COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

11/2013 VOL.56 NO.11

Technologies for Caring for People with Dementia

Association for Computing Machinery

28th IEEE INTERNATIONAL PARALLEL & DISTRIBUTED PROCESSING SYMPOSIUM

IPDPS 2014 PHOENIX

www.ipdps.org

19-23 May 2014
Arizona Grand Resort
Phoenix, Arizona USA

CALL FOR PARTICIPATION

The 28th IEEE-IPDPS will be held 19-23 May in Phoenix at the Arizona Grand Resort, an all-suites, family-friendly resort and meeting place. Phoenix has an international airport and is the hub for several US airlines. It is only a six hour drive from Los Angeles and Las Vegas and the starting point for one day tours of the Grand Canyon as well as other historic communities in the area. The IPDPS event offers a rich program of workshops, a PhD Forum, and the technical program of contributed papers, invited speakers, and tutorials. Check the IPDPS Web pages for updates and information on workshops, the PhD Forum, and other events as the full program develops.

GENERAL CHAIR
Manish PARASHAR (Rutgers University, USA)

PROGRAM CHAIR
David A. BADER (Georgia Institute of Technology, USA)

PROGRAM VICE-CHAIRS
- *Algorithms*
  Anne BENOIT (ENS Lyon, France)
- *Applications*
  Padma RAGHAVAN (Pennsylvania State University, USA)
- *Architecture*
  Yan SOLIHIN (North Carolina State University, USA)
- *Software*
  Ron BRIGHTWELL (Sandia National Laboratories, USA)

WORKSHOPS CHAIR
Ümit V. ÇATALYÜREK (Ohio State University, USA)

WORKSHOPS VICE-CHAIR
Yogesh SIMMHAN (University of Southern California, USA)

PhD FORUM CO-CHAIRS
Luc BOUGÉ (ENS Cachan, France)
Bo HONG (Georgia Institute of Technology, USA)

COMMERCIAL & INDUSTRY CHAIR
Kalyana CHADALAVADA (NCSA, USA)

IPDPS extends a special invitation to Phoenix and Southwest area companies to consider participation in IPDPS 2014 to share in the benefits of associating with an international community of top researchers and practitioners in fields related to parallel processing and distributed computing. Visit the IPDPS Website for details.

IPDPS 2014 WORKSHOPS

Held on the first and last day of the conference, workshops are a major part of the IPDPS week-long compendium of events and provide the IPDPS community with an opportunity to explore special topics and/or present work that is more preliminary or cutting-edge than the more mature research presented in the main symposium. Proceedings of the workshops are published by the IEEE Digital Library and are distributed at the conference. Please note that each workshop has its own requirements and schedule for submissions, and all the workshops are linked from the workshops page on the IPDPS Website. **Also, the submission deadline for most workshops is after the main conference author notification date of December 9th.**

HCW	Heterogeneity in Computing Workshop
RAW	Reconfigurable Architectures Workshop
HIPS	High-Level Parallel Programming Models & Supportive Environments
NIDISC	Nature Inspired Distributed Computing
HiCOMB	High Performance Computational Biology
APDCM	Advances in Parallel and Distributed Computing Models
HPPAC	High-Performance, Power-Aware Computing
HPGC	High-Performance Grid and Cloud Computing
ASHES	Accelerators and Hybrid Exascale Systems
PLC	Programming Models, Languages and Compilers Workshop for Manycore and Heterogeneous Architectures
EduPar	NSF/TCPP Workshop on Parallel and Distributed Computing Education
PDSEC	Parallel and Distributed Scientific and Engineering Computing
DPDNS	Dependable Parallel, Distributed and Network-Centric Systems
MTAAP	Multi-Threaded Architectures and Applications
LSPP	Large-Scale Parallel Processing
PCO	Parallel Computing and Optimization
ParLearning	Parallel and Distributed Computing for Machine Learning and Inference Problems
HPDIC	High Performance Data Intensive Computing
CloudFlow	Workflow Models, Systems, Services and Applications in the Cloud
JSSPP	Job Scheduling Strategies for Parallel Processing
ViPES	Virtual Prototyping of Parallel and Embedded Systems

IPDPS 2014 PhD FORUM & STUDENT PROGRAM

The annual IPDPS PhD Forum event will continue with the traditional poster presentations by students working toward a PhD in broadly defined areas related to parallel and distributed processing. However, at IPDPS 2014, we will have an enhanced student program to include several sessions to provide coaching in scientific writing and presentation skills and to give students an opportunity to hear from and interact with senior researchers attending the conference. Visit the IPDPS Website for details.

Sponsored By: IEEE computer society
Technical Committee on Parallel Processing

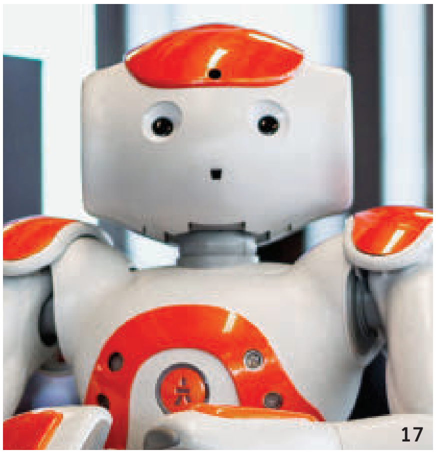
In Cooperation With:
ACM SIGARCH
IEEE Computer Society Technical Committee on Computer Architecture
IEEE Computer Society Technical Committee on Distributed Processing

COMMUNICATIONS OF THE ACM

Departments

Last Byte

News

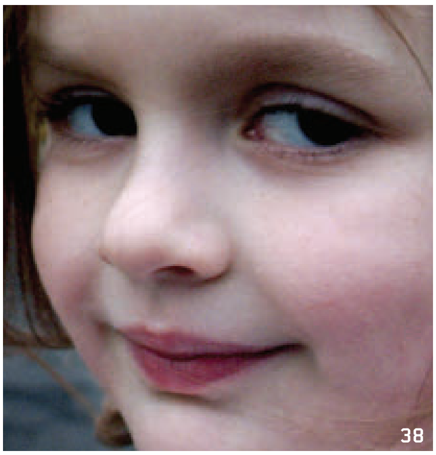

17

Viewpoints

PHOTOGRAPH BY PATRICK HIGGINS/UNIVERSITY OF SOUTHERN MAINE FREE PRESS

11/2013
VOL. 56 NO. 11

Practice

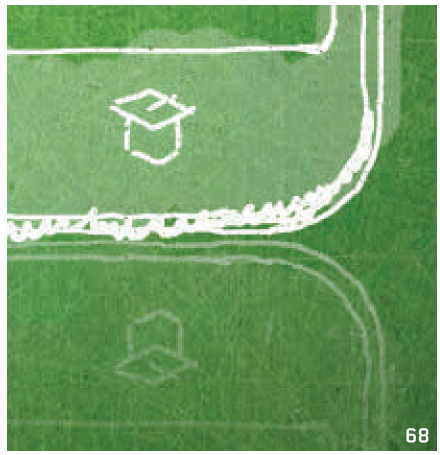


Articles' development led by acmqueue
queue.acm.org

About the Cover: While mobile technologies exist to support those who care for people with dementia, the challenge is to align these app uses with the range of different care strategies. This month's cover story (p. 60) details ways to be more sensitive to these differences when designing tools to help those who help others. Cover photo illustration by Brian Greenberg/Andrij Borys Associates, portrait photograph by Robert Bentley Harrison.

PHOTOGRAPH BY DAVE ROTH, ILLUSTRATION BY ALEKSEI GLUSHCHENKO

Contributed Articles



Review Articles

Research Highlights

Association for Computing Machinery
Advancing Computing as a Science & Profession

COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

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

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

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

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

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

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

Association for Computing Machinery

Printed in the U.S.A.

DOI:10.1145/2524713.2524714 Moshe Y. Vardi

The End of The American Network

Can the Internet be liberated from government meddling?

As details of the U.S. National Security Agency's (NSA) pervasive phone and Internet eavesdropping and surveillance operations emerged last summer, sales of George Orwell's classic novel, *1984*, were reported to have risen dramatically. Orwell described an oppressive government that continually monitors the population through ever-present "telescreens." A recent newspaper article proclaimed that "NSA surveillance programs greatly exceed anything the *1984* author could have imagined."

Indeed, what we have learned so far about the wide reach of the NSA's operations has been quite astounding. First, we learned the NSA is collecting on an ongoing basis phone records of essentially all U.S. telecommunications customers. Second, we learned the NSA monitors all Internet traffic that goes through U.S. telecommunications infrastructure. Third, we learned the NSA has intentionally weakened cryptographic standards to enable it to circumvent encryption. All of this has been authorized by secret court orders. In essence, we learned the U.S. government has stretched the meaning of "reasonable" in the U.S. Constitution Bill of Right's proscription against unreasonable searches beyond reasonableness. NSA, indeed, is certainly making Orwell's surveillance technology seem rather primitive in comparison.

The unfolding scandal reminds me of another aspect of Orwell's novel. The language spoken in Oceania, the novel's fictional country, is "Newspeak," whose grammar is based on English, but whose vocabulary is tightly controlled in an effort to limit free thought. When NSA Director James Clapper was asked, following this summer's revelations, to explain his answers given in U.S. Congress testimony earlier in the spring, he replied: "I responded in what I thought was the most truthful, or least untruthful manner." I have no doubt that Orwell would have been proud to add the phrase "least untruthful manner" to the vocabulary of Newspeak. Granted, NSA is, after all, an intelligence agency, and countries do spy on each other (at their own risk!), but the NSA is not supposed to be a domestic intelligence agency and it is not supposed to lie to the U.S. Congress!

This phrase "least untruthful manner" symbolizes for me the most disturbing aspect of the NSA scandal. The U.S. government, supposedly "of the people, by the people, for the people," to quote Abraham Lincoln's Gettysburg Address, has been untruthful to its citizens for several years and has been coercing many U.S. corporations that operate phone and Internet infrastructure to be equally untruthful. An old joke asks: "How can you tell when a lawyer is lying?" Answer: "His lips are moving." There is no need anymore to pick on lawyers; we can substitute "NSA official" in the joke. Our trust in the U.S. government and U.S. corporations has been broken. It is unlikely to be repaired in the near future.

This means, I believe, we can no longer trust the U.S. government to be the "Internet hegemon." During the 1990s, when the Internet was rising while Minitel, the French phone-based online service, was declining (it was finally retired in 2012), France's President Jacques Chirac complained about rising dominance of the Internet, which he described as "the American Internet." While many of us snickered at his provinciality, Chirac was right. The National Telecommunications and Information Administration, an agency of the U.S. Department of Commerce, continues to have final approval over changes to the DNS root zone. Thus, in spite of its being a globally distributed system, the Internet is ultimately controlled by the U.S. government. This enables the U.S. government to conduct Internet surveillance operations that would have been impossible without this degree of control.

The main argument in favor of the privileged position of the U.S. government in Internet governance is that other governments, which have been clamoring for true internationalization of Internet governance, perhaps through the International Telecommunication Union, were viewed as less trustworthy than the U.S. government. With the trustworthiness of the latter in serious decline due to the NSA scandal, voices are rising again in protest of U.S. Internet hegemony. The question being raised is "Can there be a non-U.S. Internet?" In fact, Brazil has already laid out a multipoint plan to sever ties with U.S.-controlled cyberspace.

But replacing U.S. hegemony with hegemony by other governments, who may not only have their own surveillance operations but also attempt to regulate content and restrict free expression on the Internet, hardly seems an improvement to me. The real question, I believe, is whether we can have an Internet that is free, or at least freer, from government meddling than today's Internet. In view of the Internet's centrality in our information-saturated lives, this is a question of the utmost importance.

Moshe Y. Vardi, EDITOR-IN-CHIEF

Association for Computing Machinery

Advancing Computing as a Science & Profession

membership application & *digital library* order form

Priority Code: AD13

You can join ACM in several easy ways:

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

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

Special rates for residents of developing countries:
http://www.acm.org/membership/L2-3/

Special rates for members of sister societies:
http://www.acm.org/membership/dues.html

Please print clearly

Name

Address

City | State/Province | Postal code/Zip

Country | E-mail address

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

Purposes of ACM

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

I agree with the Purposes of ACM:

Signature

ACM Code of Ethics:
http://www.acm.org/about/code-of-ethics

choose one membership option:

PROFESSIONAL MEMBERSHIP:

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

STUDENT MEMBERSHIP:

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

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

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

RETURN COMPLETED APPLICATION TO:

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

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

Satisfaction Guaranteed!

payment:

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

o Visa/MasterCard   o American Express   o Check/money order

o Professional Member Dues ($99 or $198) $ ____________

o ACM Digital Library ($99) $ ____________

o Student Member Dues ($19, $42, or $62) $ ____________

Total Amount Due $ ____________

Card # | Expiration date

Signature

DOI:10.1145/2524713.2524715

Vinton G. Cerf

Heidelberg Laureate Forum

As I write this column, I am in Heidelberg attending the first Heidelberg Laureate Forum (HLF). In short, this event is a convening of Turing, Fields, Nevanlinna, and Abel Prize honorees together with young researchers competitively selected from a large field of candidates. (For more details about HLF, see http://www.heidelberg-laureate-forum.org/.)

This inaugural Forum is sponsored by the Klaus Tschira Foundation, ACM, the International Mathematical Union, and the Norwegian Academy of Science and Letters. Six of our members, among others, were instrumental in organizing this Forum: ACM CEO John White, ACM Europe chair Fabrizio Gagliardi, ACM Europe Council member Gerhard Schimpf, Turing Award Committee chair Jennifer Chayes, and Turing recipients Juris Hartmanis (1993) and Joseph Sifakis (2007). Indeed, there are 27 Turing Award recipients here in Heidelberg participating in HLF.

The Forum, though not yet over, is already having an effect on my thinking about computability and computation. Giants in our field and in mathematics explored ideas of computability beginning many decades ago. I think of Gödel, Turing, Church, Scott, and von Neumann among many others who have wrestled with the question of decidability and its companion, computability. As a computer scientist and mathematician by training, but an engineer by profession, I struggle to bridge a cognitive gap I feel between the computing of a function and the behavior of large-scale systems such as the Internet and the World Wide Web, air traffic control systems, interactive computer games, and email services. It is one thing to ask whether there is a computation that will yield a specific number (for example, how many ways can you fold a piece of paper if limited to no more than *N* folds for some value of *N*?) and another to characterize what is being computed by an email service system.

Abstraction is a powerful tool that can reduce complexity sufficiently to render computable questions about the behavior of complex systems. Leonard Kleinrock's work on the theory of networks of queues seems a fine example. Complex store-and-forward networks are reduced to their key skeletal components: traffic statistics, buffer space, capacity of links, and topology of nodes whose analysis yields estimates of delay and throughput in the form of averages or moments. Kleinrock's famous Independence Assumption tamed complexity to yield useful results.

A recent announcement of an object called an *amplituhedron* (http://en.wikipedia.org/wiki/Amplituhedron) illustrates the power of abstraction. Without delving into details I do not understand well enough to articulate, suffice it to say this construct allows for the dramatically easier computation of the probabilities of fundamental particle interactions that had to be described earlier with a potentially unlimited number of Feynman diagrams (http://en.wikipedia.org/wiki/Feynman_diagrams). The amplituhedron abstraction simplifies the computation while apparently producing very accurate and testable results.

These observations reinforce in my mind the importance and power of well-chosen models and representations of problems. In mathematics, a change of variables can often reveal important symmetries or other properties inherent in complex problems that point the way to solution. Think about the difference in formulas that use polar coordinates rather then Cartesian coordinates for their expression. The former can often yield structural insights that, in the latter, are obscured.

These thoughts also lead me to wonder about the nature of quantum computation and whether there is more to this notion than merely the speeding up of conventionally computable results. A bit of Web searching yields the observation that classical Turing Machines can be related to Quantum Turing Machines (http://en.wikipedia.org/wiki/Quantum_Turing_machine) by means of stochastic matrices. Whether this means a classical Turing Machine can compute anything a Quantum Turing Machine can calculate is still unclear to me but I hope to find an answer at this Forum!

The HLF is a daring exploration into the relationship between computer science and mathematics, both of which have influenced one another over the course of several decades. I hope this exploration will continue at future Forums and that the insights derived by this process will stimulate new work by computer scientists and mathematicians. It may even result in the design of new kinds of computer architecture, taking advantage of the ideas exposed in this magical and historically significant city.

Oh, and if you get the chance, you should ask the attendees about the spectacular saxophone concert by four incredibly talented women presented during the opening ceremonies. The proceedings were recorded. If you like jazz, you will love the energy these performers generated!

Vinton G. Cerf, ACM PRESIDENT

DOI:10.1145/2524713.2524716

Microprocessor Architectures Follow Markets and Silicon

The viewpoint "The Value of Microprocessor Designs" by Ana Aizcorbe et al. (Feb. 2013) aimed to analyze the value of microarchitectures in isolation, as though they could be mixed-and-matched with various silicon implementation technologies over the years. This is a nonsensical proposition. For example, a Pentium III microarchitecture could not have been realized in 0.8u technology due to insufficient transistors. A 486 microarchitecture could be instantiated on, say, 90nm technology but would have been too slow to be competitive. And the design trade-offs baked into the 486 pipeline, appropriate for the silicon of the 1980s, would not leverage the much larger transistor budgets of the 1990s and later.

These microarchitectures were not independent of one another, as Aizcorbe et al. implicitly assumed. The Pentium III was the same microarchitecture as the Pentium II but with SSE instructions added. Moreover, both came from the original Pentium Pro P6 microarchitecture. The Pentium-M was also a descendant of the P6. The Pentium 4 microarchitecture was substantially different, but, as chief architect of both P6 and Pentium 4, I can testify that the Pentium 4 was not unrelated to P6.

Aizcorbe et al. also said, "The Pentium 4 design provided very little value to Intel" due to "overheating." Incorrect. The Pentium 4 was designed to the limits of air-cooling but not beyond those limits. It did not overheat nor suffer heat-related reliability issues. Had the Pentium 4 been designed with a much different microarchitecture, Aizcorbe et al. suggested Intel might have earned much higher profits. What other microarchitecture? They shed no light on this question. Neither Aizcorbe et al. nor Intel can possibly know what might have happened had the Pentium 4 been designed another way.

They also missed another important real-world issue: fab capacity and opportunity cost. Chip manufacturers can always lower the price to make demand go up, but the bottom line might suffer. Our goal in the Pentium product line was to design chips that would command high prices but also generate high demand. And we did it in the face of a market that evolved dramatically from the 486 days, when only engineers and scientists could afford PCs, through the early 1990s, when home computers were becoming common, through the late 1990s, when the Internet took off.

Aizcorbe et al. did mention such influences but did not take them into account in drawing their conclusions. Each of these market exigencies influenced the design, as well as the final results, in special ways. Microarchitectures are intrinsically bound to their markets, and their implementation technologies and are not fungible per their assumptions.

Bottom line, Aizcorbe et al. observed that Intel did not earn the same profits on all its products over the years. Some, it turns out, paid off more than others. We can agree on that much. But saying so is not news.

Robert (Bob) Colwell, Portland, OR

Authors' Response:

We welcome Colwell's comments but disagree on one major point: Our methodology did not mix-and-match different microarchitectures and silicon implementation technologies, as he suggests. We used only those combinations actually produced by Intel at a given point in time. However, our approach does hinge on the assumption that even when a new microarchitecture is available, Intel cannot retrofit all its current capacity to use it. As a result, old and new microarchitectures are used concurrently in production, allowing us to infer the incremental value of the new microarchitecture.

Ana Aizcorbe, Washington, D.C.
Samuel Kortum, New Haven, CT
Unni Pillai, Albany, NY

Computer Science Not for All

I am somewhat disturbed by ACM's effort to increase the amount of computer science covered in a general education, as laid out in Vinton G. Cerf's editorial "Computer Science Education–Revisited" (Aug. 2013). Not every student needs to train as a computer scientist. Few will ever enter the field, though all should have the opportunity. However, forcing them to learn some amount of computer science would be a mistake. They already have plenty to cover.

Computer science ultimately comes down to the various areas of mathematics, though many non-computer scientists use mathematics as well. Engineers need good command of calculus. Cryptographers need good command of discrete mathematics. Anyone who balances a checkbook must know addition and subtraction.

I would be more comfortable if ACM advocated using its resources and the Web to share its knowledge, offer online courses, and provide a social media presence so motivated students of computer science could gather and share their experiences.

Perhaps ACM should aim to establish a "Future Computer Scientists" program, coaxed, not forced.

Wynn A. Rostek, Titusville, FL

Author's Response:

Perhaps I should have been clearer. Computer science deserves equal status with mathematics, chemistry, biology, and physics in terms of "science curriculum." For most students who intend to go on to college, some number of science courses is required. To the same degree other science classes are required, computer science should count as one of those that can fulfill such a requirement. Course content should include programming, not simply use of computer-based applications.

Vinton G. Cerf, ACM President

Leverage Credentials to Limit Liability Premiums

In response to Vinton G. Cerf's editorial "But Officer, I Was Only Programming at 100 Lines Per Hour!" (July 2013), I would like to add that as long as physical damage is caused by machines, vehicles, and medical devices, their manufacturers can be held liable, whether or not a software problem is the root cause. If damage is due to add-on software (such as an app running down the batteries of a navigation device or an app crashing a car by interfering with car-to-car communication), the software manufacturer could be liable. When liability insurance premiums can be lowered by demonstrating professional qualifications, the option of acquiring personal or staff certification will be embraced by software development professionals, though, perhaps, not enthusiastically.

Ulf Kurella, Regensburg, Germany

I view myself as a "formal methodist" in the professional world of software engineering, though reality is less grand. I "manage" the application of formal methods and also (try to) sell the formal-methods concept to upper management in my organization, in part by quoting Vinton G. Cerf's fellow Turing-Award laureates E.W. Dijkstra and C.A.R. Hoare, in addition to A.M. Turing himself. (It is, of course, no reflection on these immortals that the success of my pitch is modest at best.) I also quote Cerf, context-free (July 2013): "[N]o one really knows how to write absolutely bug-free code…" and repeat the same disclaimer regarding my own (lack of) impact.

Rumor has it IEEE will soon adopt a professional-engineer license in software engineering. I hold its purported precursor, the IEEE Certified Software Development Professional (http://www.computer.org/portal/web/certification/csdp), but do not recall whether I answered the single formal-methods question, out of 180 questions, correctly in 2009 when I took and apparently passed the test.

The back-and-forth between ACM and IEEE regarding licensure and certification is interesting and probably necessary. I thank Cerf for eschewing any shrillness while providing much-appreciated humor.

George Hacken, Wayne, NJ

With the likely advances in artificial intelligence and robotics, work will eventually not be essential at all, at least not for physical sustenance.

Work Still Worth Doing

My essay "The World Without Work," included in the 1997 book *Beyond Calculation*, curated by Peter J. Denning for ACM's golden anniversary, weighed many of the issues Martin Ford addressed more recently in his Viewpoint "Could Artificial Intelligence Create an Unemployment Crisis?" (July 2013) where he wrote, "Many extremely difficult issues could arise, including finding ways for people to occupy their time and remain productive in a world where work was becoming less available and less essential." I would go further; with the likely advances in artificial intelligence and robotics, work will eventually not be essential at all, at least not for physical sustenance. Our stubbornly high levels of unemployment, likely to rise dramatically in the future, already reflect this prospect.

Certain classes of people (such as the elderly, the young, and the disabled) are generally already not expected to work. Why would anyone be required to work at all? Why not just have the government provide some basic level of support for the asking, with work something people do only to go beyond that basic level or for their own satisfaction? We are not far from this being feasible, solving the unemployment problem in a single stroke. The cost, though enormous, would be covered by harvesting the additional wealth created by society's overall increased productivity and lowered costs, the so-called Wal-Mart effect, taken to the limit. Though tax rates would have to increase, after-tax purchasing power would not have to decrease. The remaining work would be more highly compensated, as it should be.

What work is already not being done? Consider the prompt filling of potholes in our streets. If we can have self-driving cars, we can certainly have automated pothole-fillers.

Ford also said, "If, someday, machines can match or even exceed the ability of a human being to conceive new ideas… then it becomes somewhat difficult to imagine just what jobs might be left for even the most capable human workers." But some jobs will never be automated. Consider ballet dancers and preachers. With ballet dancers the audience is captivated by how gracefully they overcome the physical constraints of their bodies. An automated ballet dancer, no matter how graceful, would be nowhere near as interesting as a human one. Preachers inspire us by establishing emotional rapport; no robot, no matter how eloquent, will ever do that.

Paul W. Abrahams, Deerfield, MA

Corrections

The second author (following L. Irani) of the article "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk" (http://www.ics.uci.edu/~lirani/Irani-Silberman-Turkopticon-camready.pdf) cited in the "Further Reading" section of Paul Hyman's news story "Software Aims to Ensure Fairness in Crowdsourcing Projects" (Aug. 2013) should have been M. Silberman, not M. Silverman. Also, though not included in Hyman's article, it should be noted Silberman (of the University of California, Irvine) was a co-designer and primary builder (along with Irani) of the Turkopticon.

In Madhav Marathe et al.'s review article "Computational Epidemiology" (July 2013), the labels in Figure 2 (page 91) I(2) and I(3) were inadvertently switched; the epicurve should have been (1,1,1,2).

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

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

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

DOI:10.1145/2524713.2524717 http://cacm.acm.org/blogs/blog-cacm

Considering Privacy Issues in the Context of Google Glass

Jason Hong ponders why there has been so much negative press coverage of Google Glass with regard to privacy, considering the issue from two different perspectives.

Jason Hong
"Privacy and Google Glass"
http://cacm.acm.org/blogs/blog-cacm/167230-privacy-and-google-glass/fulltext

Google Glass, the wearable headset computer that is just starting to trickle out, has seen a lot of press, much of it negative on the issue of privacy. As someone who has studied mobile computing and privacy for about a decade, I thought it would be useful to examine why there has been such negative sentiment, looking at people's concerns from two different perspectives, and see what lessons can be drawn. (For disclosure purposes, Google has funded my research several times, and will be sending my colleagues and I several units of Glass for research purposes).

Lessons from the Ubiquitous Computing Project

Let's start out by rewinding back to 1991, when Ubiquitous Computing was first introduced to the world at large. In the seminal paper "The Computer for the 21st Century," Mark Weiser presented a grand vision in which one day computation, communication, and sensing would be enmeshed in the everyday world, and could be used to seamlessly support us in our daily lives.

Weiser led his team in developing new form factors for computation at three physical scales: tabs, pads, and boards. They also developed early forms of location-based services and context-aware computing, based on wearable badges that could be used to pinpoint one's location inside a building. While this vision may seem like old hat now, one has to keep in mind that this work was started only a few years after the first Macintosh came out, and mobile phones were still bulky affairs that were yet to be popular.

The researchers at PARC were, by and large, very enthusiastic about the potential of ubiquitous computing. However, the popular press was decidedly not. There were a number of negative news articles, with titles like "The Boss That Never Blinks," "Orwellian Dream Come True: A Badge That Pinpoints You," and "You're Not Paranoid: They Really Are Watching You."

I suspect there were two reasons why there was a backlash. First, the PARC researchers had not built any kinds of privacy protections into their system. While they knew privacy was a concern from the start, they did not have good ideas for how to address the problems. Furthermore, they had to devote a lot of effort on just making the technologies even work in the first place. However, when journalists asked the inevitable questions about privacy, the researchers did not have any good answers. To some extent, this scenario is playing out in the exact same way for Google Glass. Many people are asking reasonable questions about privacy, and I have not heard any solid responses yet about how those concerns will be addressed.

Second, there was an unclear value proposition for end users. The PARC team was comprised mostly of people from technical backgrounds, and when talking with journalists, the discussion usually centered on how the technologies worked rather than what benefit they could offer people. Interestingly, the narrative seemed to shift when the researchers started framing things in the form of "Invisible Computing," talking about how these systems could support people as they go about their everyday lives. Google is doing better on this front, by pushing the fashion aspects of Glass, by having some concept videos of what Glass can offer, and by having lots of non-developers try out the system. However, Google still has a long way to go in conveying what value it can actually offer today to everyday people.

This notion of the value proposition has been seen in the success and failure of many groupware systems as well. Jonathan Grudin, a scientist at Microsoft Research, long ago observed that those who do the work in using a groupware system have to be the same as those who get the benefits, otherwise the system is likely to fail or be subverted. My privacy corollary is that when those who bear the privacy risks do not benefit in proportion to the perceived risks, the technology is likely to fail. Right now, the implicit narrative in the popular press is that many people could be surreptitiously monitored by users of Google Glass at any time, and do not perceive any kind of value in return. Unless this perception is changed, it is likely there will continue to be negative perceptions of Glass outside of the core of early adopters.

Expectations of Privacy Change

The second perspective I will use for thinking about Google Glass and privacy is that of expectations.

One of my favorite papers about expectations is by Leysia Palen, an HCI researcher at the University of Colorado, Boulder. In 2000, Palen presented a paper at the Computer Supported Cooperative Work conference looking at the behaviors and practices of new mobile phone users. One finding was that these new users were not particularly good at predicting what their own attitudes and behaviors would be a month after getting their first mobile phone. For example, before they got their mobile phone, many participants reported being annoyed at people who used their mobile phones while driving or for casual chat in public places like restaurants and movies. However, just a few weeks later, many of the participants were exhibiting those same behaviors. Another interesting finding was that participants who had more exposure to mobile phones through friends or colleagues were better at predicting how they would be using phones.

Many other technologies have faced similar changes in expectations over time. Warren and Brandeis' famous definition of privacy as "the right to be let alone" came about in part because new cameras in the late 19th century made it possible to take photographs in just several seconds, invading "the sacred precincts of private and domestic life." Kodak cameras fared no better on the privacy front in their early days (as noted in "The Kodak Camera Starts a Craze" on the PBS.org website):

The appearance of Eastman's cameras was so sudden and so pervasive that the reaction in some quarters was fear. A figure called the "camera fiend" began to appear at beach resorts, prowling the premises until he could catch female bathers unawares. One resort felt the trend so heavily that it posted a notice: "PEOPLE ARE FORBIDDEN TO USE THEIR KODAKS ON THE BEACH." Other locations were no safer. For a time, Kodak cameras were banned from the Washington Monument. The "Hartford Courant" sounded the alarm as well, declaring that "the sedate citizen can't indulge in any hilariousness without the risk of being caught in the act and having his photograph passed around among his Sunday School children."

Similarly, in the book *America Calling*, sociologist Claude Fischer documented the social history of the telephone. In one of my favorite passages, Fischer observed that at first, many people actually objected to having landline phones in their homes, because it "*permitted intrusion…by solicitors, purveyors of inferior music, eavesdropping operators, and even wire-transmitted germs*."

While these examples may seem quaint by modern standards, they represented real concerns that people had at the time. In fact, it is worth pointing out that many of these same problems still exist (inferior music now comes in the form of the Muzak you listen to when put on hold). The main difference is that our expectations of how these technologies will be used have changed over many decades, as we have adapted via changes in our social norms and laws.

So there are two points here relevant for Google Glass. The first is that we all lack experience with how we might use wearable computers, and so it is very likely that most of our expectations will be off the mark. The second is that expectations can change over time, as we learn to adapt to the technology and its affordances, but only if we start to see real value in it.

It is also worth pointing out that expectations can also change quite rapidly and dramatically. Perhaps the best example of this is the introduction of Facebook's News Feed in 2006. Before News Feed was rolled out, you could only see a person's status updates by going to their individual profile pages. What News Feed did was aggregate all of those updates in a single place. When News Feed was first made public, people's initial reactions were predominantly negative, and often viscerally so. Many Facebook groups were formed denouncing News Feed, and Facebook CEO Mark Zuckerberg even went so far as to publicly respond to all of the negative press. Facebook stuck to its guns and continued to push News Feed. In the course of just a few months, a lot of the criticism died out as people saw value in News Feed and became used to it. Now, several years later, I seriously doubt you could find someone who would want to give up News Feed.

Now, this does not mean people's privacy concerns always change in the way you want, or the way you expect. There are also plenty of examples where products were killed or features rolled back due to serious privacy concerns. My main points here are that we all have little experience with wearable computers, expectations of privacy can change, and perceived value is a major factor in driving that change. However, there remains a very big gap in the community's understanding of the best ways of mitigating these kinds of privacy issues up front, and the best ways of managing and designing for those changes. All I can say for sure is to buckle your safety belts, because Google Glass is just one of many of these kinds of big changes in computing we will likely see in the future, and it will be a wild, scary, crazy, and exciting ride.

Jason Hong is an associate professor at Carnegie Mellon University.

Science | DOI:10.1145/2524713.2524718

Gregory Mone

Future-Proof Encryption

Researchers are relying on quantum mechanics to generate unhackable methods of communication.

THIS SUMMER, THE controversial former National Security Agency (NSA) analyst Edward Snowden answered a series of security-related questions in an online forum hosted by the *Guardian* newspaper. One worried reader asked if there was any way to hide email from the inquisitive eyes of the NSA. Snowden replied, in brief: "Encryption works. Properly implemented strong crypto systems are one of the few things you can rely on."

When these systems fail, the cause is typically human error—someone installing malware on their machine, for example—and not the result of a fundamental flaw. Yet researchers say this will not remain true if quantum computers, machines with exponentially more processing power than today's technology, become a reality. "It is reasonably clear that the classical encryption methods we are using today are going to become insecure in the long term," says physicist Vadim Makarov of the Institute for Quantum Computing at the University of Waterloo. "Once the technology to crack classical encryption becomes available in the future, all the secrets become compromised retroactively. This is just not acceptable for many kinds of secrets, like medical, political, military secrets, which have very long-term value."

Norbert Lütkenhaus, associate professor in the physics department at the University of Waterloo and a member of the Institute for Quantum Computing (IQC), is involved in research on the theory of practical quantum key distribution systems.

PHOTOGRAPH COURTESY OF THE INSTITUTE FOR QUANTUM COMPUTING

As a result, scientists have been developing systems that rely on quantum cryptography, a potentially unhackable form of communication. A group at Los Alamos National Laboratory has been operating a small but secure quantum-cryptography-based network for more than two years. The Swiss firm ID Quantique already has used quantum cryptography to secure point-to-point transactions within financial institutions.

There are a number of limitations to the nascent technology, but the potential benefits are tremendous. "It is ultra-secure," says physicist Duncan Earl, chief technology officer of Gridcom, a startup developing a quantum cryptography system for the electrical grid. "It is a security guarantee against future computer improvements. It is a future-proof technology."

A New Trust Model

The critical pieces of both quantum and standard cryptography are the keys used to encrypt and/or unscramble messages. The RSA algorithm for public-key cryptography, one of the more popular systems today, relies in part on a publicly available key that is the product of two large prime numbers. This public key is combined with a message to create a gibberish-like cipher text. Once the message is encrypted and sent, the only way to decipher it is to apply a second, private key.

It is almost impossible to derive the private key from the public one, but depending on the difficulty of that calculation could be a flaw in the long run. "It is only secure if factoring is a hard problem," explains physicist Daniel Gottesman of the Perimeter Institute. "That turns out to be very hard on a classical computer, whereas a quantum computer could run new kinds of algorithms that can efficiently factor large numbers. So if you had big quantum computers, RSA would not be secure."

Quantum cryptography takes a different approach. If two people, dubbed Alice and Bob in the crypto world, want to communicate securely, they first generate and exchange a shared, secret key. This key, which is as long as the message itself, as opposed to the relatively short 128-bit or 256-bit keys used in today's systems, is known as a one-time pad, and is only used once. Alice encrypts her message with the one-time pad, then sends it to Bob, who applies the same key to unscramble the text.

The quantum aspect of the process lies in how they generate and exchange that key. In the most common method, known as prepare and measure, Alice sends photons of light to Bob. A photon can assume a number of possible states—different spins and polarizations—that can be used to represent different bits. So, a photon with a vertical orientation might stand in for the bit 1, while a photon that is horizontally oriented could correspond to 0. Alice prepares each photon, collapsing it into a particular state, then sends it to Bob, who attempts to measure the result. They each translate what they see into key bits and compare their results.

This photon receiver nicknamed "Bob" is part of a quantum key distribution system housed at Waterloo's Perimeter Institute for Theoretical Physics. Bob communicates with photon receiver "Alice," which resides at the Institute for Quantum Computing at the University of Waterloo.

PHOTOGRAPH COURTESY OF THE INSTITUTE FOR QUANTUM COMPUTING

If someone tries to spy on the process and intercept the photons en route, then Alice and Bob will notice too many discrepancies and conclude the line of communication is insecure. But if the measurements match often enough, they are left with a matching string of random bits they can use as a shared, secret key to encrypt and then decipher a message.

The security of the key stems from the fact it relies on photons, not factoring. "The laws of physics say that if I am sending light, any attempt by an eavesdropper to make a measurement on that must create a disturbance," says quantum communications expert Jeffrey Shapiro of the Massachusetts Institute of Technology. "What Alice and Bob rely on to get their security is that law of physics. That is a different trust model than saying we know this is a computationally difficult problem and therefore we can rely on the fact that no one has a computer powerful enough to break this system."

Theory and Practice

In quantum cryptography today, the devices used to transmit and receive photons between two parties can only communicate over relatively short distances, on the scale of tens of kilometers; any farther, and the photon signals fade. Furthermore, the devices that send these pulses do not always behave precisely, sometimes generating two or more photons when the system only asks for one. "We have a gap between theory and practice," says physicist Renato Renner of the Institute for Theoretical Physics in Zurich. "The devices that are used in practice just don't do what they're supposed to do."

Hackers such as the University of Waterloo's Makarov have exploited these flaws. In 2010, Makarov and his colleagues announced they had effectively hacked a quantum cryptography system by blinding it with a bright light. Yet Makarov does not see his work as an indictment of quantum cryptography; he merely found a weakness in a particular implementation. In fact, he informed the manufacturers long before he published the work, so they were able to fix the flaws before any damage could be done.

The Swiss firm ID Quantique, which has worked with Makarov and other hackers on occasion, has been steadily improving its technology since launching its first product in 2004. ID Quantique is exploring systems that would allow quantum cryptography to work over larger distances, but the company's primary focus today is on relatively local, point-to-point communications. "Our current customers are in the financial sector and government," says ID Quantique CEO Gregoire Ribordy. "We are offering long-term confidentiality of data over a link, such as one data center to another or, in a campus network, high-security transactions between buildings."

Gridcom plans to use quantum cryptography to secure machine-to-machine communications within the electrical grid. The company, which is scheduled to roll out its first commercial system in 2015, will rely on entanglement, a phenomenon in which two photons become inextricably linked in such a way that measuring one will produce an immediate change in its twin. Gridcom will use these entangled photons to generate a stream of secure, random bits, and if anyone tampers with one of the pair, the system will immediately recognize the interference. In Gridcom's model, companies will pay a subscription per machine for access to those securely generated bits. "They get these tamper-proof keys which they use in their encryption," Earl says.

A More Secure Future

The recent advances in the field have sparked some speculation about the larger potential of quantum cryptography. When the Los Alamos group led by physicists Richard Hughes and Jane Nordholt announced they had been successfully running a small, hub-and-spoke network secured by quantum cryptography, several popular news sites called their creation a "secret quantum Internet." Hughes quickly deflated the notion, and some experts say quantum cryptography might not be ideally suited for protecting mass communications. Gottesman and other experts note there are other public key cryptography systems that may be able to do the job more efficiently and economically on their own. No one has yet proven otherwise.

Instead, quantum cryptography will more likely be used in specific cases involving small networks or point-to-point communications, when long-term secrecy is essential. Even Makarov, who has exposed glitches in today's systems, believes the future is bright. To him, the successful hacks are not a proof that quantum cryptography itself is flawed. "This is a natural step in the process to make the technology secure," he says. "Once the implementation loopholes are found and closed, then we have a really, really secure technology. Quantum cryptography is going through this process right now." C

Further Reading

Lydersen, L., Wiechers, C., Wittmann, C., Elser, D., Skaar, J., Makarov, V.
"Hacking commercial quantum cryptography systems by tailored bright illumination," *Nature Photonics*, 2010.

Gisin, N., Ribordy, G. Tittel, W. and Zbinden, H.
"Quantum Cryptography" in *Reviews of Modern Physics*, Volume 74, 2002.

Hughes, R.J., Nordholt, J.E., et. al.
"Network-Centric Quantum Communications with Application to Critical Infrastructure Protection." http://arxiv.org/abs/1305.0305

"A Multidisciplinary Introduction to Information Security." Chapman and Hall, 2011. See Chapter 5: Quantum Cryptography.

**"Quantum Cryptography"
An introductory video lecture by physicist Daniel Gottesman. http://www.youtube.com/watch?v=NejFo5hKcOg**

Gregory Mone is a Boston, MA-based writer and the author of the novel *Dangerous Waters*.

Milestones

Computer Science Awards

GOODE AWARD TO PATT

Yale N. Patt, professor of electrical and computer engineering and the Ernest Cockrell Jr. Centennial Chair in Engineering at the University of Texas at Austin, was named 2013 recipient of the Institute for Electrical and Electronic Engineers (IEEE) Computer Society Harry H. Goode Award "for nearly half a century of significant contributions to information processing, including microarchitecture insights, a breakaway textbook, and mentoring future leaders."

Patt is a Fellow of both IEEE and ACM.

IEEE LAUDS HERLIHY WITH MCDOWELL AWARD

Brown University computer science professor Maurice Herlihy was the 2013 recipient of the IEEE Computer Society's W. Wallace McDowell Award.

Herlihy, whose research focuses on practical and theoretical aspects of concurrent and distributed computing, was recognized for his "fundamental contributions to the theory and practice of multiprocessor computation."

MILLER RECEIVES 2013 KNUTH PRIZE

Computer science professor Gary L. Miller of Carnegie Mellon University received the 2013 Knuth Prize at the 45th ACM Symposium on Theory of Computing (STOC 2013) in June.

Awarded annually for outstanding contributions to the foundations of computer science, the Knuth prize is jointly sponsored by the ACM Special Interest Group on Algorithms and Computational Theory (SIGACT) and the IEEE Technical Committee on Mathematical Foundations of Computer Science.

Miller was honored for algorithmic contributions to theoretical computer science, which have had a major impact on cryptography, number theory, parallel computing, mesh generation for scientific computing, and linear problem solving. He also has made significant contributions to the theory of isomorphism testing.

He was named an ACM Fellow in 2002, and was a co-recipient of the ACM Paris Kanellakis Award for the Miller–Rabin primality test in 2003.

RCS LAUDS NEW FELLOWS

The Royal Society of Canada (RSC)'s Academies of The Arts, Humanities, and Science has announced its Class of 2013 Fellows, a group that includes Alfred Aho, the Lawrence Gussman Professor of Computer Science at Columbia University, New York City.

Aho was cited for his fundamental contributions to string searching, databases, formal languages, programming languages, compilation, and pattern matching.

Author of influential books on the analysis of algorithms and the theory of compilation, Aho's RSC Citation notes, "His research is used daily in millions of computers worldwide and has been cited thousands of times."

Aho also is a fellow of the AAAS, ACM, Bell Labs, IEEE, and is co-chair of the Contributed Articles Editorial Board of *Communications of the ACM*.

Technology | DOI:10.1145/2524713.2524719

Neil Savage

More than a Mouse

Gesture and gaze are among the newest additions to a growing family of computer interfaces.

PEOPLE INTERNET SURFING in coffee shops may soon be waving their hands in front of their laptops (if they aren't already), thanks to a new type of user interface making its way into the mainstream. In July, Leap Motion of San Francisco, CA, introduced a gesture-based controller; the device plugs into a USB port and sits in front of the computer, projecting a cone of infrared light that it uses to detect hand and finger positions with a pair of CCD cameras.

Other companies are also experimenting with new interfaces. Apple, for instance, was recently granted a patent on a system for recognizing gestures made above the surface of a touchscreen. A touch sensor in the screen reads the drop-off in capacitance as a finger moves away from the surface, and uses that drop-off to measure proximity. That allows the user to add a pulling motion to the standard swipe and pinch motions of touchscreens, and makes it possible to manipulate a virtual 3D object. Apple envisions the technology being used in computer-aided design systems, according to the patent.

That scene in the 2002 movie *Minority Report*, where Tom Cruise manipulates data on a series of large screens with waves of his hands, is moving from science fiction into reality, as computer scientists develop new types of user interfaces that go beyond the classic mouse and keyboard, from touch and voice to tracking gestures or detecting a user's gaze.

John Underkoffler, chief scientist at Oblong, of Los Angeles, CA, designed the computer interface for that movie scene, and his design is making its way into the real world. In Oblong's Boston office, Carlton Sparrell stands in front of three large display screens and makes a finger pistol with his right hand, index finger pointed straight, thumb cocked upright above it. Where he's pointing, a cursor appears and moves as his hand moves. He hovers over an object and drops his thumb to select it, pulls it over to an adjacent screen, and lifts his thumb again, releasing it.

The Leap Motion controller tracks users' hand movements.

PHOTOGRAPH COURTESY OF LEAP MOTION

When Sparrell, Oblong's vice president of product development, makes a circle with his thumb and forefinger, the objects on the screen form a circle. Swiping his hand to the left or right scrolls through a list of files. Pushing toward the screen shrinks an image while pulling back enlarges it. When he puts both arms up in an "I surrender" gesture, the screen resets to where it started.

Underkoffler says gestural input has the potential to have as significant an impact on the way humans interact with computers as when Apple popularized the graphical user interface with the introduction of the Macintosh in 1984. "There's hasn't been a really big breakthrough in user interface thinking or user interface technology in 30 years," he says. "We're still stuck with the Mac interface."

Oblong's main product is g-speak, its gestural recognition software platform. In the setup Sparrell demonstrated, he wore black gloves with small tabs on the fingers and the back of the hands. Cameras positioned around the room track light refracted from the tabs to give the computer x, y, and z coordinates for the hands and fingers.

However, the sensor technology can vary. Another system set up in the company's office uses the same Primesense camera found in Microsoft's Kinect system to track hand motions without the use of gloves, though Underkoffler says infrared tracking without gloves does not yet provide high-enough resolution to be reliable. Oblong also makes a system for office conference rooms that relies on a wand, with a gyroscope and ultrasonic sensors to track its position and buttons to select items on the screen. That system lets participants in a meeting download an app and control the display from the keyboard on their laptops or the touchscreens on their tablets or smartphones.

For Leap Motion, early reviews of its controller were mixed, with the main complaints being that not many apps accepted gestural input, and that different apps used different gestures to perform similar functions. David Holz, co-founder and chief technology officer of Leap Motion, says app developers and users will eventually figure out a gestural language that makes sense.

The challenge when introducing a new input device, he argues, is getting away from adapting actions enabled by a previous type of input and figuring out the unique abilities of the new device. “We’re leaving it open for experimentation,” Holz says. “Some of this is us building our own intuition.”

