21-25 May: ICSE 2011, IEEE/ACM Int'l Conf. on Software Eng., Honolulu; http://2011.icse-conferences.org

JUNE

19-25 Jun: CVPR 2011, IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Colorado Springs, Colorado, http://cvpr2011.org

21-24 Jun: ICDCS 2011, 31st Int'l Conf. on Distributed Computing Systems, Minneapolis; www.seas.gwu.edu/~cheng/ICDCS2011

SUBMISSION INSTRUCTIONS

The Call and Calendar section lists conferences, symposia, and workshops that the IEEE Computer Society sponsors or cooperates in presenting.

Visit www.computer.org/conferences for instructions on how to submit conference or call listings as well as a more complete listing of upcoming computer-related conferences.

0018-9162/11/$26.00 © 2011 IEEE

EVENTS IN 2011

February

9-10	ESSoS 2011
12-16	HPCA 2011
18-20	EAIT 2011

March

13-15	ISORC 2011
20-25	ACSEAC 2011
21-25	SIMUTools 2011

April

8-9	ICCP 2011
10-15	InfoCom 2011
11-16	ICDE 2011

INFOCOM 2011

The IEEE Conference on Computer Communications addresses key topics and issues related to computer communications, with emphasis on traffic management and protocols for both wired and wireless networks. Experts share recent research findings via a program of technical sessions, tutorials, panel discussions, and workshops.

Conference organizers have solicited papers on topics that include cyberphysical systems and networks, topology characterization and inference,congestion control, and network architectures.

The first InfoCom conference took place in 1982.Sponsored by the IEEE Computer Society Technical Committee on Computer Communications, InfoCom 2011 takes place 10-15 April in Shanghai. Visit www.ieee-infocom.org for complete conference details.

Can More Efficient IT Be Worse for the Environment?

Bill Tomlinson, M. Six Silberman, and Jim White, *University of California, Irvine*

In efforts to reduce energy usage, IT professionals must ensure that efficiency aligns with sustainability.

In his 5 October 2009 executive order, "Federal Leadership in Environmental, Energy, and Economic Performance," US President Barack Obama called for increased energy efficiency as part of the federal government's "integrated strategy towards sustainability" (www.whitehouse.gov/assets/documents/2009fedleader_eo_rel.pdf). While researchers generally agree that energy efficiency leads to short-term economic prosperity, there is a lack of consensus that it actually makes civilization more sustainable. Broad awareness of such ambiguity could help computing professionals and society at large reassess current priorities in this effort.

Understanding the role that energy efficiency can play in enabling IT to serve sustainability is critical, both to the computing industry and to the global ecosystem. The information and communication technology sector accounts for 2-2.5 percent of global CO_2 emissions and is growing rapidly.

IT supports a gamut of human endeavors, and as such has very broad effects. The Climate Group's 2008 report *Smart 2020: Enabling the Low Carbon Economy in the Information Age* suggests that this sector can offset five times its own carbon footprint by enabling efficiencies in other facets of society (www.smart2020.org/_assets/files/02_Smart2020Report.pdf). On the other hand, making IT more efficient can also contribute to unsustainable economic growth and related environmentally harmful activities.

ENERGY EFFICIENCY AND SUSTAINABILITY

Energy efficiency can be defined in many ways, ranging from narrow and technical to broad and inclusive. According to the US Department of Energy's Energy Information Administration, "increases in energy efficiency take place when either energy inputs are reduced for a given level of service or there are increased or enhanced services for a given amount of energy inputs" (www.eia.doe.gov/emeu/efficiency/ee_ch2.htm).

Sustainability is still more difficult to define. A sustainable system is, in the simplest terms, one that continues to exist over time. In the context of green IT, sustainability is shorthand for "global environmental sustainability," a characteristic of Earth's future in which certain essential processes persist for a period of time comparable with human lives. Exactly which processes these are, and how long they must persist, are subjects of considerable debate. However, potentially useful criteria include the continuation of the human species at a relatively high quality of life for thousands of years, the restriction of humanity's exploitation of the global ecosystem to a rate that does not exceed the rate of renewal by other factors, and a species extinction rate no greater than average across a geologic timescale.

As a practical matter, absolute sustainability is too high a bar; it is more useful to discuss whether IT systems are "aligned with sustainability" or "sustainability-directed"—that is, more sustainable than the systems they supplant. Reducing CO_2 emissions is often a proxy for moving toward sustainability because the reduction of greenhouse gasses is a key factor in mitigating global climatic disruption, an environmental problem central to all three sustainability criteria.

DIRECT AND INDIRECT EFFECTS

Many factors influence whether a given effort to increase energy efficiency aligns with sustainabil-

0018-9162/11/$26.00 © 2011 IEEE

ity. Considered narrowly, energy efficiency appears to help us live more sustainably because it enables a system to expend less energy to achieve the same end result, leading to reduced carbon emissions.

However, several indirect effects also play a role. For example, decreasing the cost of energy usage embodied in a particular product could cause consumers to buy more of that product. The savings could also end up in the pockets of shareholders or consumers, who will then use them for other purposes—to buy different goods, invest elsewhere, and so on. A decrease in energy price could also lead to entirely new industries; the Internet, for example, might not have evolved into its present form if energy had been significantly more expensive during the past several decades. Taken together, these indirect efficiency effects can sometimes account for more environmental harm than was averted by the original savings.

English economist William Stanley Jevons first recognized the potential for this energy efficiency "backfire" in 1865, when he noted in *The Coal Question* that improvements in steam engine technology paradoxically led to increases, rather than decreases, in coal usage as new industries started using such engines. In recent decades, there has been vigorous discussion of the direct and indirect impacts of increased energy efficiency. Researchers have developed an array of quantitative techniques for tracing indirect effects across different economic sectors, but there are still relatively few contexts in which the sum total of these effects has been examined, let alone established definitively.

The time lag before many indirect effects manifest themselves further complicates analysis. While the direct benefits of efficiency help reduce current energy use, the indirect effects accrue over an indefinite period of time into the future. Efficiency can thus arguably "buy us time" to develop lower-carbon energy sources and enact broad-scale change. However, if society uses such bought time to conduct business as usual, greater efficiency could ultimately make the problem worse.

Whether the time-discounted value of all the various direct and indirect effects for a particular efficiency adds up to an environmentally positive or negative result, efforts to calculate sustainability implications must consider temporal factors. Combined with the difficulty in even roughly estimating various indirect effects, the fact that many effects might not fully emerge for years or even decades suggests that, in most cases, the relationship between energy efficiency and sustainability is ambiguous.

Indirect efficiency effects can sometimes account for more environmental harm than was averted by the original savings.

Assessing sustainability impacts for IT energy efficiency is particularly difficult. As a general-purpose technology (GPT), IT influences many sectors of industry and aspects of society. For example, making server technology more energy efficient is likely to produce efficiencies across a vast array of contexts. To assess total IT efficiency accurately would entail understanding each of those contexts to a greater extent than is currently feasible. Even if such knowledge was attainable, it would likely be both publicly unpalatable due to privacy issues and directly opposed to various facets of corporate law.

However, a 2007 report by the UK Energy Research Centre, *The Rebound Effect: An Assessment of the Evidence for Economy-Wide Energy Savings from Improved Energy Efficiency,* suggests that efficiency-improvement efforts aimed at GPTs are more likely to rebound than those aimed at specialized technologies because "the opportunities offered by these technologies have such long term and significant effects on innovation, productivity and economic growth that economy-wide energy consumption is increased" (www.ukerc.ac.uk/Downloads/PDF/07/0710ReboundEffect/0710ReboundEffectReport.pdf). For example, creating a new sociotechnical system such as the Internet brings about a set of entirely new and unforeseen energy demands.

In short, whether IT energy efficiency will ultimately reduce society's total energy use is unclear. Even if does, the reductions will likely be smaller than many expect.

WHAT IS TO BE DONE?

Our objective here is not to undermine efficiency efforts. Efficiency tends to foster economic growth, which is a policy goal for all industrialized nations and often increases people's standard of living. However, many computing professionals seek to improve efficiency because they believe it is a reliable way to reduce humanity's environmental impact. Unfortunately there is no clear consensus on this relationship in the environmental science and economics literature.

If developing an energy-efficient system is no guarantee that it will align directly with sustainability, then what should we do? Not surprisingly, there is no simple answer. However, several opportunities are promising.

IT professionals interested in energy efficiency can partner with experts in economics, environmental science, and other domains to help determine which efficiencies are most likely to be sustainability-directed. In particular, those already engaged in an efficiency effort could seek to analyze their own project in this regard. Locating sustainability-aligned efficiency opportunities would enable IT professionals to continue researching efficiency in general, even if a particular efficiency effort turns out to be counterproductive in terms of sustainability.

Moving beyond efficiency, contributing to projects that reduce the demand for energy—for example, by shifting social norms away from consumption-based lifestyles—could help our civilization become more sustainable. Alternatively, if there is freedom to change direction dramatically, IT professionals could redirect their efforts to sustainability-related projects such as preserving and restoring endangered habitats or lowering global birth rates. Through such initiatives, the IT industry can continue to thrive while shrinking humanity's global footprint.

More broadly, as members of society, IT professionals can support government efforts that are likely to have positive sustainability outcomes. According to *Limiting the Magnitude of Future Climate Change* (www.nap.edu/catalog.php?record_id=12785#toc), a 2010 report by the National Research Council, "A carbon-pricing system is the most cost-effective way to reduce emissions. Either cap-and-trade, a system of taxing emissions, or a combination of the two could provide the needed incentives."

Intelligently applied efficiency is vital to achieving sustainability, but it must be coupled with a carbon tax or similar potent measure to effect meaningful environmental change; without it, we will simply get more of what we already have: increasing environmental destruction.

Identifying which particular energy efficiencies can be sustainability-directed is currently very difficult. In pursuing sustainability through efficiency, we must first determine the conditions under which efficiency aligns with sustainability and undertake only the efforts that satisfy those conditions. To borrow a phrase from University of British Columbia ecologist Bill Rees, pursuing other efficiencies will only make human civilizations "more efficiently unsustainable."

***Bill Tomlinson** is an associate professor in the Department of Informatics at the Donald Bren School of Information and Computer Sciences, UC Irvine, and a researcher at the California Institute for Telecommunications and Information Technology. He is the author of* Greening through IT: Information Technology for Environmental Sustainability *(MIT Press, 2010). Contact him at wmt@uci.edu.*

***M. Six Silberman** is a field interpreter with the Bureau of Economic Interpretation and a recent graduate of the interdisciplinary Arts Computation Engineering program at UC Irvine. Contact him at six@economicinterpretation.org.*

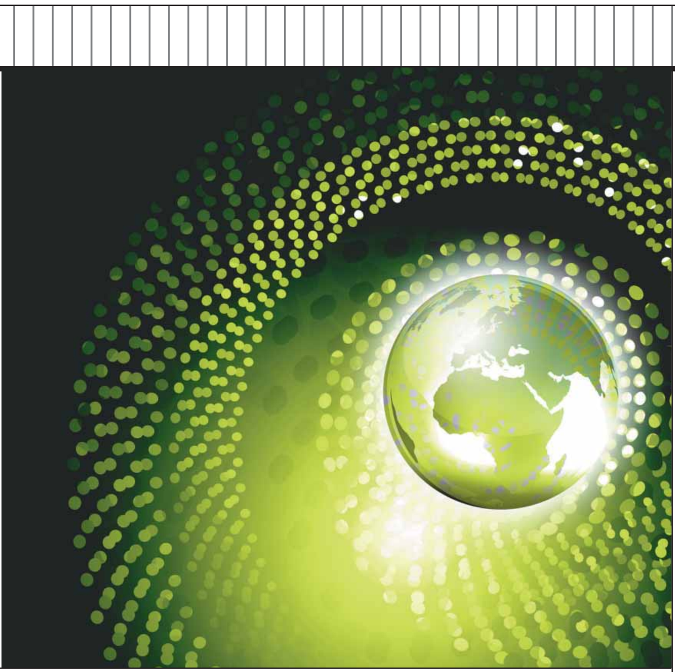
***Jim White** has been a software developer and system architect for 30 years and is currently a student in the Department of Computer Science at UC Irvine's Donald Bren School of Information and Computer Sciences. Contact him at jpwhite@uci.edu.*

The ideas discussed here were improved significantly by discussions with Kirk Cameron, Larry Smarr, G.P. Li, Stu Ross, Doug Fisher, Skip Laitner, Rebecca Black, and the members of the Social Code Group at UC Irvine. This article is based in part on work supported by the National Science Foundation under Grant No. 0644415 and by the Alfred P. Sloan Foundation.