For instance, drop-down menus exist because they are a handy way to access files via a mouse. For gesture, however, they might not be as useful as, say, making a sweeping hand motion to slide past a set of options. A lot of the actions that seem easy with a mouse and keyboard are that way just because users have been using them for so long. Holz says gesture provides a three-dimensional interaction that more closely matches how people manipulate objects in the real world than does moving a mouse across one two-dimensional plane to affect action in another two-dimensional plane. Mouse movements are an abstraction that slow down whatever the user is trying to do, he says. “Every time we sort of go one step closer to what is going on inside our minds, it lets us remove one level of abstraction that doesn’t need to be there.”

Underkoffler calls gesture “a higher-bandwidth way of getting information in and out of the head of the human.” Anything that one would want to move around and through—a structural model of a protein, a brain scan, a panoramic photograph of a city—can be handled more smoothly with gesture than with a mouse, he says. “You can see stuff, because you can fly around it, that you never could see before.”

Others agree that gestural input can take advantage of intuitive ways of dealing with the real world. “We have a lot of intuition of how to manipulate groups of objects, and a lot of that intuition is lost when you move to a graphical user interface,” says Andy Wilson, who manages the Natural Interaction Research group at Microsoft Research. For instance, someone might arrange a group of files into a pile, the same way they would pile up folders on their desk; then they would be able to grab that pile and move it somewhere else. Such a task, he says, could happen more smoothly with gesture than with a mouse.

Among the projects Wilson is working on is the Wearable Multitouch Projector. A user wears a portable projector and depth-sensing system on his shoulder that projects an image onto available surfaces, such as walls or the palm of the user’s hand. When the user touches the projected image, the sensors detect finger motion and the computer responds appropriately, creating what is in essence a virtual touchscreen.

Unlike gaze tracking systems that do not require additional peripherals, SensoMotoric Instruments GmbH’s 3D Eye Tracking system has users wear specialized eye tracking glasses.

PHOTOGRAPH COURTESY OF SENSOMOTORIC INSTRUMENTS

There are other ways to interact with computer images a user is looking at; notably, by looking at them. Tobii Technology, of Stockholm, Sweden, uses a camera to track where a user’s pupils are pointing in its Gaze system. “If you think about it, just about everything you do in front of a computer, it all starts with what you are looking at,” says Carl Korobkin, vice president of business development at Tobii.

That allows tasks to be activated by eye tracking alone. Widgets on a desktop, to provide weather reports or stock updates, can be opened with a glance. “It can tell you’re looking at it, you want some information; it provides the information,” Korobkin says.

Another way the eye tracker differs from the mouse is the ability it gives users to open hidden files just by looking off the edge of the display in a particular direction. Pages that extend beyond the screen can automatically scroll as they are read. In cases where someone needs to click on something, simply tapping a touch pad or saying “open” while the user looks at it would do.

Nobody is arguing that any one input technology will be the only one used, or that the mouse will disappear. “The future is going to be a lovely plurality, when you really will pick up the right tool for the task,” says Underkoffler. “I think the keyboard is not going to get replaced for a really long time, because it is really good at text.”

One advantage of having multiple input modalities—and systems designed to accept different types of input—is that people with physical or cognitive disabilities will have more options for interacting with computers, making them more accessible. Gaze tracking, for instance, could help people with limited hand mobility, as well as working better in situations where touch would be difficult—a sterile environment like a cleanroom or operating room, for instance. Voice might be the preferred input for blind people, as well as for drivers who don’t have their hands free.

Sensor technology is already adequate to these inputs, and is likely to improve so gloves will not be needed, says Underkoffler. And the processing demands aren’t overwhelming; g-speak runs on a $5 ARM chip. The challenge will be in developing ways to exploit the potential of new types of computer interaction. “We’re going to discover new techniques, new principles, new work patterns,” Underkoffler says, “that are going to be pretty radically transformative.” ■

Further Reading

Wearable Mulitouch Projector
http://research.microsoft.com/apps/video/default.aspx?id=160684

Aigner, R, Wigdor, D, Benko, H, Haller, M, Lindbauer, D, Ion, A, Zhao, S, Koh, J.
Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI, Microsoft Research, November 2012

Wachs, JP, Kölsch, M, Stern, H, Edan, Y.
Vision-Based Hand-Gesture Applications Communications of the ACM, 54, February 2011

Pouke, M, Karhu, A, Hickey, S, Arhippainen, L.
Gaze tracking and non-touch gesture based interaction method for mobile 3D virtual spaces Proceedings of the 24th Australian Computer-Human Interaction Conference, Melbourne, Australia, November 2012.

Neil Savage is a science and technology writer based in Lowell, MA.

Society | DOI:10.1145/2524713.2524720 Keith Kirkpatrick

Legal Issues with Robots

Who is responsible when property is damaged or someone is injured by an automated system? Existing product liability laws have not yet been tested in the context of robots.

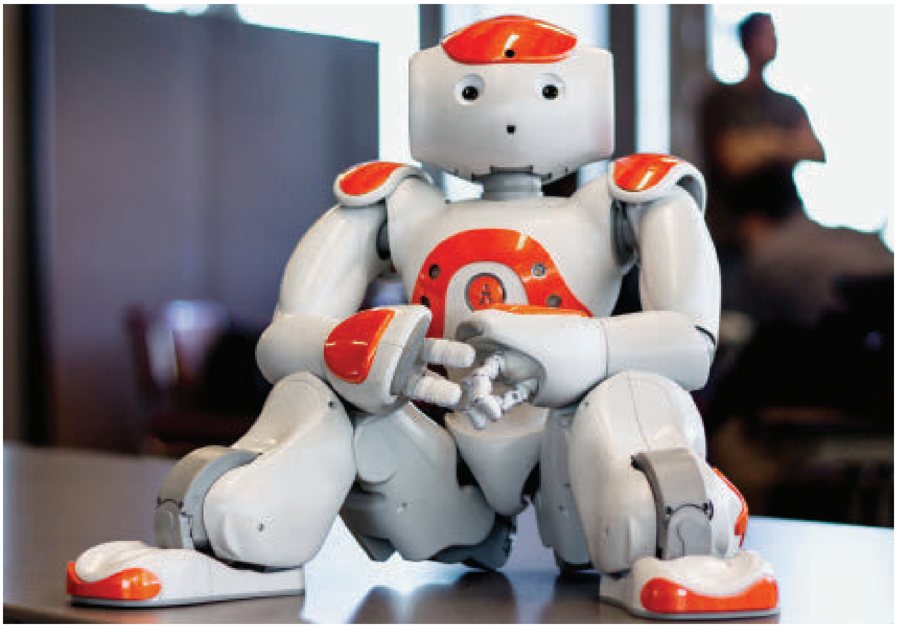
A DOCTOR EXAMINES a stroke patient located in a rural community in South America from a hospital located thousands of miles away. A stock trade is executed instantly based on real-time information. A new pair of shoes is purchased for a rock-bottom price during an Internet flash sale. A car is guided down the highway at a safe distance from others, automatically adjusting the engine speed, steering controls, and brakes based on real-time traffic information.

In each of these cases, there is no question that robotic technology is making life easier, safer, or more convenient for human beings. Despite these benefits, concerns remain about what happens when robotic technology fails, either unintentionally or by design, resulting in economic loss, property damage, injury, or loss of life.

Certainly, for some robotic systems, the issue of responsibility when an adverse action takes place largely follows traditional product liability case law, according to several legal experts, with the manufacturer bearing responsibility for a malfunctioning part, along with the operator (if it can be proven that the operator did not properly maintain the product or component according to the manufacturer's specifications).

"The way the law works, there is a structure for product liability," says Peter Asaro, a professor at New York City's New School and a scholar at the Stanford Law School Center for Internet and Society. "For the most part, it applies to these systems as it always has. There are obviously going to be cases that will be difficult to decide, and there will be new precedents that will be established during that process."

However, Asaro cautions, "A lot of companies are nervous about entering that marketplace because of the uncertainty about how some of these things will play out. There is a lot of effort right now, especially from Google, to get some structures in place in the law that would limit their liability, or shape it in a particular way that they could manage it."



Robots like this NAO 25 (nicknamed Rosie) from Aldebaran Robotics are designed as a development platform for educational researchers. If Rosie has an accident that causes damage or injury, who is responsible?

PHOTOGRAPH BY PATRICK HIGGINS/UNIVERSITY OF SOUTHERN MAINE FREE PRESS

Designing Legal Structures to Manage the Unknown

Robots are being designed with increasingly greater sophistication and decision-making capabilities, instead of simply carrying out simple instructions that have been pre-programmed by their designers or operators. As these robots are able to act and react based on external, environmental stimuli, the ultimate liability for an accident or incident caused by a split-second decision made by the robot is less clear. In these cases, the robot or autonomous machine may be functioning as intended by the designer without any clear errors or failures, and if unintended results still occur, the issues of responsibility and liability must be sorted out.

Guy Fraker, former director of business trends and foresight for insurance giant State Farm and the co-founder and CEO of consultancy Autonomous Stuff, notes that real-world scenarios may crop up that have not been foreseen by robot designers. For example, Fraker notes the complexity of a scenario in which a driverless car makes its way through a parking lot, when a shopping cart and a baby carriage simultaneously roll in front of the car. Whereas human instinct would likely cause the driver to steer closer to the shopping cart (to avoid hurting an infant), it's unclear whether an autonomous system that simply sees two objects in its path would be able to make a similar choice.

"The more we task robotics to act on our behalf, one of the first questions is, 'who is responsible' in the moment of truth," Fraker says, noting that a conscious design decision or shortcoming that results in an accident will likely result in a product liability lawsuit, whereas a human failure to intervene when a system fails to perform as intended is more likely to result in a traditional insurance scenario. "We don't have an answer for that [question] yet. Key private-sector stakeholders, trade associations, academic institutions, and public policymakers are working on frameworks for how to sort out this and other issues, but we just aren't there yet."

Another liability issue that is likely to arise involves the issue of human action—or inaction—in terms of either choosing to override the autonomous robot's actions in the name of safety, or simply letting the robot continue its intended actions. If an accident occurs, whether with a driverless car, a telepresence robot that is navigating through a hospital hallway, or a drone that is headed toward a civilian village with no military value, the questions of liability are likely to be hotly debated.

"What is going to get really complicated is if these cars appear to be doing something unsafe, and the driver overrides it," Asaro says. "Was it the manufacturer's fault, or is it the individual's fault for taking over? Because you are dealing with people's perceptions and accounts of what is going on, it is a very fuzzy area that we really have not seen a lot of legal debate about. The answer to those questions is going to come from court cases."

Regulatory Environment

As robots increasingly are designed to utilize stimuli that are taken from the same space inhabited by humans, the potential for negative consequences is enhanced, as is the potential for lawsuits seeking to assess responsibility and collect damages. This has created a scenario in which manufacturers of robots are focused on perfecting their systems for 100% reliability, which would, in effect, make liability a non-issue. However, this view is misguided, according to some risk assessors.

"The other relevant question that we have to get a handle on is one of performance expectations. For example, with vehicle operation, we know [people] cause 93% of all accidents," Fraker says. "What is an acceptable level of reliability for autonomous cars? Because if the answer is 99.99%, let's stop now, let's quit investing money, because perfection is not a realistic expectation."

Cooper says the only way open source robotics will see support from manufacturers is if they develop a framework to codify acceptable robot behavior.

Despite the push among robot or autonomous system manufacturers to perfect their systems, it is likely the government will step in to make sure robotic technology is safe for use. Thus far, four states (Nevada, Texas, Florida, and California) have signed declarations making driverless cars legal, though there are significant conditions to be met, such as a requirement that a driver is sitting in the driver's seat at all times, ready to take over in the event of an emergency, and that the driver must carry traditional accident insurance.

In the E.U., regulations governing the use of autonomous emergency braking and lane-departure warning systems have been passed, and are seen as signaling a desire to speed the introduction of driverless cars on the continent. Moreover, Fraker notes that in jurisdictions including Europe, China, and Brazil, the risk of litigation against product manufacturers is significantly lower than it is in the United States, given that these governments have taken a far more restrictive view of using the tort system to recover damages in the event of injury or death.

Furthermore, in the medical field, robots currently are and will continue to be heavily regulated in the U.S. by the Food and Drug Administration (FDA).

For example, the RP-VITA, a remote-presence robot developed jointly by InTouch Health and iRobot, is essentially comprised of two key functions: allowing remote telemedicine sessions to occur through a remote audio and video link (as well as basic sensors), and a mobility component, which uses sensing technology to allow the robot to autonomously travel between rooms. Early this year, the RP-VITA received 510(k) clearance by the FDA, which means that organization verified that the device is functionally equivalent to a device already in use. While the 510(k) clearance process does not shield manufacturers from liability claims, the process generally entails manufacturers taking steps to ensure their devices are safe.

There are no specific legal statutes surrounding the use of robots in the workplace, but the idea of robots working as agents on behalf of individuals or companies is raised by Samir Chopra, a professor of philosophy at the Brooklyn College of the City University of New York, and co-author of *A Legal Theory for Autonomous Artificial Agents*. Chopra says that as robots gain autonomy, it is likely that legal frameworks will be developed that will hold the owner or operator of a robot legally responsible for any accidents or problems that occur with the robot, due to the inability to effectively hold a machine legally liable for its actions.

"When that level of autonomy occurs, that's when it starts to make less sense to think about these robots as tools of the people who operate them, and more like agents of the employers," Chopra says, noting that this concept of agency is likely to occur as robots gain a greater level of autonomy. Furthermore, if a case of negligence or act of malice occurs as a result of the robot operating on behalf of its operator, imposing penalties on the robot itself would largely be ineffective. "What you might see down the line is that you would be forced to change [a robot offender's code or design], so they could not do it again," Chopra says.

Still, the technologies being used in driverless cars and other automated systems are far more accurate than humans, and are statistically less likely to cause harm than humans doing the same tasks.

"We don't react well enough," says Scott Nelson, CEO of MILE Auto Insurance, an automotive insurer founded to offer policies based on miles driven, rather than based on variables such as a driver's age or location. "The question is, with the regulatory hurdles and the acceptance hurdles, are we going to be willing to turn our cars over [to robots], where we don't do anything?"

That's why most lawyers interviewed for this article agree that in the near term, liability and insurance structures are going to move very slowly in addressing the new wave of autonomous technologies, simply because there is little legal precedent for these systems. Further, organizations that are currently deploying robots declined to participate in this article, perhaps wanting to avoid boxing themselves into any specific legal corner if a problem were to arise.

Open Source Opens a Host of Questions

While most of the robotic technology in use today is based on closed, proprietary technology, robots are starting to be developed using open source software. As a result, liability might not be assigned to the developers of that software if something goes wrong, since the software itself is designed to be modified. As such, it is harder to quantify what types of functionality or uses could be considered "off–label" or unauthorized, or even ill-advised, based on an open source platform.

"Liability can be addressed currently with closed robotics, because they have restricted functionality," says Diana Cooper, Articling Associate, LaBarge Weinstein LLP, and author of a paper entitled "A Licensing Approach to Regulation of Open Robotics," which was presented at the We Robot Conference at Stanford Law School in April. Cooper notes the very nature of open source robotics is that the functionality is not predefined, and there are no restrictions in place governing what actions the robot can or cannot carry out. This is creating a considerable amount of confusion for robotics developers who are worried about liability, from the components used in robots to fully completed products.

"The problem is, how will the upstream parties that create certain components that are built into these end products be sheltered from liability from any downstream modifications that might be harmful?," Cooper asks. "That is a whole different ball game, because we cannot provide warnings to the market, since we do not know what that end product will be, and we cannot rely on the defense of product misuse since open robots are intended for modification."

Legal scholars such as Cooper and Ryan Calo of the Center for Internet and Society have brought up the idea of creating a licensing system, similar to an end-user license agreement found on productivity software. In essence, the license of the robot or robotic component would stipulate that the robot or robotic component would not be used for certain actions (such as creating a weapon, or otherwise harming people, animals, or property), and would indemnify the developer and/or manufacturer of the robot or component against any claims resulting from a robot or component being used in violation of the terms of the license.

Cooper, for one, says the only way open source robotics will see support from manufacturers is if they develop some sort of framework to codify acceptable robot behavior, and create a method for indemnifying upstream suppliers of hardware and software from potential liabilities.

Says Cooper: "Hardcore open-source advocates don't really want any restrictions on the use, but now that we're inputting the software into hardware components that have actuators and a physical presence that interacts with people, I think that perhaps we should be looking at imposing some tailored restrictions." C

Further Reading

Patrick Lin, Keith Abney, and George A. Bekey (eds.)
Robot Ethics: The Ethical and Social Implications of Robotics, MIT Press, 2012, 386pp., http://ndpr.nd.edu/news/31199-robot-ethics-the-ethical-and-social-implications-of-robotics/

Richards, Neil M. and Smart, William
How Should the Law Think About Robots? (May 10, 2013). Available at SSRN: http://ssrn.com/abstract=2263363 or http://dx.doi.org/10.2139/ssrn.2263363

Legal Challenges in an Age of Robotics, Stanford University, http://www.youtube.com/watch?v=P021944l2LA

Keith Kirkpatrick is principal of 4K Research & Consulting, LLC, based in Lynbrook, NY.

ACM Member News

Wenwu Zhu's research on multimedia cloud computing is both topical and groundbreaking; he is exploring how a cloud can perform distributed multimedia processing and storage, and provide quality of service provisioning for multimedia applications and services.

This research also focuses on how to allocate computing and communications resources between the cloud and clients, which includes mobile phones for multimedia applications, and services for good Quality of Experience.

Zhu was recruited to his post in the Computer Science Department of Tsinghua University under China's "1,000 People Plan" to enlist top academics and researchers; he previously was senior researcher/research manager at Microsoft Research Asia. Zhu is the inventor or co-inventor of more than 40 patents, and has published more than 200 refereed papers on multimedia communications and networking, and wireless communications and networking.

He says he is interested in "how to use social information to reduce the intention gap in the multimedia search and how to use social information to predict emotion for the individual person."

Zhu was part of a group of Chinese computer scientists and academicians awarded the ACM SIGMM award for Best Technical Paper in 2012, for the paper "Propagation-based social-aware replication for social video contents."

"I see a convergence of cloud media, social media, and mobile multimedia," Zhu says. "Cloud can help with mobile phones, which have very limited computing and storage capabilities. The cloud can become a 'base' and a 'powerhouse' to provide content and user information for mobile phone applications and services."

—Laura Didio

DOI:10.1145/2527185

Deirdre K. Mulligan and Kenneth A. Bamberger

Privacy and Security
What Regulators Can Do to Advance Privacy Through Design

Seeking to address the challenges of privacy by design through regulatory process and formal law.

THE PERCEPTION THAT privacy is losing an arms race with technology is a constant source of public anxiety, and regulatory action. Many privacy and data protection laws directly respond to advances in technology—from cameras, to large databases, to the Internet, to cellular, to sensors. The paradigm plays out over and over again: technology erodes privacy, regulations are passed to protect it.

Ongoing revelations about the National Security Agency's massive data collection activities, and success at circumventing encryption in consumer products, at times with the cooperation—perhaps begrudging—of the corporate sector, suggest that efforts to strengthen the privacy features of technical systems are timely and necessary even in democratic countries.

But today regulators are casting a more hopeful eye at technology. They are looking to harness technical design to advance and protect privacy. Signaling the conviction with which they have embraced this new perspective, "privacy by design"—as it is commonly called—is a goal of the current draft European Privacy Regulation[4] and features prominently in U.S. privacy proposals.[a]

Bringing privacy concerns into the design of products and standards is a significant new regulatory approach. It reflects growing recognition of the substantial role that technical systems play in supporting and shaping societal values.

But advancing privacy through design presents several challenges. First, it requires regulatory strategies that encourage organizations to devote technical resources toward advancing a social imperative that is often at loggerheads with corporate interests, and is traditionally viewed as the responsibility of corporate lawyers, not technologists. Embedding privacy in technical design will likely be as expensive as having lawyers draft privacy notices, and, we would argue, more complicated. Infusing privacy into technical design will require firms to adopt new internal strategies and practices. It will also require firms to hire or create more and different kinds of privacy experts, as lawyers seem ill suited on their own to drive design.

Second, while regulators call it "privacy by design" the approaches they have developed and promoted reflect a relatively narrow view of privacy—facilitating control over personal information. This is more aptly described as "fair information practices by design" and reflects the goals of a key promoter: the data protection regulators. Data protection-driven approaches have two important limitations: they are not aimed at system architects, designers, or coders[6,7]; and, while individual control may be the touchstone of data protection, it is not the sole touchstone of privacy. Building the right "privacy" into design is critical, and today regulators are advancing too cramped a definition. A concept as complex and multifaceted as privacy is

a Executive Office of the President, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Economy* (Feb. 2012), http://www.whitehouse.gov/sites/default/files/privacy-final.pdf; U.S. Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (2012).

ILLUSTRATION BY DANIEL ZALKUS

not well served by a narrow data protection oriented approach. Regulators must adopt strategies that encourage designers to engage with multiple, context-dependent concepts of privacy. There are some indications this will happen, but ensuring it does is essential to the success of the privacy by design effort.[5]

Third, the success of this regulatory initiative turns on new privacy professionals. A field traditionally dominated by lawyers, auditors, and human resource professionals must find room for ethicists, social scientists, technical designers, architects, and engineers. Ideally, this new set of privacy professionals will be comfortable working at the intersection of law, ethics, social science, and technical design. Integrated approaches to protecting privacy are necessary if society hopes to reorient technology toward privacy's protection. While there is a growing body of academic research on values in technical design, there must be increased attention to developing curriculum to teach and train design professionals, and fostering the development of a professional community. Both are necessary to bring privacy design strategies into practice.

Over the past few years, we have looked at how corporations understand and manage privacy in different countries.[1–3] Through our research, involving almost 100 interviews of leading privacy officers, regulators, and other privacy professionals in the U.S., Germany, France, Spain, and the U.K.—we have identified a set of corporate approaches that are more aligned with integrating privacy into design, and the regulatory choices that seem to spur corporations to adopt them. Our research provides some insight into regulatory approaches likely to align corporate behavior with the privacy by design agenda.

We found similar and promising approaches to privacy management in leading German and U.S. firms.[b] The privacy work in these firms goes beyond rote compliance with data protection and privacy rules, more frequently and fully includes technical and design professionals, and is woven into the daily work of business units. Each firm had a high-level and strategic privacy lead overseeing privacy work that was integrated into business and functional units, and product lines. Personnel responsible for privacy, and technologies and processes geared to raising and incorporating privacy concerns, were distributed throughout the firms. Privacy leads stressed the importance of embedding expertise within business units and establishing specific staff personally responsible for privacy—typically through indirect reporting mechanisms—as essential to institutionalizing privacy. These distributed and decentralized systems positioned privacy as an input into the design of products and processes.

This embedded and decentralized system of organizing privacy aligns

b The privacy leads interviewed, "chief privacy officers" ("CPOs") or "data privacy officers" ("DPOs"), included those identified as field leaders by domain experts—leading privacy thinkers (both lawyers and nonlawyers) drawn from academia, legal practice (in-house and firms), trade groups, advocacy groups, and consultancies, regulators, and journalists focusing on privacy issues. Questionnaires were used to collect biographical data about the interviewees and organizational information about the firm. Additional research involved the review of internal organizational charts, process documentation, and discussions with managers and engineers responsible for policy implementation in the firms.

with privacy by design. It develops privacy experts knowledgeable about the specific technological and business choices, and situates them within the process of design where they can influence technical choices. This approach moves privacy outside the legal domain and into that of technology design and business processes.

A set of regulatory choices in Germany and the U.S. appear to encourage these sorts of embedded privacy structures and design-oriented approaches. First, in both countries a set of regulatory choices push firms to view privacy as a strategic issue. Regulatory approaches in both countries require firms to interpret and adapt regulatory goals to address firm-specific issues. While German law is more comprehensive and more detailed than the broad "unfair or deceptive practices" mandate that empowers the Federal Trade Commission—the lead U.S. privacy regulator—from the outset, Germany envisioned the private sector playing a crucial role in meeting statutory objectives. This is reflected in the requirement that firms above a certain threshold appoint an independent internal data protection officer. In the U.S., firms employ privacy professionals to interpret and navigate an ambiguous and shifting privacy landscape.

Second, in both countries privacy is connected to other ethical obligations placed on firms. In the U.S., privacy has been positioned as a contextually dependent and changing concept tied to consumer expectations and consumer protection. In Germany, privacy is tied to human and worker rights. Although the broader ethical frames differ, our interviewees credited these connections for corporate privacy responses that are more forward-looking and dynamic, rather than solely compliance focused.

Third, in both countries regulatory choices allow other constituents—privacy advocates, civil society, labor—to participate in defining a firm's privacy obligations. In Germany, our interviewees noted the influence of the independent works councils, which represent workers interests, and the independent Data Protection Officer, who represents societal privacy interests, as important constituents shaping firms' understanding of privacy obligations. In the U.S., the wide range of participatory procedures the FTC provided enabled privacy and consumer advocates to influence how firms and regulators understand privacy goals. Governance strategies that allow third parties to participate in defining the privacy obligations placed on corporations help transform privacy from an issue of legal compliance to one of "social license"—a broader constraint on firms' social standing and reputation, this in turn pushes privacy outside the legal counsel's office.

Advancing privacy through design is an important regulatory initiative.

Finally, both countries have chosen regulatory strategies that expose firm privacy activities to outside scrutiny. In particular, security breach notification reporting requirements, benchmarking activities, and public fines are used. Our interviewees claim the external shocks caused by these transparency-forcing regulations and events focus the attention of corporate executives, freeing up political and financial resources to protect privacy. In Germany, news of breaches travels quickly through the works councils and is used to improve corporate privacy management. In the U.S., reported breaches tied privacy to brand protection, increasing the attention to privacy from management, boards, shareholders, and business partners. Our interviewees report the public salience of privacy—fed by transparency—drives corporations to view privacy protection as integral to maintaining corporate good standing and therefore to invest in privacy's systematic management.

Our research suggests that collectively these regulatory strategies lead firms to view privacy as a more salient business risk. In turn, privacy professionals are viewed as strategic experts necessary to successfully interpret and mediate an uncertain external environment. Due to their control over important external resources—legal legitimacy and brand portrayal—privacy professionals gain greater influence and access to resources, and are able to break out of the role of legal compliance. In contrast, we found the regulatory approaches in France and Spain where privacy agencies interpret and fix the meaning of regulatory mandates with little consultation with regulated parties or other stakeholders result in privacy leads who are less powerful and strategic, and privacy management that is focused on compliance and auditing tasks.

Advancing privacy through design is an important regulatory initiative. Achieving it requires organizations to invest in systems to ensure privacy infuses decisions about processes, practices, and technical design. Our comparative analysis of how corporations respond to existing regulatory regimes reveals a set of often overlooked similarities in the regulatory choices in Germany and the U.S. have led firms to view privacy as a strategic issue and to create structures that bring its consideration more fully and regularly into the design phase of products and services. This is a key insight for policymakers considering regulatory reform. Our work underscores the importance of considering both regulatory process and formal law as policymakers seek to shape the mindset and behavior of private firms to advance privacy through design choices. ■

References

1. Bamberger, K.A. and Mulligan, D.K. Privacy in Europe: Initial data on governance choices and corporate practices; http://bit.ly/16I7RGk
2. Bamberger, K.A. and Mulligan, D.K. Privacy on the books and on the ground. *Stanford Law Review 247*, 260 (2011).
3. Bamberger, K.A. and Mulligan, D.K. New governance, chief privacy officers, and the corporate management of information privacy in the United States: An initial inquiry. *Law and Policy 477* (2011)
4. European Commission. Proposal for a regulation of the European parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation). COM (Jan. 25, 2012).
5. Federal Trade Commission. Protecting consumer privacy in an era of rapid change: A proposed framework for businesses and policymakers. (Dec. 2010); http://www.ftc.gov/os/2010/12/101201privacyreport.pdf.
6. Gürses, S., Troncoso, C. and Diaz, C. Engineering privacy by design. In *Proceedings of the International Conference on Privacy and Data Protection (CPDP)* (2011); http://www.cosic.esat.kuleuven.be/publications/article-1542.pdf.
7. Mulligan, D.K. and King, J. Bridging the gap between privacy and design. *University of Pennsylvania Journal of Constitutional Law 14*, 4 (2012).

Deirdre K. Mulligan (dkm@ischool.berkeley.edu) is an assistant professor at the School of Information at the University of California, Berkeley, and a faculty director at the Berkeley Center for Law and Technology.

Kenneth A. Bamberger (kbamberger@law.berkeley.edu) is a professor of Law at the University of California, Berkeley, and a faculty director at the Berkeley Center for Law and Technology.

DOI:10.1145/2527186

Pamela Samuelson

Legally Speaking
Is Software Patentable?

Assessing the shifting perspectives on patentability standards for software.

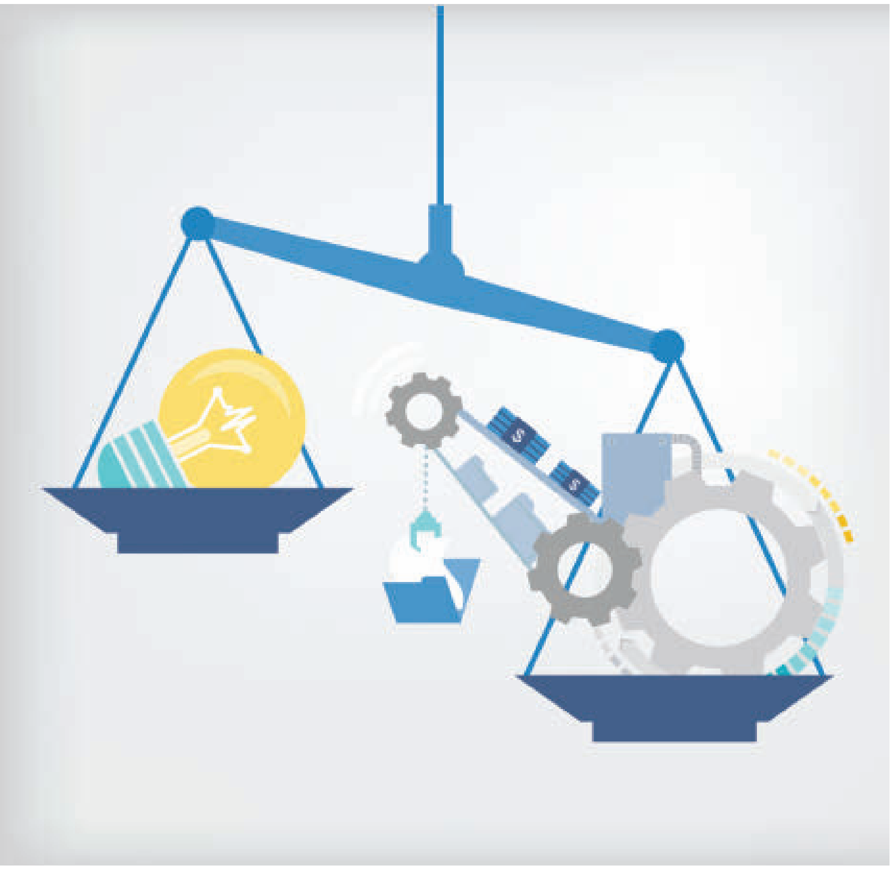


IMAGE FROM SHUTTERSTOCK.COM

THE PATENTABILITY OF computer program innovations has been controversial for nearly 50 years. In the 1960s and 1970s there was considerable doubt these innovations were or should be eligible for patent protection. By the mid-1980s the tide shifted and software innovations were increasingly deemed eligible for patents. Doubts about the patentability of software seemed to dissipate in the mid- to late 1990s. After all, programs are technological in nature and any design that can be implemented in software can equally well be implemented in hardware, which would unquestionably be patentable.

Interestingly, the tide seems to be shifting again. The German legislature recently passed a resolution calling for the cessation of patenting for most software-related innovations. New Zealand has been considering an outright ban on software patents. In the last several years, the U.S. Supreme Court has called into question many—and perhaps even most—software patents. Even the Court of Appeals for the Federal Circuit (CAFC), which has long taken a liberal view of innovations that qualify for patenting, recently struck down some software patents in the *CLS Bank International v. Alice Corp.* case.

This column will discuss the *CLS Bank* case and the deep divide within the CAFC about patentability standards. The CAFC has yet to establish a workable framework for assessing when software innovations are patentable and when they are not.

Alice's Patents

Alice obtained four patents covering a computerized trading platform for conducting financial transactions to manage the risk that one of the two trading partners will fail to meet its obligations. The trading platform relies on a trusted third party to handle settlement risks by ensuring both parties meet their obligations or the transaction fails. The specifications for these patents were substantially the same, although the claim language is different.

Some of Alice's claims are in method form. Typical is the claim for a method of exchanging obligations between parties holding a credit record and a debit record with an exchange institution, the method comprising a set of steps.

Some claims are in system form. A typical claim is for a data processing system to enable the exchange of obligations, the system comprising a data storage unit and a computer to carry out a series of stated steps.

Some claims are in computer media form. A typical claim was for a computer program product comprising a computer-readable storage medium having computer-readable program code embodied in the medium for use in a trading platform to handle settlement risks. (This is known as a *Beauregard* claim after the CAFC decision that allowed code-embodied-in-a-medium claims.)

The Lawsuit

Alice became aware that CLS Bank was conducting business that implicated its patents. Alice informed the bank about the claimed infringement. Rather than taking a license, CLS Bank filed a lawsuit asking a U.S. court to declare it had not infringed these patents and also that Alice's patents were invalid and unenforceable. Alice counterclaimed, asserting its patents were valid and infringed.

While this lawsuit was pending, the U.S. Supreme Court ruled in *Bilski v. Kappos* that a method of hedging the risk of fluctuation in prices of commodities was ineligible for a patent because it was for an abstract idea. CLS Bank likened Alice's patents to Bilski's and argued they too were ineligible for patenting. Alice asserted its patents were distinguishable from those in *Bilski*. A trial judge agreed with CLS Bank. Alice appealed from the ruling that its patents were invalid.

The CAFC Outcome

The CAFC decided to hear Alice's appeal en banc, that is, with all 10 judges participating. (Usually, appeals are heard by three-judge panels. En banc cases are rare, and typically only convened for what the court perceives as important cases.)

Seven of the 10 judges who deliberated in *CLS Bank* agreed that under Supreme Court precedents, Alice's method and computer-media claims were invalid. Two of the seven judges disagreed with the other five, however, over the reasoning that supported these conclusions.

Alice's system claims split the court even more. Five judges would have upheld these claims because they mention specific machine components and because they are too detailed to be abstract ideas. The other five regarded the system claims as functional equivalents of the method claims. Although these five acknowledged the system claims mentioned computers and data storage, they did so in such generic and functional terms that the use of any general-purpose computer would suffice. The claim details were, in their view, insignificant or conventional, which under *Bilski* do not count.

When appellate courts are evenly split, as in *CLS Bank* on the system claims, the lower court decision is affirmed, but no precedent is established without an appellate majority. The result was that Alice's system patents were invalidated as well.

Eight of the 10 judges in *CLS Bank* agreed the method, system, and media claims should rise or fall together. However, five thought all three types of claims should fall, while the other three judges thought all of Alice's claims were patentable.

Abstract Idea or Patentable Process?

Patent law provides that any new, nonobvious and useful "machine, manufacture, composition of matter, or process" is eligible for a grant of patent rights as long as the inventor applies for a patent and complies with the rules for obtaining one. Courts have interpreted this language broadly. Yet they have also said that abstract ideas, laws of nature, and natural phenomena are ineligible for patenting.

A key case interpreting these exceptions is the U.S. Supreme Court's unanimous 1972 ruling in *Gottschalk v. Benson*. It held the Patent and Trademark Office (PTO) had correctly denied Benson's application for a patent on an algorithm for converting binary coded decimals to pure binary form.

Benson made two types of claims: one for the method carried out with computer components and the other for the method more generally. The latter would have covered carrying out the method in one's mind. The Court did not distinguish between these claims; both were unpatentable.

The main explanation the Court gave was that courts had long considered processes to be patentable when they transformed matter from one physical state to another. Benson's process did not do this. This interpretation was consistent with some older cases (not cited in *Benson*) holding that mental processes were unpatentable.

The Court in *Benson* also expressed concern about the abstractness of the algorithm and how wide was the array of possible applications. A patent on Benson's algorithm would preempt all uses of the method, regardless of whether the inventor had foreseen them.

Seven years later the Court in *Parker v. Flook* ruled that a software-implemented process for updating alarm limits in catalytic converter plants was unpatentable. The only distinction between the claimed method and the prior art was Flook's algorithm.

The Court in *Flook* rejected the claims, even though they contained field of use restrictions (which would preclude wholly preempting use of the algorithm) and mentioned some post-solution acts (adjusting controls so the plant would not blow up). The Court viewed these restrictions as insignificant and worried that patent applicants would try to evade the rule announced in *Benson* if these claims were allowed.

Two years later in *Diamond v. Diehr*, the Court decided by a 5–4 vote that a process for curing rubber that utilized a computer program was a patentable process. The CAFC interpreted this decision broadly, as though *Diehr* had, in effect, overruled *Benson* and *Flook*. The CAFC began upholding software patents, which overcame the PTO's resistance to issuing software patents. While no one knows how many software patents have been issued since the mid-1980s, they probably number in the hundreds of thousands.

Bilski Resurrected *Benson*

Under the usual CAFC standard, Bilski's process for hedging the risk of price fluctuations of commodities would have easily passed muster as a process that yielded a useful, concrete, and tangible result. Yet the PTO denied a patent to Bilski, seemingly emboldened by signals from the Supreme Court about its dissatisfaction with the CAFC.

In the last several years, the U.S. Supreme Court has called into question many—and perhaps even most—software patents.

Having detected these same signals, the CAFC decided to hear Bilski's appeal en banc. This court was deeply split over the patentability of Bilski's method. Yet, a majority supported the machine-or-transformation (MoT) test. That is, a method would be deemed patentable if it was to be carried out by a machine or if it transformed something from one state to another. Because Bilski did not mention any machine in his claims, nor did he claim to transform anything, the CAFC held his method unpatentable.

The Supreme Court ultimately agreed with the CAFC that Bilski had not claimed a patentable process, but it did not find the MoT test compelling. Instead, the Court ruled the claims were too abstract to be patentable. As in *Benson*, the Court expressed concern about the preemptive effect of issuing a patent to this abstract idea. The Court in *Bilski* gave *Benson* and *Flook* as examples of unpatentable ideas, but did not define what it meant by abstract.

Implications for CLS Bank

Judge Lourie and four fellow judges picked up on the preemptive effect analysis in *Bilski* and used that lens to examine Alice's patent claims. These judges concluded that Alice's claims would have the kind of preemptive effect the Court worried about in *Bilski*. It did not matter which type of claim language was used (that is, method, media, or system) if the substance of the claims was the same.

Four other CAFC judges relied heavily on the presence or absence of references to computers and memory—in effect, applying the MoT test—to judge Alice's patents. All four regarded the systems claims as unquestionably patentable because these claims mentioned machine components. The specificity in claim language also persuaded these judges to think Alice's claims were too concrete to be abstract ideas. The tenth CAFC judge disparaged the effort to distinguish between patentable methods or systems and unpatentable abstract ideas as unproductive. She would have upheld all of Alice's patents.

Conclusion

Alice has not yet sought Supreme Court review of the CAFC ruling, but it is expected to do so. Whether the Court will decide to take the case remains to be seen. There are other software patent cases in the appellate pipeline. *CLS Bank* may not be the best vehicle to test the patentability of software as its claims are atypical.

But *CLS Bank* was the case that the CAFC selected for en banc review. As in *Bilski*, the Court may decide it has to take this case because the fractures within the CAFC do not augur well for consistency in three-judge panels on patentability questions in future cases.

One of the CAFC judges who would have upheld Alice's system patents warned that if the analysis by Judge Lourie in the *CLS Bank* prevails, this would mean "the death of hundreds of thousands of patents, including all business method, financial system, and software patents as well as many computer-implemented and telecommunications patents." This may be hyperbole—but then perhaps not. I will leave to my readers' judgment whether this would be a good or bad development.

Pamela Samuelson (pam@law.berkeley.edu) is the Richard M. Sherman Distinguished Professor of Law and Information at the University of California, Berkeley.

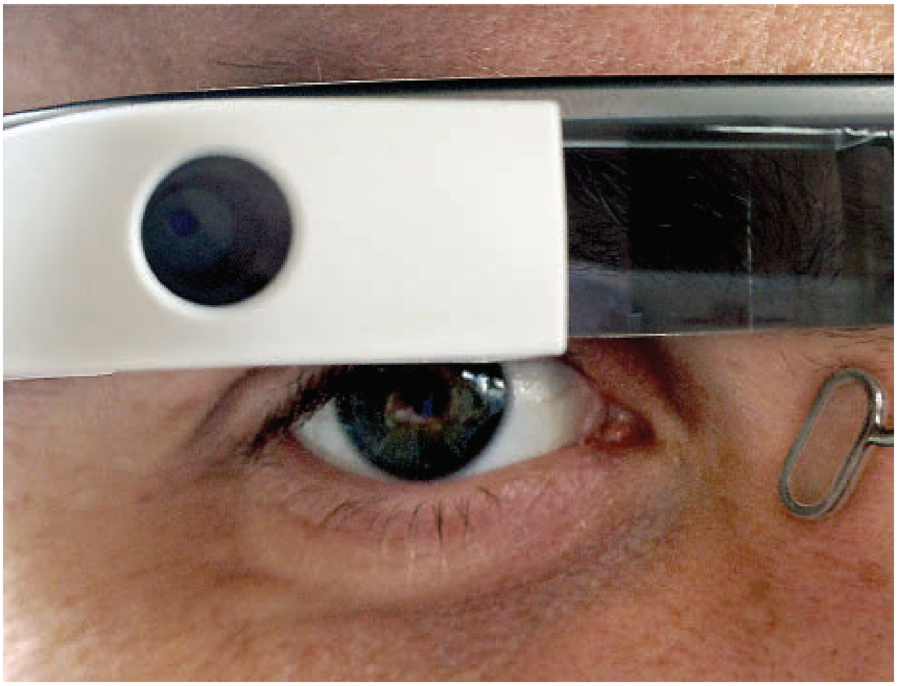
viewpoints

DOI:10.1145/2527187

Katina Michael and MG Michael

Computing Ethics
No Limits to Watching?

Considering the ethical questions raised by technologies that are moving from knowing what we are doing (and where) to knowing who we are.



PHOTOGRAPH BY MICHAEL PRAETORIUS

LITTLE BY LITTLE, the introduction of new body-worn technologies is transforming the way people interact with their environment and one another, and perhaps even with themselves. Social and environmental psychology studies of human-technology interaction pose as many questions as answers. We are learning as we go: "learning by doing" through interaction and "learning by being."[9] Steve Mann calls this practice existential learning: wearers become photoborgs,[3] a type of cyborg (cybernetic organism) whose primary intent is image capture from the domains of the natural and artificial.[5] This approach elides the distinction between the technology and the human; they coalesce into one.

With each release greater numbers of on-board sensors can collect data about physiological characteristics, record real-time location coordinates, and use embedded cameras to "life-log" events 24x7. Such data, knowingly or unknowingly collected and bandwidth permitting, may be wirelessly sent to a private or public cloud and stored, often for public view and under a creative commons license.[2] Embedded sensors on wearers can actively gather information about the world and capture details of a personal nature—ours and those of others too. These details can be minor, like embarrassing habits of less than admirable personal hygiene, or major, such as records of sexual peccadilloes or events relevant to court proceedings.

A third party might own the data gathered by these devices or the device itself. The Google Glass Terms of Service state: "...you may not resell, loan, transfer, or give your device to any other person. If you resell, loan, transfer, or give your device to any other person without Google's authorization, Google reserves the right to deactivate the device, and neither you nor the unauthorized person using the device will be entitled to any refund, product support, or product warranty."[8] Personal information stored on the Internet for ease of access from anywhere at any time raises the possibility of unauthorized access. Most wearable sleep monitors indicate when you are awake, in light sleep, in deep sleep (REM), and calculate the level of efficiency reached between your rest and wake times.[7] Monitors can tell adults how often they wake up during the night, the duration of sleep, time spent in bed, and times of awakening.[6] Sleeping patterns convey personal details about individuals, such as insomnia or compulsive obsessive disorder, sexual activity, workaholism, likely performance in stressful jobs, and other things.

Wearables can also look outward, reconstructing the world with location coordinates,[11] current speed traveled and direction, rich high-resolution photographs, and audio capture. Wearers gather data about themselves but also heterogeneous data about fixed and mobile entities, including infrastructure, living things (such as people and animals) and non-living things (such as vehicles). This is not simply derivable information, such as the "point of interest nearest you is 'x' given your position on the Earth's surface," but can be interpreted as, "Johnny is traveling at 'x' miles per hour and is a little sluggish today on his bike ride compared to yesterday, perhaps because of his late night and his consumption of one glass too many of wine while at the nearby bar."

These devices can tell us about exceptions to everyday patterns of people in or out of our social networks. Consider the potential for government surveillance beyond the Call Detail Records that caused such controversy for the National Security Agency in 2013. Relentless data capture is uncompromising. Wearers concerned only about whether the device is working as intended might not consider the consequences of unauthorized data collection. They might feel they have purchased the device and are using it to their best ability. Will they consider feelings of fraternity with strangers who do not have access to the same technology? Or will they feel they have every right to put on their wearable device and adapt their body for convenience or personal needs such as maintaining personal security, reducing liability, and increasing life expectancy? Might wearers figure that any problems are the *other* persons' problems as long as the wearers believe they are not breaking the law? Whether the device is doing what it is supposed to do (for example, work properly), might occlude more meaningful questions of societal consequences from using such devices.

Bystanders are likely to be as oblivious to data collection from wearable devices as they are from data collection of private investigators using covert devices. Yet many people vehemently oppose being a subject of someone else's recording.[1] The disappearing difference between covert and overt devices makes it possible for surveillance to become so ubiquitous that it is rendered "invisible." Anything right in front of us and ever present is in our "blind spot," hardly noticeable because we are enveloped by it like the fog. Novelty wears off over time, industrial/human factor design can help make things invisible to us, and we undergo conditioning. When surveillance cameras are everywhere, including on our heads and in our lapels, it is no longer surveillance. It is simply the human activity of "watching."

Relentless data capture is uncompromising.

CCTV cameras are arguably invasive, but we do not protest their use even though people are aware of their presence. What happens when we open the floodgates to constant watching by tiny lifelogging devices? We open ourselves to not just Big Brother, but countless Little Brothers.[15] Corporate or governmental compliance and transparency hide the fact that audiovisual collection of information will come at a cost. Multiple depictions of the same event can be stronger than a single view, and corruption can flourish even in a transparent environment. It can even be corruption on a grander scale. Crowdsourced sousveillance (watching from below)[12] might take place for authenticity or verification, but police with the authority to subpoena data for use in a court of law as direct evidence can use the data to support their "point of view" (POV), irrespective of the fact that "point of eye" (POE) does not always capture the whole truth.[a,13]

The more data we generate about ourselves, our families, our peers, and even strangers, the greater the potential risk of harm to ourselves and each other. If we lose the ability to control images or statistics about personal or public behaviors how do we make the distinction between becoming a photoborg and becoming the subject matter of a photoborg?

a Hans Holbein's famous painting *The Ambassadors* (1533) with its patent reference to anamorphosis speaks volumes of the critical distinction between PoE and PoV. Take a look (http://www.nationalgallery.org.uk/paintings/hans-holbein-the-younger-the-ambassadors), if you are not already familiar with it. Can you see the skull? The secret lies in in the perspective.