Editor: Kirk W. Cameron, Dept. of Computer Science, Virginia Tech; greenit@computer.org

Achieving Synergy in the Industry-Academia Relationship

Neil Ferguson, *Harris Corp.*



Industry needs bright, talented, well-prepared graduates to join the workforce, while academia requires insight into industry's needs to ensure that it can develop a future workforce that is prepared to meet those needs.

The synergy between industry and academia has been evident across my career in engineering as a co-op student, engineer, researcher, technologist, recruiter, business developer, and manager. The relationship is symbiotic: industry needs bright, talented, well-prepared graduates to join the workforce, while academia requires insight into industry's needs to ensure that it can develop a future workforce that is prepared to meet those needs.

We can't forget that academia is often at the leading edge of research that directly applies to industry's objectives. The relationship between industry and academia is complex, with numerous diverse aspects, from collaborating to prepare the next wave of entry-level engineers to partnering in research to solve tomorrow's problems.

EDUCATING THE FUTURE WORKFORCE

Cynics might say that the main reason for the relationship between industry and academia is so that we can all go back to our alma maters and relive the glory days of our youth. Okay, maybe not. But it is true that we return to academic institutions because they are the main source for the next generation of engineers, which in turn fuels the industrial engine. Companies use several methods to create visibility on campus and to reach students, including engaging with student organizations, providing lectures or industry talks, sponsoring student projects, and conducting mock interviews, to name a few. To strengthen the relationship, companies often also provide corporate sponsorships, donations, and scholarships. It is no coincidence that many buildings, labs, and lecture halls are going the way of college football bowls by being adorned with corporate logos.

The relationship becomes more of a partnership when the focus of both industry and academia turns to ensuring that graduating students are as thoroughly equipped as possible to enter the workforce. There is excellent dialog going on today toward this end. Many schools have industry advisory boards that work directly with them to communicate evolving industry needs and to help shape the direction of curricula and programs. It is also common for schools to proactively approach industry with surveys and working groups to gain additional perspective and input. The best way for universities to meet industry's needs is to be engaged with industry and vice versa. It is a two-way street, with both benefiting.

There is additional opportunity for greater, more focused collaboration on the educational content that schools deliver. Universities and companies are increasingly working together to define new degree and certificate programs. For example, Harris Corp. worked closely with the Florida Institute of Technology to develop a master of science in systems engineering degree with a complementary enterprise architecture certificate. This degree program was developed in response to a growing need in our industry for such expertise.

Published by the IEEE Computer Society

Similarly, the University of Central Florida proactively engaged local companies to address talent shortages in engineering by helping to define a new degree program—an MS in engineering management—that would fill the identified gap. Schools are also using alternative delivery methods such as remote degree programs or local cohorts delivering main-campus courses to expand their reach and work alongside industry to meet evolving business needs.

DESIGNING EFFECTIVE CO-OP AND INTERN PROGRAMS

Cooperative education and intern programs enable employers to hire undergraduate students prior to graduation. This "try before you buy" model benefits both parties.

In addition to gaining real-world experience, students also develop social skills, validate career decisions, begin to establish their professional network, and receive income to help offset the cost of their education. Additionally, after degree and academic performance, relevant professional experience is the next most important criterion in obtaining a job after graduation. By the time students graduate, they may have gained a year or more of valuable work experience in their field of study.

For the employer, intern or co-op programs translate into an employee entering the organization with a significant amount of highly relevant work experience, compared to that of a typical graduate. In addition, the company has already benefited from the student's assignments with the organization, gaining new ideas and increased opportunity for technology transfer. The most obvious benefit is the chance to evaluate these temporary employees while deciding whether to make a permanent offer.

Since companies want to retain students as permanent employees after graduation, it is critical that they offer quality intern and co-op programs. What makes a strong program? First, meaningful, well-thought-out job assignments that closely align with the students' studies and provide challenging tasks that will enrich their academic studies.

Another key component is a constructive feedback mechanism. Students want to excel, and to do so they need reinforcement and feedback. In turn, companies benefit by soliciting feedback from students. At Harris, we have interns create a video skit that both chronicles what they learned during their internship and also provides feedback to the company—all while having a little fun. Feedback from the students helps the company improve its program. More importantly, it demonstrates that the company values the students' input.

After degree and academic performance, relevant professional experience is the next most important criterion in obtaining a job after graduation.

Providing financial assistance to cover the costs associated with moving and housing expenses is another important concept. Other attributes of successful programs include providing the opportunity to work in rotations on a variety of assignments, offering corporate overview education, and providing exposure to the company's products and services in operation.

Work programs are a mutually beneficial partnership among schools, students, and companies. On the heels of a good work experience, the student becomes a true advocate for the company back on campus, increasing the company's visibility among students. Often, testimonials and referrals are the best marketing tools.

Early learning opportunities

In some cases, the relationship between industry and academia begins even before students enter college. Many companies and schools are proactively reaching out to the community through K-12 programs geared toward increasing interest in science and engineering among primary and secondary school students. Unfortunately, little is being done jointly in these cases, resulting in a missed opportunity.

Geographic proximity is a key factor affecting such collaborative efforts. My company has a strong K-12 program and a desire to grow it. As a corporate representative on campus, I see the same desire on the part of universities and work to explore the possibility of joint programs. What usually stops us is determining how to work together when we are hundreds of miles apart. For companies and schools to successfully collaborate in this endeavor, they need to partner when they are in the same vicinity and can focus on the same set of K-12 students.

Joint Engineers Week celebrations also create the opportunity for greater sharing and synergy, promoting the engineering profession and reaching out to current and future generations of engineering talent. As an element of Engineers Week at Harris, we invite professors to share ongoing research with our engineers and invite students from local schools and universities to gain insight into our technical focus. These forums provide excellent avenues for sharing information and cultivating interest in engineering.

Looking to the future

Is there an opportunity to combine all three partnerships into a fully integrated, collaborative program that maximizes the synergy between a student's academics, the intern/co-op experience, and joint research? I believe there is.

In some circumstances a fully integrated program that combines all three types of partnerships may provide the greatest yield. One such example is Harris's collaboration with the Florida Institute of Technology in establishing the Harris Information Assurance Institute. The Institute provides technical collaboration opportunities for all involved, as well as real-world work opportunities for students. It also helps the Florida Institute of Technology gain an understanding of corporate technical needs, which in turn flows back into curriculum planning, helping both students and industry.