Calendar of Events

November 17–20
Web Intelligence and Intelligent Agent Technology Conference
Atlanta, GA,
Sponsored: SIGART,
Contact: Yi Pan,
Email: yipan@gsu.edu

November 17–21
SC Conference on High Performance Computing, Networking, Storage and Analysis
Denver, CO,
Sponsored: SIGARCH, SIGHPC,
Contact: William D. Gropp,
Email: wgropp@illinois.edu,
Phone: 217.244.6720

November 18–21
The International Conference on Computer-Aided Design
San Jose, CA,
Sponsored: SIGDA,
Contact: Joerg Henkel,
Email: henkel@kit.edu

November 21–22
Twelfth ACM Workshop on Hot Topics in Networks
College Park, MD,
Sponsored: SIGCOMM,
Contact: Dave Levin,
Email: dml@cs.umd.edu

November 26–28
The 6th International Conference on Security of Information and Networks
Aksaray, Turkey,
Contact: Atilla Elçi,
Email: atilla.elci@gmail.com

December 2–5
International Conference on Utility and Cloud Computing
Dresden, Germany,
Sponsored: SIGARCH,
Contact: Rajhurnar Buyya,
Email: raj@cs.mu.oz.au

December 7–10
Information and Communication Technologies and Development
Cape Town, South Africa,
Sponsored: SIGCAS,
Contact: Gary Marsden,
Email: gaz@cs.uct.ac.za

December 7–11
The 46th Annual IEEE/ACM International Symposium on Microarchitecture
Davis, CA,
Sponsored: SIGMICRO,
Contact: Matthew Farrens,
Email: farrens@cs.ucdavis.edu

There is a stark asymmetry between those who use wearables and those that do not. There is much confusion over whether sousveillance[12] is ethical or unethical. The possible perils from lifelogging devices that capture the world around them are only now getting attention.[4] To what extent is it ethical to create the records of the lives of others without prior consent or cognizance? Maker or hacker communities—"prosumers" and "producers"—create personalized devices for their own consumption and become trailblazers for what is possible. But they do not speak for everyone. What is initially made to serve individual needs often is commercialized for mass consumption.

Data from others can generate a great deal of money. Consider the story of Henrietta Lacks, a poor black tobacco farmer whom scientists named "HeLa."[14] According to Rebecca Skloot, HeLa's cells were "taken without her knowledge in 1951" and became a vital medical tool "for developing the polio vaccine, cloning, gene mapping, in vitro fertilization, and more."[14] Until this year, when the family came to a mutually satisfactory arrangement with the NIH, HeLa cells were "bought and sold by the billions," without compensation or acknowledgment. Who profits from wearable devices? The company that owns the device or the data? The wearer? The person in the field of view? Historical evidence suggests it will likely be everyone else but the user wearing the device or the citizen in the field of view. Marcus Wigan and Roger Clarke suggest a private data commons as a potential way forward in this big data enterprise.[16]

Widespread diffusion and data manipulation can require more than an ordinary consumer decision about use and acceptance. Trust and adoption are key to societal conversations that will shape guidelines and regulations about what is and is not acceptable with respect to wearable computing. At what stage of the game are the "rules" to be determined and by whom?

New technologies can bring wonderful benefits, but also disputed, unintended, and sometimes hidden consequences. Technologies should aid and sustain humankind, but we cannot limit technologies to just positive applications. We should not claim benefits without admitting to the risks and costs. Wearable devices create moral and ethical challenges, especially if they are widely used. We must look beyond findings from previous studies of emerging technologies because new technologies often help create new socio-technological contexts. We cannot afford unreflective adoption. "Play" can have real repercussions. The novelty, fun, and "wow" factors wear off and we are left with the fallout. We must be vigilant in our new playhouse, and not negate the importance of moral or ethical standards alongside market values.

Wearable devices create moral and ethical challenges, especially if they are widely used.

Philosophers have contemplated the question of technology and its impact on society. Martin Heidegger, Ivan Illich, Jacques Ellul, and those of the Frankfurt School, have argued that the worst outcome from technology gone wrong is dehumanization of the individual and the loss of dignity, resulting in a "standardized subject of brute self-preservation."[10] A fundamental insight of such literature is that technology has not only to do with building: it is also a *social process*. Any social process resulting in unreflective adoption of technological marvels is profoundly deficient. More is not always better, and the latest is not always the greatest.

Charlie Chaplin's culturally significant film *Modern Times* (1936) shows the iconic Little Tramp caught up in the cogs of a giant machine. The unintended consequences of modern and efficient industrialization are clear. Chaplin's classic builds on Fritz Lang's futuristic film *Metropolis* (1926), which depicts a mechanized underground city in a dystopian society. Both films left indelible marks as prescient summaries of what was to follow. When technology becomes a final cause for its own sake, teleology and technology become confused. The old saw that "The person with the most toys wins," reflects this. What about the rest of us?

References

1. Abbas, R., Michael, K., Michael, M.G. and Aloudat, A. Emerging forms of covert surveillance using GPS-enabled devices. *Journal of Cases on Information Technology 13*, 2 (2011), 19–33.
2. Creative Commons: Attribution 2.0 Generic, n.d. http://creativecommons.org/licenses/by/2.0/.
3. Electrical and Computer Engineering, ECE1766 Final Course Project. *Wearcam.org* (1998) http://www.wearcam.org/students.htm.
4. ENISA. To log or not to log?: Risks and benefits of emerging life-logging applications. *European Network and Information Security Agency.* (2011) http://www.enisa.europa.eu/activities/risk-management/emerging-and-future-risk/deliverables/life-logging-risk-assessment/to-log-or-not-to-log-risks-and-benefits-of-emerging-life-logging-applications.
5. Gray, C. H. Cyborgs, Aufmerkamkeit and Aesthetik [transl. Cyborgs, Attention, & Aesthetics]. *Kunstforum* (Dec.–Jan. 1998); http://www.chrishablesgray.org/CyborgCitizen/kunst.html.
6. Henry, A. Best sleep tracking gadget or app? (2013); http://lifehacker.com/5992653/best-sleep-tracking-gadget-or-app.
7. Henry, A. Sleep time alarm clock for android watches your sleep cycles, wakes you gently (2012); http://lifehacker.com/5942519/sleep-time-alarm-clock-for-android-watches-your-sleep-cycles-wakes-you-gently.
8. Kravets, D. and Baldwin, R. Google is forbidding users from reselling, loaning glass eyewear. *Wired: Gadget Lab* (Apr. 17, 2013); http://www.wired.com/gadgetlab/2013/04/google-glass-resales/.
9. Mann, S. Learn by being: Thirty years of cyborg existemology. *The International Handbook of Virtual Learning Environments* (2006), 1571–1592.
10. Marcuse, H. Social implications of technology. *Readings in the Philosophy of Technology 5*, 71 D.M. Kaplan, Ed., 2009.
11. Michael, K. and Clarke, R. Location and tracking of mobile devices: Überveillance stalks the streets. *Computer Law and Security Review 29*, 2 (Feb. 2013), 216–228.
12. Michael, K. and Michael, M.G. Sousveillance and the social implications of point of view technologies in the law enforcement sector. In *Proceedings of the 6th Workshop on the Social Implications of National Security*. Sydney, NSW, Australia, 2012; http://works.bepress.com/kmichael/249.
13. Michael, K. and Miller, K.W. Big data: New opportunities and new challenges. *IEEE Computer 46*, 6 (2013), 22–24.
14. Skloot, R. *The Immortal Life of Henrietta Lacks*, Crown, New York, 2011; http://rebeccaskloot.com/the-immortal-life/.
15. Weil, J. Forget big brother. Little brother is the real threat. (Sept. 22, 2010); http://www.thenextgreatgeneration.com/2010/11/forget-big-brother-little-brother-is-the-real-threat/.
16. Wigan, M.R. and Clarke, R. Big data's big unintended consequences. *Computer 46*, 6 (June 2013), 46–53.

Katina Michael (katina@uow.edu.au) is an associate professor in the Faculty of Engineering and Information Sciences at the University of Wollongong, NSW, Australia.

MG Michael (mgm@uow.edu.au) is an honorary associate professor in the School of Information Systems and Technology at the University of Wollongong, NSW, Australia.

The authors thank Rachelle Hollander and John King for their observations and insightful comments that helped make this column more robust.

DOI:10.1145/2527188

Marshall Van Alstyne

Economic and Business Dimensions
Why Not Immortality?

A question of biology, technology, and society.

GOOGLE IS SEARCHING for ways to extend life.[5] Ray Kurzweil thinks immortality will be possible by 2045. Exponential advances in either biology or technology could make this feasible. Perhaps we will reprogram malfunctioning cells the way we rewrite old code, achieving biological "escape velocity" once we can repair cells faster than they age.[2] Or, perhaps advanced information technologies will make us robust by capturing all the information representing a person, including genotype and phenotype. Kurzweil thinks this will happen for at least some people alive today.

To live forever would be amazing. Think of the wisdom we could accumulate and the human contact we could enjoy. Despite slowing physically, neuroscientist Oliver Sacks wrote of the joys of old age on the occasion of turning 80 years old. With innumerable friends and the passage of time, he feels "not a shrinking but an enlargement of mental life and perspective."[6] If we could be freed of time's physical ravages, life could be even better. We can imagine our junior selves endowing our senior selves to pursue great works, exploration, and (not or) life's pleasures as we each define them. So what is the catch?

Death is very likely the single best invention of life. It is life's change agent. It clears out the old to make way for the new. —*Steve Jobs*

PHOTOGRAPH BY JAYMIE KOROLUK

Scientists once believed *E. coli* to be immortal. Using cellular mitosis, these hardy cells regenerate and split, regenerate and split. If they can divide indefinitely, then they might live eternally. Absent some calamity, *E. coli*'s example might show us how to elegantly and recursively program infinite genes. But this is not how nature works. Old cellular machinery accumulates in one half of a cell while new cellular machinery accumulates in the other. Cell division causes the oldest mechanics to become isolated in the same cell during cell division, and cell death isolates and clears these pockets of damage. Long-lived parts of the system, like genes, even manage short-lived parts of the system, like cells, for systemic health. Mortality clears accumulated cruft from cell lines, but it does not end there. Mortality removes accumulated problems of all kinds, leading to population vitality and renewal.

I posed the question of why we have not evolved immortality to a biologist friend of mine, Carl Bergstrom. He answered, "Biological systems fail because at some point it's cheaper to make new ones." This is nature's form of cost-benefit analysis. Optimal design balances the marginal benefit of investing in the old versus investing in the new.[1] In a poignant remark eight years before his death, Steve Jobs told Stanford graduates "Death is very likely the single best invention of life. It is life's change agent. It clears out the old to make way for the new."[a]

If we apply the question of immortality to technology itself, the cost-benefit trade-off is even clearer. Why buy a new toaster? It is cheaper than fixing the old one that wore out. Why buy a new car? The cost of a new one has become a better value than the continuing costs of repair. Why buy a new computer? The old one has accumulated so much cruft that it runs too slowly. Either parts have stopped working or it cannot run the latest software. The system is breaking down and, what is worse, it is maladapted to changes in the software environment where we need it to perform.

Here too, in technology as in biology, long-lived portions of the system tend to control short-lived portions for the health of the population. Computer networks survive nodes that come and go and they control behaviors of nodes that interfere with one another. Computers themselves survive apps that come and go, and computer operating systems control apps that interfere with one another. Longevity controls brevity. And, in order to work well, upgrading theses systems requires careful planning.

Legacy systems pose a fascinating challenge especially when obsolescence is unplanned. Consider the Y2K problem, the "millennium bug" where programs using two-digit dates could not handle the switch to a four-digit cycle. Alan Greenspan, then chair of the U.S. Federal Reserve, spoke to Congress in advance of Y2K and said, "I'm one of the culprits who created this problem. I used to write programs back in the 1960s and… It never entered our minds that those programs would have lasted more than a few years."[7] All legacy systems face problems of remaining current. Designers try to make systems long-lived, but it is impossible to anticipate *all* future needs. Could a system capable of handling a near infinite

a Stanford commencement speech. Palo Alto, CA. June 12, 2005.

Fast and easily renewed systems can be preferable to cumbersome systems that lock in static technology.

number of cases execute efficiently, being built using methods that would not become obsolete? Fast and easily renewed systems, even with occasional hiccups, can be preferable to cumbersome systems that lock in static technology.

Oliver Wendell Holmes Sr.'s poem "The Deacon's Masterpiece: The Wonderful One-Horse Shay," captures this insight. The deacon-turned-inventor uses only the finest materials and designs his shay so it has no weakest break point. It lasts sure enough. No part breaks down ahead of another. Yet the end is inevitable. Everything happens in an instant.

Colts grew to horses and beards turned to gray.
Deacon and deaconess dropped away.
Children and grandchildren—where were they?
But there stood the stout old one horse shay.
…
You see of course if you're not a dunce,
how it all went to pieces all at once.
All at once and nothing first,
just as bubbles do when they burst.

Cost-benefit analysis of immortality shows that, at some point, renewal creates more social and system value regardless of whether the context is biology or technology.

The argument need not stop there. This same principle holds true for society if we realize that *all* systems require renewal. Our limited life spans make it difficult to imagine the mortality of the institutions around us, or the need to renew those institutions. We see institutions as *continuing*, creating context for our choices. Yet we know historically that rules governing interactions among people sometimes stop working. Thomas Jefferson wrote, "We might as well require a man to wear still the coat that fitted him when a boy, as civilized society to remain ever under the regimen of their barbarous ancestors."[b] "The earth belongs to the living...Every constitution then, and every law, naturally expires at the end of [a generation]."[c] Jefferson believed legal regimes should renew every 20 years.

Laws are good. Like healthy genes or beautiful code, good laws make our systems run more smoothly. Yet they too accumulate scar tissue and system patches from redressing wrongs and righting injustices. As part of any working system, rules are subject to renewal. When not renewed, laws accumulate cruft leading to breakdown and to maladaptation. When does institutional renewal make sense? It seems ever more salient as use of the filibuster reaches levels never seen before in the U.S. Senate, and as the U.S. House passes more than 42 symbolic bills to repeal the same legislation when no such bill could have survived a presidential veto.[4] Sequestration, inability to pass budgets, downgraded credit ratings—all seem part of a broader social sclerosis. A Gallup poll of priorities has signaled that government corruption ranked number two among concerns of the general public.[3] One sitting senator remarked: "Something has gone terribly wrong when the biggest threat to our American economy is the American Congress."[d] Irrespective of party politics, perhaps the system is maladapted. Perhaps our social institutions need renewal.

Jefferson might have been hasty with his suggestion of renewing every 20 years. In fact there are few simple metrics to judge when renewal becomes necessary. In biology, species have very different life spans and, in technology, systems have very different life cycles. One economic test for renewal offers hope: When do sovereigns or elites spend more resources on themselves or their activities than on their people? When do system controllers consume wealth rather than create it? Nature removes systems whose cells have run amok. Networks isolate nodes that consume too many resources. And, in each, the control mechanisms seek to improve their own detection and recovery methods in order that controls themselves also evolve. Renewal then means improving the players *and the rules of play*, since new players under old rules just reproduce old plays. An argument that improving governance is infeasible or that control does not need to evolve is naïve: nature replacing individual cells has its cost-benefit analysis, so does consumer replacement of cars and computers. Historically, citizens replace unruly rulers and maladjusted rules.

Why not immortality? It is expensive to be immortal. The risk of immortality is the risk of non-renewal and, until now, mortality has simply been the best cost-benefit form of remedy. Biological, technological, and social systems will always need to renew. A key point of leverage will remain the controllers of each system. Fast-changing fields will amplify this need. Indeed, fast-changing biology and faster changing technology will interact with slow-changing society implying that our institutions will need to adapt faster than they have in the past.

We must also ask how governance changes as individuals within a system become longer lived. If long-lived portions of a system control short-lived portions, the grant of immortality is a grant of greater control. Survival of a control mechanism into future generations implies not just adapting but also creating value for those under the influence of control. As we invent the technology of immortality, we must invent also a society of renewal. We should not seek one without the other. Resources must benefit all contributors to a system, everyone, and not just those in control.

Renewal can be painful, yet it can be joyful. Renewal is not quite immortality but, as any parent can tell you, birth too can be amazing. C

b Letter to Samuel Kercheval, July 12, 1816.

c Letter to James Madison, September 6, 1789, in Jefferson Writings, 963.

d Sen. Joe Manchin III in "A showdown long foreseen," J. Steinhauer, *The New York Times* (Dec. 30, 2012); http://www.nytimes.com/2012/12/31/us/politics/fiscal-crisis-impasse-long-in-the-making.html

References

1. Bergstrom, C. Do we expect the body to be a "one hoss shay?" (May 16, 2010); http://evmedreview.com/?p=240.
2. Global Forum 2045. Towards a new strategy for human evolution. (June 15–16, 2013, New York, NY); http://www.gf2045.com/.
3. Lessig, L. Why Washington is corrupt. CNN Opinion (Apr. 8, 2013).
4. Mann, T.E. and Ornstein, N.J. *It's Even Worse than It Looks: How the American Constitutional System Collided with the New Politics of Extremism.* Basic Books, 2012.
5. McCracken, H. and Grossman, L. Google vs. death. *Time* (Sept. 30, 2013); http://content.time.com/time/magazine/article/0,9171,2152422,00.html.
6. Sacks, O. The joy of old age (no kidding). *The New York Times* (July 7, 2013), SR12; http://www.nytimes.com/2013/07/07/opinion/sunday/the-joy-of-old-age-no-kidding.html?_r=0
7. Testimony of Alan Greenspan, Chairman of the Federal Reserve, before the Senate Banking Committee, Feb. 25, 1998. ISBN 978-0-16-057997-4.

Marshall Van Alstyne (mva@bu.edu) is an associate professor in the department of management information systems at Boston University and a research scientist at the MIT Center for Digital Business; Twitter: InfoEcon.

DOI:10.1145/2527189

Cameron Wilson

Education
Making Computer Science Count

Combining efforts and expertise, ACM and Code.org are partnering to address a rapidly changing education landscape.

It is amazing how the discussion around computer science education at the K–12 level (primary and secondary schools) continues to evolve. In the U.S. and around the world, we have shifted from simply discussing exposing more students to computer science to actually doing it. Education and political leaders, major corporations, and even average citizens realize the ubiquity of computing in our everyday lives creates a big unanswered question: Who is going to write all the software we need for the digital age? And as computer science evolves to support all fields, how are we going to expose students to a fundamental new literacy? Those are big questions. And one clear answer is that education systems—particularly at the K–12 level—must change and adapt to the new reality of the digital age.

In the U.S., the education system has the unique characteristics of being both monolithic (big and clunky) and highly diverse (highly distributed decision making). Change does not come quickly, and when it does it is often only in pockets of reform. In general, the existing U.S. education system considers computer science education as something for the few, not a fundamental need for all. So how are we going to change this system to ensure rigorous and engaging computer science education is something all students have access to? A new effort led by Code.org in partnership with ACM, the National Science Foundation, and a number of other corporate and non-profit stakeholders is now working to provide the answer for the U.S.

ACM has been deeply involved in leading and partnering in many of the community efforts to see rigorous and engaging K–12 computer science education introduced into U.S. schools. Almost 10 years ago, ACM created the Computer Science Teachers Association to bring a professional body focused solely on understanding the issues and needs of computer science educators from around the globe. It also founded an Education Policy Committee and published a landmark study, *Running on Empty*, examining the state of K–12 computer science education. When the National Science Foundation started to build a new AP Computer Science Principles course, ACM partnered with NSF and evangelized the new course in the community. And ACM led the foundation of Computing in the Core—the community's multi-stakeholder (corporate and non-profit) coalition to advocate for K–12 computer science education. It is gratifying to see these efforts bear fruit into a new undertaking led by Code.org with substantial community and industry backing to build off these efforts.

Education systems must change and adapt to the new reality of the digital age.

Code.org is a relatively new organization that burst onto the education scene in February 2013 driving massive consumer awareness through videos of tech luminaries speaking about how students should get exposure to coding in grade school. Since then, its founder and CEO, Hadi Partovi, and I have been working in partnership with ACM, the National Science Foundation, the Computer Science Teachers Association (CSTA), the National Center for Women and Information Technology, Microsoft, and Google on developing a new strategy that leverages this consumer awareness while building upon years of efforts by the computing community.

Code.org's vision is that every student in every school should have the opportunity to learn computer science. Our plan emphasizes scaling cost-effectively and sustainably, using online curriculum to ensure consistent quality of education, and growing participation in CS among students from underrepresented groups (women, people of color, and low-income). To accomplish this, we have developed a three-pronged approach to these areas:

▸ **Educate:** the heavy lifting of bringing CS into schools: this involves vetting third-party curriculum, developing our own curriculum, and preparing teachers to teach these approaches.

▸ **Advocate:** changing policies in states so computer science courses can satisfy graduation requirements and can access the same funding sources as other "core" courses and addressing other policy-related issues.

▸ **Celebrate:** inspiring students, parents, and schools through high-level marketing (videos, events, and celebrities), providing very high impact for relatively low cost.

As of this writing, Code.org's education pillar is still under development. Our other areas, however, are well under way and are ready for the computing community to engage in immediately.

In the U.S., the "advocate" pillar has largely been handled by the multistakeholder, non-partisan advocacy coalition Computing in the Core, which, with the support of ACM, is merging with Code.org. As part of that merger, Computing in the Core and Code.org's agenda is shifting more toward state policy, where most education decision making is vested. Our headline goal is to "make computer science count," which means allowing existing courses to satisfy a core math or science.

There are three key reasons why this is important. First, Advanced Placement Computer Science courses are about 50% larger in states that "count" AP CS compared to ones that do not. Second, participation by students of color is 37% higher in states where computer science "counts." In fact, CSTA released a recent study that shows a key driver for whether underrepresented groups participate in computing courses (females and students of color) is whether the CS course they are taking counts toward a graduation credit or not. Third, this policy can leverage other policies to increase access.

The most notable example of this is a recent change the National Collegiate Athletic Association (NCAA) made to eligibility requirements for incoming student athletes. If a CS course contains significant programming elements and counts toward a mathematics and/or science credit in a school district, it will *automatically* meet NCAA student-athlete entrance requirements. This change updates the NCAA's previous policy where CS courses did not count toward these requirements.

We need to shift students from being simply consumers of the technology to creators of it.

And the "make CS count" campaign can extend to four-year institutions. If you are faculty, you can reach out to your admissions office to see about allowing computer science courses count toward admission requirements for incoming students. As increasing numbers of school districts start treating rigorous computer science as a math or science course, the lift of shifting admission requirements should be easier.

Our work at the state level can also be supplemented at the local level by asking your local school board to make computer science count toward a math or science credit. In many areas of the country, such decisions are local. Computing in the Core (http://computinginthecore.org) has tools under its "get involved" page for making this happen, and we are developing even more online petitioning tools that will help parents and professionals that want to bring this issue to their local community.

When courses are taught in this area or are part of the curriculum, they are too often focused on teaching students simply how to use technology (basic technology literacy) instead of how to create technologies. To answer the big questions noted earlier in this column, we need to shift students from being simply consumers of the technology to creators of it. Imparting students with these computational thinking skills starts with ensuring access to engaging and rigorous K–12 computer science education. This is the goal of numerous efforts that have gained new momentum in 2013.

Under its "celebrate" pillar, Code.org along with Computing in the Core's partners and many others are planning to make Computer Science Education Week, Dec. 8–14, (see http://csedweek.org) much bigger this year. Code.org is organizing a massive "Hour of Code" campaign during CSEdWeek. Our goal is to recruit 10 million students or adults to take an Hour of Code. This will be a self-guided online tutorial anybody can do, with just a Web browser or smartphone. No experience is needed as the content will target students from third grade and up. There will also be offline and third-party approaches hosted on the CSEdWeek website.

You can help by recruiting teachers, schools, and students to participate in the Hour of Code. Or you can promote the Hour of Code yourself through social media or on campus at your organization or institution. If you are an employer you can have employees participate in the Hour of Code, and if you are an employee you can organize others to participate with you as a social experience.

This new vision comes with a clear opportunity. All of our efforts have generated broad awareness with education decision makers that education systems must change to allow for more computer science instruction. The question is what will we, as a community, do to help shape a new landscape where computer science is something all students have access to and are expected to learn. 2013 should be seen as the start of something big—where K–12 computer science education went from something for the few, to access for all. My hope is that this is a broad engagement across numerous actors throughout our community. And you can be a part of this massive effort by engaging locally with your school boards and four-year systems with the simple message of ensuring computer science courses count toward core credits or entrance requirements.

Cameron Wilson (cameron@code.org) is on assignment from ACM as Code.org's chief operating officer and vice president of government relations.

DOI:10.1145/2527190

Tobias Langlotz, Jens Grubert, and Raphael Grasset

Viewpoint
Augmented Reality Browsers: Essential Products or Only Gadgets?

How lessons learned from the evolution of the Web and Web browsers can influence the development of AR browsers.

Augmenting the real environment with additional information is an idea prospected for over two decades within the augmented reality (AR) research community. When deployed in outdoor environments, virtual information overlays enable a wide range of applications, from tourist guides and pedestrian navigation to urban gaming. The number of people aware of augmented reality is increasing, in part due to growing media coverage.

Commercially, mobile AR is experienced primarily through augmented reality browsers that augment the physical environment with digital information associated with geographical locations or real objects by using smartphones with camera, GPS, and compass sensors. While still relatively small in the mobile applications landscape, AR technology has nevertheless become a noticeable player. AR browsers have achieved more than 20 million downloads from mobile app stores, and some are even preinstalled on smartphones.

The first example of a mobile AR browser running in an outdoor environment was the Touring Machine, developed in 1997 by Feiner et al.[1] Like the Touring Machine, the first generation of commercial AR browsers that came out in 2008 provided graphical augmentation based on geographic location and viewpoint information delivered from integrated sensors such as GPS, compass, and gyroscopes. The second generation of AR browsers are now able to link virtual content to physical objects of our everyday life (for example, magazines, posters, advertisements) using computer vision-based recognition and tracking techniques.

Will people use an AR browser for navigating to a restaurant, or a mobile map application?

The next generation of AR browsers may use a head-mounted display, like a future, more immersive version of Google Glass, which offers new possibilities for personal information systems that move beyond handheld devices.

Only Gadgets?

Despite this evolution it is still unclear how AR browser technology has been received or adopted by end users and whether this technology will move beyond being a gadget to become an important everyday product. While AR browsers should enable users to easily relate digital information to a real-world context they still must prove they can actually fulfill this promise, especially as competing solutions such as location-based services using digital maps are well established and rather intuitive. Will people use an AR browser for navigating to a restaurant, or a mobile map application?

In the few studies that have investigated the usage of current AR browsers[2,3] the overall findings have been rather negative. Users perceived the technology more as a gadget than a must-have application. While these studies highlighted the critical lack of interesting content, they also identified other major issues. Failing to address these shortcomings may prevent AR technology from achieving mass-market adoption.

In this Viewpoint, we will discuss some of these issues that, in our opinion, should be addressed by companies and researchers in the field. Our observations are based on examining how lessons learned from the evolution of the Web and Web browsers can be reflected in the development of AR browsers.

Content as the Key to Success?

For any type of browser the most important criteria are the amount and variety of information that can be accessed. Some of the most popular AR browsers (for example, Layar,[a] Junaio,[b] and Wikitude[c]) only support content created for that specific browser or converted from popular online databases (for example, Wikipedia). The resulting browser-specific content is then generally grouped into hundreds of information "layers." Similar to the early days of the Web that started with hundreds of HTML pages, the amount of AR-accessible content is still low compared to the information provided through current Web services (for example, Twitter, Flickr, YouTube). The available content is even sparser when you consider it geographically distributed over the entire planet.

One of the reasons for this is the content accessible through AR browsers is still widely separated from existing content accessible through standard Web browsers. The business plan of most companies in the field has them acting as gatekeeper for the available information, and this status is manifested by using proprietary formats that differ between the various AR browsers—unlike desktop browsers. From a content provider's point of view this requires them to author the content differently for each of the various AR browsers.

Another negative aspect of current AR browsers is the basic nature of displayed information: primarily text with a limited amount of pictures and even less 3D content. This is in sharp contrast to the current richness of the non-AR Web experience in terms of media form, dynamics, design, and quality.

In our view, lessons should be learned from the rise of Web 2.0. Firstly, standardized formats, architectures, and protocols should be developed to describe the content structure. Standards organizations such as Khronos, X3D, OGC, and W3C[d] have already made an effort toward conceptualizing and defining AR, but this work is still in a preliminary phase. Some research groups have tried to establish formats such as ARML (Augmented Reality Markup Language) or KARML (an AR extension to KML), yet there is no agreed-upon and widely implemented standard for describing content in AR. The KHARMA architecture implemented in the Argon browser[4] is an early, but not yet commonly adopted, model for an AR web infrastructure.

Secondly, the availability of accepted standards enables the combination of content from various sources, forming a more valuable information product: the AR mashup. This would narrow the gap between the information in the Web and the content produced for AR browsers.

Finally, Web 2.0 has revealed the advantages of user-generated content. Currently, end-user authoring of content is only supported through a desktop interface, but there will be significant benefit for AR browsers if content generation can happen spontaneously and in-situ.

More than a Browser

In addition to improvements in the AR browser software, there are changes that could be made in low-level software and hardware that will significantly improve the user experience. To pre-

a See http://www.layar.com
b See http://www.junaio.com
c See http://www.wikitude.com

d See http://www.w3.org/community/ar/

PHOTOGRAPH BY CDL HANDHELD AUGMENTED REALITY, GRAZ UNIVERSITY OF TECHNOLOGY

An AR browser application links geographical location of the real world (point of interest) with digital content (text labels): the information is spatially registered in the environment when seen through the camera view of a smartphone.

An AR browser view of a New York City street intersection with subway information and distance overlays displayed.

cisely augment the environment with digital information AR needs reliable tracking technology. Depending on the application, a wide range of different tracking methods has been developed. AR browsers currently support sensor-based tracking and, more recently, vision-based tracking. However, these tracking techniques are generally used side-by-side but not tightly combined. This results in digital information that is registered in different reference-spaces (for example, geo-referenced or object-referenced).

Easier access to software components providing fused sensor results is needed to make information access more practical for the developers. This will also allow better integration of more precise six degrees of freedom tracking technology that will improve the AR browser experience. This is an important area of research being explored by a number of AR research labs.

This tracking flexibility can be delivered either by AR browser companies or by hardware manufacturers who are showing increasingly interest in tracking technology.

Another aspect that could benefit from hardware manufacturer engagement is power management for AR. So far, AR browsers are rarely used for longer than several minutes. Due to the use of multiple energy-intensive hardware components (3D graphics, video, sensors), an AR application will drain a smartphone battery very quickly. To support an increasing use of AR technology, more efforts are needed from phone manufacturers to develop high-capacity batteries and energy-efficient sensors and algorithms.

Future Outlook

What if enough AR content would become available everywhere? Think back to when the number of Web pages dramatically increased due to the growing interest in the Web by companies and consumers in the late 1990s. What if standards are established and future tracking technologies allow for accurate augmentation in arbitrary environments? Would an AR browser then be the ultimate digital interface to the real world? Doubts remain that this would be the case. Preliminary research indicates users like that contextually relevant information is available at a certain locations, but could not agree on a single optimal technique for presenting it.[3] When displaying mainly 2D content that is registered to points, users might prefer 2D map interfaces over AR browsers.

Companies have been mainly focusing on solving technical issues and hoping designers and end users produce the content and use-cases for their application. However, AR browsers still lack support for a broad range of applications or tasks within the browsers. While the Web has evolved and is now supporting a wide range of tasks and applications through dedicated Web interfaces (for example, communication, business, social networking), AR browsers so far only allow a single task: passive information browsing.

Conclusion

Current AR browsers are mainly used by people who want to try out the technology. For these current users, AR browsers may be more like a gadget. However, as AR researchers, we still see a huge potential and interest in the technology, which is shown by the increasing number of downloaded AR browsers.

In this Viewpoint, we discussed some of the issues we see in the current implementations of augmented reality browsers. Some of these limitations—such as the lack of rich content or obvious use cases—can be addressed by the AR browser companies and a growing community of users. This will hopefully permit this technology similar widespread adoption as the Web 2.0. Other problems such as energy efficiency or improved tracking will affect the usage of the future generation of AR browsers, but are more difficult to solve and require the involvement of hardware manufacturers.

Finally, further research is needed on how to seamlessly connect any real object in our environment with digital content to bring us closer to the ultimate digital interface to the real world. C

References

1. Feiner, S., MacIntyre, B., Höllerer, T., and Webster, A. A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Personal Technologies 1*, 4 (1997), 208–217.
2. Grubert, J., Langlotz, T., and Grasset, R. Augmented reality browser survey. Institute for Computer Graphics and Vision, University of Technology Graz, Technical Report Number 1101, 2011.
3. Olsson, T. and Salo, M. Narratives of satisfying and unsatisfying experiences of current mobile augmented reality applications. In *Proceedings of CHI 2012* (2012), 2779–2788.
4. MacIntyre, B., Hill, A., Rouzati, H., Gandy, M., and Davidson, B. The Argon AR Web browser and standards-based AR application environment. In *Proceedings of ISMAR 2011*, (2011), 65–74.

Tobias Langlotz (tobias.langlotz@icg.tugraz.at) is a senior researcher at Graz University of Technology, Austria.

Jens Grubert (jens.grubert@icg.tugraz.at) is a Ph.D. candidate at Graz University of Technology, Austria.

Raphael Grasset (raphael.grasset@icg.tugraz.at) is a senior researcher at Graz University of Technology, Austria.

PHOTOGRAPH BY ANDRIJ BORYS

HEIDELBERG LAUREATE FORUM

Association for Computing Machinery

Inviting Young Scientists

Meet Some of the Greatest Minds of Mathematics and Computer Science

Young researchers in the fields of mathematics and/or computer science are invited to participate in an extraordinary opportunity to meet some of the preeminent scientists in the field. ACM has joined forces with the Heidelberg Laureate Forum (HLF) to bring students together with the very pioneering researchers who may have sparked their passion for science and math. These role models include recipients of the Abel Prize, the ACM A.M. Turing Award, and the Fields Medal.

The next Heidelberg Laureate Forum will take place September 21–26, 2014 in Heidelberg, Germany.

The week-long event will focus on scientific inspiration and exchange through a series of presentations, workshops, panel discussions, and social events involving both the laureates and the young scientists.

Who can participate?

The HLF invites new and recent Ph.D's, Ph.D. candidates, other graduate students involved in research and undergraduate students with solid experience in and a commitment to computing research to apply.

How to apply:

Young researchers can apply online:

https://application.heidelberg-laureate-forum.org/

The materials required for a complete application are listed on the site.

What is the schedule?

The deadline for applications is **February 15, 2014**.

We reserve the right to close the application website early should we receive more applications and nominations than our reviewers can handle.

Successful applicants will be notified by **April 15, 2014** and will receive full support to attend the Forum.

DOI:10.1145/2524713.2524721

Article development led by acmqueue
queue.acm.org

Merging the art and science of software development.

BY JAMES ROCHE

Adopting DevOps Practices in Quality Assurance

SOFTWARE LIFE CYCLE MANAGEMENT was, for a very long time, a controlled exercise. The duration of product design, development, and support was predictable enough that companies and their employees scheduled their finances, vacations, surgeries, and mergers around product releases. When developers were busy, quality assurance (QA) had it easy. As the coding portion of a release cycle came to a close, QA took over while support ramped up. Then when the product released, the development staff exhaled, rested, and started the loop again while the support staff transitioned to busily supporting the new product.

From a development team's perspective, the product release represented a closed-loop process that was repeatable and formulaic. Outside of a bug that might need expert attention from developers, most of the staff was repurposed to focus on the next version as soon as the product was made available to customers.

In this world, the QA team's responsibilities consisted largely of mimicking customers' usage patterns and minimizing the discovery of unpleasant surprises. In the waterfall model, a QA team's badge of honor was its ability to anticipate, find, and deliver reproducible scenarios that could result in problems for users.

Building out random hardware configurations, simulating puzzling network landscapes, and wiring up seemingly orthogonal factors in underlying platform architecture are all facets of a great QA engineer's knowledge. The very best ones can look at a product's design and predict where the problems might occur. With stand-alone desktop software, shipping bugs is very expensive since the fixes involve service releases, so QA was given more time and staff to verify everything was working as expected.

I remember when this era of predictable variables and finite schedules ended. I remember it very specifically, in fact. I was working for a company trying to launch a prototype of an eBay-toppling person-to-person marketplace product. We grinded for months testing the service from the user interface down. The night of the big release, the biggest conference room in the building was packed, champagne was served in cardboard coffee cups, and the whole extended team anticipated the release, eventually chanting the countdown, "5, 4, 3, 2, 1!" Hugs and toasts were everywhere as the CEO used the new site to buy silly items we joked about in the product's catalog. Then I got a tap on the shoulder. It was the engineering lead, and his words were simple: "We think you should be downstairs with us."

PHOTOGRAPH BY DAVE ROTH

Upstairs, everyone was smiling with anticipation at our product's debut. But, in a much smaller room downstairs, with door-desks rimming the walls, no one was smiling. The website "worked," but it was running too hot considering the tiny amount of traffic.

It was about 7:00 P.M. My job was primarily to sustain load on the service via a hastily prepared shell script to generate traffic. Meanwhile, half a dozen engineers hung over a single screen with logs scrolling and graphs bouncing up and down. After we tweaked the load balancers, the troublesome graph lines started pointing downward, signaling relief from the issue we were facing. The whole room breathed a cautious sigh of relief.

The party upstairs ended hours before I left the office that night. My dog and I walked home sober and tired at 11:30 P.M., satisfied but anxious to hear the fate of the product after the announcement the following day.

We made the front page of the *New York Times* the next morning, huge recognition for us, but my primary memory of that time was a newfound appreciation for the afterlife of a software product.

What products live solely on desktops anymore? How many product cycles have even 24 hours between the last bug fix and exposure to customers? More importantly, how does this change the people and the process involved in releasing, maintaining, and supporting applications?

Over the Wall

I recently came across a funny online meme called "Disaster Girl" that, admittedly, I might have been the last person on earth to encounter. In the background of the photograph, a home smolders while neighbors and firefighters stand a safe distance away. In the foreground, there is a young girl with a smarmy smile, together with loud text reading, "Worked fine in dev. Ops problem now."[5]

Part of the reason I suspect this is not a newly created joke is that the world really does not work that way anymore for most software creators. Where waterfall development once dictated the process of handing off from developers to QA to operations, the responsibilities now have hazy borders, and most engineers' skill sets span multiple disciplines.

The conflict between contributors is now best laid to rest. Cantankerous QA engineers have almost no place in a scrum team. Products grow from prototypes. The design phase has been replaced by a less cyclic, democratic evolution from development to demo.

There is no forum for contentious stand-downs within daily stand-ups. Gone too is the mystique of the operations engineer. Where there once was the Hollywood image of the nocturnal know-it-alls divining the status of a product in a room of glowing screens with indecipherable logs, the culture of DevOps has removed much of the magic and caffeine dependency from those roles.

One refreshing outcome is a steep upswing in expertise and professionalism within the industry. At the same time, engineering prowess allows teams to automate much of the tracking that once required 24/7 staffing of human system monitors, and this has improved the work/life balance of engineers across the globe. For each group—including developers, quality assurance, and operations—significant changes have happened across their roles to highlight aspects of the modern software architecture template.

So, What Exactly Has Changed?

Think back to the mid-1990s when most of what was installed onto machines came via floppy disks or CD-ROM media. The bulk of software seldom interacted with any resources outside of its own host's storage, and software that did interact with external resources did so via a limited array of enterprise products.

The development-to-support chain of command was as defined earlier. Developers worked from a spec, and the "Minimum Requirements" document was a rigid set of constraints. The testing effort consisted mainly of populating a matrix of hardware/feature pairings, driven by a daily installation, and a mix of manual and automated tests. Automation was primarily driven to gather performance metrics for benchmarking and stress tolerance. Operations were largely the responsibility of customer-support technicians who had day-to-day contact with those installing the product. Suffice it to say, there was no need for a dedicated operations effort because the landscape was limited mostly to each user's local environment.

Between this era and the present, the numerous flip-flops between thin vs. thick clients have yielded a broad spectrum of client capacity for applications. On the thick side, gaming and data-management software require very powerful clients with strong support for the myriad network options in the modern landscape. The thin-client end of the spectrum is populated with browser-based apps. Bridging the thick/thin outliers are a massive number of applications where the local processing environment is paired with data and asynchronous processing over a network.

These new offerings run far slimmer than anything discussed in the early days of the thick/thin trade-off, and now far outnumber those on any other platform. With mobile platforms growing in processing power, that thick/thin balance has a new spectrum of its own. Mobile is an interesting category, since apps are constrained by both screen size and processing capacity, and users' expectations are much lower than for a similar app run on a desktop environment.

When this migration away from isolated clients became the norm, the roles within software development teams moved in tandem. Unfortunately for many, this wasn't an immediately obvious requirement. Software cemeteries are overflowing with products whose management did not react quickly enough to the newly networked environments. Those that did adapt were forced to move roles out of the traditional dev/QA/support framework and build teams that were more collaborative in their approach to building fault-tolerant systems with a functionally infinite life span.

The waterfall is now a river; development and operations flow together. When a product launches, there are seldom plans to upgrade or deprecate functionality, and support is ongoing. The industry has generally done a good job adapting to the expectations of this new landscape.

Most notably, engineering efforts now serve multiple masters. Where the focus used to be purely on precision, now longevity and scalability hold equal footing in product requirements discussions. Your application might still move mountains, but now it needs to move mountains for everyone on earth forever.

Where QA had traditionally held the line on risk definition and maintainability, the development and operations people (shortened to DevOps) now have an equal, if not dominant, role. Bargaining for features now involves feasibility and sustainability beyond the limits of PC chipsets and media capacity. DevOps own the profile for system capacity and expandability within platform resources internal and external to the organization.

DevOps and Development, Concepts, and Misunderstandings

No standard definition exists for DevOps. Blog posts on the topic abound, but the best one can derive is that the sands are still shifting on what the term actually means. One side of the argument identifies a specific job description for DevOps, well summarized in a blog post stating that developers work mostly on code, operations work mostly with systems, and DevOps is a mix of those two skill sets.[6] The other side of the table is not specifically opposed to that notion but argues that DevOps is not really a job (in that you do not hire a "DevOp," per se), but rather that the spirit of DevOps speaks to an emerging need in the modern software development and support landscape.[3]

This issue has divided the community for some time. Engineers who proudly describe themselves as DevOps are clearly not on the same page as those who think there is no such thing, but the existence of thousands of open job listings specifically seeking "DevOps engineers" proves that many people at prominent companies believe the term describes a specific role. Others in the industry believe the term describes new criteria for development, testing, release, support, and metrics gathering. In this article I am taking no sides but use *DevOps* to describe the new criteria rather than a specific job.

In my experience, I have observed the DevOps notion emerge from many of the same industry forces that have brought the job description of QA engineer closer to that of developer. The more one knows about the code and platform for an app, the better one will be at building and fixing that app. DevOps, whether in a situation that has operations engineers picking up development tasks or one where developers work in an operations realm, is an effort to move the two disciplines closer.

This merging of software development and support came about as the need arose for a better toolset to detect and measure problems in networked systems. Hundreds, if not thousands, of different companies created homegrown solutions to common tasks, and then these became tedious to build, scale, and maintain. Not long ago, a number of products became available that gave development organizations a way to address common needs such as problem diagnosis, deployment management, task automation, and process standardization. In this way, development organizations, whether large or small, could consolidate the resources dedicated to building such mechanisms.

The biggest benefit derived from this new infrastructure is the ability to quantify aspects of the development and support effort. Without uncertainty, the development process can be described with figures that perfectly communicate the data to heads-down developers, as well as to nontechnical project and product owners. This is a huge value to teams, and one thing is clear from the aforementioned debate within the community: there is definitely a pre-DevOps and post-DevOps era of software development.

Across the entire art form, the influence of DevOps puts efficiency and process into perspective. There are specific areas, though, where the DevOps wave has served to completely and permanently improve product development and ownership, thanks specifically to a sharper focus on metrics and process standardization.

There are specific areas where the DevOps wave has served to completely and permanently improve product development and ownership, thanks specifically to a sharper focus on metrics and process standardization.

Process Standardization

The industry is once again consolidating its platform options, and system management is becoming increasingly standardized as industry outliers build shims for popular products such as Chef, Splunk, and Jenkins, among others. The handful of time-tested products that embody the hammers and screwdrivers of the DevOps toolkit are so ubiquitous that administrator-level experience with them is part of many job descriptions. While this might not seem new or odd in a small company where the infrastructure uses these products as the cornerstones for their systems, large companies with multiple teams developing and deploying separately are finding this standardization to be a valuable alternative to forcing homegrown solutions, allowing devil-may-care independence, or anything in between.

For many companies, this standardization sanitizes the release and support processes. Many organizations grew their processes around immobile objects. Desktop software could really only release around CD/DVD manufacturers and publishers, but now they have online app marketplace standards setting the guidelines. Similarly, online products with critical uptime requirements have led to costly maintenance contracts with data-center service providers. With the ability to now bring this work in-house comes a better understanding of the process and less of a dance for those responsible for publishing a new product.

True validation for any software organization's development, deployment, and support process is the concept of continuous deployment—which is, in effect, the harmonizing of all three tasks to the extent that customers are not notified, or even concerned, about software upgrades. Continuous deployment requires a development organization to achieve sufficient quality so no stone is left unturned in the short path between a source-control submission and a release. It means deployment must be undetectable to users, some of whom will have the software version changed out from under them mid-session. It also means the response to any problems encountered in the course of such a surreptitious release is both proactive and immediate.

A definition of DevOps in a 2010 blog post by Dmitriy Samovskiy[8] names one aspect of the job description "QA in Production." This undoubtedly raises the hackles of many in the industry, but thinking about how this helps define a new standard for everyone in the software industry makes it a valuable bullet point in an otherwise controversial post. Again, to the degree that the most valuable contribution from the DevOps culture surfaces in the software support role, embedding the associated culture throughout the whole product life cycle pushes valuable improvements all the way back to the design phase of an application.

Moving Software Development Metrics to a DevOps World

The aspect of the modern software landscape that most benefits creators is the availability of both thin and thick client solutions at design time. Processing power on most clients, now including most mobile devices, is on par with server-side processing once the costs of data transit are considered. DevOps makes it possible to realize the promise. When the complexity of a feature on the client brings it close to exclusion, acknowledging this trade-off is a necessity. Furthermore, accurate depiction of the give-and-take is the difference between shipping something that succeeds versus shipping something that becomes a punch line for the rest of the industry.