Another excellent example is the Integrated Product and Process Design (IPPD) Program at the University of Florida. A collaborative partnership between industry and the university, this program gives students the opportunity to work in small, multidisciplinary project teams, where they gain important practical experience in teamwork and communication while developing their leadership, management, and people skills.

Other excellent examples are the DARPA Grand/Urban Challenges and the Association for Unmanned Vehicle Systems International (AUSI) Autonomous Underwater Vehicle Competition, where industry, government, and academia partner to address real-world problems. While such endeavors require committed stakeholders and a significant financial investment, as well as introducing concerns about intellectual property, contracts, and compliance, the benefits are compelling.

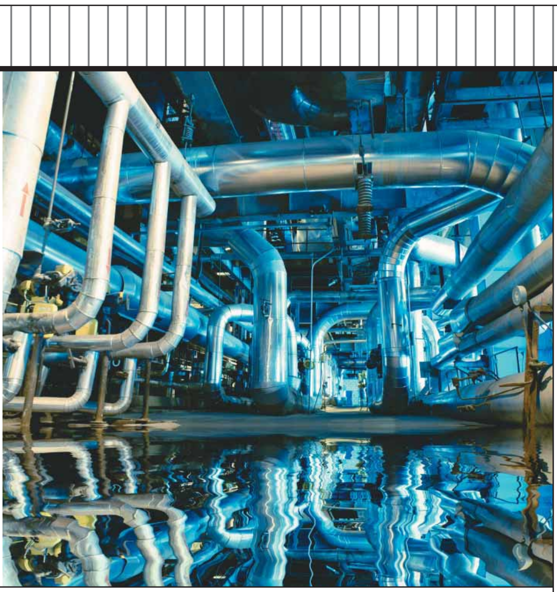
The complex relationship between academia and industry covers a wide spectrum from college recruiting to joint research projects. A tremendous amount of excellent work is being accomplished today, and we have the opportunity to achieve even more synergy in the future. Such endeavors have the opportunity to significantly reinforce what is learned in the classroom, produce real-world solutions with impact beyond the industry-academia partnership, and prepare the best possible engineering workforce for the future.

***Neil Ferguson** is the senior engineering manager for the Enterprise Architecture Core Technology Center and is the Harris Healthcare Solutions engineering leader at Harris Corp.'s Government Communications Systems Division. As Harris's Georgia Tech campus brand manager, Ferguson heads up all recruiting and student involvement activities at Georgia Tech. Contact him at nferguso@harris.com.*

Editor: Sumi Helal, Department of Computer and Information Science and Engineering, University of Florida; helal@cise.ufl.edu

cn Selected CS articles and columns are available for free at http://ComputingNow.computer.org.

The Promise and Peril of Social Computing



John Riedl
University of Minnesota

Social computing has the potential to fundamentally change the structure of human relationships. Will it succeed?

Welcome to the new Social Computing column. This column, to appear bimonthly in *Computer*, will highlight interesting developments in the world of social computing. In addition to exploring interesting phenomena in the social computing world, these articles will strive to peek into what the future holds by examining recent research.

WHAT IS SOCIAL COMPUTING?

Social computing always involves groups of people interacting in some way, whether working together, playing together, or simply enjoying each others' company. Moreover, this interaction always involves computers. Many, if not most, social activities succeed perfectly well without computing. Humans organized into groups to learn mathematics, create the pyramids, and invent soccer, all without using computers. However, computers are fundamentally changing the way we interact today. One of this column's goals is to understand the ways in which computer-mediated interaction is succeeding and failing, and what the consequences are of both successes and failures.

The column will maintain a broad perspective on what constitutes social computing. Articles will examine collaborative projects like Wikipedia, consumer product review and rating systems like that offered by Amazon, social networking sites like Facebook, content-sharing sites like Flickr, and even microblogging services like Twitter—though with a healthy dose of skepticism about the depth of socializing 140 characters at a time.

At the technological margins, definitions get murky. For example, smartphone calls do not count as social computing even though a call is initiated on one computer, is connected through a packet-switched computer network, and ends on another computer. On the other hand, videocasts count, at least if the viewers can interact with the presenter and one another. At least one column will examine why some videos on YouTube go viral while others are scarcely viewed.

Social computing can also include social organizations that perform actual computing. These "collective intelligence" systems include futures markets like the Iowa Electronic Markets, which predict everything from the outcomes of US congressional elections to box office receipts for the latest *Twilight* movie. Such markets collect and process the opinions of hundreds or thousands of people about some future event to form a prediction of that event. An effective prediction market requires sampling large numbers of participants with divergent views and creating structures that motivate participants to provide honest information.

Although collective intelligence systems are an important aspect of social computing, this column aims construes the term more broadly, and will include many other types of social computing systems as well.

WHO AM I?

As the Social Computing column editor, I should tell you a little about myself. I am a professor in the Department of Computer Science and Engineering at the University of Minnesota, where I codirect the GroupLens Research group (www.grouplens.org). GroupLens began in 1992 as one of the first recommender system research projects.

Our group, which currently consists of three faculty and about 20 students, continues to study these recommender systems, but in the past decade it has branched out to explore

0018-9162/11/$26.00 © 2011 IEEE

SOCIAL COMPUTING AND DUNBAR'S NUMBER

In his provocative book *Grooming, Gossip, and the Evolution of Language* (Harvard Univ. Press, 1998), anthropologist and evolutionary biologist Robin Dunbar popularized the view that language might have originally evolved as a more efficient way to manage social relationships. In a nutshell, his theory is that primate societies depend on strong relationships among individuals, and that maintaining those relationships is an important and expensive activity.

Members of primate societies spend many hours grooming one another—picking out debris and parasites from fur. Studies have shown that if chimpanzee Esther regularly grooms chimpanzee Debbie, then Debbie is much more likely to give a banana to Esther later. But grooming does not scale. If Esther wants to form an equally strong relationship with Fred, she must spend as much time grooming Fred as she does grooming Debbie. Further, grooming is an evolutionarily expensive activity. The more time Esther spends grooming Fred, the less time she has to find food for herself and take care of her offspring. Ultimately, the time required for grooming creates an upper limit on the size of the population in a chimpanzee society: only so many pairwise bonds can be created, limiting the maximum size of a stable society to about 30 chimpanzees.