This brings up an interesting aspect of DevOps, in that we have now finally moved beyond the point where QA engineers' insights were consistently coupled with questionable data. DevOps at this point own the server-side numbers in the sense that load, capacity, and scalability have become easily derived figures.

When dealing primarily with client-side code, the variability of client permutations and a lack of accurate behavior measures make QA input far less objective and verifiable. For example, many QA discussions start with, "We have customers complaining about this feature running slow on their phones." Easy follow-up questions include, "What else is running on those phones?", "Are they running on Wi-Fi, 3G, or 4G?", "Is this a problem with a particular phone?", and then go on from there. A well-informed DevOps engineer can say, "Our product now uses 40% of our capacity with our 10,000 active subscribers, so if we can't double the number of instances between now and the next release, we can't take on more features or more customers." No one is going to second-guess such a data-driven DevOps ruling, whereas QA engineers were traditionally unable to support their arguments without doing a lot more research. In a way, DevOps have an easier job now in terms of defining and addressing problems with a client/service app landscape. At the same time, it is a daunting responsibility for QA to provide that same level of validity when describing issues in the application ecosystem.

There is already an emphasis on delivering much of the on-the-ground experience via client application metrics. Crash reporting is something many applications suggest users allow at install time. Browser-based applications have the ability to track just about every aspect of each interaction a user has with an app, with mobile devices delivering a similar crumb trail. The thinner the client, the more data is lost on an application crash. Without explicit permission, most browsers and all mobile operating systems prevent access to the data necessary to evaluate the cause of an application crash.

When dealing primarily with client-side code, the variability of client permutations and a lack of accurate behavior measures make QA input far less objective and verifiable.

Ironically, the source of QA's client data comes at the cost of throughput for the app. There are certainly many ways to collect data from clients running an application, but mitigating the amount of data and frequency of delivery negatively affects the performance of the client and the availability of the service. Again, browser-based apps and those running on mobile platforms experience the worst effects. While apps installed to an operating system can run sentinel processes that emit data at regular intervals whether or not the targeted application is active, browser-based applications lose all identity as a client unless they are active. So, yes, it's a frustrating circumstance for QA engineers. Innovation is constantly improving the situation for client-side data, but the trade-off between the quantity of data and the impact to the application is never going to go away.

Perhaps the best representation of the state of the art is the number of products vying for customers in the mobile platform space. Within the iOS landscape, Apple provides its native crash log reporting to all apps sold in the iTunes App Store.[4] As just described, users need to opt in at install time to provide this data to the application owner, so there is not as much data provided as with third-party providers. In the third-party space, the providers truly proliferate.[7] The products provide a variety of specialties and each comes with a DevOps-inspired graph-based user interface to track day-to-day frequency, stack-trace weak points, and provide for the global pinpointing of exception-bearing clients.

For applications running on Android, a similar proliferation of products outside of the native platform abound. The native Android SDK gives a far better venue for reporting than seen in native iOS,[2] including reporting for crash events, as well as caught and uncaught exceptions. Extensions make integration of these reports easier,[1] primarily for those keeping to the Google infrastructure. Largely, products that have the most usage in the iOS space are also available on the Android platform. These products are likely most attractive to application developers with multiplatform offerings.

Where Do QA and DevOps Converge?

After analyzing the techniques available to QA and DevOps, the comparison still begs for details on how, within the engineering phase of a development cycle, the client experience can be captured with the validity and insight that's available to an operations engineer in the DevOps era.

The currency of QA influence is manifest in the schedule, in feedback to design decisions, in prioritization of bugs, and in the decision to release despite open bugs, with the degree of convergence being largely a function of risk assessment. Using app metrics gathered from production, engineers have long been able to project severity onto discovered bugs. The task is not as easy, though, when working on new features or when a new design ends up forcing users into some completely new workflow.

The schedule for testing must move from the antiquated waterfall model to one that acknowledges the importance of testing and release simulation. This is obvious, and it constitutes a platitude of the software-development ideal. The new concept introduced by the influence of the DevOps culture promotes programmatic analysis of client scenarios. Expansion of proactive analysis of client scenarios should essentially automate the variety of use-case paths that a user could and, maybe more importantly, is likely to encounter. Issues discovered that fall more closely to the user's critical path will obviously be graded with higher severity than those that impact users falling outside of this usage pattern. The difference between the common assessment and the same analysis through DevOps lenses is the presence of data. Usage data obviously must be the basis for severity assessment.

Assessing design decisions requires expert insight into all guidance. A hallmark of the DevOps culture is a working knowledge of the overall system with the ability to profile and make predictive issue assessments based on a solvent knowledge of the entire application landscape. From the QA perspective, this insight gives credence to questions about precision, performance, and probabilities of delivering the proposed outcome of a design decision. This is where the insight of a developer is largely sufficient, but the DevOps aesthetic requires the presentation of information in the form of metrics that cannot be otherwise questioned or second-guessed.

Prioritizing bug fixes provides make-or-break influence over a software release. Mandating bug fixes prior to a release is a different responsibility. Getting bugs fixed in specific windows of the development cycle opens the opportunity for more and more interesting testing. It is obviously very important to unblock testing whenever possible, but there are frequent conflicts within the development team about when bug fixing officially begins. This is where priority must override the schedule.

When testing is blocked, testing goals are compromised. It is important to make testing goals an immobile force and to use metrics to define when two opposing forces may collide. To ensure the health of a release, testing must succeed over a certain number of days. If critical bugs are blocking that testing, the release can and should be delayed. Bug priority is required to ensure a sufficient amount of work is done to guarantee an acceptable—and successful—release.

The bottom line in the decision to release really is the bottom line, and the popularity and use of a software application is what will define its value. Whether or not this drives the financial or popular success of an effort, it is unquestionably the validator.

Release decisions must always acknowledge the presence of known customer-facing issues, and QA is frequently charged with assessing the readiness of the release. There will always be a lot of debate, but the data should speak for itself. This is where the bullet hits bone for application ownership. The data should identify what customers will see, when, and how often they will see it. An obscure bug in "Sign Up" is a lot different from an obscure bug in "Login" since the latter is a precursor to every session. Issues seen only in a dying operating system may limit exposure but also increase support costs. The goal for QA here should be to present the viable data appropriately and note issues for tracking after release.

Whether viewing DevOps as a specific role on a team or as a culture, the focus on engineering and quantified metrics gives teams the ability to make projections and trade-offs that empower product architects and management. In a way, it is a maturation of the software-development process that has been a long time coming.

No other science and few other industrial efforts make decisions without a large volume of data to help predictably validate the outcome. Where DevOps has connected the system administrative skill set to the development effort, moving that same ground shift up the chain into the quality engineering effort helps teams make more informed decisions on the development and release of their products. ■

Related articles on queue.acm.org

The Antifragile Organization
Ariel Tseitlin
http://queue.acm.org/detail.cfm?id=2499552

Traipsing Through the QA Tools Desert
Terry Coatta
http://queue.acm.org/detail.cfm?id=1046955

Quality Assurance: Much More than Testing
Stuart Feldman
http://queue.acm.org/detail.cfm?id=1046943

References

1. Acra; http://acra.ch/.
2. Google Analytics. Crashes and exceptions—Android SDK, 2013; https://developers.google.com/analytics/devguides/collection/android/v2/exceptions.
3. Jones, R. How to hire DevOps. Gun.io, 2012; http://gun.io/blog/how-to-hire-devops/.
4. Mac Developer Library; https://developer.apple.com/library/mac/navigation/.
5. Meme Generator; http://memegenerator.net/instance/22605665.
6. Mueller, E. What's a DevOp? The Agile Admin, 2010; http://theagileadmin.com/2010/10/08/whats-a-devop/.
7. Rocchi, C. Overview of iOS crash reporting tools, Part 1. Ray Wenderlich, 2013; http://www.raywenderlich.com/33669/overview-of-ios-crash-reporting-tools-part-1.
8. Samovskiy, D. The rise of DevOps, 2010; http://www.somic.org/2010/03/02/the-rise-of-devops/.

James Roche is a software developer and engineering manager focusing on testing frameworks. His work spans 16 years on projects including Amazon's retail platform, Amazon Web Services; Adobe InDesign; and Adobe's Digital Publishing Suite. He is contributing to a prototype project for Adobe, looking forward to a release soon.

DOI:10.1145/2524713.2524722

Article development led by acmqueue
queue.acm.org

Eliminating memory hogs.

BY NEIL MITCHELL

Leaking Space

A SPACE LEAK occurs when a computer program uses more memory than necessary. In contrast to memory leaks, where the leaked memory is never released, the memory consumed by a space leak is released, but later than expected. This article presents example space leaks and how to spot and eliminate them.

Let's first consider two "real-life" space leaks. Xena, the xylophone enthusiast, buys a 26-volume printed encyclopedia, but she only wants to read the article on xylophones. The encyclopedia occupies a lot of space on her bookshelf. Xena could throw away all but the X volume, reducing the shelf requirements; or she could cut out the xylophone article, leaving only a single piece of paper. In this example, Xena is storing lots of information but is interested in only a small subset of it.

Xena's friend Shawn, the statistician, is curious about how many redundant pages Xena is storing. To determine the total number of pages in the encyclopedia Shawn buys a copy of the 26-volume encyclopedia, even though he is interested in only the number of pages per volume. Actually, Shawn does not need to know the sizes of 26 separate volumes but only the total size—information that could be written on the back of a stamp.

In this example, Shawn is storing lots of information, and while each volume contains useful information, the result could be stored more compactly.

Figure 1 depicts the memory layout Xena and Shawn might represent if they were computer programs. In both cases a solid blue arrow points to the encyclopedia, representing the memory being retained. A dotted red arrow points to the information that is actually useful.

ILLUSTRATION BY JUKKA NIEMINEN

A space leak would occur if a program loaded the encyclopedia but did not immediately reduce it to the interesting part, resulting in the encyclopedia being kept in memory longer than necessary. Eliminating space leaks is about controlling when evaluation occurs, reducing the time between allocating memory and discarding it. Unsurprisingly, features that complicate evaluation order are particularly vulnerable to space leaks. The two examples this article focuses on are *lazy evaluation* (where evaluation of an expression is delayed until its value is needed) and *closures* (a function value combined with its environment). Both of these features are found in lazy functional languages such as Haskell.

Example 1: Delete

How does lazy evaluation cause a space leak? Consider the following Haskell definition:

```
xs = delete dead [alive, dead]
```

This fragment creates a variable `xs` and a two-element list using the `[_,_]` notation, containing both `alive` and `dead`. Then the element `dead` is removed from the list using `delete`. A call to `length xs` returns 1, indicating there is only one element in `xs`. In the absence of lazy evaluation, the memory layout would look like Figure 2a, where `xs` references a list containing alive as the only element; dead is not referenced and thus can be garbage collected.

Haskell uses lazy evaluation (also known as call-by-need), however, so after `xs` is defined, the memory would look like Figure 2b. Instead of pointing at a *value*, `xs` points at an *expression*, which may be replaced with an actual value later. There are still two paths from `xs` to `dead`; thus, `dead` cannot be garbage collected, even though we know it will never be used. The variable `dead` is part of a space leak because `delete` is being evaluated later than desired.

As previously mentioned, `length xs` will return 1, but as a consequence of computing the length, it will evaluate `delete`. The act of evaluating length `xs` reduces `xs` to a value, which eliminates the space leak. A program using

Figure 1. The information of interest to Xena and Shawn.

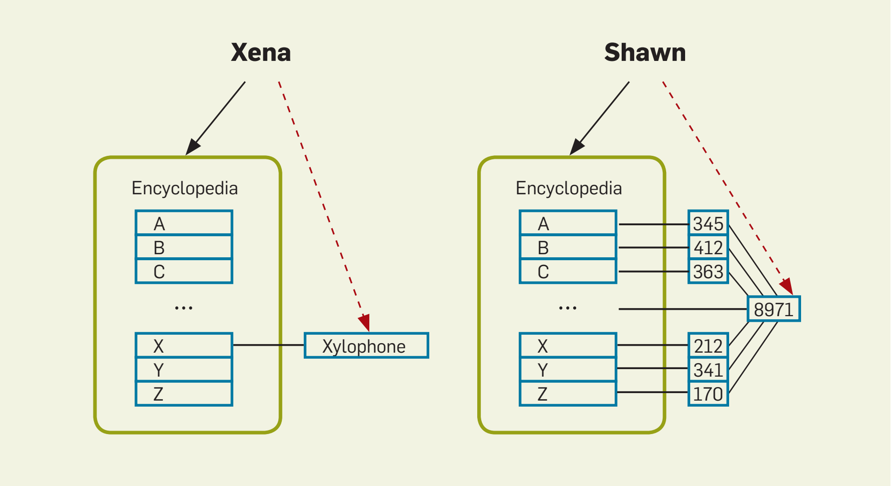
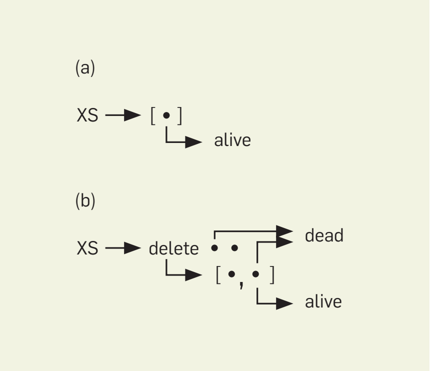


Figure 2. Lazy evaluation.



lists can end up with a space leak if it frequently adds and deletes elements but never uses the list to compute the length or look up values.

More generally, a space leak can occur when the memory contains an expression—where the expression grows regularly but where the evaluated value would not grow. Forcing the evaluation usually solves such leaks; this makes evaluation of some variables *strict* instead of lazy.

Forcing Evaluation

Eliminating space leaks often requires forcing the evaluation of an expression earlier than it would normally be evaluated. Before describing how to force evaluation, it is necessary to define how an expression is evaluated:

- An expression is in *normal form* if it cannot be evaluated further. For example, the list `[1,2]` is in normal form. Lists are constructed from `[]` (pronounced "nil") for the empty list, and `(:)` (pronounced "cons") to combine a head element to the tail of a list, so `[1,2]` can equivalently be written `1:2:[]`.
- An expression is in WHNF (weak head normal form) if the outermost part does not require further evaluation. For example, `(1+2):[]` is in WHNF since the outermost part is `(:)`, but it is not in normal form since the `(+)` can be evaluated to produce `3`. All values in normal form are by definition also in WHNF.

To force evaluation to WHNF, Haskell provides strictness annotations with the commonly used *bang patterns* extension.[3] You can define a function `output`, which prints "Output" to the console, followed by its argument:

```
output x = do
    print "Output"
    print x
```

Printing `x` evaluates `x` to normal form, so the function `output` will first print "Output," then evaluate `x` to normal form and print it. Adding an exclamation mark as a strictness annotation will force evaluation of `x` sooner:

```
output !x = ...
```

Now evaluating `output x` will first evaluate `x` to WHNF, then print "Output," then evaluate `x` to normal form and print it.

Why Lazy?

Given that strictness avoids the space leak in Example 1—and (as shown later) several other space leaks—why not make all values strict? Certainly most languages have strict values, and even some variants of Haskell default to strict evaluation.[1] As with all language design decisions, lazy evaluation is a trade-off—space leaks are a disadvantage—but there are also many advantages. Other articles discuss the advantages of lazy evaluation in depth,[2,6] but here, briefly, are a few reasons why it is a good choice:

- Defining new control structures in strict languages often requires macros or building them into the compiler, while lazy evaluation allows such patterns to be expressed directly.
- Laziness allows variable bindings to be made without considering which bindings are evaluated in which code paths, a great simplification when combined with complex conditionals.
- Simulating laziness in a strict language is usually more difficult than forcing strictness in a lazy language, so laziness can be a better default.
- When combining functions, a strict language often requires the simplistic composition to take more memory than a lazy language, as will be demonstrated in Example 2.

Example 2: Sum

Consider the following code:

```
sum [1..n]
```

In Haskell, this expression creates a list containing the numbers 1 to n, then adds them up. In a strict language, this operation takes $O(n)$ space: it would first generate a list of length n, then call `sum`. In a lazy language, however, the items in the list can be generated one at a time as they are needed by `sum`, resulting in $O(1)$ space usage. Even if you replace `[1..n]` with numbers read from a file, the program can still run in $O(1)$ space as laziness automatically interleaves reading numbers from a file and computing the sum.

Unfortunately, this code when compiled with Glasgow Haskell Compiler (GHC) takes $O(n)$ space as the result of a space leak, but when using the -O1 optimization flag takes $O(1)$ space. More confusingly, for some definitions of `sum` the code takes $O(1)$ at all optimization settings, and for other definitions the code always takes $O(n)$.

Why does the space leak arise? Consider the following definition of `sum`:

```
sum1 (x:xs) = x + sum1 xs
sum1 [] = 0
```

The first equation says if the list has at least one item in it, bind the first item to `x` and the list containing the remaining items to `xs`. The sum is then defined recursively by adding the first element to the sum of the remaining elements. The second equation expresses the base case, and the sum of the empty list is 0. Let's consider evaluating `sum1 [1..n]` for some large value of `n`, which proceeds as shown in Figure 3. You can trace the evaluation by looking at what the program will require next, working from the top left part of the expression. For example, initially `sum1` looks at the list to determine which expression to match: which one requires evaluating `[1..n]` to produce `1:[2..n]`. As evaluation proceeds it builds up the term `1 + 2 + 3 + 4 ...`, taking $O(n)$ space. While the program never has the whole list in memory at once, it instead has all the items of the list joined with "+" operations.

After the space leak is identified, strictness can be used to eliminate it. Given the expression `1 + 2`, it can be reduced to `3` immediately; and provided the program keeps performing the addition as the computation goes along, it will use only constant memory. Alas, with the definition of `sum1`, the expression is actually `1 + (2 + (3 ...`, meaning that `1` and `2` cannot be reduced. Fortunately, addition is associative, so sum can be redefined to build up `((1 + 2) + 3) ...`:

```
sum2 xs = sum2' 0 xs
    where
        sum2' a (x:xs) = sum2'
        (a+x) xs
        sum2' a [] = a
```

Defining `sum2` in terms of an auxiliary function `sum2'` takes an additional accumulator `a`, which is the value of all elements of the list processed so far. Tracing the evaluation looks more promising:

```
sum2 [1..n]
sum2' 0 [1..n]
sum2' 0 (1:[2..n])
sum2' (0+1) [2..n]
sum2' (0+1) (2:[3..n])
sum2' ((0+1)+2) [3..n]
```

Figure 3. Evaluation of `sum1`.

```
sum1 [1..n]                -- initial value
sum1 (1:[2..n])            -- sum1 requires the list
1 + sum1 [2..n]            -- sum1 reduces per the equation
1 + sum1 (2:[3..n])        -- + requires both its arguments
1 + (2 + sum1 [3..n])
```

Figure 4. Computing mean without a spave leak.

```
mean xs = mean' 0 0 xs
    where
        mean' !s !l (x:xs) = mean' (s+x) (l+1) xs
        mean' !s !l [] = s 'div' l
```

Now literal numbers are applied to addition, but the space leak is still present. Fortunately, there is now a suitable target for a strictness annotation. You can define:

```
sum3 xs = sum3' 0 xs
    where
        sum3' !a (x:xs) = sum3'
        (a+x) xs
        sum3' !a [] = a
```

The strictness annotation on the accumulator argument `a` results in the accumulator being evaluated before the next element of the list is processed. Revisiting the trace:

```
sum3 [1..n]
sum3' 0 [1..n]
sum3' 0 (1:[2..n])
sum3' (0+1) [2..n]
sum3' 1 [2..n]
sum3' 1 (2:[3..n])
sum3' (1+2) [3..n]
sum3' 3 [3..n]
```

shows that `sum3` takes $O(1)$ space and does not have a space leak. The definition of `sum` in the standard Haskell libraries is defined equivalently to `sum2`; but with optimizations turned on, the compiler infers the strictness annotation, making it equivalent to `sum3`.

Example 3: Mean

Consider another example:

```
mean xs = sum xs 'div' length xs
```

This function computes the `mean` of a list `xs` by taking the `sum` and dividing by the `length` (the backticks around `div` allow the use of a function as an infix operator). Assuming a space-leak-free definition of `sum`, how much space will `mean [1..n]` take?

Using lazy evaluation—namely, reducing the top left expression first—the answer is $O(n)$. Evaluating `sum xs` requires evaluating the entire list `xs`, but since that list is also used by `length xs`, `xs` must be retained in memory instead of being collected as it is produced.

In this example a smarter evaluation strategy could eliminate the space leak. If the program evaluated the first element of `xs`, then applied both `sum` and `length` to it, the function would take constant space. Another approach to computing `mean [1..n]` is to remove the sharing of the list:

```
sum [1..n] 'div' length [1..n]
```

Here the list has been duplicated, and both arguments to `div` run in constant space, allowing the entire computation to run in constant space. Unfortunately, any work required to compute the lists will be duplicated.

The real solution is to take the pattern used for `sum3` and extend it so instead of accumulating just the sum, you also accumulate the length. The full definition is illustrated in Figure 4.

This accumulates the sum (`s`) and length (`l`) as local parameters, which are strict arguments to the helper function. The resulting definition has no space leak and runs in $O(1)$.

Example 4: Space Leaks and the Garbage Collector

The previous examples have inserted strictness annotations to eliminate space leaks. Not all space leaks can be removed by strictness annotations, however[5]; sometimes, special behavior is required from the garbage collector.[10] As an example, let's improve the impact of an academic paper by placing an exclamation mark at the end of the title as shown in Figure 5.

The `improve` function takes the source of the paper and produces a new paper. It splits the text into a variable `pair`, consisting of the first line and the remaining text, using the auxiliary `firstLine`. The function then takes the first element of the pair using `fst`, and the second element using `snd`, and uses the string append operator "++" to insert an exclamation mark between them. The first equation of `firstLine` matches strings with a leading newline character and produces an empty first line, followed by the text. The second equation recursively calls `firstLine` with everything but the first character, then creates a result where the first character is at the front of the first line. The final equation ensures the empty input produces empty outputs.

It should be possible for `improve` to run in *O*(*1*) space, producing an output character after examining each input character, and requiring only a small amount of memory. In the second equation of `firstLine`, after matching `y:ys` (that is, consuming an input character), the program immediately produces `(y:_, _)`, making an output character available via lazy evaluation before making the recursive call. Unfortunately, using the obvious implementation techniques, this function requires space proportional to the first line of `xs`, so *O*(`fst pair`).

To understand the space usage, consider the evaluation of `improve "abc..."` depicted in Figure 6.

In each step of `firstLine` a pair is produced where the second component of that pair is simply the second component of the recursive call. The result is both a linear chain of `snd` calls and all the characters being retained by references to the first component of each `rest` variable.

If the `snd` functions were forced, then this space leak would be eliminated to produce:

As with all language design decisions, lazy evaluation is a trade-off—space leaks are a disadvantage—but there are also many advantages.

```
let rest4 = firstLine "..."
'a':'b':'c':fst rest4 ++ "!\n" ++
snd rest4
```

Unfortunately, there is nowhere to put a strictness annotation to perform the appropriate reduction. Although you want to force the evaluation of `snd`, you are also relying on the laziness of the pair in the recursive call of `firstLine` to achieve *O*(*1*) space. Fortunately, the garbage collector can solve this problem. The function `snd` is a selector—given a pair, it selects the second component. It does not compute any new values, does not allocate memory, and is cheap to compute. As such, the program can *evaluate* `snd` *during garbage collection*, which eliminates the space leak. The reduction of selector functions during garbage collection is now a standard feature of lazy functional languages, automatically removing space leaks that would otherwise be impossible to eliminate.

Example 5: Space Leaks and Closures

All the examples so far have been in Haskell, but other garbage-collected languages are also susceptible to space leaks. While few languages are lazy by default, many support *closures*—a lambda expression or function, plus some variables bound in an environment. One popular language that makes extensive use of closures is JavaScript.

The JavaScript code in Figure 7 uses the Web Audio API[8] to retrieve an MP3 file and compute its duration:

This function uses the `XMLHttpRequest` API to load an MP3 file, then uses the Web Audio API to decode the file. Using the decoded `audio` value, you can add an action that tells the user the MP3's duration whenever a status button is clicked.

The implementation uses three local functions, two of which reference variables defined locally to `LoadAudio`. Those variables will be captured inside a closure when the local functions are referenced. As an example, the first function is assigned to `onreadystatechange` and captures the `request` variable defined three lines before.

After `LoadAudio` has run, the "status" button has an onclick event that runs the following code:

```
alert("MP3 is "
  + audio.duration +
  " seconds long");
```

This code references the `audio` object, which stores the audio data—taking at least as much memory as the original MP3. The only thing ever accessed, however, is the `duration` field, which is a number, taking a mere eight bytes. The result is a space leak.

This space leak has many aspects in common with the lazy evaluation space leaks. The code references an expression `audio.duration`, which keeps alive a significant amount of memory, but when evaluated uses only a small amount of memory. As before, the solution is to force the evaluation sooner than necessary as shown in Figure 8.

Now the duration is computed before the `onclick` event is registered, and the `audio` element is no longer referenced, allowing it to be garbage collected.

JavaScript Selectors

While you can modify the code to eliminate the space leak, could the garbage collector have eliminated the space leak? The answer is yes, provided that `audio.duration` is cheap to compute, cannot change in the future, and will not cause any side effects. Since there are no other references to `audio`, the value to which `audio` refers cannot change; and since `audio.duration` is a read-only field, it was likely computed when the `audio` value was constructed. This optimization would be an instance of the selector evaluation from Example 4.

Unfortunately, the selector optimization is less applicable in JavaScript than in Haskell, because most values are mutable. Consider the example in Figure 9. This code defines a dictionary containing both `pi` (a number) and `fiveDigitPrimes` (a large array), then adds an event handler that uses `pi` only. If constants were immutable, then the garbage collector could reduce `constants.pi` and remove the reference to constants. Alas, the user can write `constants = {pi : 3}` to mutate `constants`, or `constants.pi = 3` to mutate the `pi` field, meaning evaluation in advance is unsafe.

Figure 5. Definition of `improve`.

```
improve xs = fst pair ++ "!" ++ snd pair
    where pair = firstLine xs

firstLine ('\n':ys) = ([], '\n':ys)
firstLine (y:ys) = (y:fst rest, snd rest)
    where rest = firstLine ys
firstLine [] = ([], [])
```

Figure 6. Evaluation of `improve`.

```
let rest4 = firstLine "..."
let rest3 = ('c':fst rest4, snd rest4)
let rest2 = ('b':fst rest3, snd rest3)
let rest1 = ('a':fst rest2, snd rest2)
'a':'b':'c':fst rest4 ++ "!" ++ snd rest1
```

Figure 7. Retrieving an MP3 file.

```
function LoadAudio(mp3)
{
    // Load 'mp3' file into 'request.response'
    var request = new XMLHttpRequest();
    request.open('GET', mp3);
    request.responseType = 'arraybuffer';

    request.onreadystatechange = function(){
        if (request.readyState != 4) return;

        // Decode the audio data
        window.AudioContext = window.AudioContext ||
                              window.webkitAudioContext;
        var context = new AudioContext();
        context.decodeAudioData(request.response, function(audio){
            document.getElementById("status").onclick = function(){
                alert("MP3 is " + audio.duration + " seconds long");
            }
        });
    };
    request.send();
}
```

Figure 8. Reordering evaluation to reduce memory usage.

```
var duration = audio.duration;
document.getElementById("status").onclick = function(){
        alert("MP3 is " + duration + " seconds long");
};
```

Figure 9. Unnecessary memory usage.

```
var constants = {pi : 3.142, fiveDigitPrimes : [10007,10009,10037,...]};
document.getElementById("fire").onclick = function(){
        alert(constants.pi);
};
```

While the difficulties of mutation mean that JavaScript does not reduce such functions in practice, it is not an insurmountable barrier. Consider a memory layout where you know which references are being used as read-only (such as, `alert(constants.pi)`) and which are not (that is, `constants.pi = 3`). This information can help determine which variables are used only as read-only and thus are guaranteed to be constant. If `constants` and `constants.pi` are both determined to be immutable, then the field lookup could be performed by the garbage collector, freeing both `constants` and `fiveDigitPrimes`.

In Haskell, lazy evaluation is common (the default) and space leaks caused by selectors are unavoidable, making the decision to apply selector optimization obvious. In languages such as JavaScript, adding code to solve fixable space leaks at the cost of making the normal code slower or more complex may not be a sensible trade-off.

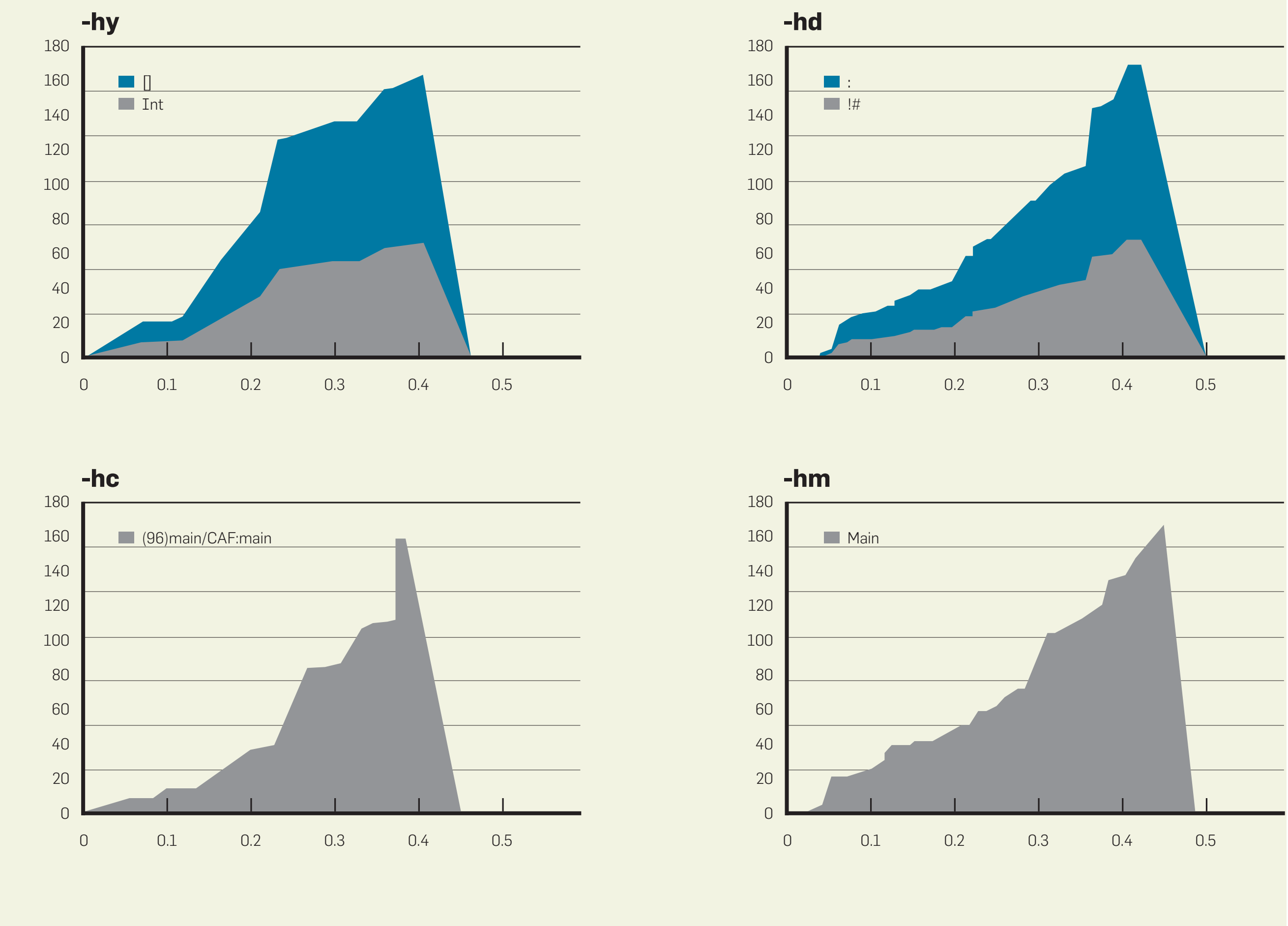
Detecting Space Leaks

The five examples of space leaks presented here provide some guidance as to where space leaks occur and how they can be fixed. All the examples, however, have consisted of only a handful of lines; for space leaks in big programs the challenge is often finding the code at fault. As Haskell is particularly vulnerable to space leaks, the compiler provides a number of built-in profiling tools to pinpoint the source of space leaks. Before looking at which tools are available, let's first consider which might be useful.

Space leaks are quite different from memory leaks—in particular, the garbage collector still knows about the memory referenced by the space leak and will usually free that memory before the program terminates. Assume a definition of `sum` contains a space leak; as soon as `sum` produces a result, the garbage collector will free any intermediate space leak. A program with a space leak will often reach its peak memory use in the middle of the execution, compared with memory leaks that never decrease. A standard technique for diagnosing memory leaks is to look at the memory after the program has finished, to see what is unexpectedly retained. This technique is not applicable to space leaks.

Instead, it is often useful to examine the memory at intermediate points throughout the execution, looking for spikes in the memory usage. Capturing the entire memory at frequent intervals is likely to require too much disk space, so one solution is to record summary statistics at regular intervals, such as how much memory is allocated by each function.

Figure 10. Profiles for the mean example with different memory groupings. The *x*-axis is time in seconds and the *y*-axis is memory in MB.



Haskell tools. The Haskell compiler provides several profiling modes that generate plots summarizing memory usage. To generate a profile, first compile the program with the following flags:

```
ghc --make Main.hs -prof -fprof-
auto -fprof-cafs -rtsopts
```

These flags are:

▸ `ghc --make Main.hs`. Compile the file `Main.hs` into an executable, as normal.

▸ `-prof -fprof-auto -fprof-caf`. Turn on profiling in the executable and make sure it is able to record information about top-level definitions.

▸ `-rtsopts`. Allow the resulting executable to accept profiling options.

The resulting program can run as normal, but with additional flags it can also generate profile information:

```
main +RTS -xt -hy
hp2ps -c main.hp
```

Using the `mean` example presented earlier produces the first plot shown in Figure 10. The first command runs the resulting `main` executable with some flags to the runtime system (anything after `+RTS`). The `-xt` flag includes any stack in the profile output (this author believes `-xt` should be on by default), and `-hy` generates a report summarized by type. The first command generates a file `main.hp`, and the second command turns that into a PostScript file `main.ps` (in color, due to the `-c` flag). In the plots shown I also passed `-i0.01` to sample the memory more frequently, which is usually necessary only when trying quick-running toy examples.

Haskell has a number of profiling modes, and the simplest approach is to try them all and see which produces the most useful information. The four standard types of profiles, shown in Figure 10 are:

▸ `-hy`. Summarizes the memory by type. The example has some lists (`[]`) and numbers (`Int`). This summary answers the question of *what* is in the memory.

▸ `-hd`. Summarizes by description, showing a more refined version of `-hy`. In the example there is a close correspondence to `-hy`, with all `Int` entries matching `I#` (which is the internal constructor of `Int`) and lists matching `(:)`. Any group below a threshold is hidden; otherwise, there would likely be a single `[]` denoting the end of the list.

▸ `-hc`. Summarizes by cost center, a named area of the source code automatically inserted on all top-level definitions. It can also be manually inserted with an annotation in the code. In Figure 10 `main` has been attributed all the memory, probably a result of optimization inlining `mean` inside of it. This summary answers the question of *where* was the memory created.

▸ `-hm`. Summarizes by module, which is a more granular version of a cost center.

From a combination of these plots you can see the function main in the module `Main` allocates a large list of numbers. It allocates the list over 0.4 seconds, then quickly consumes the list over 0.1 seconds. This memory usage describes what would be expected from the original definition of `mean`.

For larger programs the plot will often contain a lot of memory usage that is expected—and not relevant to the space leak. To simplify the plot you can filter by any of the four types: for example, passing `-hc -hy[]` will generate a plot grouped by cost center but only for memory where the type is a list.

As seen in the `sum` example, compiling with different optimization settings may cause space leaks to appear or disappear, and, sadly, compiling for profiling can have similar effects (although it is relatively rare). As a fallback, any Haskell executable can be run using `+RTS -hT`, which produces a plot summarized by type without compiling for profiling. This causes fewer changes to the behavior of the program.

Before using the profiling tools, read the Profiling section of the GHC manual, which covers several additional flavors of profiling. For a better idea how the profiling tools can be applied to large programs and how to interpret the results, I recommend the following two "tales from the trenches" from me and Edward Yang:

▸ http://blog.ezyang.com/2011/06/pinpointing-space-leaks-in-big-programs/

▸ http://neilmitchell.blogspot.com/2013/02/chasing-space-leak-in-shake.html

Compiling with different optimization settings may cause space leaks to appear or disappear, and, sadly, compiling for profiling can have similar effects (although it is relatively rare).

JavaScript tools. One tool Haskell lacks is the ability to pause execution at a certain point and explore the memory. This feature is available in some JavaScript implementations, including in Chrome as the heap profiler.

The Chrome heap profiler allows a snapshot of the memory to be taken and explored. The profiler displays a tree of the memory, showing which values point at each other. You can summarize by the type of object, see statistics about how much memory is consumed and referenced by a certain value, and filter by name. A feature particularly useful for diagnosing space leaks is the ability to see what references are keeping a value alive. The two JavaScript space leaks in this article produce heap snapshots that easily pinpoint the problem.

Are Space Leaks Inevitable?

Garbage collection frees programmers from the monotony of manually managing memory, making it easier for languages to include advanced features such as lazy evaluation or closures. These advanced features lead to more complex memory layout, making it more difficult to predict what memory looks like, potentially leading to space leaks.

Compilers for lazy functional languages have been dealing with space leaks for more than 30 years and have developed a number of strategies to help. There have been changes to compilation techniques and modifications to the garbage collector and profilers to pinpoint space leaks when they do occur. Some of these strategies may be applicable to other languages. Despite all the improvements, space leaks remain a thorn in the side of lazy evaluation, providing a significant disadvantage to weigh against the benefits.

While space leaks are worrisome, they are not fatal, and they can be detected and eliminated. The presence of lazy evaluation has not stopped Haskell from being used successfully in many projects (you can find many examples in the conference proceedings of the Commercial Users of Functional Programming). While there is no obvious silver bullet for space leaks, there are three approaches that could help:

- Some complex problem domains have libraries that eliminate a large class of space leaks by design. One example is Functional Reactive Programming, which is used to build interactive applications such as user interfaces and sound synthesizers. By changing how the library is defined you can both guarantee certain temporal properties and eliminate a common source of space leaks.[7] Another example is stream processing, which is used heavily in Web servers to consume streams (for example, a JavaScript file) and produce new streams (for example, a minimized JavaScript file) without keeping the whole stream in memory. Several competing stream libraries are available for Haskell. All of them ensure memory is retained no longer than necessary and the results are streamed to the user as soon as possible.
- Space leaks are often detected relatively late in the development process, sometimes years after the code was written and deployed, and often only in response to user complaints of high memory usage. If space leaks could be detected earlier—ideally, as soon as they are introduced—they would be easier to fix and would never reach end users. Certain types of advanced profiling information can detect suspicious memory patterns,[9] and some experimental tools can annotate expected heap usage,[4] but nothing has reached mainstream use. The Haskell compiler does partition memory in such a way that some space leaks are detected—the `sum` example fails with a message about stack overflow for lists of length 508146 and above, but the other examples in this article use all available memory before failing.
- The tools for pinpointing space leaks are powerful but certainly not perfect. An interactive viewer can explore existing plots,[11] but users are still required to specify how the memory is grouped before running the program. It would be much easier if all four groupings could be captured at once. A feature missing from Haskell programs is the ability to take a snapshot of the memory to examine later, which would be even more powerful if combined with the ability to take a snapshot when memory exceeded a certain threshold. Pinpointing space leaks is a skill that takes practice and perseverance. Better tools could significantly simplify the process.

Related articles on queue.acm.org

NUMA (Non-Uniform Memory Access): An Overview
Christoph Lameter
http://queue.acm.org/detail.cfm?id=2513149

Software Transactional Memory: Why Is It Only a Research Toy?
Calin Cascaval, Colin Blundell, Maged Michael, Harold W. Cain, Peng Wu, Stefanie Chiras, and Siddhartha Chatterjee
http://queue.acm.org/detail.cfm?id=1454466

You Don't Know Jack about Shared Variables or Memory Models
Hans-J Boehm and Sarita V. Adve
http://queue.acm.org/detail.cfm?id=2088916

References

1. Augustsson, L. Pragmatic Haskell. Presentation at the Commercial Users of Functional Programming Conference (2011); http://www.youtube.com/watch?v=hgOzYZDrXL0.
2. Augustsson, L. More points for lazy evaluation. Things that Amuse Me; http://augustss.blogspot.co.uk/2011/05/more-points-for-lazy-evaluation-in.html.
3. Glasgow Haskell Compiler Team. The Glorious Glasgow Haskell Compilation System User's Guide, Version 7.6.3 (2013); http://www.haskell.org/ghc/docs/latest/html/users_guide/index.html.
4. Hofmann, M. and Jost, S. Static prediction of heap space usage for first-order functional programs. In *Proceedings of the 30th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (2003), 185–197.
5. Hughes, J. The design and implementation of programming languages. Ph.D. thesis. Oxford University, 1983.
6. Hughes, J. Why functional programming matters. *Computer Journal 32*, 2 (1989), 98–107.
7. Liu, H. and Hudak, P. Plugging a space leak with an arrow. *Electronic Notes in Theoretical Computer Science 193* (2007); 29–45.
8. Rogers, C. Web Audio API; http://www.w3.org/TR/2012/WD-webaudio-20120802/.
9. Röjemo, N. and Runciman, C. Lag, drag, void and use—heap profiling and space-efficient compilation revisited. In *Proceedings of the 1st ACM SIGPLAN International Conference on Functional Programming*, (1996), 34–41.
10. Wadler, P. Fixing some space leaks with a garbage collector. *Software: Practice and Experience 17*, 9 (1987), 595–608.
11. Yang, E. hp/D3.js (2013); http://heap.ezyang.com/.

Neil Mitchell is a Haskell programmer who works for Standard Chartered Bank. His open source projects include a search engine (Hoogle), a code suggestion tool (HLint), and a library for writing build systems (Shake).

DOI:10.1145/2524713.2524723

Article development led by acmqueue
queue.acm.org

Combining agile and SEMAT yields more advantages than either one alone.

BY IVAR JACOBSON, IAN SPENCE, AND PAN-WEI NG

Agile and SEMAT—Perfect Partners

TODAY, AS ALWAYS, many different initiatives are under way to improve the ways in which software is developed. The most popular and prevalent of these is the agile movement. One of the newer kids on the block is the Software Engineering Method and Theory (SEMAT) initiative. As with any new initiative people are struggling to see how it fits into the world and relates to all the other things going on. For example, does it improve or replace their current ways of working? Is it like lean, which supports and furthers the aims of the agile movement; or is it more like waterfall planning, which is in opposition to an agile approach?

Many have wondered whether SEMAT and agile are complementary or competitive initiatives, and if they are complementary, how do they fit together? In this article we demonstrate how these two initiatives support each other and we discuss the benefits of using them together.

Generally speaking, both initiatives promote non-prescriptive value-based philosophies that encourage software development teams to select and use whatever practices best fit their context and, more importantly, continuously inspect, adapt, and improve their ways of working. These two initiatives complement one another, providing the perfect foundation for teams that want to master the art of software development.

The agile movement has provided a new way of looking at the day-to-day activities of software development—how teams are built and work is organized. This has led to the empowerment of development teams and the prominence of agile practices (such as Scrum and test-driven development) as the modern developer's practices of choice.

SEMAT is a new way of looking at the domain of software engineering (here the term is used interchangeably with software development), providing an understanding of the progress and health of development efforts and how practices can be combined into an effective way of working. SEMAT adds to the power of agility by providing a common reference model all teams can when continuously inspecting, adapting, and improving their ways of working.

The two initiatives when used together, truly empower teams to innovate, experiment, and improve the results they deliver.

This article focuses on how SEMAT can help existing and future agile teams. It is designed for those already familiar with agility.

What SEMAT Adds to Agile

Agile provides a set of values that influence and shape the way software developers go about their daily work and interact with one another, their customers, and their stakeholders.

It has also given us many methods that share common principles but differ in practice. These are methods that developers must be able to inspect and adapt as circumstances change. The agile methods give teams a great starting point on their agile journey but they need to evolve to meet the team's changing needs and reflect the lessons they learn. This is reflected in the growing number of agile teams that assemble a bespoke method from the available set of practices rather than taking a made-to-measure method off the shelf.

The use of SEMAT can help agile teams do the following:

Detect systemic problems early and take appropriate action. Agile teams continuously inspect and adapt using fast feedback and close collaboration to avoid problems and provide direction to the team. To support and encourage this way of working, SEMAT provides a number of simple checklists to help teams understand their progress and health, and to help them in the early detection of problems with their way of working. The checklists that SEMAT provides are akin to those used in other professions. For example surgery teams in U.K. hospitals reduced death by surgical errors by 47% by using a simple 19-question checklist that had questions such as "Do you know the names of the other members of the surgical team?" In the same way the use of the SEMAT checklists reduces the risk of teams failing catastrophically by helping them avoid many of the common mistakes that lead to failure such as ever increasing technical debt, loss of stakeholder support, inefficient ways of working, unrealistic expectations and dysfunctional teams.

Measure the team's progress and health regardless of the method or practices selected. The key measure of progress for all agile teams is the amount of working software they produce and the speed with which they produce it. SEMAT complements these measures by providing another view of the progress and health of the team and its work—a view that can help teams maintain their speed as they and the systems they produce mature. By using SEMAT and the simple checklists it provides to assess their current state, teams can easily understand where they are, where they should go next, and how their efforts fit within any organizational governance practices they need to support.

Compare and contrast practices and select the best ones for the team. Agile teams are perpetually looking for new practices to help them improve their way of working and evolve their methods. SEMAT provides the mechanisms to understand the extent, purpose, and content of practices, helping teams understand their coverage and where they overlap, conflict, or complete. It also allows teams to plug-and-play practices, safely mixing and matching them within the context of their favorite agile framework—for example, Scrum or Kanban.

Evaluate the completeness of the set of practices selected, and understand the strengths and weaknesses of the way of working. In the rush to adopt new practices, teams sometimes leave holes in their way of working, the consequences of which often do not become apparent until the team's speed starts to drop and it consistently falls short in achieving its objectives. This does not mean the way of working needs to be predefined or complete; in fact, it is probably better if it is not. What is important is the team members are aware of what they have agreed on, where they are aligned, and where they might need help. The use of SEMAT helps teams reason about the way of working and make fact-based decisions about the breadth and depth of their selected set of practices. Having mechanisms to help understand the strengths, weaknesses, and completeness of their way of working is invaluable for those teams truly committed to continual improvement.