Dunbar argues that humans invented language because it is superior to grooming—especially if you do not have a taste for nits. Language lets people communicate with many others simultaneously and, crucially, about other people. If Debbie, Esther, and Fred were human, Fred might give the banana to Esther just because he has heard that she is so good at caring for her children. The effect is that human societies can be stable up to much larger sizes. According to Dunbar, tightly connected human social groups tend to include about 150 members, a number limited by the size of our neocortex, the brain structure responsible for, among other things, keeping track of our social relationships.

Social computing systems could eventually lead to an expansion of Dunbar's number. What if we develop social computing systems that are so efficient in maintaining relationships that they support richer and more complex social structures than our poor neocortexes can maintain on their own? Would such a society be better in fundamental ways than existing societies?

social computing issues as well. Lately, our focus has been on Wikipedia, tagging, the movie recommender system and virtual community Movie Lens, online question-answering systems, and Cyclopath, a geowiki for bicylists in the Minneapolis-Saint Paul area.

In addition to my academic research, I was the cofounder and CTO (later chief scientist) of Net Perceptions, a recommender systems startup company that operated from 1996 to 2004.

In this column I will draw on both my experience in industry to select interesting and important issues to discuss and my experience in academia to present intriguing recent research results that speak to those issues. When possible, I will also discuss the results' implications for various types of organizations, including companies.

SOCIAL COMPUTING'S EVOLUTION

Social computing has existed for decades. Many existing forms were present in the PLATO (Programmed Logic for Automated Teaching Operations) system at the University of Illinois in the early 1970s, including multiuser chat rooms, group message boards, and instant messaging. However, during the next two decades, social computing grew relatively slowly. There were many reasons for this, but two of the most important were that many people did not have easy access to the Internet and that the user interfaces for accessing these social tools were confusing and difficult to use.

After a long period of relative hibernation, social computing exploded in the early 1990s and has now become ubiquitous—to the point that in 2006 *Time* magazine named "you" its Person of the Year in recognition of individual content contributions on the Internet. During the past decade, social computing has helped elect presidents and topple governments, and created the largest encyclopedia in history. After such a strong start, what will happen next?

One reason researchers are so interested in social computing is that it has the potential to change the type and structure of human relationships, as the "Social Computing and Dunbar's Number" sidebar explains. Maintaining and growing these relationships, which have been nurtured in much the same way for more than 100,000 years, is one of the most basic human activities.

There is good evidence that computers can help our brains be much more efficient at certain tasks, like computing square roots, and that being better at those tasks can help us create a stronger society. On the other hand, there are many reasons to be skeptical that the social computing systems we are developing can fundamentally increase our social effectiveness. Yes, Facebook makes it easier to remember a distant friend's birthday, but does that kind of sharing really nurture the rich social network that will support us during a hospital stay?

Some researchers worry that social computing might damage the structure of society in important ways—for example, by increasing balkanization. In the physical world I will still have frequent contact with my neighbor even if I disagree with him on politics, while in the virtual world I can easily avoid contact with anyone whose opinions I do not like.

Figure 1, which shows the links among political blogs during the 2004 US election cycle, powerfully illustrates this phenomenon. Nearly all of the links were liberal bloggers linking to other liberal bloggers or conservative bloggers linking to other conservative bloggers. The risk is that if people only talk to others with whom they agree, they will never challenge their own opinions or be open to new ideas. Over time, the ideas of each community will become more extreme and its members less open to interaction with other communities.

Happily, social computing also has demonstrable positive effects. For example, a future column will look at research by Moira Burke and her colleagues at Carnegie Mellon University that correlates having a rich social network on Facebook with feelings of well-being. Students who have more directed communication with their Facebook friends feel stronger bonds to them, and correspondingly experience less loneliness. (On the other hand, students who spend more time reading undirected communication on Facebook feel less connected and lonelier.)

A major question about social computing is whether it is here to stay or just a flash in the pan. For example, as Figure 2 shows, recently there has been a downturn in Wikipedia's incredible exponential growth. Research suggests that new Wikipedia authors find the community inhospitable and often leave after their first entry. The next Social Computing column will examine the possible reasons for this change and what it means for Wikipedia's future.

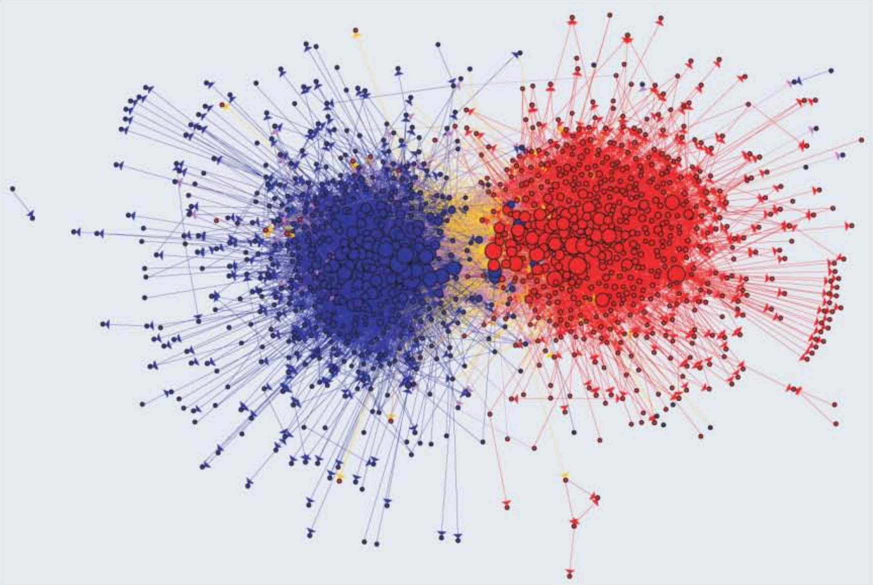


Figure 1. Nearly all of the links among political blogs during the 2004 US election cycle were liberal bloggers (blue) linking to other liberal bloggers or conservative bloggers (red) linking to other conservative bloggers. Figure source: L.A. Adamic and N. Glance, "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog," *Proc. 3rd Int'l Workshop Link Discovery* (LinkKDD 05), ACM Press, 2005, pp. 36-43.