Keep an up-to-date record of the team's way of working, and share information and experiences with other teams. The agile community thrives on collaboration and interaction. The sharing of practices and experiences helps individuals, teams, organizations, and the industry as a whole improve and evolve. SEMAT provides mechanisms to help teams accurately record their way of working in a lightweight, agile fashion, which they can share in real time with their colleagues and collaborators. This provides transparency with respect to the team's way of working, and it helps everyone understand what the team is doing without getting confused by out-of-date descriptions of what the team is supposed to be doing or what the team members thought they would be doing before they actually gained experience doing it.

Be agile with methods, easily and safely evolving the team's set of practices as it inspects and adapts its way of working. Inspecting and adapting the way of working is essential for any agile team that truly wants to continuously improve. Its effectiveness can be hindered when teams: become too wedded to the current set of practices, effectively freezing their way of working; select different but less-effective practices that introduce more problems than they address; or do not understand where they are in the evolution of the software system and therefore which practices they should change. SEMAT provides the frameworks and thinking tools to help teams more effectively inspect and adapt their way of working, understand the consequences of their decisions, and continuously improve their way of working.

SEMAT for Agile Organizations

SEMAT provides additional support that helps entire organizations become agile without compromising the agility of the teams that form them. In particular, it helps:

Establish the ground rules for software development within the organization, and capture organizational values and principles in a practice-independent fashion. Software development does not happen in isolation. Development teams must always be cognizant of the culture, values, and principles important to the organizations they work with. They need to establish some common ground and shared understanding with the other teams and areas of the organization they interact with. SEMAT provides a simple definition of the common ground shared by all software-development teams. This forms a firm foundation for organizations wanting to integrate software development into their businesses and value flows. Organizations can extend the SEMAT definitions to capture any additional rules or advice that applies to the specific kind of software they develop or the specific environment within which they develop it. Establishing the common ground is a prerequisite to organizational agility, but it is not sufficient. It should be complemented with an organizational practice exchange where the teams can share the practices they use.

Define practice-independent governance procedures and quality gates. For business, legal, and safety-critical reasons, many organizations feel the need to apply governance to their software-development efforts. Most large organizations are legally required to perform financial and/or technical governance on their software teams and the software they produce. Unfortunately, many organizations define their governance as a series of sequential phases, each with a predefined set of required artifacts that must be completed and signed off before the next phase can be started.

It is impossible for agile teams to achieve their full potential in this kind of rigid, prescriptive environment. Governance is there to provide checks and balances and ensure the quality of the results produced. Governance procedures and quality gates should be aligned to the natural evolution of the software systems produced, focused on the key results required rather than artifacts to be produced, and manifested as simple practice-independent checklists. They would then provide a framework to support, rather than inhibit, agile and lean ways of working. This is the approach that SEMAT takes, allowing governance procedures and quality gates to be defined in a lightweight and practice-independent fashion. The agile teams can then mix and match whichever agile practices they desire, and they can continuously inspect and adapt without ever having to fall out of governance.

Track and encourage the use of practices within the organization. Agile teams love to learn and share new practices; it is a fundamental part of the approach to continuous improvement. By basing all software-development effort around a common ground, teams can more readily and easily share their practices. By setting up a practice exchange to facilitate the sharing and distribution of practices, an organization can gain insight into which practices are being used where, and which sets of practices are producing the best results. This helps organizations to become true learning organizations continuously evolving their set of recommended practices, withdrawing those that are past their sell-by date, and promoting new practices when needed.

More readily and easily form teams and mobilize teams of teams. Although agile teams are intended to stay together, the reality is they are regularly changing team members, even when they do not work for organizations that insist on matrix management approaches and constant reorganization. Context switching in this way can often reduce velocity, increase friction, and waste time. SEMAT provides teams with a common language for software engineering that will help them understand one another, clearly express themselves, and share the practices they know—all of which will help them collaborate quickly and effectively—minimizing wasted time, pointless discussions, and unnecessary misunderstandings. It also provides mechanisms for modeling the competency required by teams and attained by individuals. This can help organizations find the right people to join the right teams and then observe their development as software professionals.

Scale agile approaches across teams of teams and systems of systems. Scaling agility is one of the biggest challenges currently facing organizations that want to become more agile. The SEMAT approach helps organizations scale agility in a number of ways:

- It establishes a common ground for all the teams involved. Scaled agility requires many teams to collaborate, working on the same systems and improving the same value flows. In this situation it is even more essential that all the teams have a shared understanding of what they are doing and a shared language to help them communicate.
- It allows teams to be flexible about their practices. Scaling agility requires even more flexibility in the set of practices that teams can use. Teams collaborating on the same system will need to share practices with one another. Teams working on certain systems will need to use some of the practices originally used to develop the system. SEMAT's ability to mix and match practices, swapping them in and out of play as needed, provides the flexibility in the way of working that teams need to succeed in a scaled agile environment.
- It helps teams understand their interaction points with other teams, the boundary of their responsibilities, and how their progress and health affects the teams they work with. If everybody is using a common ground to indicate their responsibilities and how they are progressing, then inter-team working is easily monitored and improved.

Select enterprise-level tooling. By providing a common ground for software development, SEMAT also provides a common ground for enterprise-level tooling. The separation of the shared common ground from the various practices used helps organizations understand which practice-independent tooling they need, which practice-specific tooling they need, and how these are related. SEMAT also helps teams understand how to integrate the tools they use by providing definitions of the common elements they will share.

Figure 1. Software development as a multidimensional endeavor.

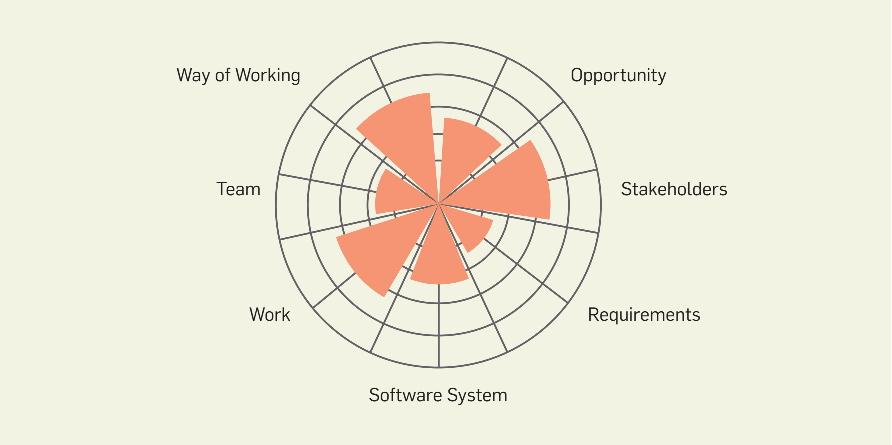
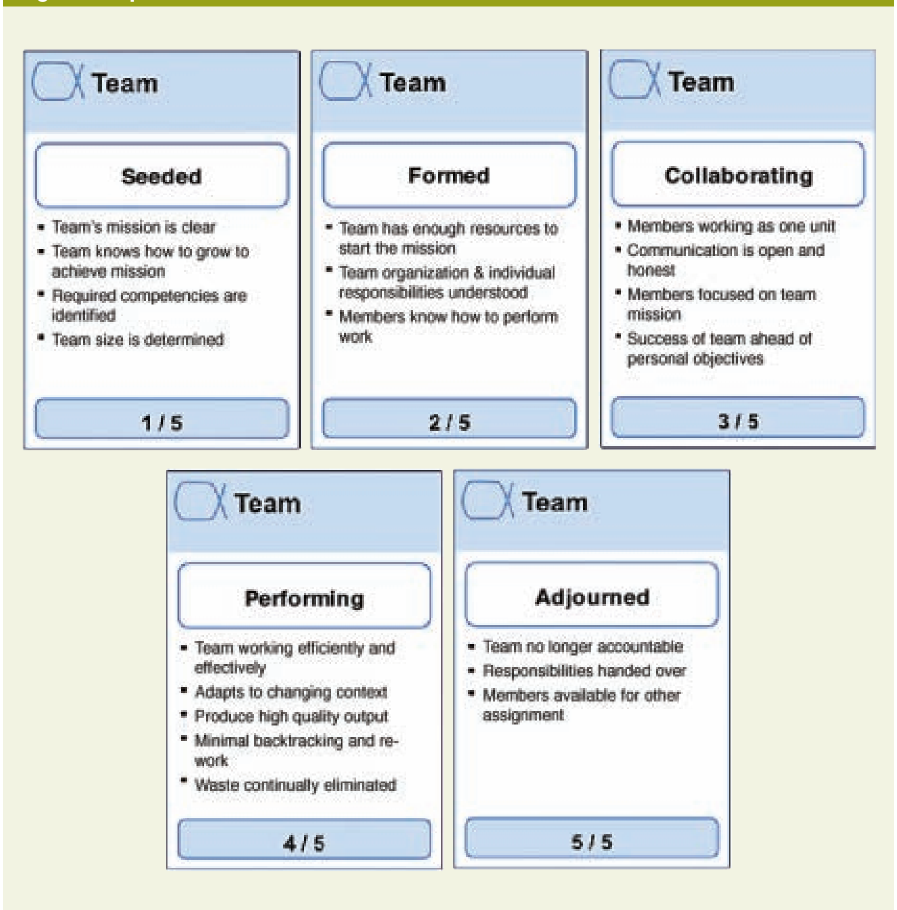


Figure 2. Alpha state cards with checklists.



What Agile Adds to SEMAT

SEMAT is nonprescriptive to such an extent it does not even insist upon adopting an agile approach. It does not care what approach a team adopts as long as it produces "good" software in an effective and healthy fashion.

Adopting agile values brings many benefits to teams and organizations—too many to go into in this brief article. For organizations adopting SEMAT, agility adds a number of important elements in the area of software-engineering methods, including:

▸ *Principles and values.* The addition of agile values to the SEMAT framework provides a necessary qualitative dimension to the evaluation of progress and health.

▸ *Many practices.* The agile community is a hotbed of new and innovative practices, all of which could be codified and made available as SEMAT practices for teams safely to compare, contrast, and mix and match.

▸ *A driving force for improvement.* Agility embeds the inspect-and-adapt cycle into every aspect of the team's work.

Before adopting SEMAT, a team should establish the principles and values it would like its new way of working to embody; otherwise, it will be very difficult to select the right practices or break out of what can at first appear to be an academic process-building exercise.

The brevity of this section, when contrasted to the earlier "What SEMAT Adds to Agile" section, is a reflection of the broader acceptance and knowledge of agility than SEMAT. It does not represent the relative value or impact of the two initiatives.

How Does SEMAT Do All This?

The goal of the SEMAT initiative is to provide software developers with a sound practical and theoretical foundation to improve their performance. (For more detailed information, see Jacobson et al.[3])

The first step in the SEMAT initiative is to establish a common ground for software professionals (developers, testers, among others) to stand upon when they talk about what they do. This common ground manifests itself in Essence, a kernel of universal elements in software development—elements prevalent in every development endeavor. Essence includes these elements: requirements, software system, work, team, way of working, opportunity, and stakeholders.

These elements have states, which can be used to measure progress and health. For example, a team can take the following states: seeded, formed, collaborating, performing, and adjourned. To achieve a particular state, a number of checkpoints must be fulfilled, representing real achievements. To achieve state collaborating, for example, the following checkpoints have been fulfilled: the team works as one cohesive unit; communication within the team is open and honest; the team is focused on achieving the team mission; and the team members know each other.

Traditionally, checkpoints have been used to measure the completion of an activity or a document, but the SEMAT checkpoints measure outcome. Thus, the universal elements represent achievements rather than documents or artifacts. This makes them agnostic to any particular method—agile or not. These elements are called *alphas*.

Software development is multi dimensional, and alphas identify the typical dimensions every software-development endeavor must consider to progress in a healthy manner. A radar chart, as depicted in Figure 1, gives a view of the current progress along each dimension.[1] Each line originating from the center represents an alpha, and the radials on that line represent the current state for that alpha.

Essence also provides a lightweight approach to describe practices on top of the kernel and thus extend the kernel. From a library of practices, teams can select appropriate ones and compose them to get the way of working they are satisfied with. In this way, they can evolve their way of working over time by replacing existing practices with newer and better ones. Practices can be of different kinds—for example, business, social, or technical. Each practice can add guidance for moving an alpha from one state to another, or it can add alphas not included in the kernel. In this way the endeavor will add more dimensions. It can also add work products to each alpha it touches. For example, the use-case-driven development practice might add a use case as an alpha and use-case specifications and realizations as work products.

Cards and checklists. The Essence specification provides a detailed description of the kernel alphas, including the definition of their checkpoints. In its daily work, however, a team would not carry the Essence specification with it. A more concise and practical representation in the form of a deck of cards suffices. Figure 2 shows the state cards for the team alpha.

Each card has the name of the alpha at the top, followed by the state name and a concise list of checkpoints. These act as useful reminders for developers.

Boards and visuals. In addition to state cards, there are alternative ways of working with alphas—for example, an alpha abacus, as shown in Figure 3. An abacus is a Chinese calculation device with beads (counters) on a wire (representing digits). In the alpha abacus, each wire represents an alpha, and each bead an alpha state.

This visual board can be used for a variety of purposes. One possible use is for a team to evaluate its current state (where it is) and discuss its next objective (where it wants to go next). This is easily visualized by drawing imaginary lines and positioning the beads as shown in Figure 3.

Games. Once cards and visuals are available, it becomes natural to have games. For example, Progress Poker, a game that evaluates progress and health, is similar to the Planning Poker game used in agile methods. In Progress Poker, each member of the team selects a state card for each alpha to represent the current state of development. If they all choose the same state card, it means they have a common understanding of the progress. If they choose different cards, they probably have different understandings of where their development stands, and different expectations of what needs to be done. This misunderstanding usually signifies the presence of risks. Once it is discovered, team members can have further discussions to reach a consensus. Other games—Objective Go, Chasing the State, and so on—can be found at http://www.ivarjacobson.com/alphastatecards.

Case Studies

The case studies described in this section are good examples of how software-development teams can make good use of SEMAT and Essence.

Equipping Coaches in a Large Telecommunications Company

We worked with a large Chinese telecommunications-product company that had a number of internal coaches. The capabilities of these coaches were critical to each team's ability to improve. Equipping the coaches to detect development problems early was important. In our first contact with one of the coaches, we asked how his team was doing. He felt that progress was good. We then asked him to evaluate progress using a deck of alpha state cards. He laid the cards on the table and started shifting them and quickly identified that progress of the stakeholders alpha was slow. He recognized this was a risk and made it a point to work out a plan to address the risk, which was basically to achieve the first four states of the stakeholder alpha. The initial discussion with this coach took only 15 minutes. A further discussion found the coach came from a development background rather than a business-analysis background, which was probably the reason he neglected the stakeholders dimension.

In this particular case, the coach in question was weak in one area. In other cases, coaches had neglected other dimensions represented by the software-system alpha such as design and quality. In yet other cases, there were disagreements among team members about the way-of-working alpha. Whatever the case, the Essence alphas were simple, intuitive, and effective tools for evaluating progress and health.

Running Development in an Internet Media Product Line

The next case study involves several development teams in Beijing collaborating to deliver an Internet media server. This was a new product line, and the team members and leaders were relatively junior. They had much to learn, not just about how to work, but also about their problem domain. In addition, they were transitioning from a traditional stovepipe organization where testers and developers worked separately to one in which developers and testers collaborated as a cross-functional team.

Our approach involved using the kernel and the use-case-driven development practice[2] to design the team visualization board shown in Figure 4. This team visualization board provided visualization from three different perspectives:

- *Process.* This made the alphas visible to team members so they would know their current iteration objectives (that is, which kernel alpha states they needed to get to). This also included a section showing the current states for the use-case slices they are working on. A use-case slice is a piece of use case that represents a unit of work. The states of a use-case slice were described using state cards similar to those in Figure 1. This made the cri-

Figure 3. Alpha abacus.

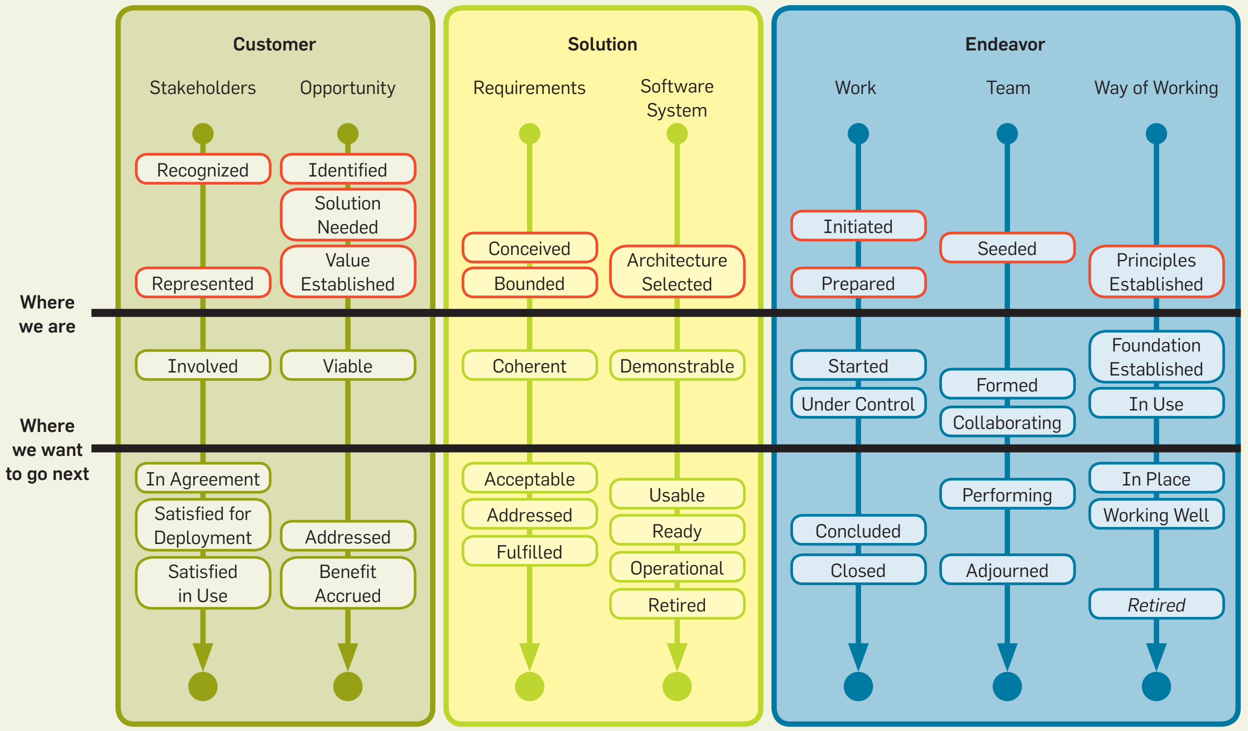
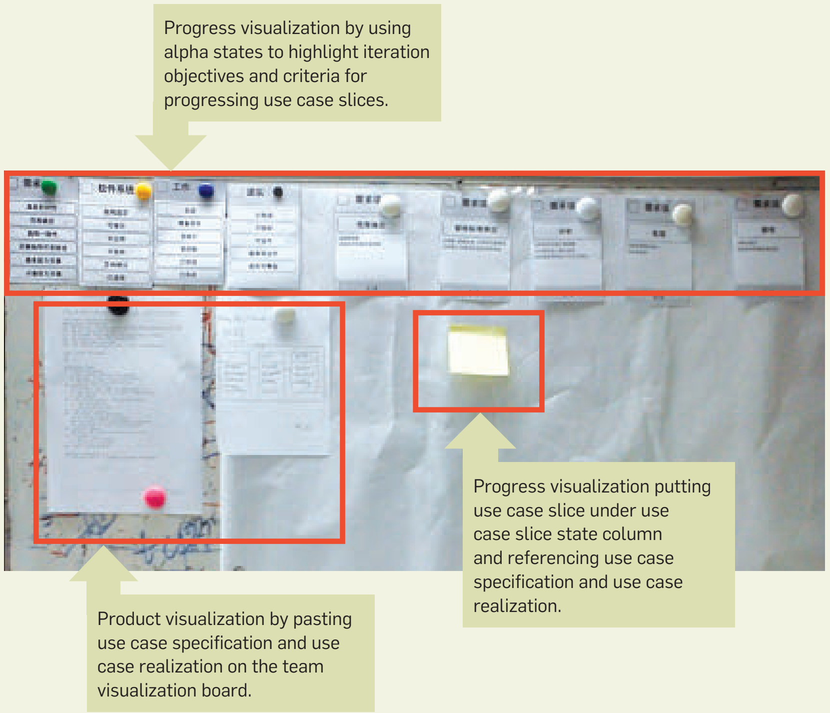


Figure 4. Team visualization board.



teria for achieving each state visible to team members during their daily work.

▸ *Product.* Each team was assigned a use case. The use-case specification and realization, represented by a UML diagram, were pasted on the team board and always represented the current agreement. Changes were scribbled onto the use-case specification and realization. If they became eligible (after some significant changes), someone on the team had to create a clean version.

▸ *Progress.* Post-it notes representing use-case slices were pasted on the board. During daily meetings, members working on the use-case slices would talk about their work in progress by referencing their slices against the requirements and design, as well as the "definition of done" from the process visuals.

Having process, product, and progress visuals readily available not only helped the junior members to understand quickly what they needed to do, but it also helped the team detect any misconceptions quickly.

Improving Collaboration Among Teams

This case study occurred in a Japanese consumer electronics product line of e-book readers. The company had three models, each with different capabilities such as Wi-Fi or 3G access, touchscreen, and so on. It had three product teams (one for each model) and three development teams, as well as an acceptance test team, user experience team, and hardware team. Each team had about four members. Because each team worked separately, coordination was poor, leading to bottlenecks.

We helped the teams make their development work visible through the use of alphas. Specifically, they identified two alphas: a use-case slice and a user-experience element. They defined states for these alphas and the checkpoints for achieving those states. They made the current states visible on a product-line visualization board similar to that in the previous case study, with two exceptions: it had a section for user-experience elements; and it encompassed the entire product line rather than a single team. This was possible because the number of members in each team was relatively small.

Team leaders used the product-line visualization board to plan and discuss progress. With the visualization board, they were able to look ahead and make necessary preparations. In this way each team could make the effort to complete their parts for each integration event, thus eliminating bottlenecks.

Quick Start for Offshore Collaboration

The final case study involved a Japanese company that started a new product line with the help of a Chinese offshore vendor providing development and testing. The product line evolved from an initial eight-person team, with whom we worked primarily, into 50 (local Japanese) plus 200 (offshore Chinese) members. These numbers excluded hardware development and local contractors (working on device drivers) who were an integral part of the overall development. This all occurred in the span of about two years.

This Japanese company had no described way of working, and the Chinese vendor's norm was to follow its client's approach, so there was no starting point. Using Essence, we were able to help the Japanese company describe a way of working that included these practices: iterative development, use-case-driven development, continuous integration, and test-driven development.

The next challenge was determining how to allocate parts of the development to the Chinese vendors. The Japanese company wanted this to be gradual so that as the Chinese members grew in their understanding, they could take on larger responsibilities. The allocation of responsibilities was based on both architecture and process. In terms of architecture, the Chinese vendors could work on the user interface and mid-tier areas, whereas the device drivers and processing closer to hardware specifics remained within the Japanese developers' responsibilities because this required highly specialized skill and the hardware was changing.

In terms of the development process, the alpha states provided a convenient way of discussing responsibilities and involvement. The development process involved several streams of work represented by the alphas. The progress through the requirements alpha states represents the main development. Two other alphas were added to represent work on architecture and acceptance.

In the beginning the Japanese client had primary responsibility over most of the alpha states. As the Chinese vendor grew in knowledge, it assumed greater responsibilities. The alpha states provided a simple means of agreeing on the collaboration. It is important to note that when the Chinese vendor assumed responsibility over one alpha state, it did not mean the Japanese shook off all involvement. The Japanese developers were still involved, but as assistants to the Chinese members.

Using Essence, the Japanese product-line organization could describe their processes, responsibilities, and involvement. It helped the teams get started. It also helped team leaders (both Japanese and Chinese) understand their areas of responsibilities quickly as development grew from eight people to 250.

A Firm Foundation for Sustainable Improvement

SEMAT and agile are two complementary—and perfectly aligned—initiatives. They are both nonprescriptive frameworks that help you think about and improve your software-development capability.

If you are serious about making sustainable improvements in your software-engineering capability, either within your team or across your whole organization, then the combination of Agile and SEMAT offers many benefits above and beyond those gained from either initiative alone. **C**

Related articles on queue.acm.org

The Essence of Software Engineering: The SEMAT Kernel
Ivar Jacobson, Pan-Wei Ng, Paul McMahon, Ian Spence, and Svante Lidman
http://queue.acm.org/detail.cfm?id=2389616

UX Design and Agile: A Natural Fit?
Terry Coatta and Julian Gosper
http://queue.acm.org/detail.cfm?id=1891739

Breaking the Major Release Habit
Damon Poole
http://queue.acm.org/detail.cfm?id=1165768

References

1. Graziotin, D. and Abrahamsson, P. A Web-based modeling tool for the SEMAT Essence theory of software engineering. *J. Open Research Software 1*, 1 (2013), e4; http://dx.doi.org/10.5334/jors.ad.
2. Jacobson, I. and Spence, I. Use case 2.0: Scaling up, scaling out, scaling in for agile projects. Ivar Jacobson International, (2011); http://www.ivarjacobson.com/resource.aspx?id=1225.
3. Jacobson, I., Ng, P-W., McMahon, P., Spence, I. and Lidman, S. The Essence of software engineering: The SEMAT kernel. *ACM Queue* (Oct. 24, 2012); http://queue.acm.org/detail.cfm?id=2389616.

Ivar Jacobson, chairman of Ivar Jacobson International, is a father of components and component architecture, use cases, the Unified Modeling Language, and the Rational Unified Process. A contributor to modern business modeling and aspect-oriented software development, Jacobson is one of the leaders of SEMAT, working to renew software engineering as a rigorous discipline.

Ian Spence is Chief Scientist at Ivar Jacobson International, and has been involved in many large-scale agile adoptions over the years reaching thousands of people. He also led the work of the Essence kernal and has co-authored three software development books.

Pan-Wei Ng is the Asia Pacific CTO at Ivar Jacobson International. He is an advisor and coach to software development organizations of all sizes and has helped teams apply SEMAT ideas since its inception. He also invented the alpha state cards while coaching in Japan.

DOI:10.1145/2500495

Mobile apps manage data on individual residents to help carers deliver more person-centered care.

BY NEIL MAIDEN, SONALI D'SOUZA, SARA JONES, LARS MÜLLER, LUCIA PANNESE, KRISTINE PITTS, MICHAEL PRILLA, KEVIN PUDNEY, MALCOLM ROSE, IAN TURNER, AND KONSTANTINOS ZACHOS

Computing Technologies for Reflective, Creative Care of People with Dementia

DEMENTIA IS A condition related to aging, with symptoms ranging from memory loss to decreased reasoning and communication skills.[8] The number of people with dementia was estimated in 2010 at 35 million worldwide, a figure expected to double by 2030.[23] People with dementia in economies where attaining great age is increasingly the norm are cared for in residential homes by professional (though not highly paid) carers, typically women, often mothers and housewives, under pressure to balance care and administrative duties.[12] Their work is often viewed as having low social status, contributing to high staff turnover and high numbers of inexperienced carers.[24] Improving the quality of care in such an environment is a pressing concern.[19]

In residential homes, digital technology could potentially improve the quality of care, reduce paperwork, and raise the social standing of care work. However, many care homes in the U.K. have at most one or two desktop computers for managing both their finances and their residents' records. Wireless networks are uncommon, and residents themselves only rarely have access to email or social media. Indeed, technology sometimes has been perceived as putting undo pressure on carers.[17] Moreover, managers often lack the skills needed to introduce and advocate for digital technology,[17] although the situation is changing as Internet access and mobile computing become commonplace. Recent initiatives (such as the U.K.'s Get Connected program, http://www.scie.org.uk/workforce/getconnected/index.asp) have further increased the technological readiness of homes. For the first time, a foundation for the use of digital technology in dementia care in residential homes is available.

Person-Centered Care

Digital technology has the potential to support the care sector's move to more person-centered care, or an individualized approach recognizing the uniqueness of each resident, seeking to understand the world from the resident's

» key insights

- **New types of mobile apps can help support the physical care of people in residential care.**
- **Computerized creativity support contributes to person-centered care of people with dementia.**
- **Computerized support helps carers reflect on past work experience in search of new knowledge.**

PHOTOGRAPH BY ROBERT BENTLEY HARRISON, WWW.ROBERTHARRISONPHOTO.COM

perspective and providing a social environment that supports the psychological needs of each resident.[4] The sector recognizes digital technology has the potential to provide the right information about the right resident at the right time, as well as deliver more cost-effective training to new carers.

Digital technology helps people with dementia maintain a sense of self; for example, in 2007, Hanson et al.[9] reported positive results with a digital support service for people with early dementia as long as core usability problems could be resolved. In 2012, Wallace et al.[21] reported the use of digital devices designed as furniture to provide notions of home, intimacy, and possessions to support a sense of personhood. While evidence suggests the technology yields improved communication between carer and resident, it is designed primarily as support for carers delivering the required person-centered care.

Electronic care systems give carers direct access to resident data, though do not necessarily guarantee carers understand the world from an individual resident's point of view. Indeed, a carer can learn something new about a resident by observing highly personal resident behavior, reflecting on it and looking for possible explanations.[18] Carers must often reevaluate the care experience, attend to feelings about it, then generate new perspectives in order to adjust the care to be given. Such behavior is consistent with established general models of reflective learning.[3] Although digital technology can support such event-driven reflective learning, little research has been done, at least until recently.

Since 2010, the European Union-funded Mirror Integrated Project[13] has investigated new forms of digital technology to enable reflective learning about residents (http://www.mirror-project.eu). Here, we report how it aims to bridge the gap between the increasing volume of available digital data about individual residents and the learning carers need to deliver more person-centered care.

The tablet-based Digital Life History app supports collaborative explorations of past life events by carers and residents, as well as subsequent learning about residents by carers through reflections on that collaboration.

Care Homes

The successful uptake of digital technology in residential homes is a challenge; for example, in 2012, Muller et al.[14] reported that parachuting existing technologies into residential homes is unlikely to be effective. New designs must instead be framed in thoughtful socio-technical themes (such as sociality and trust). To discover the themes relevant to dementia care in residential homes, we observed and interviewed carers, then led co-design activities at pilot homes.

The observations and interviews revealed most dementia care is mobile and physical. Carers traditionally use paper documents to record access information about residents. Indeed, these documents are often used as workarounds to obstacles posed by existing digital technology; for example, carers often supplement their observations with written notes on information sheets located next to desktop computers they would use to enter more detailed notes into the electronic care systems later on. However, the resulting delays recording care notes, exacerbated by queuing with other carers to use the computers, can lead to poor recorded data quality in the homes we studied.

Therefore, as part of the co-design activity, we facilitated carers in one pilot home to role-play care activities with various mobile objects. Outcomes revealed mobile computing has the potential to improve several important aspects of care:

Support. Provide more effective support for carers than for residents whose interactions with mobile devices are often limited by physical and cognitive impairment;

Information source. Deliver a single source of information about residents, as carers often struggle to retrieve and communicate information about residents from disparate paper sources;

Distance. Reduce the distance between personalized care and information, as care work is still generally event-driven and frequently interrupted, and carers often need to update resident information at unpredictable times;

Memory. Reduce memory load on carers, as they often rely on memorized information to deliver individual care, while mobile solutions can reduce the amount of information about residents and care tasks they need to remember;

Collaboration. Coordinate collaborative work through shared external representations; social interactions alone are insufficient, while external representations on mobile technology have the potential to enhance coordination of care; and

Information volume. Reducing the amount of information to be memorized during care could free up the cognitive resources needed for reflective learning, while more immediate access to resident data can trigger and enable this reflection.

As part of the Mirror project, we investigated new mobile digital technology for supporting person-centered care and reflective learning about people with dementia. These solutions are available to carers through the dementia App Sphere (http://www.mirror-project.eu/showroom-a-publications/mirror-apps-status), a collection of interoperable mobile apps supporting dementia care and reflective learning in residential homes (see Figure 1).

Mobile Solutions

Some of the apps communicate with a digital life history repository that makes all available data on each resident accessible to carers through mobile devices. The history contains data about a resident's past, as well as data (such as about social contact with carers) recorded and captured directly from the other apps in the App Sphere. A separate bespoke Web application enables relatives to upload information about residents' histories from a range of devices, differentiating the Digital Life History app from its equivalents (such as Digital Life Story from http://mylifesoftware.com, as described in Webster et al.[22]).

During care work, a carer can use two apps on personal mobile devices carried around during a shift. Earlier pilot studies found that, in spite of management concerns, carers were willing to carry and interact with mobile iPod Touch devices to capture and share observations about residents.[11] The first app developed for this purpose (in 2010) was an adapted version of the micro-blogging Yammer app client that provides a single source of information on each resident's behavior, health status, and well being, helping reduce a carer's memory load while enabling more personalized care through reflective learning; it interoperates with a second app called Carer. Carers encountering challenging behavior can invoke it to help provide resolution through creative, reflective thinking.

Figure 1. Architecture of the Mirror dementia App Sphere.

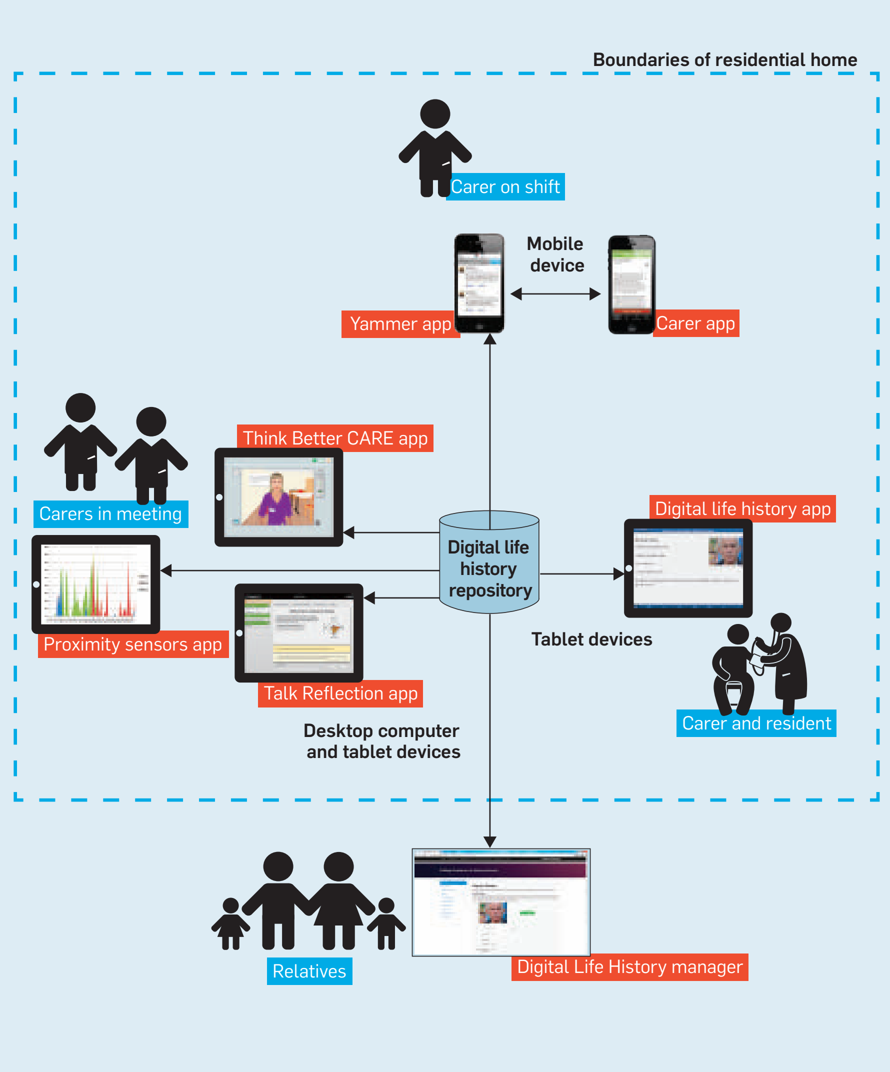


Figure 2. Apps (left to right): Digital Life History, proximity sensor for capturing care data, and Think Better Care.

Figure 3. Features of the Carer app for recording challenging behavior, browsing retrieved cases, viewing creativity prompts, and generating care-plan enhancements.

Use of the Carer app in two residential homes is covered in more detail later.

During team meetings, carers can use other apps to support reflection on resident reminiscing,[6] behavior, and care given; for example, they can reflect on observed behavior through the portable Digital Life History app, the digital equivalent of physical scrapbooks, including photographs and written notes (see Figure 2). While immersing digital content in familiar artifacts (such as televisions,[21] shoeboxes, and time-card devices[2]) have been shown to improve interaction with residents, they do not explicitly support the kind of reflective learning needed by carers to personalize care to individual residents. The tablet-based Digital Life History app supports collaborative explorations of past life events by carers and residents, as well as subsequent learning about residents by carers through reflections on that collaboration.

Carers can use a second app to reflect on their own care patterns based on data from proximity sensors that capture interactions between themselves and residents, as in, say, exploring motivations for unusually long contact sessions with a resident. Using sensors to provide information about dementia care is not new,[5] but our residential care sensor design, inspired by the SocioMetric Badge,[16] is smaller and consumes less power, allowing routine use. Embedded in the wristwatch worn by residents and carers in Figure 2, it broadcasts a unique ID over a 1.5m radius to capture close carer contact with residents.

Carers can also use a third app we developed called Talk Reflection (described later in more detail) to learn through reflection about distressing conversations that can undermine a carer's emotional state.

Less-experienced carers can train on a desktop-based 3D app called Think Better Care to resolve and reflect on dementia-care scenarios in a virtual and hence safe environment. During a typical session, carers receive tutorial guidance from a virtual learning companion called Maria to help reflect on the strengths, weaknesses, and effectiveness of the care given individual residents. The design of the companion is based on Lev Semyonovich Vygotsky's social learning theory.[20]

To address ethical constraints on resident data use, the App Sphere enables privacy through anonymity and encryption of important resident information. In some apps (such as Yammer and Carer), identifiers (such as resident room numbers) known only to carers in a residential home are used to document information about individual residents. In others (such as proximity sensors), the source of resident data is not recorded. Effective reflective learning on individual residents necessitates the recording and sharing of resident information; the residential homes that participated in the co-design of App Sphere emphasized information recording and use over privacy of resident information.

Carer App

Carers regularly encounter challenging behavior from residents, including refusal to eat and take medications, even physical and verbal aggression, which, for carers, is difficult to diagnose and resolve. No two residents are alike, and resolution effective for one is often not effective for another. The resulting need for person-centered care often means an effective resolution is new to both residents and their carers. Carers must exhibit creative, reflective thinking to generate new resolutions. We therefore developed an iOS app called Carer (http://www.mirror-project.eu/showroom-a-publications/mirror-apps-status/86-carer) to support creative thinking based on studies of the effectiveness of such techniques with carers[25] and reflective learning based on new approaches to care.

Carers are able to retrieve resolutions of cases involving challenging behavior in dementia care, as well as in analogical domains (such as policing, schooling, and parenting), automatically retrieving the previous cases from a server-side XML database using one of two services in response to natural language entries typed and/or spoken by carers into the app. One supports case-based reasoning with similar cases based on information-retrieval techniques like those applied to people with a chronic disease.[10] The other supports analogical reasoning with cases from different domains based on a computational model of analogical matching.[7] A third supports the "other worlds" technique, providing less-constrained domains in which to generate ideas, and more generally for resolving the challenging behavior.[1] And a fourth service automatically generates creativity prompts from retrieved case content, letting carers record new ideas resulting from creative thinking in audio form, then reflect on them by playing them back, generating further ideas, composing them into a care plan, and sharing the plan with other carers (see Figure 3).

We evaluated the Carer app in two different unconnected U.K. residential homes we called A and B. We gave carers in each an iPod Touch running the Carer app and Yammer, including seven carers in home A who had them for 28 consecutive days and eight carers in home B for 42 consecutive days. We provided personal Carer training, as well as the creativity techniques it supports. We collected evaluation data from two main sources: a data log implemented in the Carer app that automatically records the date and time each app feature on each device is used, and a focus group with carers in each home at the end of the evaluation period, with audio recorded, transcribed, and analyzed through predefined themes.

All seven carers in home A carried their devices throughout their shifts, using Carer to generate plans for new care enhancements. Each used an average of six to 23 separate app features (such as retrieve past cases and request new creativity prompts) over a seven-day period. Most used two app services: case-based reasoning with previous cases of good practice in dementia care and creativity prompts automatically generated from them to create 10 separate new care enhancement plans for their residents. They used at least one of these 10 plans to increase the quality of life of one resident by reducing the resident's violent outbursts during medical treatment based on the novel idea of having two carers present to provide reassurance. We viewed this as a successful outcome of the initial 28-day trial.

Our analysis of app log data found over 70% of app feature use occurred outside shifts. The focus group found most carers had only the time needed to create and reflect on new care ideas outside their shifts. App use in shifts was also reduced due to a short-lived technical network problem that disproportionately affected carer confidence, as the carers incorrectly took responsibility for app errors when the network was unavailable.

Although the carers did not use the Carer app every day, the frequency of app use should be understood in the context of home A, which did not specialize in dementia care but did care for existing residents who developed dementia. As a consequence, challenging dementia-based behavior was

Figure 4. Features of the Talk Reflection app, including carer self-assessment of emotions during talks and documented conversations.

occasional and not well understood by carers, and the app's guidance led them to resolve unfamiliar challenging behaviors in new ways in the absence of a prescriptive care strategy. Even so, we found obstacles to such creative, reflective thinking. The home's workflow thus had to be redesigned to allow more in-shift time for creativity and reflection, as well as more systematic support outside carer shifts, and carers needed more effective training in creativity techniques and in the mobile technologies being used.

Meanwhile, home B was an acknowledged quality provider specializing in dementia care. Seven of its eight carers carried their devices throughout their shifts, using the Yammer app, but stopped using the Carer app after just nine days of the trial. A reason we identified during the focus group was the app's support for creative thinking to generate new knowledge about residents did not align with the home's strategy of providing specialist dementia care at different stages of the condition based on carers' in-depth knowledge of the characteristics and needs of individual residents.[18] Not only did the app fail to stratify its support along the stages of dementia it did not provide knowledge about individual residents. As a consequence, the carers rejected the app due to lack of perceived benefit, a decision reinforced by a principled stance following the personalization theme pursued at the home. In hindsight, rolling out the Digital Life History app and exchanging data with carers might have yielded greater app acceptance.

Home B's rejection of the Carer app underscored the need for app support for creativity and reflective learning to align with a home's dementia-care strategy, and was not an obstacle to app use in home A. Successful app use also appeared to require more flexible care working practices and training in new techniques and technologies.

Care homes not only support residents with and without dementia differently, different and inconsistent care strategies are implemented across homes.

Talk Reflection App

The tablet-based Talk Reflection app helps carers share and reflect on strategies for holding difficult conversations with residents and for managing their emotional responses to them. It can be used in different work settings to allow carers to document, share, and give feedback on conversations during a shift, after work, or in meetings, online or off. The app is connected to a central repository of sharable documented conversations that can be user-commented with notes and outcomes from prior reflection sessions to facilitate sharing. Conversation assessments are presented in several visual forms (such as the spider graph in Figure 4) to enable quick browsing. Moreover, the app offers simple creativity techniques to carers for discovering new ways of managing difficult conversations. Each documented conversation can also be linked to the relevant resident in the digital life history so other carers are made aware of prior difficult conversations associated with that resident.

We also tested the Talk Reflection app in another U.K. residential home where the manager and five carers used two iPads for 33 days. Data included log files from app use, pre- and post-questionnaires, and feedback given in a workshop. All six used Talk Reflection on 62 separate occasions to document difficult conversations and situations, view a total of 99 documents, create 19 new ones, and comment on existing ones on 20 occasions, though only five such comments were made on documents previously written by another carer. As documentation of difficult conversations and use of tablet technology were new to the carers, we viewed this amount of app use a success. However, it also triggered reflection by carers, though only three reflective outcomes were documented. The workshop found carers usually communicated and discussed the comments verbally during their shifts, as it was quicker and more immediately beneficial. The obvious downside was insufficient numbers of comments of difficult conversations documented in outcomes in the Talk Reflection app.

Carers reported using the Talk Reflection app improved their own well being and handling of difficult situations; one example involved a conversation with relatives requesting a particular resident go to a hospital, contrary to the view of the carers. The deterioration of the resident's condition in the hospital led them to reflect and agree to changes in similar future conversations.

However, the carers also reported technical and organizational obstacles to app use during their shifts; for example, a wireless network was not available in all resident rooms due to the home's concrete walls. The home's manager restricted the times available for app use so it would not intrude on other care duties, even though the carers explicitly asked for more time to use it to discuss difficult conversations. Some residents' relatives even complained about app use; use of such technologies in residential homes is rare, and the relatives erroneously assumed the carers were playing rather than working.

These obstacles again made clear the need to adapt care workflows, this time to mandate the recording of and reflection on difficult conversations and provide sufficient resources both during and outside shifts. As in home A, they highlighted the need to align the app with a home's prevailing care strategy to motivate app use as an integral aspect of care. Care-home managers also need to do more to gain approval for app use. Simple changes being considered include a poster campaign to inform residents of the roles of new technologies and use of white tablet covers with blue crosses to indicate they are exclusively intended for care work.

Conclusion

Our development of new types of mobile apps as part of the Mirror program to support reflective, creative thinking needed by carers delivering person-centered care to older people with dementia contrasts with historical use of digital technologies retrieving information for carers and triggering reminiscing by residents. While prior evaluations revealed the potential of such apps, appreciating it necessitates addressing significant obstacles, mainly the need to align app use with the range of care strategies in residential homes. Homes not only support residents with and without dementia differently, different and inconsistent care strategies are implemented across homes; for example, person-centered care that rejects the disease model (such as explored in Stokes[18]) differs from support for best-practice care themes (such as a positive culture, as in My Home Life Movement[15]). Rolling out even a single app must be sensitive to these differences, and new technologies must be mixed and adapted to different care models. While the App Sphere provides a baseline for mixing technologies, our next step is to configure them for different care models as a starting point for more effective uptake.