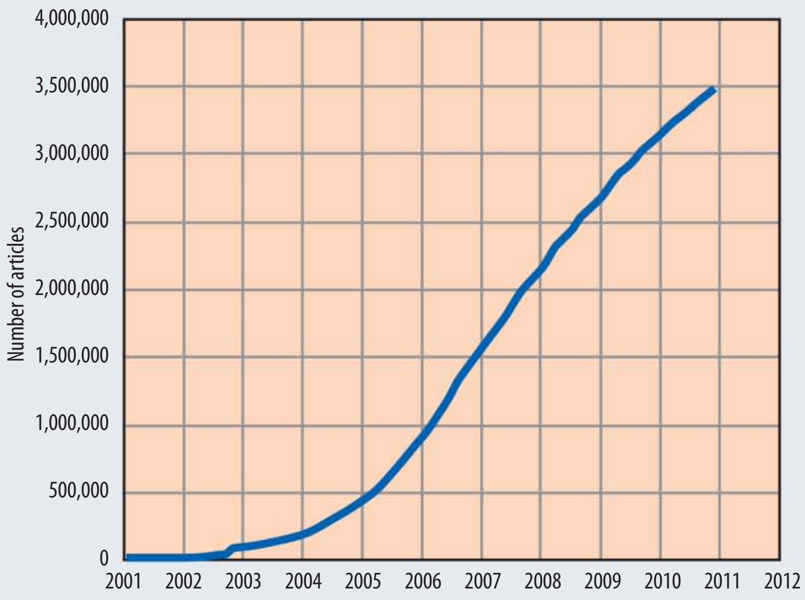


Figure 2. Number of articles on English-language Wikipedia from its creation in 2001 through June 2010. Recently there has been a slow down in the online encyclopedia's incredible exponential growth.

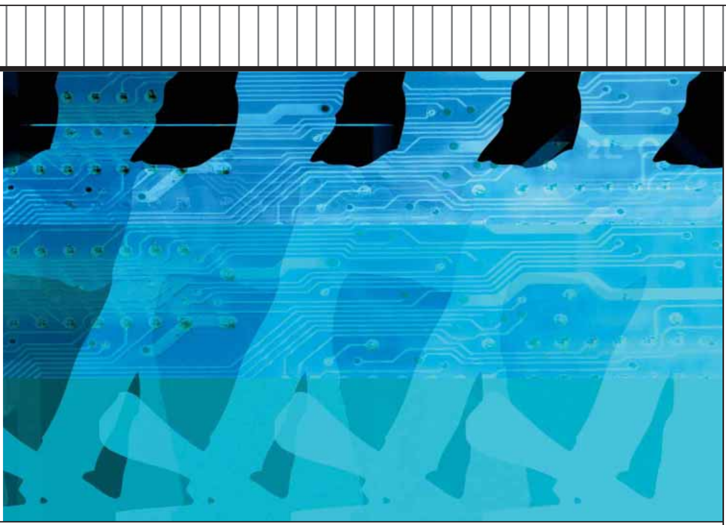
Feedback and suggestions for article ideas and guest authors are always welcome. I will author or coauthor some future columns, but I will also invite colleagues with interesting perspectives to submit their own pieces. Ideally, suggestions should include a concept, a few bullets to outline related ideas, and optionally a recommended author (including self-nominations). I have a queue of upcoming articles and cannot promise that your idea will be published immediately, but I will make an effort to get articles on current events out as quickly as possible.

I look forward to hearing from you—the only way this column can stay relevant is for its readers to actively contribute. C

John Riedl *is a professor in the Department of Computer Science and Engineering at the University of Minnesota. Contact him at riedl@cs.umn.edu.*

Dissecting the Human Identity

Karl Ricanek Jr.
University of North Carolina Wilmington



Biometrics has emerged as the key technology for personal identification, with India leading the way in full-scale deployment of a biometric-based universal ID system.

The terrorist attacks on 9/11 brought about the need for rapid and precise personal identification. Since then, biometrics—which comprises methods for uniquely recognizing humans based on physiological or behavioral traits—has emerged as the key technology for personal ID. For example, fingerprint and face recognition are well-known biometric techniques that exploit physical traits, while handwriting and speaker recognition are popular behavior-based biometric methods.

Researchers are actively exploring new biometric cues, such as the periocular region (soft tissue around the eye) and heartbeat rhythm patterns, as well as working to improve the performance of existing modalities.

MARKET TRENDS

The past decade has seen rapid growth in the development and deployment of personal ID technologies for various market segments including

- *civil identification*—e-passports and e-IDs for border and immigration control;
- *criminal identification*—primarily fingerprints, but face and iris recognition are on the rise;
- *access control*—face, iris, and fingerprint recognition;
- *attendance*—hand geometry recognition;
- *surveillance*—face and gait recognition for tracking on closed-circuit TV;
- *consumer identification*—vascular and vein pattern recognition for purchasing consumer goods and remote transactions; and
- *device/system access*—face, vascular, and vein pattern recognition.

As Table 1 shows, the international market for biometric technology is projected to increase from $3.4 billion in 2009 to $9.3 billion by 2014, a compound average annual growth rate of 22.3 percent. Iris, face, and vascular recognition systems are expected to lead this substantial growth.

Government security initiatives will be the primary drivers for face and iris recognition, while companies will exploit the compact nature of vascular recognition for commercial applications such as user verification on mobile devices.

INDIA'S BIOMETRIC-BASED UNIVERSAL ID

India is the first large country to implement a biometric-based universal ID (UID) and, as such, provides a useful illustration of the technology's promise as well as some of the challenges facing its large-scale deployment.

With a population second only to China, and the world's fourth largest economy, India has the largest number of social service programs, which provide $30 billion in relief to 150 million families each year. However, nearly 600 million Indians, mostly the poor, do not have a definitive ID and thus either cannot access benefits or services or must show some proof of identity for each new benefit or service. The lack of a UID also encourages fraud, which consumes 20 to 40 percent of total social service expenditures (R. Mashruwala and S. Prabhakar, "Multi-Modal Biometrics for One Billion People," presentation, IEEE Computer Society Workshop on Biometrics in association with CVPR, 2006).