Acknowledgment

This research is supported by the European Union-funded Mirror Integrated Project grant 257617, 2010-1. C

References

1. Allan, D., Kingdon, M., Murrin, K., and Rudkin D. *Sticky Wisdom.* Capstone Publishing Company Limited, Chichester, U.K., 2002.
2. Banks, R., Kirk, D., and Sellen A. A design perspective on three technology heirlooms. *Human-Computer Interaction 27*, 1–2 (2012), 63–91.
3. Boud, D., Keogh, R., and Walker, D. Promoting reflection in learning: A model. In *Reflection: Turning Experience Into Learning*, D. Boud, R. Keogh, and D. Walker, Eds. Kogan Page, London, 1985, 18–40.
4. Brooker, D. *Person-centred Dementia Care: Making Services Better, Bradford Dementia Group Good Practice Guides.* Jessica Kingsley Publishers, London and Philadelphia, 2007.
5. Carswell, W., McCullagh, P.J., Augusto, J.C., Martin, S., Mulvenna, M.D., Zheng, H., Wang, H.Y., Wallace, J.G., McSorley, K., Taylor, B., and Jeffers W.P. A review of the role of assistive technology for people with dementia in the hours of darkness. *Technology and Health Care 17*, 4 (July 2009), 281–304.
6. Damianakis, T., Crete-Nishihata, M., Smith, K.L., Baecker, R.M., and Marziali, E. The psychosocial impacts of multimedia biographies on persons with cognitive impairments. *The Gerontologist 50*, 1 (Feb. 2009), 23–35.
7. Falkenhainer, B., Forbus, K.D., and Gentner D. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence 41*, 1 (1989), 1–63.
8. Graham, N. and Warner, J. *Understanding Alzheimer's Disease and Other Dementias.* British Medical Association, Family Doctor Books, Poole, Dorset, U.K., 2009.
9. Hanson E., Magnusson, L., Arvidsson, H., Claesson, A., Keady, J., and Nolan M. Working together with persons with early-stage dementia and their family members to design a user-friendly technology-based support service. *Dementia 6*, 3 (Aug. 2007), 411–434.
10. Houts, P.S., Nezubd, A.M., Magut Nezubd, C., and Bucherc, J.A. The prepared family caregiver: A problem-solving approach to family caregiver education. *Patient Education and Counselling 27*, 1 (Jan. 1996), 63–73.
11. Karlsen, K., Zachos, Z., Maiden, N.A.M., Jones, S, Turner I., Rose, M., and Pudney, K. Supporting reflection and creative thinking by carers of older people with dementia. In *Proceedings of the First International Workshop on Pervasive Care for People with Dementia and Their Carers* (Dublin, Ireland, May 23). IEEE, New York, 2011.
12. Meyer, J. Help the aged. *My Home Life: Quality of Life in Care Homes, A Review of the Literature.* London, 2007; http://myhomelifemovement.org/downloads/mhl_review.pdf
13. Mirror Project. *Reflective Learning at Work.* Saarbrucken, Germany, 2012; http://www.mirror-project.eu
14. Muller, C., Neufeldt, C., Randall, D., and Wulf, V. ICT development in residential care settings: Sensitizing design to the life circumstances of the residents of a care home. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Austin, TX, May 5–10). ACM Press, New York, 2012, 2639–2648.
15. My Home Life Movement. London, 2013; http://myhomelife.org.uk
16. Olguin, D., Gloor, P.A., and Pentland, A. Capturing individual and group behavior with wearable sensors. In *Proceedings of the AAAI Spring Symposium on Human Behavior Modelling* (Palo Alto, CA, Mar. 23). AAAI, Palo Alto, CA, 2009, 68–74.
17. Social Care Institute for Excellence. *E-readiness in the Social Care Sector for SCIE: Final Report.* Social Care Institute for Excellence, London, Jan. 2010; http://www.scie.org.uk/workforce/getconnected/files/ereadinesssurvey2010.pdf
18. Stokes G. *And Still the Music Plays.* Hawker Publications, London, 2008.
19. U.K. Department of Health. *Caring for Our Future: Reforming Care and Support.* Presented to Parliament by the Secretary of State for Health by Command of Her Majesty, July 2012; https://www.gov.uk/government/publications/caring-for-our-future-reforming-care-and-support
20. Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press, Cambridge, MA, 1978.
21. Wallace, J., Thieme, A., Schofield, G., and Oliver, P. Enabling self, intimacy, and a sense of home in dementia: An enquiry into design in a hospital setting. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Austin, TX, May 5–10). ACM Press, New York, 2012, 2629–2638.
22. Webster, G., Fels, D.I., Gowans, G., and Hanson, V.L. Portrait of individuals with dementia: Views of care managers. In *Proceedings of the 25th British Computer Society Conference on Human-Computer Interaction* (Newcastle, U.K., July 4–8). British Computer Society, Swindon, U.K., 2011, 331–340.
23. Wimo, A. and Prince M. *World Alzheimer Report 2010: The Global Economic Impact of Dementia.* Alzheimer's Disease International, London, 2010; http://www.alz.co.uk/research/worldreport/
24. Wright, J. et al. All-Party Parliamentary Group on Dementia. *Prepared to Care: Challenging the Dementia Skills Gap.* U.K. Government, London, June 2009; http://alzheimers.org.uk/site/scripts/download_info.php?fileID=735
25. Zachos, K., Maiden, N.A.M., Pitts, K., Jones, S., Turner, I., Rose, M., Pudney, K., and MacManus, J. A software app to support creativity in dementia care. In *Proceedings of the Ninth ACM Creativity and Cognition Conference* (Sydney, June 17–20). ACM Press, New York, 2013, 124–131.

Neil Maiden (N.A.M.Maiden@city.ac.uk) is a professor of systems engineering and co-founder of the Centre for Creativity in Professional Practice at City University London.

Sonali D'Souza (sonali.dsouza5@gmail.com) is a principal information architect/UX designer at the British Council, London.

Sara Jones (s.v.jones@city.ac.uk) is a senior lecturer in the School of Informatics and co-founder of the Centre for Creativity in Professional Practice at City University London.

Lars Müller (lmueller@fzi.de) is a researcher at the FZI Research Center for Information Technology in Karlsruhe, Germany.

Lucia Pannese (lucia.pannese@i-maginary.it) is managing director of Imaginary Società a Responsabilità Limitata, an Italian enterprise specializing in serious games, virtual worlds, and interactive digital technologies.

Kristine Pitts (Kristine.Pitts.1@city.ac.uk) is a research fellow in the Centre for Creativity in Professional Practice at City University London.

Michael Prilla (michael.prilla@rub.de) is a senior researcher in computer-supported cooperative work and human-computer interaction in the Information and Technology Management work group of the Institute for Work Science at the Ruhr University of Bochum, Germany.

Kevin Pudney (kevin.pudney@ntlworld.com) is a social science researcher and former head of IT of the Nottinghamshire Social Services Department.

Malcolm Rose (mal@clara.net) is an economist and former director of regeneration and head of policy in District and County Councils, Nottinghamshire, U.K.

Ian Turner (ian.turner@virgin.net) is chairman of the U.K.'s Registered Nursing Home Association.

Konstantinos Zachos (kzachos@soi.city.ac.uk) is a research fellow in the Centre for Creativity in Professional Practice at City University London.

DOI:10.1145/2500494

Tablets offer hope for improving learning and collaboration but only if truly integrated into learning settings.

BY EVGENY KAGANER, GABRIEL A. GIORDANO, SEBASTIEN BRION, AND MARCO TORTORIELLO

Media Tablets for Mobile Learning

AS CONSUMERS CONTINUE to shift everyday activities onto personal mobile devices, organizations seek to provide similar capabilities for their employees. Introduction of the iPad in April 2010 and ensuing explosion of the worldwide media tablet market was yet another impetus to the rising importance of mobility in the enterprise. Tablets offer a sweet spot for mobile workers looking for media, collaboration, and basic personal productivity capabilities on the go. Forrester Research expects almost one-third of tablets to be sold directly to businesses by 2016.[7]

Enterprise initiatives involving media tablets span multiple contexts: as notepads for executives; showcase and presentation tools for sales professionals; decision-support devices for technical staff in the field; and education support for organization training and development. Mobile learning, a strategy to facilitate, enhance, and extend teaching and learning through mobile devices, is often touted as a way to address the challenges of the world's increasingly global, mobile, technologically savvy work force.[9]

Recent industry reports, drawing on anecdotal evidence and casual case studies, point to anticipated benefits from tablet-enabled mobile-learning initiatives, including making learning more personalized and continuous,[12] increasing its relevance by providing timely access to context-specific information,[11] and helping learners maintain a sense of community and collaboration.[3] They also identify potential drawbacks involving mainly the technological limitations (such as an inadequate typing experience for input-intensive tasks) of today's tablets and broader implementation concerns (such as data security and cost).[5] In general, however, there is common sentiment among industry analysts that tablets will improve the overall learning experience once key technical and organizational wrinkles are worked out.

A significant number of mobile-learning studies have been carried out in higher education, including several set up as research studies aimed at understanding the effects of mobile tablets on learning, tablet usability and adoption patterns, and best practices for tablet deployment and implementation (see Table 1).

Besides generating insight, their applicability beyond higher education is limited. In particular, all have focused on full-time undergraduate and graduate students whose learning needs could differ significantly from

» key insights

- **Learners are comfortable using tablet solutions but struggle incorporating them into the way they learn and collaborate.**
- **Success with mobile-learning initiatives requires a close fit between course design and the mobile functionality being offered.**
- **Mobile learning initiatives should be viewed as long term; expected benefits may not come quickly, and management must guide and support learners through the process of evolving their practices.**

ILLUSTRATION BY ALEKSEI GLUSHCHENKO

Table 1. Main research-based iPad studies in higher education.

Site	Participants	Design/Data Sources	Key Findings*
University of Minnesota	450 freshmen in College of Education and Human Development.	No control group; No custom apps; Data via survey, focus group.	**LI.** iPad increased media production; personal productivity; helped extend learning beyond classroom;
Pepperdine University	Undergraduate and full-time graduate students in business, health, law, math, and religion.	Control group in each course; No custom apps; Data via surveys, focus groups, interviews, participant observations, course grades.	**IS.** Success depends on assuring fit between teaching style and app, creating iPad-friendly materials, integrating iPad into course. **LI:** >60% felt iPad use not directly beneficial to learning; iPad effect on grades (quiz scores) mixed.
Reed College	Approximately 40 undergraduate students in political science.	No control group; No custom apps; Data via surveys, email feedback, group interview.	**LI.** Economic effect of iPad depends on availability and cost of e-content. **UP.** Unlike Kindle, iPad involves advantages with respect to legibility, touchscreen, form factor, battery life, durability, paper savings, multipurpose use, referring to text in class, annotation; iPad limitations include .pdf handling, file system, and keyboard.
Oklahoma State University	123 undergraduate students in communications and marketing.	No control group; No custom apps; Data via participant observation, surveys.	**IS.** To assure success, iPad must truly be integrated into course; iPad increased perceived course workload; integration of iPad in learning requires extra effort. **LI.** 75% of students believed academic experience improved; iPad increased student reading, depth of class discussion, early assignment submission.
Indiana University	A faculty learning community used iPads in several courses.	No control group; No custom apps.	**IS.** Successful iPad use requires faculty change course material. **LI.** iPad increased student collaboration between students and faculty.
University of Notre Dame	71 undergraduate and seven graduate students in business and law.	No control group; No custom apps; Data via surveys, interviews.	**IS.** Students felt more technical courses might be less suitable for integration with iPad. **LI.** Majority of students believed they were learning more with iPad; Perception of usefulness was steady over time; Undergrads felt iPad was more useful than graduate students. **UP.** >70% of students never or almost never printed course materials.

* IS = implementation success; LI = learning impact; UP = use patterns; see http://community.pepperdine.edu/it/tools/ipad/research/similarstudies.htm

those of working professionals. Moreover, no study we know of from higher education has involved mobile apps custom-built for a particular learning environment, an approach common in corporate learning. Finally, no study, except one at Pepperdine University, incorporated a control group to ensure more robust analysis of iPad effects.

The research we report here aims to overcome these limitations and give technology and learning experts a more nuanced understanding of how media tablets might influence key dimensions of learning and collaboration by working professionals.

Data Collection Methodology

We conducted surveys in three rounds:

Survey round 1 (both sections). **Before the term began;**

Survey round 2 (both sections). **12 weeks into the term; and**

Survey round 3 (iPad section only). **16 weeks into the term (end of term).**

For the surveys, we drew all questionnaire items from previously established scales; for example, for learning we used four items about how recent cases promoted new knowledge and skill development.[14] We used six items from the same study to measure students' perceived satisfaction with iPads. We also used the scales to measure perceived team collaboration effectiveness,[1] team coordination effectiveness,[6] team cohesion,[2,10] team commitment,[13] and team conflict.[8]

We also conducted interviews in three rounds:

Interviews round 1 (iPad section only). **Two weeks into the term;**

Interviews round 2 (iPad section only). **10 weeks into the term; and**

Interviews round 3 (with iPad professors). **At end of the term.**

The Study

Our 2011 study was based on a three-month field experiment with 124 students enrolled in an executive-MBA (EMBA) program at IESE Business School, a top-ranked international business school with campuses in Barcelona, Madrid, Munich, New York, and Sao Paulo. We conducted the experiment at the school's main campus in Barcelona, Spain. Participants (average age 32, with 7.6 years of professional experience) were from a variety of educational backgrounds and worked in a range of industries.

The EMBA program is based on the case-study method, where learning takes place through discussion and students are required to prepare case analyses before class. As most students have full-time jobs, classes are held on weekends. Teamwork and collaboration are key parts of the learning process, with students assigned to small teams (roughly six students each) to prepare the cases, as well as course projects. Hour-long team meetings are held mornings before class; additionally, most teams employed a combination of email, videoconferencing, and face-to-face meetings during the workweek to facilitate collaboration. The setting and participant profiles in the study were thus like many organizational environments, especially those with semidistributed collaborative teams where mobile devices for supporting learning might be considered.

A separate 2011 study commissioned by the International University Consortium for Executive Education (UNICON, http://uniconexed.org/) identified five main use areas where mobile technologies could enhance learning: digital content, support and coordination, collaboration, assessment, and gaming/simulations.[12] Accounting for the nature of the learning environment in the EMBA program (emphasizing cases and teamwork), our study focused on digital content and collaboration. To examine how mobile learning affected them, we distributed iPads to 62 students in a randomly selected section of the EMBA program for one academic term and also distributed cases, readings, and preparation questions for all courses in digital format through the iPads. These were first-generation iPads with a prepaid 3G connection and preinstalled software suite, including a custom-developed application for digital-content delivery and review (the Case app),[a] a social-collaboration tool, general productivity apps (such as a preconfigured iOS mail client), and an office-productivity suite (iWork). Not included were printed course materials, though we did allow students to print any of the digital materials provided through the Case app.

Students in the other section (62 students) did not receive iPads but were instead given binders with printed versions of the cases and course materials (typically several thousand pages). This non-iPad section served as the study's control group.

To collect data, we conducted three rounds of surveys in both sections, as well as multiple rounds of interviews with the iPad group and participating professors. Our focus was on understanding the adoption and usage patterns of the mobile learning solution and their influence on key dimensions of learning: outcomes, communication/collaboration, and team dynamics (see Figure 1 and the sidebar "Data-Collection Methodology").

Key Findings

Digital content. The interviews showed us that the students perceived greater accessibility and portability of digital content among the main benefits of participating in the study. Being able to carry around all course materials and download new cases on demand was especially appreciated. Also appreciated was more efficient browsing and searching within documents, especially during team meetings and class discussions. One student said, "By browsing annotations, I can quickly and easily find the part of the case concern-

Figure 1. Study design.

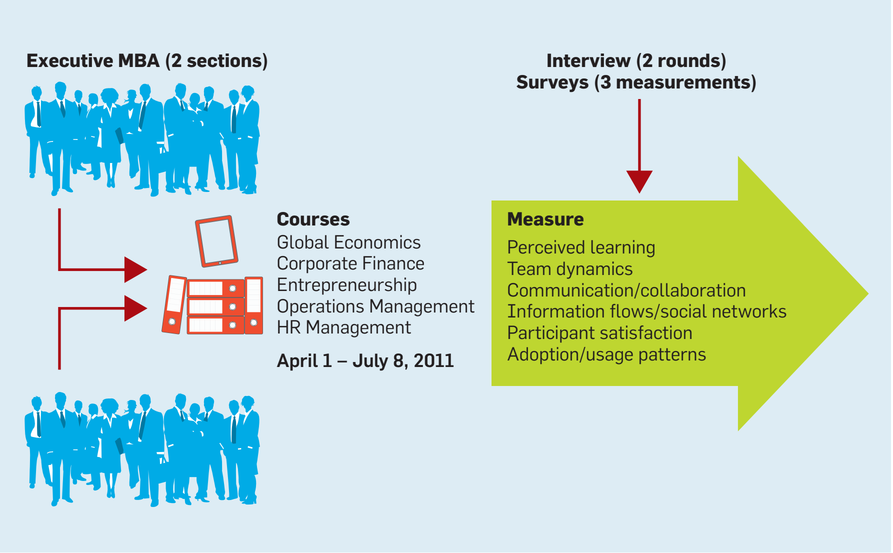
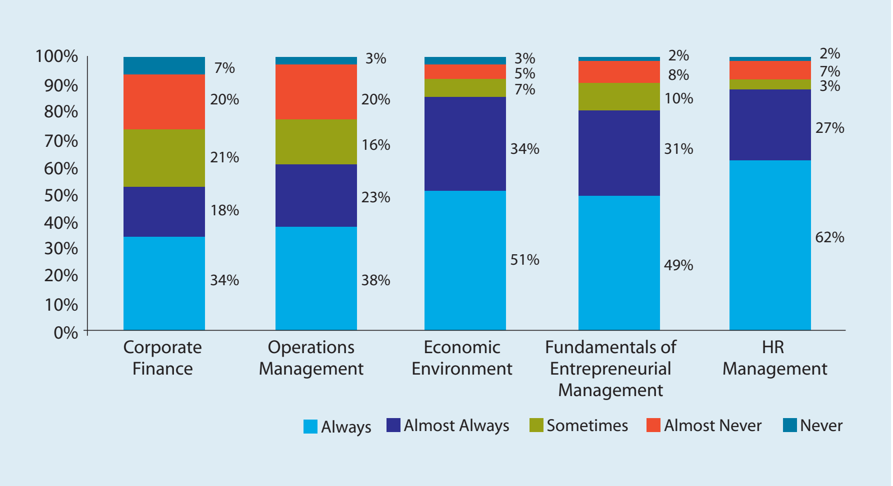


Figure 2. For each of the following courses, how often did you readcases and other assigned materials on the iPad?



a The IESE Case app, built on top of the iAnnotate software library, is used for managing, viewing, reading, and annotating cases and other course materials in .pdf format. Features include: navigate and search through cases (by schedule, name, code, subject, and professor name); annotate with notes, highlighting, and free drawing; select and share annotations with other users of the app; and update and download the latest cases and course materials (can be synced with iTunes); see demonstration video at http://podcasts.iese.edu/website/flv/insight/REPV-45-E.mp4. The app was custom-built for the iPad study by an external provider and pretested over several weeks by a team of eight EMBA students from the second year. We incorporated feedback from the pretest into the app before deploying it to study participants.

ing the current discussion. I can look up numbers or facts and plug those into the discussion on the fly. This is very powerful and something I could not do before with the paper cases."

We found adoption and usage behavior was generally consistent with this observation. Depending on the course, 52% to 89% of the students in the iPad section always or almost always read cases and other assigned materials on an iPad. Moreover, they relied on an iPad more for reading materials in more qualitative courses (such as entrepreneurship and human-resources, or HR, management) and less in more quantitative courses (such as finance and operations) (see Figure 2). Our follow-up interviews showed this occurred due to limitations of iPad hardware and the .pdf document format; another student said, "Cases that include many exhibits and appendixes with numbers (such as those in corporate finance, for instance) are more difficult to work with on the iPad. You have to constantly go back and forth between the text and the tables, which is a bit of a pain."

We observed a reverse pattern with respect to printing. Students printed the case documents more often in quantitative than in qualitative courses (see Figure 3). In addition, 2% to 20% of the participants (depending on course) reported reading materials on the iPad while also printing them, with overlap greatest in HR management and economics. When we probed the reasons for printing, participants cited (in descending order of importance) the need to see multiple pages at a time, a preference for reading and taking notes on paper, and the desire to back up digital documents/notes. This finding is consistent with the University of Notre Dame's 2010 mobile computing study in which participating students viewed e-content offered through tablets as less viable in more technical courses.

Communication and collaboration. Whereas digital-content features won significant adoption among the iPad-using students, collaboration-related features went largely unused. With anytime/anywhere mobile Internet connectivity and a user-friendly online collaboration platform, we gave the students a chance to decrease the frequency of their face-to-face meetings by shifting a portion of it online. With busy schedules and plenty of travel during the week, they had a strong incentive to take the opportunity. Despite the fact that seven of the eight teams established workspaces online, none used the platform regularly. Likewise, the "share annotations" feature of the Case app (viewed by Case app developers as a major improvement over paper-based cases) went largely unused, with 75% of the participants never or almost never using it.

One student said, "We just didn't know how to make productive use of this feature. In the beginning our team tried. But everybody was highlighting so much, and people tended to focus on different things; so in the end, once you put it all together it really was not helpful."

Another intriguing, if counterintuitive, finding concerns student satisfaction with team communication and collaboration practices, as well as with the perceived contribution of these practices to course quality, learning, and team coordination. We found a consistent pattern in which the control group showed significant improvement on all such dimensions, from the time of the first survey (pre-iPad) to the second survey (post-iPad), whereas the iPad group's performance was unchanged.

Consistent with differences in team communication and collaboration, survey data also revealed differences in how members of the two sections interacted with one another. For example, we asked them to identify class-

Figure 3. For each of the following courses, how often did you print cases and other assigned materials?

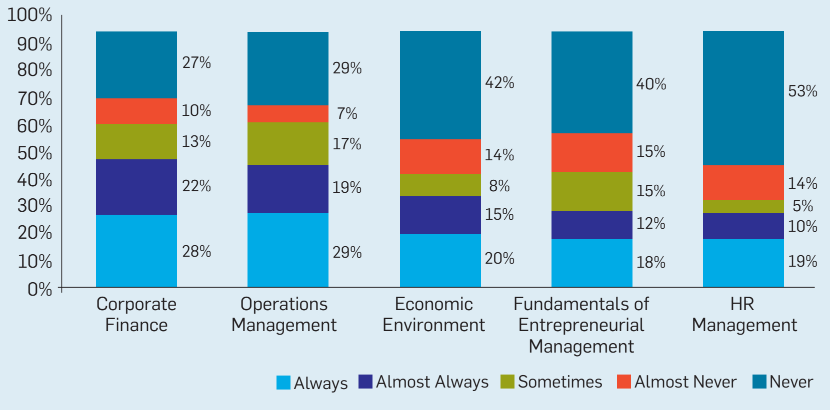


Table 2. Effects of mobile learning solutions on the learning experience.

	Change from Survey 1 to Survey 2	
Dimension of learning experience	**iPad**	**Control**
Team dynamics		
Team cohesion	+	+
Team satisfaction	0	+
Team commitment	0	+
Team conflict	0	0
Team communication/collaboration		
Collaboration effectiveness	0	+
Coordination effectiveness	0	+
Satisfaction with team communication and collaboration	0	+
Number of interactions with classmates in the section	–	0
Learning		
Satisfaction with cases' contribution to learning	–	+

+ = statistically significant improvement; – = statistically significant deterioration; 0 = no change mates with whom they interacted most frequently to discuss topics and materials seen in class, with respondents able to choose any number of classmates from a complete list of section members. Surprisingly, the number of classmates chosen by the iPad group decreased from the first to the second survey, while the number for the control group was the same.

Team dynamics. Introduction of the mobile learning solution also had a mixed effect on team dynamics. We collected survey data about students' perception of team conflict, cohesion, satisfaction, and commitment. Concerning team conflict, neither group showed significant differences, though they did differ in terms of team satisfaction and team commitment. Here, the two groups reflected a pattern similar to that for team communication and collaboration; that is, the control group showed significant improvement between the first and the second surveys, while the iPad group failed to show improvement.

We also saw general improvement in team cohesion over time in both sections, a finding consistent with what is typically observed in the EMBA program. By the end of the first year, teams tended to function more effectively, as students spent time together, completing a number of group assignments and projects. The mobile learning solution did not seem to change this dynamic.

Learning. Surprisingly, we found a negative iPad effect on how students perceived the outcomes of their learning experience. The iPad group's satisfaction with the contribution of cases to learning decreased during the pilot; that is, the students came to believe the cases contributed less to helping them understand course topics and create new knowledge. On the other hand, students in the control group reported increased satisfaction with case-based learning. This is an important, if somewhat surprising, finding, as one of the main objectives of the study was to enhance students' case-based learning through digital content; Table 2 outlines the study's main insights with respect to how the mobile learning solution affected the students' learning experience. The general pattern was the control group showed significant improvement on a number of learning and collaboration dimensions, whereas the iPad group showed no improvement.

Being able to carry around all course materials and download new cases on demand was especially appreciated.

Implications

Important implications apply not only to educational institutions but to a much broader range of organizations looking into the training and development needs of their increasingly mobile workers.

No quick payoff. Despite the overall satisfaction and general enthusiasm shown by study participants, the mobile learning solution seemed to have a negative effect on several important dimensions of the learning experience. Team communication and collaboration were affected, as was student satisfaction with case-based learning. This should raise a flag for organizations launching mobile-learning initiatives, even while it is important to understand why these effects take place. Our data and interviews with study participants showed a number of factors were in play. First, the iPads caused disruption in personal work habits and teamwork practices by shifting student learning curves. The students seemed to no longer focus on how to work as a team and analyze cases but were forced to learn to accomplish these goals through technology that some likely never encountered before; that is, the students had to cope with the pressure of maintaining their academic performance while creating new learning and collaboration practices to incorporate the iPads. In many cases, the result was frustration and heightened tension within the teams. The significant learning curve tablets bring to a learning setting was also recognized by other studies, including those at Pepperdine University and at Indiana University.

The tension was further exacerbated by the fact that, at least in some iPad-using teams, different team members opted for different ways of working with cases and course materials. Some relied exclusively on the iPad; others exported the cases out of the case application, reading them on PCs and laptops; and the third group would print cases and use them on paper. As a result, multiple use practices coexisted within the same team, adding complexity to sharing materials and coordinating work among team members. Added complexity contributed to decreased student satisfaction with collaboration and learning. This disruption mechanism was not reflected in any of the previous studies yet will play an important role, as many organizational settings are built around small self-led teams.

Note, too, our findings regarding deterioration in learning satisfaction and team collaboration are not necessarily consistent with the earlier studies; for example, studies at the University of Notre Dame and at Oklahoma State University found iPad use improved student perception of the learning experience in undergraduate settings. This difference might have been due to the team-based nature of the learning environment in our study, leading to additional tension among students. Another possible explanation reflects differences between undergraduate students and the working professionals who participated. The former were digital natives whose main focus was education; the latter were digital immigrants balancing professional and family responsibilities with academic activities.

While alarming, the initial deterioration in performance should not be viewed as a trigger to discontinue or scale down mobile learning initiatives. The main causes of deterioration—disruption to individual habits and work practices and diversity of user behavior within a team—are short term and can be managed proactively. Organizations should shift their mind-sets and view mobile learning initiatives as long term. The expected benefits may not accrue right away, and learners may need significant help to refine their habits and practices. Management is responsible for guiding and supporting them through this journey.

Support and education. When it comes to support and education in technology implementations, the focus is too often on troubleshooting technical difficulties and helping individual users master functional features of the solution. Relying on this approach in the context of mobile learning is risky; for example, learners do not struggle with how to use the solution (tablets support an intuitive user experience) but rather how to incorporate it into the way they learn and collaborate. Insofar as learning and collaboration practices are deeply ingrained in one's mind, this challenge is much greater than retraining employees to switch from an old workflow to a new one. The most effective approach to creating a winning mobile learning strategy is to understand what can promote such a change in practice.

Perhaps mobile learning will also help organizations finally bridge the gap between learning in the classroom and learning in the workplace.

In our study, use cases with immediate and visible benefits (such as digital content) enjoyed significant adoption among study participants without much effort from the project team. Being able to readily access, annotate, and search digital cases and course materials allowed them to be more productive quickly. To enjoy this visible increase in productivity, most study participants willingly overlooked the wrinkles of the technological platform and adjusted their reading habits. The other iPad studies also found significant reading benefits from tablet use, with two—Oklahoma State University and the University of Notre Dame—finding, at times, tablets increased the amount of reading done by undergraduate students. In our study, use cases offering attractive but less immediate benefits (such as collaboration) were generally overlooked. The students in the iPad section were intrigued by the possibility of moving part of their team collaboration online but never followed through. Yet, it is this second category of use cases—collaboration support—that often promises the most dramatic long-term improvement in the learning process.

Management can employ various mechanisms to facilitate and motivate desired user behavior on the part of learners, as in, say, the "quick win" use cases, helping them be more productive quickly and build up "good will" that can be leveraged to push for broader adoption of "difficult" use cases (such as collaboration). In our study, enthusiasm for digital content delivery led study participants to express openness toward experimenting with Socialcast (http://www.socialcast.com), a social-collaboration platform, in their teamwork.

Project team management can also encourage adoption of "difficult" use cases by embedding them into learners' everyday activities; for example, in order to expose the students to Socialcast, first-line technical support during the study was offered exclusively through that platform. In our final round of interviews, many participants expressed satisfaction with Socialcast, suggesting they would, given more time, use it to facilitate teamwork.

Another reason the traditional approach to training, focusing on the individual user, might not be effective in mobile learning involves the team nature of this learning environment. As discussed, the clash of use practices within iPad-section teams made task coordination and knowledge sharing more difficult. One way to address it would be to focus on educating the team rather than just individual users. This can be done through, say, documenting best practices and use cases discovered by the learners, proactively disseminating it back to the learner teams. It could also be helpful to leverage competitive dynamics among teams by measuring their ongoing performance on key learning indicators, making it transparent to the broader learner community, along with the use practices adopted by each team.

Redesign. One main conclusion of two earlier studies—at Oklahoma State University and at Pepperdine University—was that achieving success in mobile-learning initiatives requires a close fit between course design and the mobile functionality being offered, a conclusion supported by our results. Moreover, organizations seeking to shift learning and collaboration onto mobile platforms must not only integrate the device into the course but revisit key assumptions behind the design of the learning process and content. Replicating the traditional "analog" experience in a mobile solution, even if integrated in a traditional course design, might work as a first step, but longer term, such replication will fail to take full advantage of the opportunities due to mobility or live up to learners' expectations. In our study, participants quickly caught on to the limitations of using the old .pdf format on the iPad. One student said, "Now that we have the capabilities, it would be nice to see more interactive features built into the cases (such as short videos or interactive graphs). In the future, what would be really great is to ability to 'ask' the main characters of the case follow-up questions. That's how it works in real life…"

Other previous studies—most notably those at Indiana University and at Oklahoma State University—also found that having new types of content (such as video and interactive) was important in engaging students with tablet-based learning materials.

Post-PC mobility lets organizations rethink how learning and collaboration take place. Producing new types of engaging content, delivering it when and where learners need it most, and embedding rich social interaction directly into learning scenarios are just a few approaches to consider. Personal mobile devices are becoming universal hubs that integrate the everyday activities of individuals, cutting across traditional divides of personal vs. professional and online vs. offline. Perhaps mobile learning will also help organizations finally bridge the gap between learning in the classroom and learning in the workplace.

Conclusion

Market penetration of post-PC devices (such as smartphones and tablets) is expected to dwarf traditional desktop and laptop computers by a factor of four by 2016,[4] transforming consumer behavior and expectations and reshaping industries and business models along the way. Education and workplace learning will be affected, and educational institutions and corporate training organizations alike must proactively prepare for the upcoming shift. As we found, the road to a sustainable mobile learning environment will likely involve many challenges and disappointments; learning outcomes might not improve immediately, team dynamics might suffer, and learners might simply refuse to adopt critical functionality. Organizations must take a long-term view, developing a roadmap rooted in their learning settings and objectives. That roadmap should account for the cultural change learners must cope with and detail support mechanisms to facilitate the transition for both individuals and teams. It should also include a fundamental rethinking of the design principles behind learning content and process to take full advantage of post-PC mobility. Organizations that manage to do so will be the places where people strive to work and study, redefining the idea of personal development. C

References

1. Alavi, M. Computer-mediated collaborative learning: An empirical evaluation. *MIS Quarterly 18*, 2 (June 1994), 159–174.
2. Barry, B. and Stewart, G.L. Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied Psychology 82*, 1 (Feb. 1997), 62–78.
3. Bonig, R. *How iPads, Media Tablets and other Mobile Devices Challenge Higher Education CIOs.* Gartner Report, Stamford, CT, June 2011.
4. Business Insider Intelligence. *The Mobile Industry: In-Depth* (Mar. 2012); https://intelligence.businessinsider.com
5. Chess Media Group. *Future Workplace Survey*, 2011; http://www.thefutureworkplace.com
6. Galegher, J. and Karaut, R. Computer-mediated communication for intellectual teamwork: An experiment in group writing. *Information Systems Research 5*, 2 (June 1994), 110–138.
7. Gillett, F.E. *Tablets Will Rule the Future Personal Computing Landscape.* Forrester Research. Cambridge, MA, Apr. 23, 2012; http://www.forrester.com/Tablets+Will+Rule+The+Future+Personal+Computing+Landscape/fulltext/-/E-RES71581
8. Jehn, K.A. and Mannix, E.A. The dynamic nature of conflict: A longitudinal study of intragroup conflict and group performance. *Academy of Management Journal 44*, 2 (Apr. 2001), 238–251.
9. Meister, J., Kaganer, E., and Von Feldt, R. The year of the media tablet as a learning tool. *T+D [Training and Development] Magazine* (Apr. 2011).
10. Rosenfeld, L.B. and Gilbert, J.R. The measurement of cohesion and its relationship to dimensions of self-disclosure in classroom settings. *Small Group Behavior 20*, 3 (Aug. 1986), 291–301.
11. Schooley, C. *Time to Get Serious About Mobile Learning.* Forrester Report, Forrester Research, Cambridge, MA, Sept. 2010; http://www.forrester.com/Time+To+Get+Serious+About+Mobile+Learning/fulltext/-/E-RES57576
12. UNICON. *Going Mobile in Executive Education: How Mobile Technologies Are Changing the Executive Learning Landscape.* Research Report, London, Nov. 2011; http://www.ashridge.org.uk/website/content.nsf/FileLibrary/C67FAE5265440F728025798000404608/$file/ABS_MobileLearning.pdf
13. Van Der Vegt, G., Emans, B., and Van De Vliert, E. Team members' affective responses to patterns of intragroup interdependence and job complexity. *Journal of Management 26*, 4 (2000), 633–655.
14. Yeo, S., Taylor, P., and Kulski, M. Internationalizing a learning environment instrument for evaluating transnational online university courses. *Learning Environments Research 9*, 2 (2006), 179–194.

Evgeny Kaganer (ekaganer@iese.edu) is an associate professor of information systems at IESE Business School, Barcelona, Spain.

Gabriel A. Giordano (giordano@ohio.edu) is an assistant professor of management information systems at the College of Business of Ohio University, Athens, OH.

Sebastien Brion (sbrion@iese.edu) is an assistant professor of managing people in organizations at IESE Business School, Barcelona, Spain.

Marco Tortoriello (mtortoriello@iese.edu) is an associate professor of strategic management at IESE Business School, Barcelona, Spain.

DOI:10.1145/2500117

The competitive nature of AT, the scarcity of expertise, and the vast profits potential, makes for a secretive community where implementation details are difficult to find.

BY PHILIP TRELEAVEN, MICHAL GALAS, AND VIDHI LALCHAND

Algorithmic Trading Review

Stock market data, as visualized by artist Marius Watz, using a program he created to represent the fast-paced "flows" of data as virtual landscapes.

STOCKSPACE (2009), MARIUS WATZ. CUSTOM SOFTWARE ART BASED ON STOCK MARKET DATA. COMMISSIONED BY KNIGHT CAPITAL GROUP.

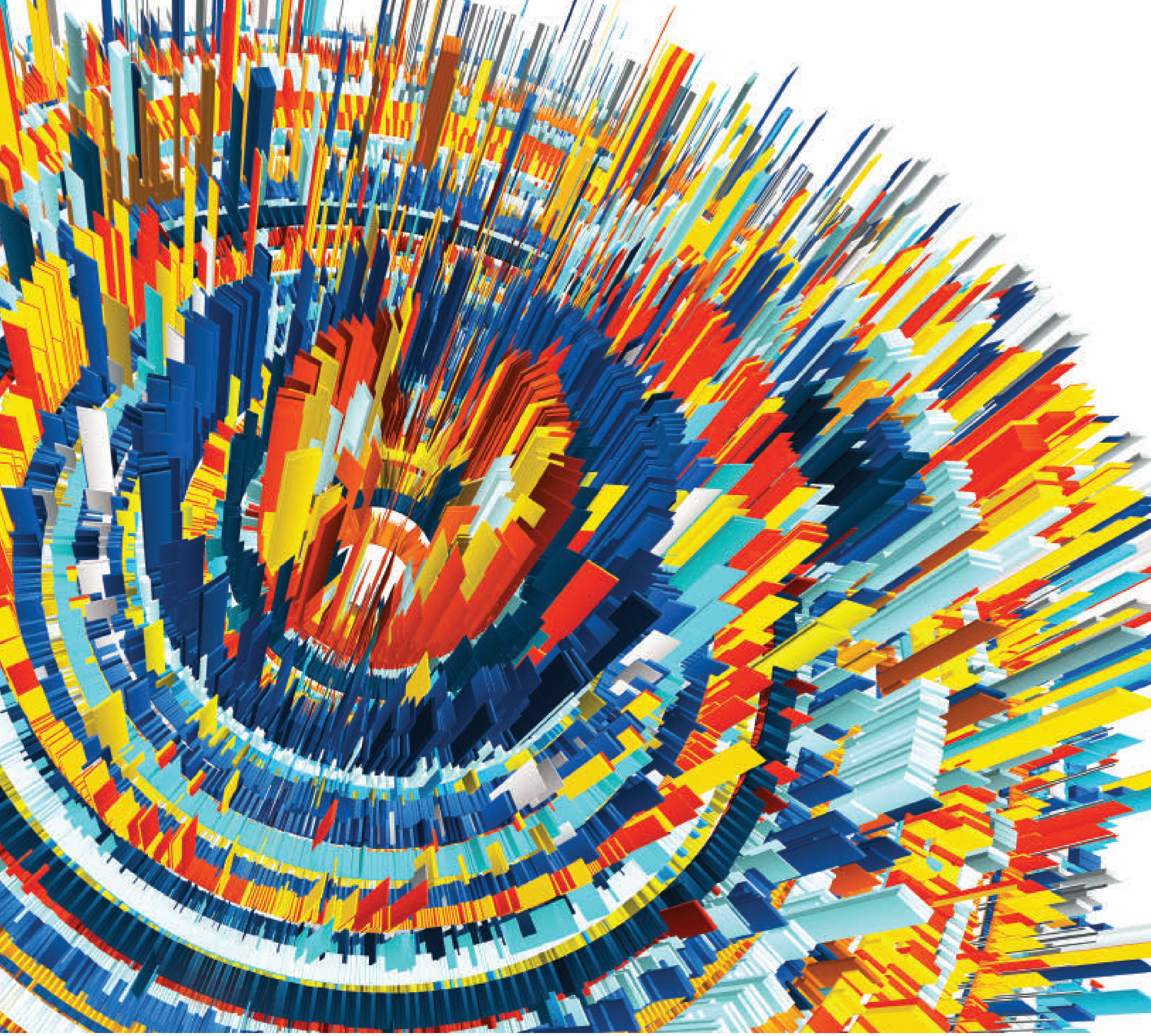
ALGORITHMIC TRADING[1,9,13,14] is growing rapidly across all types of financial instruments, accounting for over 73% of U.S. equity volumes in 2011 (Reuters and Bloomberg). This has been a fascinating research area at University College London, where for eight years algorithmic trading systems and an Algorithmic Trading/Risk platform[7] have been developed with leading investment banks/funds.

To tell this story, first we must clarify a number of closely related computational trading terms that are often used interchangeably:

- **Algorithmic trading** (AT) refers to any form of trading using sophisticated algorithms (programmed systems) to automate all or some part of the trade cycle. AT usually involves learning, dynamic planning, reasoning, and decision taking.
- **Systematic trading** refers to any trading strategy that is a "rule-based" systematic/repetitive approach to execution trading behaviors. This is often achieved

» key insights

- Algorithmic Trading (AT) refers to any form of trading using sophisticated algorithms (programmed systems) to automate all or some part of the trade cycle. AT is data-driven and usually involves learning, dynamic planning, reasoning, and decision taking.
- The key stages in AT are the pre-trade anaylsis, signal generation, trade execution, post-trade analysis, risk management, and asset allocation.
- Developers use various types of simulations, including backtests and optimizations, to evaluate and improve utility of their algorithms.



through utilization of an expert system that replicates previously captured actions of real traders.

▸ **High-frequency trading (HFT)** is a more specific area, where the execution of computerized trading strategies is characterized by extremely short position-holding periods in excess of a few seconds or milliseconds.

▸ **Ultra high-frequency trading** or low-latency trading refers to HFT execution of trades in sub-millisecond times through co-location of servers and stripped down strategies, direct market access, or individual data feeds offered by Exchanges and others to minimize network and other types of latencies.

AT presents huge research challenges, especially given the economic consequences of getting it wrong, such as the May 6, 2010 Flash Crash[11,18] in which the Dow Jones Industrial Average plunged 9% wiping $600 billion off the market value and the Knight Capital loss of $440 million on August 1, 2012, due to erratic behavior of its trading algorithms. Current research challenges include:

▸ **Data** challenges cover the quantity/quality of the data, processing data at ultra-high frequency and increasingly incorporating new types of data such as social media and news.

▸ **Algorithm behavior**: An AT system may have a few or many hundreds of variables; often a minor change to a variable can have a catastrophic impact on performance.

▸ **Algorithms' interaction**: Very little is known about the interaction of AT systems within a market (or even within a firm) leading for example to flash crashes.

▸ **High-performance computing**: Low-latency trading is a major driver for performance, and also algorithm research in Graphics Processing Units (GPUs) and Field-Programmable Gate Arrays (FPGAs).[2]

Market Microstructure, Data, and Research

To understand AT, it is useful to understand how a trade is executed by an Exchange, the different types of trading, and the objectives and challenges.

Trade execution: Dealers generally execute their orders through a shared centralized order book that lists the buy and sell orders for a specific security ranked by price and order arrival time (generally on a first-in-first-out basis). This centralized order-driven trading system continuously attempts to match buy and sell orders. As shown in Figure 1, the order book is divided into two parts: buy orders on the left ranked highest price at top and sell orders with lowest price at top. Orders are generally listed by price and time priority, which means most Exchanges prioritize orders based on the best price and first-in-first-out framework.

There are many variants to the order book model described here. Different Exchanges will accept different order types (market orders, limit orders, stop orders, among others).[9] In developing an AT system, understanding the market microstructure—the detailed process that governs how trades occur and orders interact in a specific market—is of paramount importance.

The "ideal" AT prerequisites include:

- **Centralized order book**: Shared centralized order book, listing the buy and sell orders.
- **Liquid markets**: A highly liquid market and typically one suitable for high-frequency trading of securities (for example, equities, FX).
- **Order book depth** provides an indication of the liquidity and depth (the number of buy and sell orders at each price) for that security or currency.
- **Financial information protocols** (for example, the Financial Information eXchange protocol, or FIX) for the computer communication of information.

Figure 1. Trade order book.

Order Book – ABC Inc.

Buy		Sell	
Quantity	Price	Price	Quantity
5,000	**99**	**99**	**4,000**
8,000	98	100	10,000
10,000	97	101	1,000
15,000	95	103	15,000

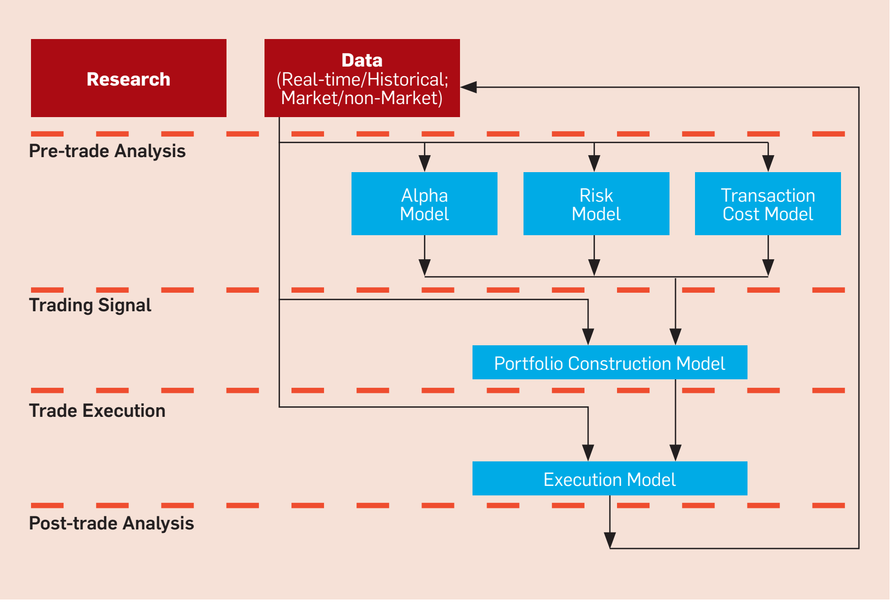
(a) Before matching a trade

Order Book – ABC Inc.

Buy		Sell	
Quantity	Price	Price	Quantity
1,000	**99**	100	10,000
8,000	98	101	1,000
10,000	97	103	15,000
15,000	95	104	3,000

(b) After matching a trade

Figure 2. Components of an AT system.



AT System Components

The trading process (as illustrated by Figure 2) may be divided into five stages:

- **Data access/cleaning**: Obtaining and cleaning (financial, economic, social) data that will drive AT.
- **Pre-trade analysis**: Analyzes properties of assets to identify trading opportunities using market data or financial news (data analysis).
- **Trading signal generation**: Identifies the portfolio of assets to be accumulated based on the pre-trade analysis (what and when to trade).
- **Trade execution**: Executing orders for the selected asset (how to trade).
- **Post-trade analysis**: Analyzes the results of the trading activity, such as the difference between the price when a buy/sell decision was made and the final execution price (trade analysis).

Although we might think of AT as automating all stages, much of the activity in the industry is devoted to the data access/cleaning and pre-trade analysis stages, with the latter stages of the trading process being supervised by humans.

Figure 2 illustrates the major components of each stage of an AT system. The pre-trade analysis comprises computing analytical features through three main components: Alpha model, the mathematical model designed to predict the future behavior of the financial instruments that the algorithmic system is intended to trade; Risk model that evaluates the levels of exposure/risk associated with the financial instruments; and Transaction Cost model that calculates the (potential) costs associated with trading the financial instruments.

At the Trading Signal stage, the Portfolio Construction model takes as its inputs the results of the Alpha model, the Risk model, and the Transaction Cost model and decides what portfolio of financial instruments should be owned and in what quantities, in the next time horizon.

At the Trade Execution stage, after the trading signal has been generated and the portfolio constructed, the Execution model performs several decisions with constraints on (actual) transaction costs and trading duration: the most general decision is the Trading Strategy followed by the Venue and Order type, decisions handled by smart order routing.

Finally, at the Post-trade analysis stage, the results of the trading activity—for example, the difference between the expected price and the final execution price and P&L—are evaluated.