To reduce fraud and facilitate access to services, India will issue a unique 12-digit ID number to all citizens, including infants. While the current system requires reauthorization of benefits based on reidentification, the UID program guarantees portability of service as people move from one area to another. Participation is voluntary

Published by the IEEE Computer Society

Table 1. Market volume of biometric technology, 2009-2014.
(Source: IBG, Biometrics Market and Industry Report 2009-2014)

Source	(Unit: M$ USD)					
	2009	2010	2011	2012	2013	2014
Fingerprint	971.0	1,380.0	1,740.1	2,064.1	2,422.9	2,827.2
AFIS/live scan	1,309.1	1,489.9	1,816.5	2,064.1	2,422.9	2,965.8
Iris	174.4	287.8	360.8	480.5	578.3	730.3
Hand geometry	62.0	62.8	63.7	68.2	76.0	85.0
Middleware	275.0	327.7	413.8	525.2	625.2	732.6
Face	390.0	510.8	675.4	848.5	1,097.3	1,417.8
Voice	103.8	109.3	113.5	136.3	167.5	189.7
Vascular	83.0	102.1	132.2	172.2	199.5	235.7
Others	54.0	85.6	107.5	131.8	154.2	184.9
Total	$3,422.3	$4,356.9	$5,423.6	$6,581.2	$7,846.7	$9,368.9

to make it more politically palatable, but people cannot obtain government assistance without a UID.

India has chosen to use biometrics for automatic identity authentication—in particular, face, iris, and 10-print fingerprint recognition—rather than a physical card or token. The use of multimodal biometrics will ensure that everyone can be enrolled in the system, regardless of cultural or religious practices or occupations. For example, Muslim females who wear traditional face and head coverings cannot be enrolled in a face-based biometric system due to heavy occlusion of the face; however, biometric systems can capture their irises and 10-print fingerprint.

As Figure 1 shows, taking fingerprints can be problematic for manual laborers due to the smoothing of their fingers' friction ridges. Iris enrollment can likewise be difficult for very dark-eyed people and even impossible for those with aniridia (absence of an iris). Every biometric cue has problems, but it is rare for a single person to confound a multimodal biometric system.

Because India is a large country with minimal infrastructure outside metropolitan areas, capturing, storing, and securing biometric data present major challenges. In deploying the UID system, the government must address questions such as:

- How do administrators acquire biometric data in rural areas?
- What steps can they take to ensure the integrity of hand-carried biometric data and to prevent theft and forgery? What happens if such data is compromised or stolen? Is it lost forever or can it be reissued?
- What data storage architecture should be used? Can a centralized data repository be made impregnable?

Biometric researchers around the world will be watching to see how India, which ambitiously hopes to enroll 600 million people in the system by 2014, responds to these challenges.

Identity sciences will play a pivotal role in the interactions we have with real-world objects, humans, and digital devices. Researchers and practitioners are integrating identity technology into robotics, human computer interfaces, gaming systems such as Xbox Kinect, mobile devices, door locks, automobiles, and so on. Therefore, I invite all *Computer* readers to actively participate in this column by submitting ideas for future articles or by providing feedback on published articles.

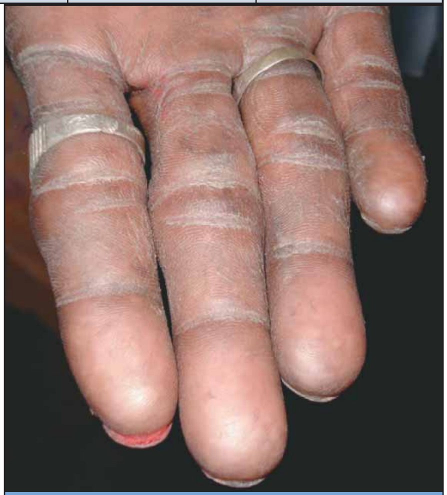


Figure 1. India's universal ID program integrates multiple biometric cues to overcome the limitations of using a single modality. Manual-labor occupations tend to whither fingerprints to the point that they cannot be used for identification. Source: R. Mashruwala and S. Prabhakar, "Multi-Modal Biometrics for One Billion People," presentation, IEEE Computer Society Workshop on Biometrics in association with CVPR 2006.

***Karl Ricanek Jr.**, the Identity Sciences column editor, is director of the Face Aging Group at the University of North Carolina Wilmington. Contact him at ricanekk@uncw.edu.*

Continued from page 100

This last case was interesting because one of the two examination papers contained an extensive set of multiple-choice questions. The other paper required essays to be written. There was no problem getting teachers to mark the essays under contract, but marking the multiple-choice answers was tedious and error-prone, as was the calculation of the overall result.

The quotation for the programming required to do the job on an IBM 650 computer in Sydney was far too high, but fortunately I was able to actually wire the programs by plugging the control panels with connectors to two IBM 604 calculator panels to do the job in Melbourne at a vastly lower cost (tinyurl.com/ibm604). So much for the computer revolution.

A significant aspect of digital technology is its support for and encouragement of ever-increasing complexity.

In all such cases there was no doubt about who the user was, and so no doubt about where my responsibility lay. More often, especially when I worked in the field as a systems engineer, there were many users in several roles, even if only as managers and subordinates.

DIRECT USERS

The challenges of multiple user roles was brought home when I still worked in the service bureau, which by this time had installed a stored program computer, an IBM 1620, that allowed tackling more complex problems.

The management of a large manufacturing firm had a room with quite a few people, each with an Odhner calculator on their desks. The calculator was used each month to derive a schedule of parts and subassemblies needed from a month-by-month schedule of a year's assembly production. The calculation was complex because parts might have multiple suppliers—and because of factors like optional extras and the need for buffer stock—so good judgment was needed.

Designing and writing the program was quite time-consuming. When finished, the customer's management approved the test results, and the program went into service. However, panic broke out in the service bureau when the customer contact called to cancel the job after only two months. A visit to the plant and discussion with management there revealed that the schedulers had complained of the increase in workload from their having to check the details on the printouts from the computer.

Management proved to be the problem, with the job's planning being supervised by management—the indirect users. The direct users—the schedulers—had only been involved in explaining the details of their calculation to me. This gave me a striking lesson in system implementation.

We saved the situation after lengthy discussions with the schedulers, particularly their union representatives, bringing home the point that their job would be much more interesting and better done if they would take the printouts as being close enough, without detailed checking. This would free them to be in much closer touch with their suppliers, by onsite visits as well as by phone, so that they could find out what was actually happening and be able to cope with any supplier hitch before it hit the production line.

INDIRECT USERS

In the 1960s, working as a systems engineer outside the service bureau meant spending most of my time in customers' offices, mostly in an advisory mode with programmers and their managers (The Profession, Nov. 2004, pp. 126-128). This often involved discussing details with direct users. Sometimes these details revealed a lack of understanding of the direct users' work by the indirect users. In such cases, office politics often made it effective for me or my salesman partner to act as a mediator.