Narang[13] and other industry experts emphasize the importance of the (quantity and frequency of) "clean" data available for analysis to the success of the trading system. Data sources broadly comprise:

- **Financial data**: Price data on financial instruments from Exchanges and electronic communication networks (ECNs), and also financial information from services, such as Bloomberg and Thomson Reuters.
- **Economic data**: Fundamental economic data, such as the overall state of the countries' economies (for example, unemployment figures), interest rates, gross domestic product, and national policy.
- **Social/news data**: Sentiment data "scraped" from social media (Twitter, Facebook), RSS feeds, and news services; subdivided into:
- **Real time**: Live data feeds from Exchanges, ECNs, news services, or streamed social media.
- **Historic**: Previously accumulated and stored financial, economic, and social/news data;

and into:

- **Raw data**: Unprocessed computer data straight from source that may contain errors or may be unanalyzed.
- **Cleaned data**: Correction or removal of erroneous (dirty) data, caused by contradictions, disparities, keying mistakes, missing bits, and so on.

Analyzed data: Context-specific features of raw and cleaned data that emphasize principal components of underlying data. These data sources, real time and historic, are the cornerstone of the research, design, and back testing of trading algorithms and drive the decision making of all AT system components. Buying (raw and especially cleaned) data is hugely expensive and cleaning data is highly time consuming, but essential due to the sensitivity of trading algorithms.

Pre-trade analysis, specifically the Alpha model, analyzes (for example, data mines) real-time and historic data to identify potential trade opportunities. The three principal techniques are:

Buying (raw and especially cleaned) data is hugely expensive and cleaning data is highly time consuming, but essential due to the sensitivity of trading algorithms.

- **Fundamental analysis**: Evaluates variables that can affect a security's value, including macroeconomic factors (like the overall economy and industry conditions) and company-specific factors (like financial reports and management).
- **Technical analysis**: The approach is mainly concerned with trend analysis as well as identification and interpretation of chart pattern formations (for example, head and shoulders) in price/volume. It summarizes trading history of securities (for example, company stocks) by analyzing changes in their price or traded volumes.
- **Quantitative analysis**: Applies wide range of computational metrics based on statistics, physics, or machine learning to capture, predict, and exploit features of financial, economic, and social data in trading.

Alpha Models

For the Alpha model, as illustrated by Figure 3, there are essentially two alternative basic strategy approaches:

- **Theory-driven**: The researcher/programmer chooses a hypothesis for the way the securities are likely to behave and then uses modeling to confirm their hypothesis.
- **Empirical**: The researcher/programmer uses an algorithm to data mine and identify patterns in the underlying data, without any preconceived views on the behavior of a security.

Constructing the Alpha model and more specifically setting the variables can be a highly complex task. Numerous factors influence the actual algorithm implementation: forecast goals, like direction, magnitude, duration, probability); forecast time horizon, such as millisecond, day, week; the mix of instruments; the data available; actual setting of the model's variables; and the frequency of running the model.

Risk Models

Risk models[1,13] focus on the risks associated with a target portfolio of financial instruments and the relevant factors that affect the economic climate and hence the current/future value of the financial instruments. It does so by attempting to quantify both the risk associated with individual instruments and with the portfolio. Figure 4 illus-

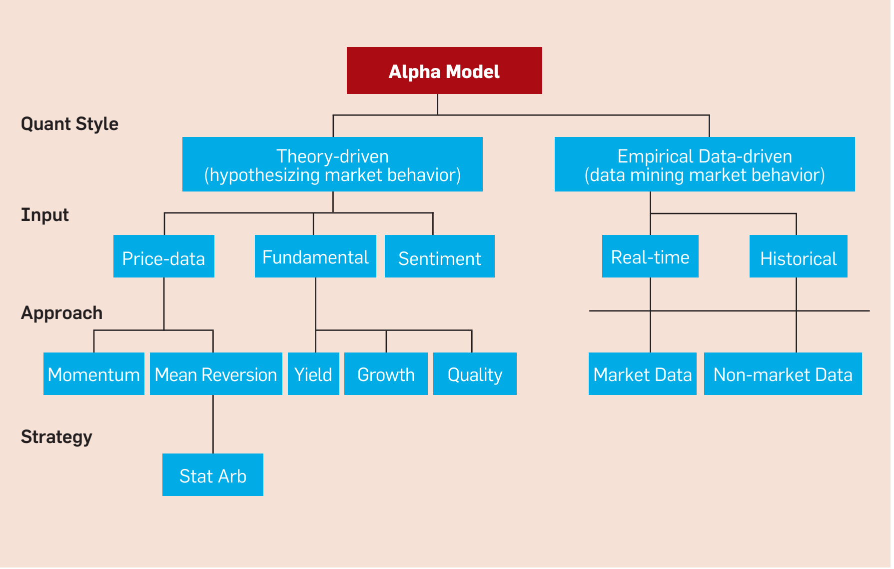


Figure 3. Alpha models—predicting the future.

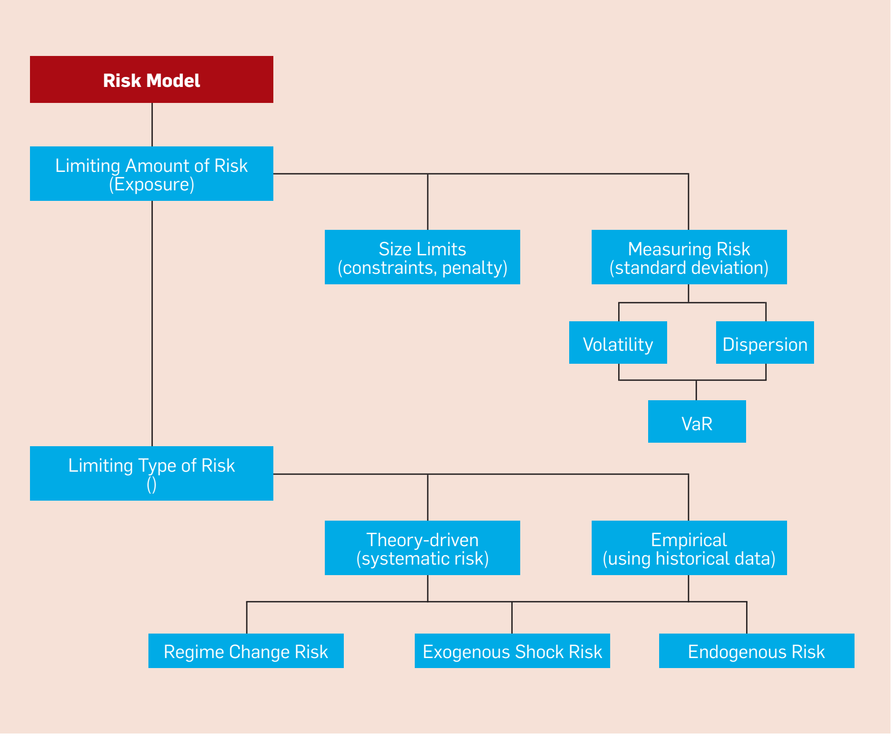


Figure 4. Risk models—sizing of exposures.

trates the two principal approaches to risk that we refer to as:

- **Limiting amount of risk**: Limiting the size of risk is managing the exposure through limiting by penalty or constraint, such as leverage; or calculating volatility and dispersion; and
- **Limiting type of risk**: Eliminating whole types of exposure. As with other models, the approaches subdivide broadly into Theory-driven and Empirical using historical data.

In limiting these two, we take care of optimizing the risk-reward ratio.

Transaction Cost Models

Lastly, the Transaction Cost model calculates all possible transaction costs, providing the Portfolio Construction model with the ability to estimate the theoretical spectrum of transaction costs associated with the trading options when constructing the portfolio. These include commissions, slippage and market impact, and so on.

At the **Trading Signal Generation**, the Portfolio Construction model takes as inputs the results of the Alpha model, Risk model and Transaction Cost model and optimally selects the "best" portfolio of financial instruments to hold to maximize potential profits, while limiting risk and minimizing trading costs.

Portfolio Construction Models

Portfolio Construction models, as illustrated in Figure 5, broadly subdivide into:

- **Rule-based models** are heuristic specifications defined by the Quant concerning how to assemble the portfolio of instruments. For Rule-based models there are essentially four classes of models.[13] For example, with Equal

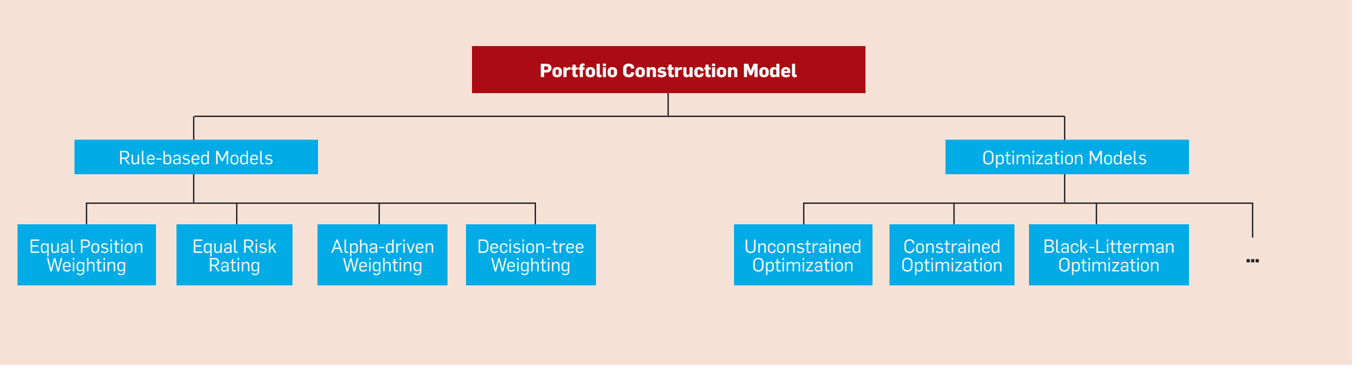


Figure 5. Portfolio construction models—assembling portfolio.

Position Weighting all chosen instruments are given equal weighting.

▸ **Optimization models** use algorithms with an objective function that iterates to a portfolio that gives, for example the highest rate of return. Examples include the Black-Litterman—a sophisticated optimizer popular with AT.[13]

The trade execution stage decides where, what, and when to trade, requiring several decisions with constraints on (actual) transaction costs and trading duration.

Execution Models

The execution model (Figure 6) decides:

▸ **Trading venue** selects the Exchanges where the orders are to be placed, with many securities being potentially tradable on multiple Exchanges.

▸ **Execution strategies** cover scheduling of trades, such as the popular Volume Weighted Average Price (VWAP) that executes a buy order in a financial instrument (for example, stock), as close as possible to its historical trading volume.

▸ **Order type** subdivides in market orders, where a buy or sell order is to be executed by the system immediately at current market prices; and limit orders, where an order to buy a security is specified at no more (or sell at no less) than a specific price.

Finally, **post-trade analysis** compares the expected and actual performance of the executed trades. This involves cost measurement and performance measurement. For example, cost can be measured between the actual realized execution price achieved and the so-called benchmark price.

AT Strategies

When we use the term an "algorithmic trading strategy" we are typically referring to the precise nature of the entire spectrum of activities employed by a software system starting from pre-trade analysis, model selection, signal generation, trade execution, and post-trade handling. Two common strategies are detailed in Dunis et al.[4] and Kaufman:[10]

▸ **Momentum**: A trend following trading strategy that aims to capitalize on the continuance of existing trends in the market. The algorithm assumes large increases in the price of a security will be followed by additional gains and vice versa for declining values.

▸ **Mean Reversion**: A trading strategy assuming prices and returns eventually move back toward the mean or average. A popular strategy is mean reversion (pairs trading) where two historically correlated securities that have diverged are assumed to converge in the future.

To illustrate trading strategies, we discuss so-called market-neutral and risk-neutral AT,[13,16] defined as:

▸ **Market-neutral** strategies are trading strategies, widely used by hedge funds or proprietary traders that go "long" on certain instruments while "shorting" others in such a way that the portfolio has no net exposure to broad market moves. The goal is to profit from relative mispricings between related instruments.

▸ **Risk-neutral** strategies are trading strategies insensitive to risk, meaning a strategy, which is indifferent to the risk involved in an investment and is only concerned about expected return. This strategy type is used by market makers concerned more with continuous financial flow rather than risk aversion/seeking.

For example, in equity markets, market-neutral strategies take two forms.

▸ **Equity market-neutral** emphasizes fundamental stock picking, where a portfolio of 'long' and 'short' positions is maintained to be beta neutral. (The beta of a stock or portfolio is a number describing the correlation of its returns with those of the financial market as a whole.) 'Long' and 'short' positions are also managed to eliminate net exposure through diversification.[3,16]

▸ **Statistical arbitrage** ("stat arb") is an equity trading strategy that employs time series methods to identify relative mispricings between stocks. One popular technique is pairs trading; pairs of stocks whose prices tend to move together; that is they are identified as correlated.[5,12]

The purpose of market-neutrality of a portfolio is hedging away risk, created by large collective movements of the market. The idea is that such a portfolio should not be affected by any type of such price movement, neither up nor down. To achieve such neutrality, one must simultaneously manage (buying) long and (selling) short positions in the market.

Algorithm System Architecture

We now look at the trend following or momentum strategy in more detail using Equities as an example.

Momentum strategy. A typical momentum trading strategy for stocks aims at capturing trends in stock prices. It uses the contention that stocks with an upward momentum will continue rising and should be bought and stocks with downward momentum will continue falling and should be sold. We use the nomenclature of "winners"

Figure 6. Execution models—venue, strategy, order type.

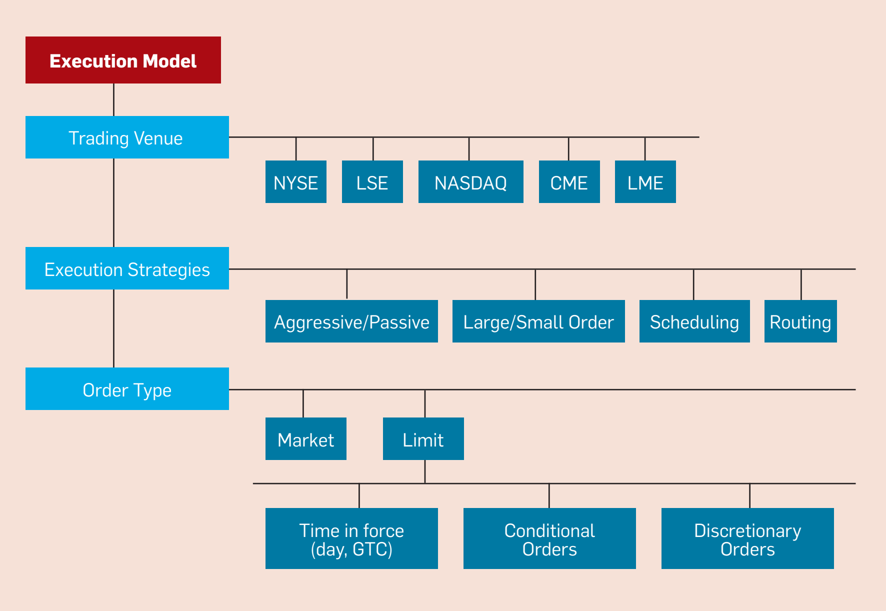


Figure 7. Download Yahoo! finance data in Python.

```
-> import ystockquote
-> Input ticker
-> Input start date
-> Input end date
-> Store data in variable "data"
```

Actual Code Fragment from Python:

```
import ystockquote

ticker = 'GOOG'
start = '20080520'
end = '20080523'

data = ystockquote.get_historical_prices (ticker, start, end)
```

Figure 8. Alpha model pseudo code.

```
-> Choose a look0back period for a specific security [example: 1 hr, 30
mins, 10 mins, 1 min, 30 seconds]
-> Access tick data for the look-back period
-> Compute a simple moving average metric [SMA]
        -> In order to calculate SMA we need to select
                 1) Time-scale for historical data (e.g. 1 min)
                 2) Number of data points to be used in the SMA
(e.g. 120 data points)
                 -> SMA [60} = {current price + (price 1 min before) +
(price 2 mins before) + ..... (price 59 mins before)}/60

-> Repeat for all stocks [i.e. all 406 stocks in the S&P500]

-> We now have enough data to enter the signal generation stage of the
Alpha Model
```

for the former and "losers" for the latter. In our implementation of the momentum strategy, we use the Simple Moving Average for the last 60 minutes (SMA) metric to capture upward/downward trends.

The strategies commence by conducting pre-trade analysis on stock price data of a specific granularity. Pre-trade analysis usually entails computing the value of one or more technical indicators.

For the purpose of the strategy presented here we have used a single technical indicator SMA. The value of SMA is used for signal generation.

If the stock price for one of the shortlisted stocks exceeds the value of the SMA, the system generates a buy signal and vice versa. This stage is usually called signal generation. During signal generation the stocks that need to be traded get picked via the signals.

The next phase of the strategy consists of Portfolio construction, where we determine the amounts of each stock to hold. A simple portfolio construction rule is to invest equal dollar amounts in each stock that needs to be traded since we do not know how each stock would perform. After we run the strategy for at least a day, we can compute stock-specific performance metrics to invest aggressively in specific stocks that performed well.

Our Trade Execution model assumes there is a fixed strategy update cycle where the strategy does the pre-trade analysis for the look-back period, scans for trading signals and places trades and this sequence of actions is repeated at regular intervals (the most optimal frequency is chosen during optimization over multiple back-tests, as we will discuss later) Each trade has an investment horizon—the maximum duration one can hold a position (a parameter of the strategy) and we may also temporarily withdraw from trading any stock whose number of incurred consecutive losses or the total unrealized loss exceeds a predetermined value (another parameter).

Real-time strategy implementation and optimization. The basic building block of our implementation of the momentum strategy is the unit generating specific realizations of the trading scenario for any particular set of controlling parameters. Its responsibility is to traverse all the concurrent strands of stock data day-by-day, calculate the scores, grade the stocks, and carry out the resulting trades. At the same time it keeps track of its own stock holding inventory and current cash balance. Its output is the time development curve of the total value of its holdings (stock and cash), for which the Sharpe Ratio is then calculated. (The Sharpe Ratio is a measure of the excess return (or risk premium) per unit of deviation in an investment asset or a trading strategy.)

The strategy runs, described here, must be repeated for every combination of parameters' values (different scoring thresholds, investment horizons) via exhaustive enumeration. Out of this set of all possible realizations, the optimal strategy, with best Sharpe Ratio, is then selected.

Trading System Implementation

We now present a simple AT system based on momentum, which is market neutral and risk neutral, together with pseudo-code fragments to illustrate its implementation. We have removed all possible extraneous details with the goal of making the implementation easy to understand. For a more intricate discussion of momentum trading, read Wang and Kochard.[17]

Our simple implementation is based on the concept of price momentum; that is a tendency of rising stocks, the winners, to keep rising, and falling stocks, the losers, to fall further. Momentum, as a property of the Stock Market, is somewhat controversial, but its existence in stock returns has been confirmed by empirical studies. The strategy discussed here will consist of trading stocks from a broad fixed collection selected to represent the S&P500 universe. At any time we will hold a portfolio of long positions in top winners and short positions in bottom losers[8] selected from these stocks.

It is important to note the description of the strategy does not assume any particular frequency. Essentially this strategy can be applied at different frequencies (a parameter chosen by the trader and typically optimized during multiple back-testing runs, as we will discuss later).

Financial data. The first step toward building an AT system is building an infrastructure to handle huge amounts of real-time data, and the accumulation of historical data of specific granularity to compute the values of fundamental, technical, and quantitative metrics that fuels the pre-trade analysis. Figure 7 provides Python pseudo-code for data access from sources, such as Yahoo! Finance, using *ystockquote*, to import Google ticker data.

As discussed, the strength of the pre-trade analysis is partially reliant on the quality and granularity of data being pushed into the system. Granularity specifies whether the data is tick, or one-minute candle, three-minute candle, five-minute candle, daily candle, and so on. A candle specifies the highest, lowest, open, and close price in the period specified.

Pre-trade analysis. Having gathered and cleaned the data, the two components of the pre-trade analysis stage discussed here are the Alpha model and the Risk model. We start with the pseudo-code for the Alpha model stage (Figure 8) for the S&P500.

During pre-trade analysis, the trading system digests the data input and computes statistical metrics that indicate or highlight different features in the data. For equities these features could be volatility, trend, co-integration, mean-reversion. The "indicativeness" of the metrics is crucial to the success of any trading strategy. Standard metrics include: simple moving averages or proprietary metrics.

Risk Model (Pre-trade and Signal generation). Next the Risk model code (illustrated in Figure 9) calculates the risk metrics for the various instruments that are candidates for the portfolio to be calculated at the Trading Signal generation stage. As shown by the pseudo-code this uses a standard VAR (Value at Risk) calculation.

The risk model of the trading system periodically computes risk metrics for each security and for the

Figure 9. Risk model pseudo code.

```
-> Compute Historical VAR per security (10%, 1 -day VAR)
        -> Collect price data for the security for the
past 500 days                                              Pre-trade:
        -> You have 500 data points                        Selecting
        -> Look at the 50th worst outcome                  entrants to
        -> This is the 10% 1 day VAR for the security      the portfolio
-> If Current Price of Security < Historical VAR
        -> Do not include in portfolio                    Risk signaling
-> Else
        -> Include in portfolio
-> Compute portfolio VAR
        -> Historical VAR (10%, 1 day-VAR)
                  -> Steps:
                          1) Go back 500 days
                          2) Re-value the portfolio on market prices
                          on the past 500 days
                          3) You now have 500 data points for the last
                          500 days which indicate how the portfolio
                          must have behaved on the last 500 days
                          4) For 10% VAR 0.1*500 = 50th worst value is
                          the 10%, 1-day VAR
-> Monitor portfolio for risk violations
        -> If Portfolio value < VAR                       Risk signaling
                   -> Close all positions in portfolio     Risk Feedback
                                                            to execution
                   -> Suspend Algorithm until manual intervention
```

Figure 10. Trade signal generation stage—Alpha model.

```
-> Carefully monitor the value of SMA [60] for each security {1, 2, 3….n}
where n = 406
-> For: each security {1, 2, 3.....n} where n = 406
        -> If SMA [60] score < current price [classify security
for bullish divergence – "Winner"]
                  -> Flag buy signal
        -> If SMA [60] score > current price [classify security
for bearish divergence – "Loser"]
                  -> Flag sell signal
```

Figure 11. Trade execution.

```
-> Upon Buy signal for a security
   -> Place market buy order for security [such that (quantity*price)
 = dollar allocation for security]      Apply same logic to all 406 stocks
-> Upon Sell signal for a security
   -> Place market sell order for security [such that (quantity*price)
 = dollar allocation for security]
-> Upon forming a portfolio of greater than 1 stock
-> Compute portfolio beta [the degree of sensitivity of the portfolio
value to movements in the market index]
-> If beta > 0 [this means positive exposure w.r.t the index]
        -> Short/Sell the right amount of the S&P 500 index [this will
generate a negative beta which will cancel
    your portfolio's positive bet              [Hedging: Market neutrality]
-> If beta < 0 [this means negative exposure w.r.t the index]
        -> Buy/Long the right amount of the S&P500 index [this will
generate a positive beta which will cancel
    your portfolio's negative beta]
-> For: Each security {1, 2, 3….406} -Track existing positions
periodically [example: every 15 mins]
        -> If Profit threshold reached for a specific security
                -> Liquidate profits & add to cash
        -> If Loss threshold reached for a specific security
                -> Close position, take in losses
        -> If Investment Horizon reached
                -> Close position irrespective of profits/loss
```

portfolio of securities. If a VAR break occurs (a VAR break is when the portfolio value falls below the Value at Risk threshold), a set of predefined rules kick in. In this pseudo-code, a VAR break triggers the algorithm to close all positions, that is, liquidate the portfolio and suspend its running until again manually restarted. In a similar way, customized rules can be predefined for each risk metric that states how the system must behave in the event any of the risk metrics breaching thresholds.

At Trade Signal Generation (illustrated by the code in Figure 10) we use the features computed in the pre-trade analysis for the Alpha model and define a set of heuristic rules for signal detection. A classic example is described here:

If the current price hovers above the 60-minute SMA, it is expected to be trending upward and vice versa. This logic is implemented in the trade execution.

Next we look at the Trade Execution stage. The Trade Execution code (illustrated by Figure 11) typically decides the trading venue of the order to be issued based on an assessment of liquidity. However, here we stick to the assumption of a single trading venue. Once a buy signal is flagged by the system, it reacts by placing a buy order worth a certain amount of money on the market. (For example: if we start with 10 stocks and $1 million in cash and equal allocation, a buy signal for GOOG Google Inc. when the price is 591.93 would trigger a buy order for 168 stocks of GOOG amounting to 168*591.93 = $100,000).

Market neutrality is maintained by beta-hedging (holding a portfolio such that beta is as close to zero as possible) by buying or selling appropriate amounts of the stock index to which the equities belong. In this case, since we have shortlisted equities from S&P500 index market neutrality is maintained by going long or short the S&P500 index.

Post-trade analysis. As discussed, in trading the so-called Implementation Shortfall is the difference between the decision price and the final execution price (including commissions, taxes, among others) for a trade. This is also known as the "slippage." The post-trade analysis calculates Implementation Shortfall, P&L, among others.

While back-testing does not allow one to predict how a strategy will perform under future conditions, its primary benefit lies in understanding the vulnerabilities of a strategy through a simulated encounter with real-world conditions of the past.

Once the trades have been executed, the results need to be monitored and managed. The P&L generated from the trades must be analyzed such that stocks that perform badly are allocated less money in allocation than the ones that do well.

Back-testing and Performance

Before implementation, the whole algorithmic trading strategy is thoroughly back-tested.[15] Back-testing is a specific type of historical testing that determines the performance of the strategy if it had actually been employed during past periods and market conditions. While back-testing does not allow one to predict how a strategy will perform under future conditions, its primary benefit lies in understanding the vulnerabilities of a strategy through a simulated encounter with real-world conditions of the past. This enables the designer of a strategy to "learn from history" without actually having to make them with actual money.

The implementation of the AT system might be back-tested on daily price data from the S&P500, taken from the dates of July 1, 2005 to July 1, 2011 (six years) from Yahoo! Finance. It is important to note the granularity of the data the strategy is back-tested and optimized on determines the minimal frequency level of the strategy. In this case, since the strategy is back-tested on daily data, it should have one run per day—a run refers to calculation of pre-trade, scanning for signals, and placing trades.

On data cleaning: We have based our trading portfolio case study on the definition of the S&P500 from April 16, 2007. It is the last reasonably complete list we could find in the public domain. (S&P forbids free circulation of the current list of the S&P500 components.) After removal from this list of all the stocks for which there are no complete coverage of the entire six-year stretch of our experiment, we are left with 414 stocks. With regard to cleaning, we found that for eight of these stocks the price data contains abnormal daily returns, with absolute values exceeding three. As realistic trading systems would have prevented these stocks from being traded during those extreme price jumps, we have decided, for simplicity, to remove them entirely. The remaining 406 stocks will form the initial contents of our trading portfolio.

It must be noted here that, as the composition of the S&P Index keeps changing, not all of the stocks we picked will necessarily still be part of the Index at the end of our experiment. However, by insisting on the completeness of the pricing information for all the stocks in our portfolio we ensure they were all actively traded throughout. As the adjustments of the Index are impossible to monitor, we will ignore them, as if the Index was frozen.

Our initial trading account starts with a unit of wealth, representing $1 million.

Back-testing. We can divide the back-testing simulation into two phases. The first five years constitute the in-sample period, when we train the algorithm, in order to choose the parameters (for example: the optimal look-back period, investment horizon, risk thresholds) giving the best P&L performance. Then, the following year, the out-of-sample period, is the test run used to validate the selected parameters.

In-sample training: Optimization. During this phase, different variations of the strategy are applied to past data. Variations of a strategy can be defined as the same underlying model with different time parameters for computation of pre-trade metrics. For example: the look-back period for computing the metrics could range from few seconds to few hours. A variation can also mean applying the same model with different time frame for investment horizons of trades. For example: a strategy could be tested where the maximum holding time for positions is three minutes or 30 minutes or an hour. After each variation is applied to past data, the P&L performance is recorded and the Sharpe ratios computed. The strategy variation with the best P&L performance and Sharpe Ratio is chosen to be applied to the out-of-sample period.

There are two adjustable parameters controlling the trading in this strategy:

- Time-window used to calculate the pre-trade metrics.
- Time-frame between portfolio adjustments-investment horizon.

Their optimal values are established by looking for the best in-sample strategy performance, as measured by the Sharpe ratio.

Calculation of Sharpe Ratio

If we denote by R_t the daily returns, that is, the daily relative (%) changes of value of our total wealth (the sum of the trading account, the net value of the portfolio and the current value of the SPX-hedge), then the associated Sharpe Ratio (SR) is given as

$$SR = \text{mean}(R_t) / \text{std}(R_t);$$

where mean(R_t) and std(R_t) are respectively the average and the standard deviation of the returns calculated over the entire training period of our strategy. As the mean value represents the gains and the standard deviation conveys the risk, the maximum value of the Sharpe Ratio offers the best compromise between the performance and risk.

Back-testing results. The strategy variation with the best Sharpe Ratio and out of sample performance is chosen to be implemented in real time. Back testing is a key step in any algorithmic strategy construction as without back-testing we would not know which combination of parameters gives the highest profit for a given strategy.

Conclusion

As noted at the outset, the research challenges (and the consequences of getting it wrong) are still poorly understood. Indeed, challenges persist and span data handling—cleaning, stream processing, storage, and using new types such as social data and news for forecasting; algorithm behavior—selecting the best computational statistics and machine learning algorithms, and setting their variables to optimize performance; algorithms' interaction—understanding the interactions of AT systems within a firm and within a market; and high-performance computing—support of low-latency trading and the use of novel hardware such as GPUs and FPGAs.

We hope this article serves to encourage computer scientists to pursue research in the AT area. Our goal was to provide a comprehensive presentation of AT systems, including pseudocode fragments, and as a case study, a simple market-neutral trend following AT strategy.

Acknowledgments

This article was produced by the U.K. Centre for Financial Computing (www.financialcomputing.org). We thank Dan Brown and Jan Grochmalicki, and students contributing to the AT and risk platform, and the extensive research conducted with the banks and funds. We also thank Trading Technologies, Chi-X, and Knight (HotSpotFX) for granting us access to their datafeed streams for research; Barclays, Citigroup, BNP Paribas Fortis, and BAML for their advice and support, Clive Hawkins (Barclays), Nick Edwards (Trading Technologies), Martin Frewer (BNP Paribas Fortis), Kevin Gillespie (HotSpotFX), and Piotr Karasinski (EBRD).

References

1. Chan, E.P. *Quantitative Trading: How to Build Your Own Algorithmic Trading Business.* Wiley Trading, 2009.
2. Che, S., Li, J., Sheaffer, J., Skadron, K. and Lach, J. Accelerating compute-intensive applications with GPUs and FPGAs. In *Proceedings of the IEEE Symposium on Application Specific Processors* (June 2008).
3. Chincarini, L. and Kim, D. *Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management.* McGraw-Hill Library of Investment & Finance, 2006.
4. Dunis, C. L., Laws J. and Naïm P. *Applied Quantitative Methods for Trading and Investment.* Wiley Finance Series, 2003.
5. Ehrman, D.S. *The Handbook of Pairs Trading: Strategies Using Equities, Options, and Futures.* Wiley Trading, 2006.
6. Financial Information eXchange (FIX) Protocol; http://en.wikipedia.org/wiki/Financial_Information_eXchange or http://fixprotocol.org/
7. Galas, M., Brown, D. and Treleaven, P. A computational social science environment for financial/economic experiments. In *Proceedings of the Computational Social Science Society of the Americas* (2012).
8. Jegadeesh, N. and Titman, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *J. Finance 48*, 1 (Mar. 1993), 65-91.
9. Johnson, B. Algorithmic trading & DMA: An introduction to direct access trading strategies, 2010.
10. Kaufman, P.J. *The New Trading Systems and Methods.* Wiley Trading, 2005.
11. Lo, A. Finance is in need of a technological revolution. *Financial Times*, Aug. 27, 2012.
12. Meucci, A. Review of Statistical Arbitrage, Cointegration, and Multivariate Ornstein-Uhlenbeck. SSRN preprint 1404905, 2010.
13. Narang, R.K. *Inside the Black Box: The Simple Truth About Quantitative Trading.* Wiley Finance, 2009.
14. Nuti, G., Mirghaemi, M., Treleaven, P. and Yingsaeree, C. Algorithmic trading. *IEEE Computer 44*, 11 (Nov. 2011), 61–69.
15. Pardo, R. *The Evaluation and Optimization of Trading Strategies*, Wiley Trading 2008.
16. Stokes, E. *Market Neutral Investing.* Dearborn Trade Publishing, 2004.
17. Wang, P. and Kochard, L. Using a Z-score Approach to Combine Value and Momentum in Tactical Asset Allocation. SSRN preprint 1726443, 2011.
18. Zigrand, J.P., Cliff, D. and Hendershott, T. Financial stability and computer based trading. *The Future of Computer Trading in Financial Markets, U.K. Gov.*, 2011.

Philip Treleaven (p.treleaven@ucl.ac.uk) is a professor of computer science and head of the U.K. Ph.D. Centre in Financial Computing at University College London, U.K.

Michal Galas is a Research Fellow and chief programmer in the U.K. Ph.D. Centre in Financial Computing at University College London, U.K.

Vidhi Lalchand is a research student at University College London, U.K.

DOI:10.1145/2500726

Coupling content adaptation with context awareness is a promising approach for improving the user's experience.

BY MAMADOU TOURAD DIALLO, HASSNAA MOUSTAFA, HOSSAM AFIFI, AND NICOLAS MARECHAL

Adaptation of Audiovisual Contents and Their Delivery Means

WITH THE ADVENT of technology characterized by heterogeneity and the increasing offers of services and applications, a user's Quality of Experience (QoE) has become a crucial factor for the success or failure of new applications and services. *Content adaptation* is the process of selecting, generating, or modifying content to suit the user's preferences, consumption style, computing and communications environment, and usage context. This can be applied within media types, such as reducing image size or resolution, and across media types, like converting video from one codec to another. Content adaptation can be implemented through three general approaches: Client-side approach (user); server-side approach (media source); or, intermediate-proxy approach between the client and the server.[10]

The client-based approach is performed by the client's device, therefore providers can create one version of content and the same content can always be displayed on different devices. The terminal selects the content, which is matching the most of its capacities and the available bandwidth. Examples are Scalable Video Coding (SVC), where the delivered content is coded in different versions in the form of different layers and the client terminal selects the layers that best match the terminal capacity and network status; and HTTP Adaptive Streaming (HAS), where content is coded considering different quality levels and the terminal requests the best matching content based on device capacity and available bandwidth.

In the server-based approach, the content is adapted before being delivered to the client, thus reducing the transmission time and the bandwidth consumption since the delivered content is in an adapted form (only the content users need will be sent). This

» key insights

- **Today, online video services like mobile TV, IPTV, or WebTV, are providing content adaptation techniques under the form of adaptive streaming protocols, which mostly rely on network status, CPU usage, and screen resolution.**
- **The proliferation of connected devices stresses the need to address the increase in the end-user's expectations on quality of experience together with network-friendly approaches. For this purpose, adaptation of content and services are required to meet users' preferences, application expectations, and network characteristics, among others.**
- **To improve user satisfaction, we need to consider contextual information about the user, the terminal, the networks, and the content.**

TV BY ANDREA CRISANTE. PROGRESS BAR IMAGES FROM SHUTTERSTOCK.COM

also reduces the processing time at the client side.[11] Examples include Windows Media Services, Adobe Flash Media Server, and QuickTime Streaming Server, where they store multiple variants of the same content in the server and the content best matching to the client content is selected through the client identification sent in the client request to the client profile.[1]

In the intermediate proxy approach, content adaptation is done by an intermediate element. For example, in Wegner et al.,[19] the Media Aware Network Element (MANE) is a proxy between the server and the client that gathers the terminal capacity in the client side and the network characteristics and selects the best adapted content based on the terminal and network status. On the other hand, several research contributions exist on content adaptation methods that adapt the content and its delivery based on information related to network congestion, terminal capacity, measured QoE. However, to our knowledge, there is no method that combines all these factors for both content adaptation and delivery. In our work, we consider a multidimensional approach that adapts the content and its delivery means for enhancing the user's experience. Adapting the content is mainly in terms of changing bit rate, changing codec, while adapting the content delivery means is related to how content will be delivered (unicast, multicast, through which access network). The adaptation process should consider the measured QoE coupled with context information on the user, devices, network, and the content itself to take the adequate adaptation decision. Our notion of QoE is explained in Diallo et al.,[4] where the user experience is evaluated as a function of different context concerning the user (preferences, content consumption style, level of interest, location), his environment including the terminal used (for example, capacity or screen size) and network (available bandwidth, delay, jitter, packet loss rate).

In this article, we examine the existing video streaming techniques and compare the HTTP-based techniques with other techniques. We then focus on the content delivery means, present the content adaptation, and discuss the research contributions for content adaptation and the adaptation of content delivery means.

The adaptation process should consider the measured QoE coupled with context information on the user, devices, network, and the content itself to take the adequate adaptation decision.

Video Streaming Techniques

Today there are several streaming methods to deliver audiovisual content, including: HTTP streaming, Real Time Streaming Protocol (RTSP, developed by IETF), Microsoft Media Services (MMS), and Real Time Messaging Protocol (RTMP) for Adobe Systems. Here, we review these different techniques.

HTTP-based streaming techniques carry out a dynamic content adaptation before and/or during the session following a client-based approach. The adaptation is managed by the player. The different HTTP-based streaming techniques are:

▸ *HTTP download.* Audiovisual content is delivered through the HTTP protocol just as classical Web objects/pages. To view the content, the media player on the terminal must wait until the video is downloaded in the full disk of the terminal.

▸ *HTTP progressive download* is based on HTTP and allows the player to begin playback of the media before the download is complete. The key difference between media streaming and the Progressive Download is in how the digital media content is received and stored by the end user's device that is accessing the digital media. The media player for the Progressive Download playback makes use of metadata located in the header of the file and a buffer in the user device. When a specified amount of content becomes available in the buffer, the playback is started. This specified amount of buffered content is embedded in the media file by the content producer in the encoder settings and is reinforced by additional buffer settings imposed by the media player. It can also be computed by the player using encoding profile (codec, bitrate, and so on).

Adaptive streaming based on HTTP (HAS)[13,15] allows a multi-rate client-based streaming for multimedia content, where different bitrates and resolutions are available for the same content and the streaming is managed by the client. As explained in Oyman and Singh,[16] HAS deployment presents opportunities for services and content providers when compared with traditional adaptive streaming techniques. The server sends a manifest file containing the description of content pieces namely chunks (supported codec, minimum bandwidth, resolution, bitrates, URL). The Media Player then selects the most adequate fragment based on the available bandwidth and device capacity (CPU load, supported codec, bitrate). The following are examples of Adaptive Streaming methods based on HTTP.

▸ *Apple HLS (HTTP Live Streaming)*: This solution was very quickly adopted by Over The Top (OTT) players and is currently available on all Apple devices (iPhone, iPad, iPod) as well as some set-top boxes (AirTies, Netgem, Amino) and software players (VLC media player release 1.2.0, QuickTime X Player).

The native codec chosen for HLS is MPEG H.264 for video and Advanced Audio Coding (AAC) for audio. HLS is implemented as follows: Encoding video in H.264/TS format at different bitrates. For each encoding profile, a Stream Segmenter cuts the content into pieces called chunks, typically 10 seconds each, and generates a play list file (m3u or m3u8) with a URL for each chunk of this encoding profile. Generating a general index file (manifest) indicating each available encoding profiles (bitrate, codec) and the URL of the corresponding playlist files. Distributing content chunks, playlists, and manifest through an HTTP server (origin or cache). Looking for the most suitable bitrate by the receiving device, which can be a laptop, set-top boxes, or mobile phone and selects the most suitable chunk based on the device capabilities and CPU status.

▸ *Google WebM* is the Google HTTP-based solution for content adaptation proposed in 2010 to provide an OTT solution that is royalty-free. It uses VP8 video codec for video and Vorbis for audio and does not require segmentation of the media into chunks. However, one media stream is seen as one file. To stream a video through WebM, the following procedure takes places. Encoding the video and audio content in VP8 and Vorbis respectively, in different bitrates (such as, quality profiles). Multiplexing them into a WebM file, which must be automatically refreshed in real time in the case of a live video. Using a Web server (origin or cache) to deliver the WebM files. The adaptive bitrate process mainly relies on the server that selects the audio/video streaming bitrates before multiplexing and pushes in an output buffer all the content packets ready to be sent. While sending the buffer content to the network, the server detects if there is enough bandwidth toward the client, otherwise it scales down to a lower-quality profile (lower bitrate).

▸ *Microsoft Smooth Streaming* is the protocol released by Microsoft in 2009 as an extension of Silverlight 3.0,[a] an application for writing and running rich Internet applications. It supports the H.264 and AAC codecs respectively for video and audio. The general principle behind Smooth Streaming is quite similar to HLS streaming: that is, encoding video and audio in different bitrates (such as quality profiles); using a Stream Segmenter to generate content fragments (chunks) and multiplexing them into a container; distributing video content through an HTTP server (origin or cache); generating and distributing a manifest file that lists the available profiles (bitrates, resolution), languages, and corresponding URLs for chunks.

Once the client receives the manifest, it is able to request some indexed fragments according to its environment (available bandwidth, screen resolution, supported codec). Like HLS, Smooth Streaming leaves the client the choice of the relevant bitrate.

▸ *Adobe HTTP Dynamic Streaming (HDS).* Adobe's solution for streaming media over HTTP is a comprehensive open source video delivery. The Adobe HDS principle is similar to Microsoft Smooth Streaming. The differences, however, include: creation of manifest files (.f4m); creation of segmented files (.f4f) that correspond to chunks (fragments); creation of index files (.f4x) containing specific information about the fragments inside the segmented files (available bitrates, codecs, URLs to stream content). All these files are multiplexed into a single stream and sent to the client device. The supported codecs for HDS are H.264 and VP6 (video), and AAC or MP3 (audio). The terminal in the manifest file has many quality choices and selects the most suitable.

▸ *Dynamic Adaptive Streaming over HTTP (DASH).* MPEG DASH is a promising ISO Standard for video streaming services over HTTP published in April 2012 and is gaining popularity.[18] DASH has the potential to replace existing proprietary technologies like Microsoft Smooth Streaming, Adobe Dynamic Streaming, and Apple's HLS. A unified standard is needed because it will help rationalize cost storage, development, maintenance, support, and evolution for all DASH devices. All HTTP-based adaptive streaming technologies have two components: the pure encoded A/V streams, and manifest files that indicate to the player which streams are available (bitrates, codecs, resolutions, among others) and how to access them (for example, chunk URL). For DASH, the AV streams are called the Media Presentation, while the manifest file is called the Media Presentation Description that is encoded in an XML format.

Like other adaptive streaming techniques, this manifest identifies alternative streams, their respective URLs, network bandwidth, and CPU utilization. On this basis, the player chooses the most adapted stream. Two types of file segment types are allowed in DASH: MPEG2 TS (currently used by HLS), and ISO Base media file format (ISO BMFF, currently used by Smooth Streaming and HDS). This simplifies potential migration of existing adaptive streaming platforms to MPEG DASH, as the media segments can often remain the same, and only the index files need to be migrated to the Media Presentation Description (MPD) format.

Adaptive streaming techniques not based on HTTP. RTSP,[b] RTMP,[c] and MMS[d] streaming techniques are not HTTP based and the adaptation (if it is enabled) is managed by the server following a server-centric approach. In RTSP, the content adaptation can take place making use of the Real Time Control Protocol (RTCP) reports sent between the clients to the server. These reports contain information such as packet loss, jitter, RTT (measured/estimated at the client side) that can help the server adapt the content to network conditions. For example, if the connection deteriorates and the transfer rate decreases, the content is streamed with a lower quality so that playback interruptions are avoided, and stream quality is increased if the connection becomes more fluid.

An advantage of this type of streaming technique is the fast start ability, that is, to start content streaming without delay. On the other hand, the drawback mainly lies in the need of a dedicated server with non-negligible license cost (example are Xiph, Icecast, Real Helix Streaming Server, Windows Media Services, Adobe Flash Media Server, QuickTime Streaming Server, and so on) and in difficulties of passing through firewalls. Consequently, this type of streaming is more adapted for users who are connected in managed networks where operators/content providers have a mastery of end-to-end links and application needs in terms of resource consumption. Then they can anticipate the network congestion because most of these streaming techniques use the User Datagram Protocol (UDP) and this transport protocol does not retransmit lost data. Here, we discuss these streaming techniques.

Audiovisual delivery based on RTSP: RTSP was developed by Real Networks, Netscape, and Columbia University. It is available on the products of these companies. The RTSP is a network control protocol designed for use in entertainment and communications systems to control media streaming. The stream consists of cutting the data into packets (fragments) whose size is adapted to the available bandwidth between the client and the server. When the client receives enough packets (buffering), the client application begins to play a package.

Audiovisual delivery based on RTMP: Initially a proprietary protocol developed by Macromedia, RTMP

a http://www.microsoft.com/silverlight/

b http://www.cse.wustl.edu/~jain/papers/ftp/rtp.pdf

c http://www.adobe.com/devnet/rtmp.html

d http://msdn.microsoft.com/en-us/library/cc251059.aspx

Table 1. Adaptive streaming and non-HTTP adaptive streaming.

TECHNIQUES PARAMETERS	HTTP ADAPTIVE STREAMING				NON-HTTP ADAPTIVE STREAMING		
Different methods	HLS	Smooth streaming	HDS	WEB M	RTSP	MMS	RTMP
User or server centric	User centric	User centric	User centric	Server centric	Server centric	Server centric	Server centric
Standard player	iOS for mobile and QuickTime for PC	Silverlight for PC/Xbox360/ WinMobile7	Flash Player10.1	Chrome navigator	Depend on server that implemented the solution	Windows media player	Adobe Flashplayer
Origin server	Web server HTTP 1.1	IISv7	Flash Media Server 3.5	Web server HTTP 1.1	Streaming server	Windows media server	Adobe Flash server
Recommended chunk duration	10s	2s (fixed)	2–4s	No chunk	No chunk	No chunk	No chunk
Standard content protection	AES	Playready	Adobe Flash Access 2	No protection	CAS	CAS	CAS
Proprietary or not	Apple propriety	Microsoft propriety	Adobe propriety	Google propriety	Standardized by IETF	Microsoft propriety	Public use
Video/audio codecs	H.264 (video) and AAC (audio)	H.264 for Video and AAC (Audio)	H.264/VP6 (video) and AAC/ MP3 audio	VP8 (video) and vorbis(audio)	Depend on server that implemented the solution	Windows media video and Windows media audio	H264 (video) and AAC (audio)
Advantages	Adapted to bandwidth variation, the user application manages the client bitrate, fast content switching The required resource is a HTTP server, Firewalls/NATs traversal			Royalty free and open solution	Adapted to real time. It can be easily controlled by the operator as it is a server-based approach.		It is in public use
Disadvantages	The operator has no control over bandwidth consumption as it is fully client-centric; that is, the client switches from one flow to the other if network conditions allow it..			The client expectations are not taken into account; it is the server that manages the bitrate.	Packet losses cause artifacts. Dedicated server is required		

AAC: Audio Advanced Coding, CAS: Conditional Access System, AES: Advanced Encryption Standard

Table 2. Defined profiles based on access type.