A more interesting aspect of working with indirect users sometimes arose when we were asked to help plan for introducing a digital computer into a company with little prior experience using such machinery. The difficulty lay in getting agreement to a coherent company-wide approach to project objectives and priorities.

IBM's practice in such cases was to persuade top management to attend a two-day meeting, ostensibly to set up such objectives and priorities. The meeting would be held away from their offices and homes so that they would not be distracted by day-to-day issues. On the first day, they would each explain to us how their department worked so that, on the second day, we would jointly discuss and agree on the objectives and priorities.

I particularly remember one such meeting. It was run, as usual, by an IBM specialist. He used a stand for pads of what we called butchers' paper and a supply of colored felt-tipped pens to write with. On the first day, the specialist questioned the managers one by one about their functions and responsibilities, while fielding occasional questions from other attendees.

As the questions were answered, the specialist wrote the key words as bullet points on a sheet of butchers' paper displayed on the stand. The meeting room walls had rails with clips along them for the filled sheets to hang from.

The first day ended with the attendees completely surrounded by bullet points, which they studied and discussed informally before going to dinner. On the second day, these bullet points were reorganized into categories of company-wide significance and used for compiling a list of general objectives with dependencies and priorities.

The striking thing about this meeting was the number of times managers would demonstrate by their questions an ignorance of how other departments worked. Indeed, it was IBM's covert objective for such meetings to break down the barriers between departments at the level of top management, the ultimate indirect users. The formal and informal thanks given to the IBMers usually stressed the learning experience as much as the worth of the plans drawn up.

THE CYCLE

The preceding description relates to the 1960s. The 1970s were quite different, at least in my experience, because of the Data Processing Department's rise.

As the DP department grew and took on more and more responsibilities, it became less and less responsive to its users and increasingly powerful politically within the organization. As the applications and data files became more complex, they became ever-more mysterious to the various users and their managers. This allowed the DP system analysts and managers to dictate what applications would be implemented and how they would affect the users, especially the direct users who often became automatons driven by DP software.

IBM systems engineers became ever-more distant from their customers, becoming increasingly invested in putting together responses to DP department specifications while helping programmers and analysts deal with crises resulting from over-ambitious projects. The more interesting work with users was often forbidden to us.

By the 1980s, I was out of normal fieldwork, and stayed out. But I watched with interest the effect of cheap personal computers on the traditional DP politics and the accompanying culture as it evolved into the more pretentious IT industry.

In its early stages, IT was like the DP of the 1960s. PCs and their generic software enabled direct users to take back more of the responsibility for their own work from the DP department. However, as PC software became increasingly complex, the direct users' work became constrained by what the software allowed, in much the way DP operated in the 1970s.

Then came the Internet and the complex applications based on it, and on the massive centralized databases the Internet made worthwhile. This returned political power to the old DP departments, now called IT departments.

A significant aspect of digital technology is its support for and encouragement of ever-increasing complexity. This burgeoning complexity is hard for the IT departments of business and government to handle individually. Hence, the rise of outsourcing, which is in effect a return to the service bureau paradigm.

These personal and extreme generalizations relate to large-scale data processing and suggest that in such areas the interests of the direct user are sadly and irresponsibly neglected. Large organizations greatly reduce the significance of direct users, requiring them to interact with their software as that software dictates.

Personal computing is not all that different. Much videogaming requires unthinking reaction to the stimulus of the software, for example. The difference here is that the suppliers of personal software benefit from enticing the direct user into persisting in using it and its successors. Corporations, on the other hand, seek to use software to reduce their need for employees at the level of direct users, which is partly how a demand for outsourcing arose.

The important question is whether the computing profession and the teachers of computing are supporting such attitudes or not.

Neville Holmes*, editor of the The Profession column, is an honorary research associate at the University of Tasmania's School of Computing and Information Systems. Contact him at neville.holmes@utas.edu.au.*

Computers and Their Users

Neville Holmes, *University of Tasmania*

Computer uses and users multiply in ever greater variety.

Employees who have worked for IBM for 25 years join the IBM Quarter Century Club, an autonomous social group. Having recently moved back to mainland Australia, I have been attending local QCC dinners with great enjoyment.

At last November's dinner, I had an interesting conversation with a current IBM employee in which we compared the IBM of today with the IBM I started with more than 50 years ago. The gist of this was that systems engineers like myself once went out to customers, but customers now come to IBM for systems solutions.

The reason for the difference would seem to be that around the 1960s, digital computers, being very costly, were usually leased rather than bought, though system software was free. It was therefore important that customers got value for their money so that they would renew their leases each year. The key to satisfied customers was not to solve their problems for them but to teach them how to solve them on their own. Customers now come to IBM seeking services.

Curiously, I recently came across an abstract of a book on IBM Australia's changing nature that "traces the move of IBM from Australia's leading exporter of elaborately transformed manufactured goods in the 1990s to becoming Australia's largest exporter of IT services today" (http://qccaustralia.org). From my discussion at the QCC dinner, I got the strong impression that the services are directed at solving customers' problems for them.

OUTSOURCING

Much of the IT service industry focuses on outsourcing. This seems like an avoidance of responsibility by managers as much as a search for the usually spouted cost savings. Add to this the frequent failures and cost overruns of digital megaprojects, and the question of professional responsibility comes to the fore.

Such responsibility is complex (The Profession, July 2003, pp. 98-100). Beyond the broad social responsibility of any formal professional, just as for doctors and lawyers, for example, the computing professional's primary responsibility is to the client user. But who is that?

Often this is obvious. Shortly after I joined IBM, I was assigned to a service bureau for training, moving from machine operation to plugging programs. The service bureaus were the forerunners of outsourcers—or should that be *insourcers*? The main customers were small enterprises that couldn't afford their own data processing equipment, larger businesses that needed help with their monthly and yearly processing peaks, or lessees preparing for the installation of new equipment.

The larger businesses took responsibility for all but the physical processing. The new lessees were trained to take responsibility. But the small businesses and individuals were the most interesting because I had to work directly in partnership with them to design and implement a solution to their problem. The user in such cases was normally the person I worked with, the person who would use the printed output.

Most memorable were a researcher at a medical institute who needed an analysis of data on sleeping sickness in New Guinea, a manager in a small manufacturing firm who needed job cost details summarized monthly, and a government officer who needed help with processing the results of a university entrance examination in chemistry.

Continued on page 98

Published by the IEEE Computer Society

0018-9162/11/$26.00 © 2011 IEEE