Device	iPhone		iPad	
Access type	3G	Wi-Fi	3G	Wi-Fi
Encoding rate	150kbps	280kbps	380kbps	1Mbps
Frame rate	15fps	25fps	15fps	25fps
Codec	H264 BP	H264 BP	H264 BP	H264 BP
Resolution	QVGA	HVGA	QVGA	HVGA

is now Adobe proprietary and free to use. This protocol is used for streaming audio, video, and data over the Internet, between Adobe player (Flash) and a server. Audiovisuals delivery based on MMS, Microsoft's proprietary network streaming protocol used to transfer data in Windows Media Services. MMS can be transported via UDP or TCP.

Table 1 compares the advantages and disadvantages of streaming techniques through the standard player and the origin server.

Content Adaptation Means for Enhanced Delivery

Here, we describe some means of performing content adaptation for mobile TV, IPTV, and WebTV services.[14]

Content delivery adaptation in mobile TV. The most common way to adapt content in mobile TV relies on using different encoding profiles according to the Radio Access Technology (or RAT, such as EDGE, UMTS, HSPA, LTE, or Wi-Fi) and to the client's device (form factor, OS, hardware specificities, screen resolution). While device information can be obtained indirectly through the standard User-Agent HTTP header, an enrichment process adds an extra header containing the client's identity (MSISDN) and its RAT information. Generally, most of the operators are constrained to use different video profiles for addressing software, hardware, and legal limitations (profiles for 3GPP mobiles, HLS, Windows media). Each TV channel is coded in a different profile and each profile has its own characteristic (video encoding rate, resolution, video codec, frame rate, audio encoding rate, audio codec, which corresponds to different quality level. For example the following video profiles could be proposed as a function of the device and the access type.

Table 2 illustrates fictive profiles that depend on the terminal screen and the access type.

Content delivery adaptation in WebTV and IPTV. For IPTV Live and Video on Demand (VoD) service, the user generally has two choices: Standard Definition (SD) streams and High-Definition (HD) streams. Nevertheless, the operator may enforce a given content quality based on the available network bandwidth and on the end user's subscription type. WebTV contains TV and/or VoD services offered by a third party available from any Internet access. This method is thus by default available on unmanaged networks, where the best-effort is the unique possible QoS traffic class. Some content providers like French TF1 or France 24 adapt content to the network conditions by using HTTP Adaptive Streaming for their live TV channels. With this technology, users do not care about whatever video quality to choose in order to match their available bandwidth: the video player will automatically request content pieces that are the most adapted to its network status and the device capacity. On the contrary, some OTT content providers for the VoD let the client choose the type of delivery as follows: Streamed mode available in SD (for example, 620kbps stream) for "Instant Viewing;" Progressive Download mode available for both HD (for example, 1500kbps/1GB/WMV file); and in SD (for example, 620kbps/450MB/WMV file), with a possible non-negligible start-up delay for HD that depends on the client's bandwidth.

Limitation of content adaptation means in current deployment. The current content adaptation for mobile TV, IPTV, and Web TV does not consider a sufficiently large set of parameters to fully enable optimal QoS and QoE. The user context is considered in a limited manner through mainly considering the characteristics of the device used and the network used. The user's localization (physical/geographical location, but also location in terms of proximity to the access network entities) and the user's consumption style are not considered. In addition, network context is considered only in terms of bandwidth availability while ignoring the cost of using this bandwidth instead of allocating it to monetized services. Neither is considered the matching degree of the content to the users' preferences. Consequently, context awareness and user centricity need more consideration in content adaptation through studying users' localization including their physical location and in terms of proximity to the access network entities; or considering the users' consumption style for a given type of video service through, for instance, a trade-off between start-up delay and quality content. For example, a fast start is the most important factor for news and live streams, while content quality is the most important one for a movie.

The current content adaptation for mobile TV, IPTV, and Web TV does not consider a sufficiently large set of parameters to fully enable optimal QoS and QoE.

Related Research Contributions

The classification of existing research contributions on content adaptation show three main categories: *Content adaptation*: Which version of a given content shall we transmit? (codec, bit rate, video resolution). This aspect is related to the encoding of the content information. *Content delivery adaptation*: How the content is delivered through the network (unicast, multicast, from which server(s), from a cloud, through which access network)? This aspect is related to the service and network aspect of the content transmission. And *adaptation of content and delivery*: Integrating adaptation of both content and its delivery.

Content adaptation. Several research contributions exist about adapting content based on terminal capacity, network congestion, user profile, and service requirements. For instance, the method in Kim et al.[6] provides a QoE-guaranteed service that maximizes the visual expectation of the viewer by considering the size of the LCD panel size on his device. In Mohan and Agarwal,[12] users are allowed to define their preferences (user profile) during service subscription, according to some categories based on QoS requirements (streaming, conversational, interactive, background). For example, the streaming traffic class is sensitive to packets losses. The user can also select different types of subscription (Bronze, Silver, Gold, Platinum) for each profile and traffic class. There are maximum and minimum QoS parameters where an adaptation is needed for each type of service and user profile.

The method introduced in Koo et al.[8] adjusts the quality level and transmission rate of video streaming on the basis of the wireless channel status (Modulation Coding Scheme, Signal to Interference-Ratio level), location, and client buffer status. The transmission rate is determined as a function of the network context (for example, packet loss, jitter) and some player buffer ratios. In Guo et al.,[8] the adaptation of the transmission rate is done on the basis of the pre-buffering time and of the available bandwidth (network status/context), so that QoE is maximized even in case of network congestion.

Khan et al.[7] adapts content using two parameters: the "congestion" (C) and the "degradation" (D). The congestion is defined as the fraction of the number of video blocks lost (BL) divided by the total number of video blocks sent (BS) within an interval of time. After predicting the estimated quality of experience (denoted MOSt), the degradation is defined as the difference between the maximum achievable Mean Opinion Score (MOS) and the estimated MOS (MOSt). The Sender Bit Rate (SBR) is computed by on an algorithm using congestion and degradation. W3C proposed by Bakhtiar,[1] content adaptation techniques within the Composites Capabilities Preferences Profiles (CC/PP) for Web content and User Agent Profiles (UAProf) for mobile phones. These frameworks can be used to deliver devices contexts (screen size, audio/video capabilities) and users preferences (language, type of content) and allow devices to communicate their capabilities and preferences to servers. The server can then accurately adapt content according to this information.

Adaptation of content delivery. Research contributions on the adaptation of content delivery adaptation focus on two main approaches: the network-centric approach, in which decisions are made at the network side (mainly by network operators) and principally based on the network operator's benefits; and the user-centric approach, where the decision is based on the user's benefit, generally, without considering network load-balancing or other users. It should be noted that the user-centric approach has the main drawback from a load balancing perspective, since users generally consider only their own benefit while making decisions. This could result in bad performance of the overall network and service.[2] Ksentini et al.[9] considers QoE measurements over different access types. After predicting a MOS with Pseudo Subjective Quality Assessment (PSQA), a vertical handover (change in access network and/or technology) is carried out toward the network offering the best MOS; that is, the best QoE.

A new algorithm is defined in Ciubotaru and Muntean[3] called Smooth Adaptive Soft-Handover Algorithm (SASHA). Its goal is to improve the user-perceived quality while roaming through a heterogeneous wireless network environment. The score of each connection is evaluated based on a comprehensive Quality of Multimedia Streaming (QMS) score including the following metrics: QoS, QoE, cost, power efficiency, and user preferences. The idea is to adapt delivery in the network that has the best QMS score. Each metric is weighted with a proper coefficient.

Adaptation of content and delivery means. Pichon et al.[17] examine the most suitable content to be delivered to the user and selects the best delivery mean. Two decision entities are considered, namely the *service manager* responsible for the service delivery, and the *mobility manager* responsible for the network connectivity. For content adaptation, the service management entity will be notified when a terminal requests the streaming of a new video (contents encoding is done with SVC), and decides which version should be sent according to the user rights, to his preferences, to his terminal capabilities, and to the network congestion. The service management entity then provides its decision to the service execution entity that sends the corresponding signalization. For content delivery adaptation, the mobility manager is notified about the network-related events and service requirements and retrieves network-related information and decides which possible

A new algorithm is called Smooth Adaptive Soft-Handover Algorithm. Its goal is to improve the user-perceived quality while roaming through a heterogeneous wireless network environment.

network connection(s) must be used for every service based on information such as cost, network load, and user preferences.

Comparison of content adaptation and content delivery adaptation techniques. Here, we provide a comparative study regarding different issues that should be addressed in content adaptation techniques, mainly considering user context, user satisfaction, network congestion, and required resources. There are a lot of research contributions on content adaptation and its delivery. We have therefore classified them into several categories:

Terminal capacity has a lot of advantages among which we can mention the consideration of user context by using the terminal capacity. The disadvantage of this method is not considering the dynamic variation of user's needs, network resources optimization, and user satisfaction.

Network congestion. The main advantage of this method is the network resources optimization. The lack is no consideration of others context information like user context (his location, his preferences, his profile), user measured quality of experience.

User profile and service requirements. The advantage is the consideration of user context by reviewing his profile. This adaptation technique is easy to implement because the user profile and service needs are known by the server platform. The disadvantage is the no consideration of user context, network congestion in the adaptation process.

Network congestion and terminal capacity considers some context information like terminal capacity and aids resource optimization. The needed resource is a centralized server for gathering network status and terminal feedback. The user location, the measured QoE, is not considered in the adaptation technique.

Network congestion and measure QoE. This type of adaptation considers the user satisfaction (QoE) and optimizes the network use. Some context information is missed in the adaptation technique like user location or terminal capacity. In the literature we notice some limitations in the existing work as follows: The method in Kim et al.[6] does not consider the dynamic variation of users' needs and the network resources optimization. The solution presented by Mohan and Agarwal[12] could not adequately enhance the user's experience since the media source is not aware of the user context (terminal capacity, user preferences, and localization). The method presented by Koo and Chung[8] can be difficult to execute because it is not easy to ask each user to implement his profile when he subscribes to a service. Some important context information is missed in the proposed method by Guo et al.,[5] like the user localization and terminal capacity.

A general notice is the lack of contributions considering the adaptation of both the content and their delivery means.

Conclusion

This article surveyed content adaptation techniques considering the application-level adaptation and the adaptation of the delivery methods. Firstly, a review on video streaming techniques was given comparing them through several operational, performance, and user-centric criteria. Secondly, the content delivery means were reviewed and compared considering both operational solutions and research contributions.

We notice several research contributions on content adaptation based on terminal capacity, user preferences, network congestion, and so on; however there is a lack of fully contextual methods making use of context information on the terminals, network, and users. In addition, there are limited contributions considering adaptation of both contents and their delivery means. Consequently, we solicit the need to consider the user context (environment, location), terminal context, network context, content context for better content adaptation, and optimized content delivery aiming to improve the user's satisfaction. C

References

1. Bakhtiar, B. Video Content Adaptation Based on User Preferences and Network Bandwidth. Thesis report, (2007).
2. Ciubotaru, B. and Muntean, G. Quality of multimedia streaming-oriented handover management solution for mobile application. *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting.* (May 2009).
3. Ciubotaru, B. and Muntean, G.-M. SASHA—A quality-oriented handover algorithm for multimedia content delivery to mobile user (June 2009).
4. Diallo, M.Y., Moustafa, H., Afifi, H. and Laghari, K. Quality of experience for audiovisual services. EuroITV 2012 (July 2011).
5. Guo, T., Cai, J. and Foh, C.H. Scalable video adaptation in wireless home networks with a mixture of IPTV and VoD users. In *Proceedings of Globecom 2011*. IEEE.
6. Kim, J., Um, T-W, Ryu, W. and Lee, B.S. Heterogeneous networks and terminal-aware QoS/QoE-guaranteed mobile IPTV service. *IEEE Communications* (May 2008).
7. Khan, A., Sun, L., Jammeh, E. and Ifeachor, E. Quality of experience-driven adaptation scheme for video applications over wireless networks. *IEEE Communications* (July 2010).
8. Koo, J. and Chung, K. MARC: Adaptive rate control scheme for improving the QoE of streaming services in mobile broadband networks. In *Proceedings of the 2010 International Symposium on Communications and Information Technologies* (Oct. 26–29, 2010). IEEE.
9. Ksentini, A., Viho, C. and Bonnin, J.-M. QoE-aware vertical handover in wireless heterogeneous networks. In *Proceedings of the 7th Annual Wireless Communications and Mobile Computing Conference* (2011). IEEE.
10. Laakko, T. and Hiltunen, T. Adapting Web content to mobile user agents. *IEEE Internet Computing* (Mar. Apr. 2005).
11. Mahan, R. Smith, J.R., Li, C-S. Adapting multimedia Internet content for universal access. *IEEE Transactions on Multimedia* (Mar. 1999).
12. Mohan, S. and Agarwal, N. A convergent framework for QoS-driven social media content delivery over mobile networks. In *Proceedings of the 2nd Annual Conference on Wireless VITAE* (Feb. 28-Mar. 3, 2011). IEEE.
13. Moustafa, H. and Maréchal, N. and Zeadally, S. Mobile multimedia applications delivery technologies. *IT Professional 14*, 5 (Sept/Oct. 2012). IEEE Computer Society.
14. Moustafa, H. and Zeadally, S., Eds. *Media Networks: Architectures, Applications and Standards.* CRC Press, May 2012.
15. OTT Streaming, 2nd Edition. White Paper. (Sept. 2011); http://www.anevia.com/IMG/pdf/Anevia_White-Paper_OTT-Streaming_2nd_Edition.pdf /
16. Oyman, O. and Singh, S. Quality of experience for HTTP adaptive streaming services. *IEEE Communications* (Apr. 2012).
17. Pichon, D. Rieublandou, G., Bonnin, J.M. and Seite, P. Coordinated service and mobility management to maximize user QoE and optimize network resource use. In *Proceedings of the IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications* (2009).
18. 3GPP Specification TS 26.247. Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH) (Release 10).
19. Wenger, S., Hanuksela, M., Stockhammer, T., Westerlund, M. and Singer, D. RTP payload format for H.264 video. RFC 3984, IETF (2004).

Mamadou Tourad Diallo (mamadoutourad.diallo@orange.com) is working on his Ph.D. within Orange Labs, Issy-les-Moulineaux, France. His thesis work concerns new video content and delivery adaptation.

Hassnaa Moustafa (hassnaa.moustafa@intel.com) is a Research Scientist at Intel Labs, Intel Corp., Hillsboro, OR, where she works in video delivery optimization over wireless networks. The work described here occurred during her stay at Orange Labs, Issy-les-Moulineaux, France.

Hossam Afifi (hossam.afifi@mines-telecom.fr) is a professor at the Institut Mines Telecom, Saclay, France, where he works on multimedia services and applications.

Nicolas Marechal (nicolas.marechal@orange.com) is a project manager and an R&D engineer at Orange Labs, Issy-les-Moulineaux, France. His areas of interest include the optimization of content delivery over a mobile environment.

Distinguished Speakers Program

talks by and with technology leaders and innovators

Chapters • Colleges and Universities • Corporations • Agencies • Event Planners

A great speaker can make the difference between a good event and a WOW event!

The Association for Computing Machinery (ACM), the world's largest educational and scientific computing society, now provides colleges and universities, corporations, event and conference planners, and agencies – in addition to ACM local Chapters – with direct access to top technology leaders and innovators from nearly every sector of the computing industry.

Book the speaker for your next event through the ACM Distinguished Speakers Program (DSP) and deliver compelling and insightful content to your audience. **ACM will cover the cost of transporation for the speaker to travel to your event.** Our program features renowned thought leaders in academia, industry and government speaking about the most important topics in the computing and IT world today. Our booking process is simple and convenient. Please visit us at: **www.dsp.acm.org**. If you have questions, please send them to **acmdsp@acm.org**.

The ACM Distinguished Speakers Program is an excellent solution for:

Corporations Educate your technical staff, ramp up the knowledge of your team, and give your employees the opportunity to have their questions answered by experts in their field.

Colleges and Universities Expand the knowledge base of your students with exciting lectures and the chance to engage with a computing professional in their desired field of expertise.

Event and Conference Planners Use the ACM DSP to help find compelling speakers for your next conference and reduce your costs in the process.

ACM Local Chapters Boost attendance at your meetings with live talks by DSP speakers and keep your chapter members informed of the latest industry findings.

Captivating Speakers from Exceptional Companies, Colleges and Universities

DSP speakers represent a broad range of companies, colleges and universities, including:

- IBM
- Microsoft
- BBN Technologies
- Raytheon
- Sony Pictures
- McGill University
- Tsinghua University
- UCLA
- Georgia Tech
- Carnegie Mellon University
- Stanford University
- University of Pennsylvania
- University of British Columbia
- Siemens Information Systems Bangalore
- Lawrence Livermore National Laboratory
- National Institute of Standards and Technology

Topics for Every Interest

Over 400 lectures are available from 120 different speakers with topics covering:

- Software
- Cloud and Delivery Methods
- Emerging Technologies
- Engineering
- Web Topics
- Computer Systems
- Open Source
- Game Development
- Career-Related Topics
- Science and Computing
- Artificial Intelligence
- Mobile Computing
- Computer Graphics, Visualization and Interactive Techniques
- High Performance Computing
- Human Computer Interaction

Exceptional Quality Is Our Standard

The same ACM you know from our world-class Digital Library, magazines and journals is now putting the affordable and flexible Distinguished Speaker Program within reach of the computing community.

DOI:10.1145/2524713.2524724

Technical Perspective
Centip3De Demonstrates More than Moore...

By Shekhar Borkar

MOORE'S LAW WILL continue to provide exponentially increasing transistor integration capacity to integrate diverse functions in a system, enabling unprecedented compute capability, and making it ubiquitous to further enrich our lifestyles. However, the same physics that made this possible also pose some barriers, namely energy consumption and limitations of system interconnections. Supply voltage reduction has slowed down for numerous reasons, resulting in higher energy with higher levels of transistor integration. Higher integration also demands more interconnections, which have started hitting the fundamental limits of physics.

The Centip3De design is truly a marvel of an energy-efficient system demonstrating two key concepts to address these challenges, namely near-threshold computing (NTC), and 3D integration.

NTC calls for reducing the supply voltage, reducing the frequency of operation to improve energy efficiency—a little counterintuitive. But if you think about it, the frequency reduces almost linearly while the energy reduces quadratically, providing substantial gain in energy efficiency. But, hey, the critics would say, NTC would employ substantially more transistors in a system to provide the same throughput! That is exactly what Moore's Law and transistor integration capacity provides; if you do not use it with NTC then you will not get to use it because of energy consumption. Employing NTC may sound simple, but it needs careful system optimization. Supply voltage reduction reduces active energy, but the leakage energy starts becoming a substantial portion of the total energy, thus requiring a detailed system-level analysis and architecture conducive for NTC.

3D integration is realized by stacking thinned chips and interconnecting them with vias (thin insulated vertical wires); integrating chips in the third dimension, and hence the name. This 3D integration substantially reduces interconnect length, decrease signal delay, and also saves interconnect energy. Additionally, this integration allows you to combine a system with dissimilar technologies, such as DRAM, SRAM, and logic chips in one system stack—in stark contrast to integrating all these technologies into a single process that is prohibitively expensive. Once again, the system architecture and the design must comprehend 3D integration from day one.

The Centip3De system described in the following paper implements a many-core system, architected with 3D integration in mind, incorporating processor chips, cache chips, and DRAM chips into a 3D cube—all designed for NTC operation, demonstrating a wide operating range. Clearly, this is a novel approach to system design with two promising technologies, and I would not be surprised to see it catch on quickly!

The Centip3De system implements a many-core system, architected with 3D integration in mind, incorporating processor chips, cache chips, and DRAM chips into a 3D cube—all designed for NTC operation.

Shekhar Borkar is an Intel Fellow, an IEEE Fellow, and director of extreme scale research at Intel Corp.

DOI:10.1145/2524713.2524725

Centip3De: A Many-Core Prototype Exploring 3D Integration and Near-Threshold Computing

By Ronald G. Dreslinski, David Fick, Bharan Giridhar, Gyouho Kim, Sangwon Seo, Matthew Fojtik, Sudhir Satpathy, Yoonmyung Lee, Daeyeon Kim, Nurrachman Liu, Michael Wieckowski, Gregory Chen, Dennis Sylvester, David Blaauw, and Trevor Mudge

Abstract

Process scaling has resulted in an exponential increase of the number of transistors available to designers. Meanwhile, global interconnect has not scaled nearly as well, because global wires scale only in one dimension instead of two, resulting in fewer, high-resistance routing tracks. This paper evaluates the use of three-dimensional (3D) integration to reduce global interconnect by adding multiple layers of silicon with vertical connections between them using through-silicon vias (TSVs). Because global interconnect can be millimeters long, and silicon layers tend to be only tens of microns thick in 3D stacked processes, the power and performance gains by using vertical interconnect can be substantial. To address the thermal issues that arise with 3D integration, this paper also evaluates the use of near-threshold computing—operating the system at a supply voltage just above the threshold voltage of the transistors.

Specifically, we will discuss the design and test of Centip3De, a large-scale 3D-stacked near-threshold chip multiprocessor. Centip3De uses Tezzaron's 3D stacking technology in conjunction with Global Foundries' 130nm process. The Centip3De design comprises 128 ARM Cortex-M3 cores and 256MB of integrated DRAM. Silicon measurements are presented for a 64-core version of the design.

1. INTRODUCTION

High-performance microprocessors now contain billions of devices per chip.[11] Meanwhile, global interconnect has not scaled nearly as well since global wires scale only in one dimension instead of two, resulting in fewer, high-resistance routing tracks. Because of this, and increasing design complexity, the industry is moving toward system-on-chip (SoC) designs or chip-multiprocessors (CMP). In these designs, many components share the same die, as opposed to one giant processor occupying it entirely. These systems typically include processing cores, memory controllers, video decoders, and other ASICs.

Three-dimensional (3D) integration reduces the global interconnect by adding multiple layers of silicon with vertical interconnect between them, typically in the form of through-silicon vias (TSVs). Because global interconnect can be millimeters long, and silicon layers tend to be only tens of microns thick in 3D stacked processes, the power and performance gains by using vertical interconnect can be substantial. Additional benefits include the ability to mix different process technologies (CMOS, bipolar, DRAM, Flash, optoelectronics, etc.) within the same die, and increased yield through "known good die" testing for each layer before integration.[8]

Recently, a DARPA program provided the opportunity to evaluate an emerging 3D stacking technology that allows logic layers to be stacked on top of a DRAM. As part of this project our group at the University of Michigan was given the chance to explore architectures that leverage the high-bandwidth, low-latency interface to DRAM afforded by 3D integration. This paper will discuss the resulting design and test of Centip3De, a large-scale CMP that was recently presented at HotChips'24 and ISSCC[5] (Figure 1). Centip3De uses Tezzaron's FaStack® 3D stacking technology[14] in conjunction with Global Foundries' 130 nm process. The Centip3De design comprises 128 ARM Cortex-M3 cores and 256MB of integrated DRAM. Silicon measurements are presented for a 64-core version of the design in Section 4 which show that Centip3De achieves a robust design that provides both energy-efficient (3930 DMIPS/W) performance for parallel applications and mechanisms to achieve substantial single-thread performance (100 DMIPS) for serial phases of code.

2. BACKGROUND

2.1. 3D integration

The Centip3De prototype uses Tezzaron's FaStack® technology that stacks wafers of silicon (as opposed to individual dies) using copper bonding.[14] A step-by-step process diagram is shown in Figure 2. Before each wafer pair is bonded, a layer of copper is deposited in a regular honeycomb pattern that is then thinned. Due to the regularity

The original version of this paper is entitled "Centip3De: A 3930 DMIPS/W Configurable Near-Threshold 3D Stacked System With 64 ARM Cortex-M3 Cores" and was published in the *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*. San Francisco, CA, February 2012, pp. 190–191.

Figure 1. Floorplan accurate artistic rendering. Centip3De includes seven layers, including two core layers, two cache layers, and three DRAM layers.

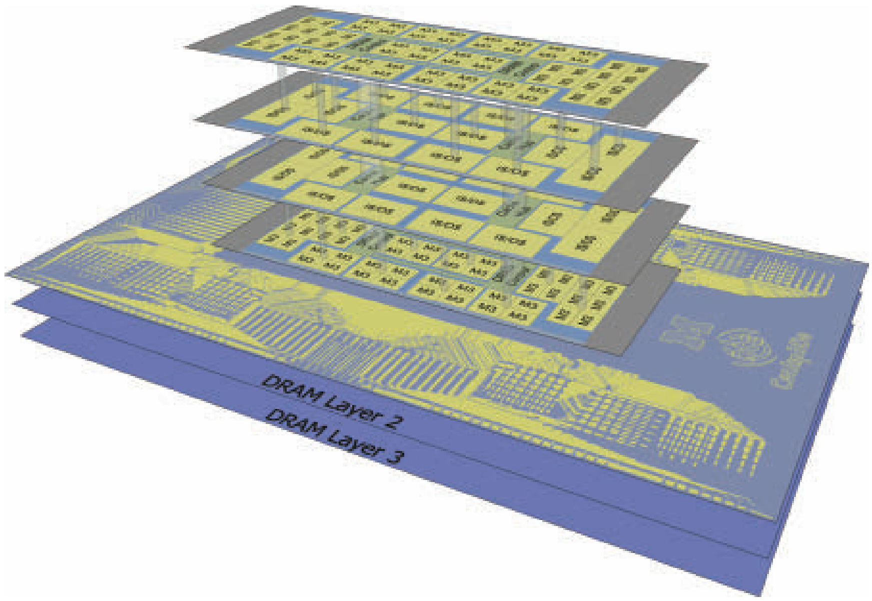
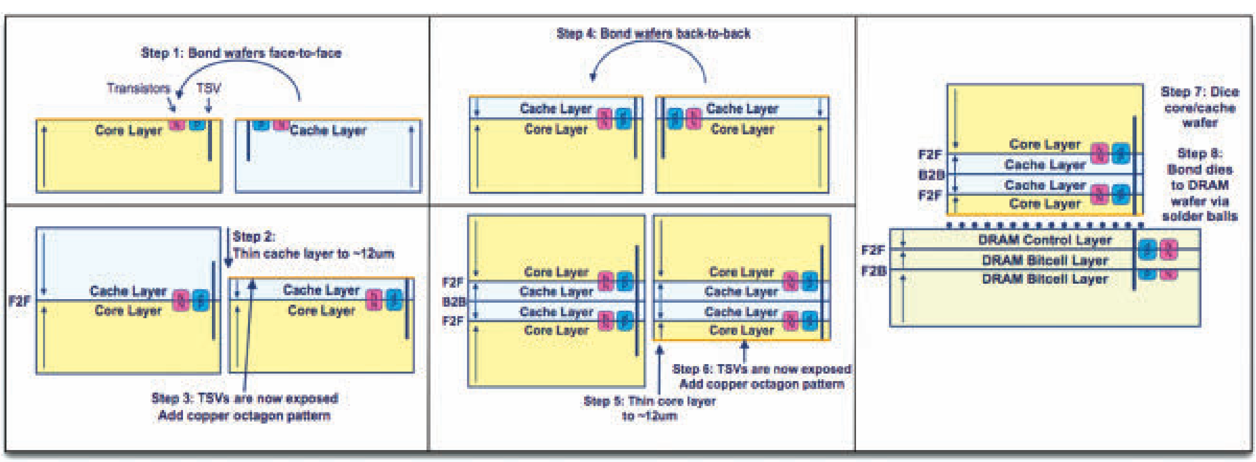


Figure 2. Centip3De's 3D stacking process. The core and cache wafers are bonded with a face-to-face (F2F) connection, then thinned to expose the TSVs. Next two pairs of wafers are bonded back to back (B2B) and thinned. Finally, the logic layers are diced and die-to-wafer bonded to the DRAM.



of the pattern, it is very flat after thinning, and when two wafers with this pattern are pressed together with heat, the copper forms a strong bond. After a bond, one side of the resulting wafer is thinned so that only a few microns of silicon remains, which exposes TSVs for 3D bonding, flip-chip, or wirebonding. The TSVs are made of Tungsten, which has thermal-mechanical properties suitable for creating vias. They are created after the transistors, but before the lowest metal layers, preventing the loss of any metal tracks. Since their diameter is small (1.2 μm), their length is short (around six microns after thinning), and they only couple to the bulk silicon, their parasitic resistance and capacitance is very small (about as much capacitance as a small gate). The finished wafer stack has a silicon pitch of approximately 13 microns with a per-interface TSV density of up to 160,000/mm^2. The combination of thin silicon layers, a high density of Tungsten TSVs, and planes of bonding copper also helps to maximize heat dissipation in this stacking technology.

A similar process is used to stack two dies of different sizes. In this design, the Tezzaron Octopus DRAM[15] is much larger than the core and cache layers. To stack these, the wafer of smaller dies is thinned, coated with copper, and then diced (Figure 2, step 7). The wafer of larger dies is also thinned and coated with copper. The smaller dies are put in a stencil to hold them in place and then are pressed together with the larger wafer to make the bond (step 8). A larger copper pattern than that used for the wafer-to-wafer bond is needed to support the less-precise alignment of die to wafer.

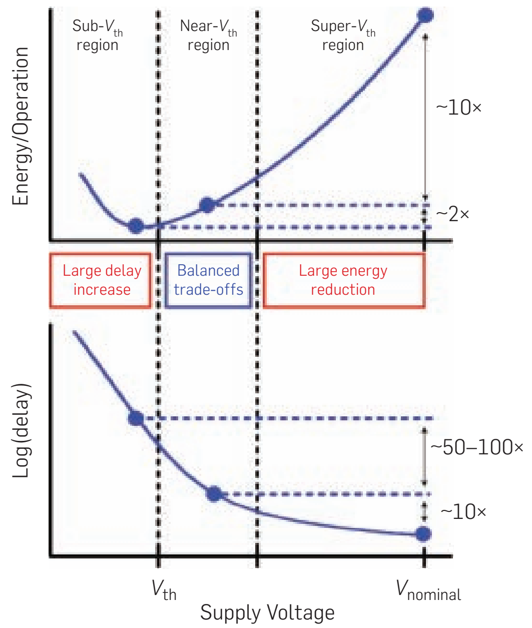
Alternative 3D and 2.5D techniques. The landscape of 3D stacking technologies is quite diverse and provides a wide range of possibilities. For the logic layers of Centip3De, we used the Tezzaron wafer-to-wafer 3D process, which provides unusually small TSVs on an extremely fine pitch (5 μm). This allows complicated designs to be implemented where even standard cells can be placed and routed across multiple layers within a synthesized block. However, this technology relies on wafer-to-wafer bonding for the logic layers which can result in yield issues for large runs. Alternative TSV technologies exist that rely on micro-bump die-to-wafer or die-to-die bonding which resolves some of the yield issues, but have a much larger TSV pitch. The die-to-wafer bonding was used by Centip3De to connect the logic layers to the DRAM layers. These types of 3D stacking are more useful for routing buses and interconnects between synthesized blocks, but are not ideal for 3D connections within synthesis blocks. A more near term integration solution is 2.5D technology that employs silicon interposers. This technique takes the micro-bump a step further by placing dies adjacent to each other and connecting them through an interposer. The 2.5D technology helps mitigate the thermal issues associated with 3D integration, at the expense of longer interconnects and lower TSV densities that are associated with micro-bumps.

2.2. Near-threshold computing (NTC)

Energy consumption in modern CMOS circuits largely results from the charging and discharging of internal node capacitances and can be reduced quadratically by lowering supply voltage (V_{dd}). As such, voltage scaling has become one of the more effective methods to reduce power consumption in commercial parts. It is well known that CMOS circuits function at very low voltages and remain functional even when V_{dd} drops below the threshold voltage (V_{th}). In 1972, Swanson and Meindl derived a theoretical lower limit on V_{dd} for functional operation, which has been approached in very simple test circuits.[6, 13] Since this time, there has been interest in subthreshold operation, initially for analog circuits[16] and more recently for digital processors,[12] that demonstrates operation with a V_{dd} below 200 mV. However, the lower bound on V_{dd} in commercial applications is typically set to 70% of the nominal V_{dd} to guarantee robustness without too significant a performance loss.

Given such wide voltage scaling potential, it is important to determine the V_{dd} at which the energy per operation (or instruction) is optimal. In the superthreshold regime ($V_{dd} > V_{th}$), energy is highly sensitive to V_{dd} due to the quadratic scaling of switching energy with V_{dd}. Hence voltage scaling down to the near-threshold regime ($V_{dd} \sim V_{th}$) yields an energy reduction on the order of 10× at the expense of approximately 10× performance degradation, as seen in Figure 3. However, the dependence of energy on V_{dd} becomes more complex as voltage is scaled below V_{th}. In the subthreshold regime ($V_{dd} < V_{th}$), circuit delay

Figure 3. Energy and delay in different supply voltage operating regions.



increases exponentially with V_{dd}, causing leakage energy (the product of leakage current, V_{dd}, and delay) to increase in a near-exponential fashion. This rise in leakage energy eventually dominates any reduction in switching energy, creating an energy minimum, V_{opt}, seen in Figure 3.

The identification of an energy minimum led to interest in processors that operate at this energy optimal supply voltage[17, 19] (referred to as V_{min} and typically 250–350 mV). However, the energy minimum is relatively shallow. Energy typically reduces by only ~2× when V_{dd} is scaled from the near-threshold regime (400–500 mV) to the subthreshold regime, though delay rises by 50–100× over the same region. While acceptable in ultra-low energy sensor-based systems, this delay penalty is not acceptable for a broader set of applications. Hence, although introduced roughly 30 years ago, ultra-low voltage design remains confined to a small set of markets with little or no impact on mainstream semiconductor products. In contrast, the tradeoffs in the NTC region are much more attractive than those seen in the subthreshold design space and open up a wide variety of new applications for NTC systems.

2.3. NTC and 3D synergy

One of the biggest drawbacks to 3D integration is the thermal density. By stacking dies on top of each other, local hotspots may effect neighboring dies, and hotspots that occur on top of each other complicate cooling.[9] Compounding this problem further is the stagnation of supply-voltage scaling in newer technology nodes leading to a breakdown of Dennard scaling theory. As a result, thermal density in 2D chips is also increasing.[3] Both of these issues are exacerbated by the fact that these hotter logic dies are now stacked on top of DRAM, resulting in the need for more frequent refreshes. On the other hand, NTC operation suffers from reduced performance because of lower supply voltages.

However, there is an important synergy between these two technologies: (1) by using NTC, the thermal density is reduced allowing more dies to be stacked together as well as on top of DRAM; and (2) by using 3D integration, additional cores can be added to provide more performance for parallel applications, helping to overcome the frequency reduction induced by NTC.

3. CENTIP3DE ARCHITECTURE

Centip3De is a large-scale CMP containing 128 ARM Cortex-M3 cores and 256MB of integrated DRAM. Centip3De was designed as part of a DARPA project to explore the use of 3D integration. We used this opportunity to explore the synergy with NTC as well. The system is organized into 32 computation clusters, each containing four ARM Cortex-M3 cores.[2] Each cluster of cores shares a 4-way 8KB data cache, a 4-way 1KB instruction cache, and local clock generators and synchronizers. The caches are connected through a 128-bit, 8-bus architecture to the 256MB Tezzaron Octopus DRAM,[15] which is divided into 8 banks, each of which has its own bus, memory interface, and arbiters. Figure 4 shows a block diagram for the architecture, with the blocks organized by layer. Centip3De also has an extensive clock architecture to facilitate voltage scaling.

3.1. Designing for NTC: A clustered approach

A key observation in NTC design is the relationship between activity factor and optimal energy point.[4] The energy consumed in a cycle has two main components: leakage energy and dynamic energy. At lower activity factors, the leakage energy plays a larger role in the total energy, resulting in higher V_{opt}. Figure 5 illustrates this effect in a 32 nm simulation, where activity factor was adjusted to mimic different components of a high performance processor. Memories in particular have a much lower activity factor and thus a much higher V_{opt} than core logic. Although NTC does not try to achieve energy-optimal operation, to maintain equivalent energy↔delay tradeoff points, the relative chosen NTC supply points should track with V_{opt}. Thus, V_{NTC_memory} should be chosen to be higher than V_{NTC_core} for them to maintain the same energy↔delay

Figure 4. High-level system architecture. Centip3De is organized into four-core clusters, which connect to eight DRAM buses. The diagram is organized by layer.

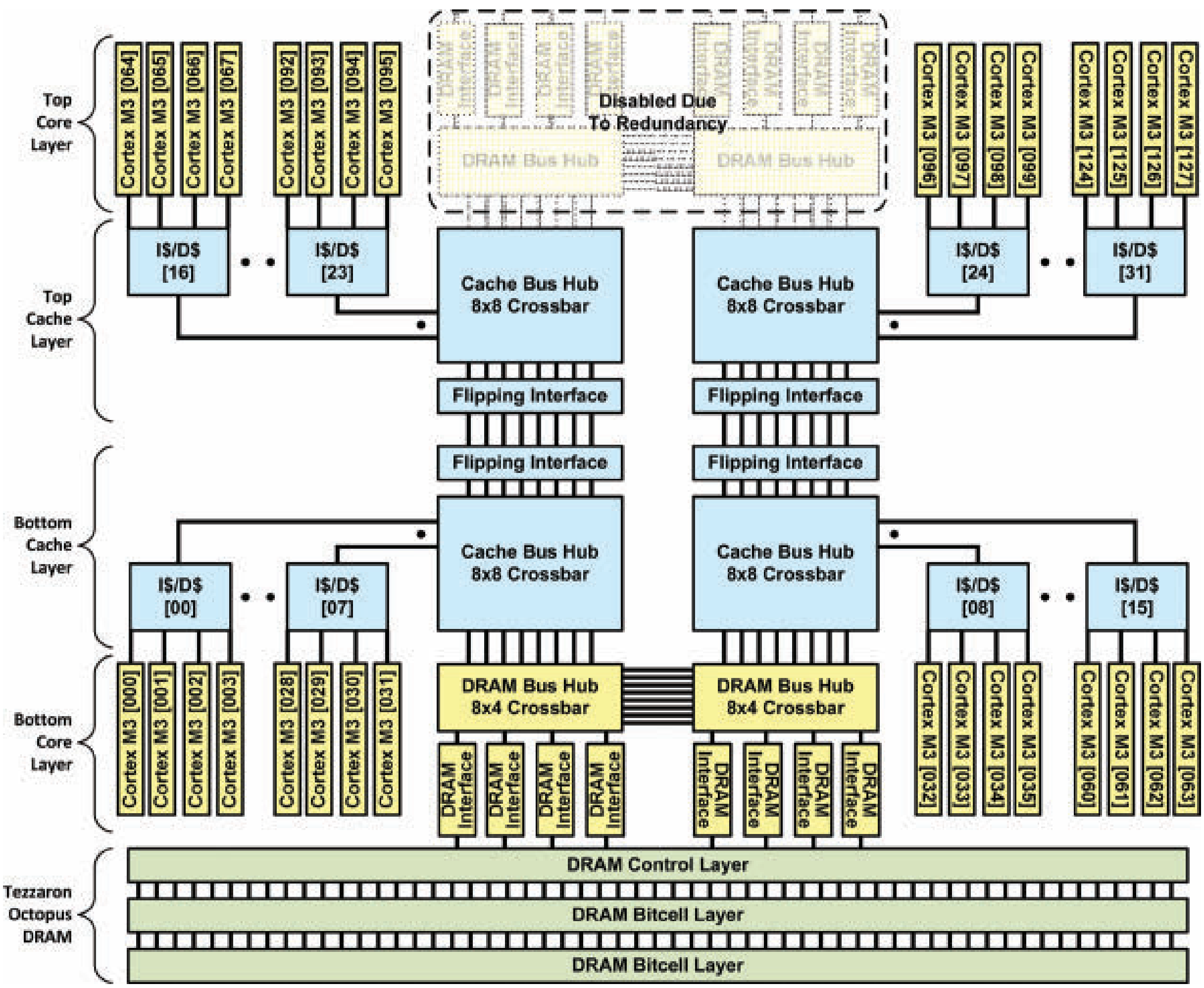


Figure 5. Activity factor versus minimum operating energy. As activity factor decreases, the leakage component of energy/operation increases thereby making the energy optimal operating voltage increase. To account for this, Centip3De operates caches at a higher voltage and frequency than the cores. An 8T SRAM design was used to ensure functionality and yield at V_{opt}.

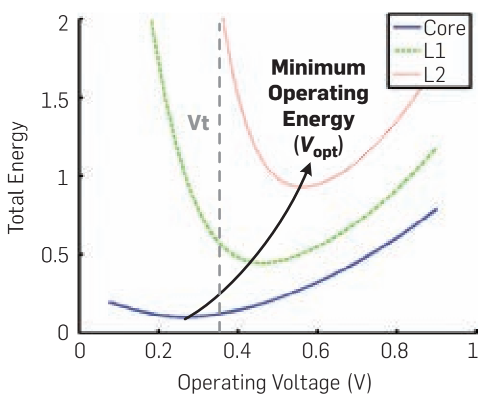
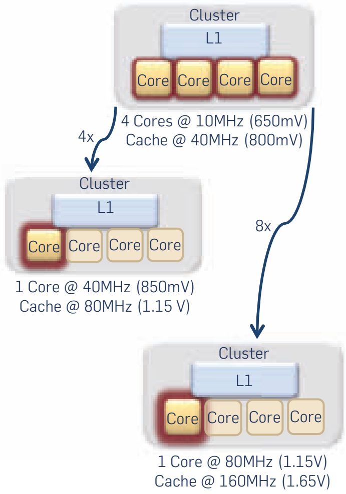


Figure 6. Boosting architecture. Diagram of how a cluster can be boosted to 4× and 8× single-thread performance by disabling some cores and increasing the voltage of the rest.



trade-off point. In addition, because V_{opt} is lower than the minimum functional voltage[a] for typical SRAM designs, a low-voltage tolerant memory design is necessary. This can be achieved by either specially designed 6T SRAM[18] or an 8T SRAM design.[10]

In Centip3De, we have used this observation to reorganize our CMP architecture. Instead of having many cores each with an independent cache, the cores have been organized into 4-core clusters and their aggregate cache space combined into a single, 4×-larger cache. This larger cache is then operated at a higher voltage and frequency to service all four cores simultaneously. To do this, the cores are operated out-of-phase and are serviced in a round robin fashion. Each core still sees a single-cycle interface and has access to a much larger cache space when necessary.

The NTC cluster architecture also provides mechanisms for addressing a key limiter in parallelization: intrinsically serial program sections. To accelerate these portions of the program, Centip3De uses per-cluster DVFS along with architecturally based boosting modes. With two cores of the cluster disabled, the cluster cache access pipeline can be reconfigured to access tag arrays and data arrays in parallel instead of sequentially and change from a 4× core↔cache frequency multiplier to a 2× multiplier. The remaining core(s) are voltage boosted and operate at 2× their original frequency, roughly doubling their single-threaded performance. A diagram of boosting-based approaches with their operating voltage and frequencies in the 130 nm Centip3De design is shown in Figure 6. The performance of the boosted cores is further improved by the fact that they access a larger cache and hence have a lower miss rate. To offset the related increase in cluster power and heat, other clusters can be disabled or their performance reduced.

[a] The operating voltage at which enough SRAM cells, in the presence of variation, function to meet yield constraints.

The Centip3De NTC cluster architecture has manufacturability benefits as well. By using lower voltages, many of the components have improved lifetime and reliability. Redundancy in the cluster-based architecture allows faulty clusters to be disabled and can be coupled with known good die techniques to further increase yield. Centip3De's boosting modes in combination with "silicon odometer" techniques improve performance while maintaining lifetime.[7]

Additional benefits of the NTC cluster architecture include reduced coherence traffic and simplified global routing. Coherence between the cores is intrinsically resolved in the cache while the top level memory architecture has 4× fewer leaves. Drawbacks include infrequent conflicts between processes causing data evictions and a much larger floorplan for the cache. In architectural simulation, however, we found that the data conflicts were not significant to performance for analyzed SPLASH2 benchmarks, and we effectively addressed floorplanning issues with 3D design. In particular, the core↔cache interface is tightly coupled and on the critical path of most processors and by splitting this interface to share several cores in a 2D design results in a slow interface.

To address the floorplan issues, we leverage the 3D interface to co-locate the cores directly above the caches and minimize the interconnect. The floorplan of the core↔cache 3D cluster is shown in Figure 7. The face-to-face connections from the cores appear in the center of the floorplan. This is particularly beneficial since the SRAMs use all five metal layers of this process, thereby causing significant routing congestion. By using 3D stacking, we estimate that cache routing resource requirements were reduced by approximately 30%. Because the parasitic capacitance and resistance of the F2F connections are small, the routing gains are not offset by significant loading (Figure 8).

Figure 7. Cluster floorplan with F2F connections represented as dots. There are 1591 F2F connections per cluster saving approximately 30% of routing and enabling the clustered architecture.

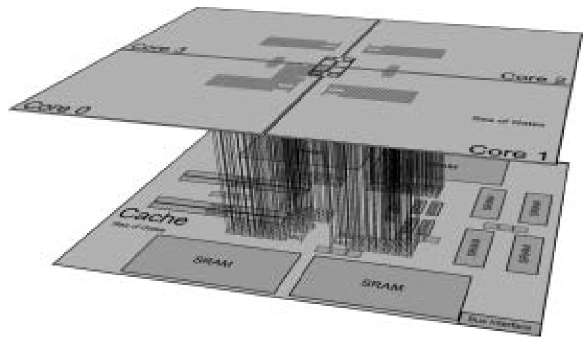
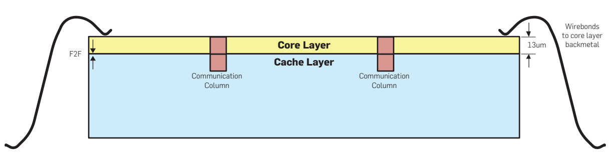


Figure 8. Side view of the measured system. The two-layer stack is formed by a face-to-face wafer bond and then the core layer is thinned to expose TSVs. Wirebonds are used to connect the TSVs to a test package. Both the face-to-face and TSV technologies are tested with this prototype.



4. MEASURED RESULTS

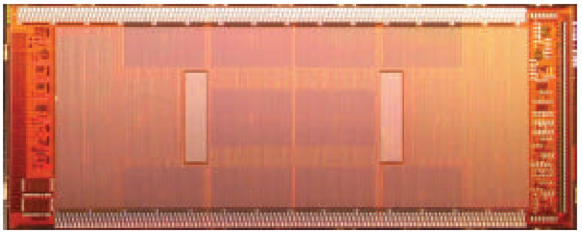
Silicon for Centip3De is coming back in three stages. Already received and tested is a two-layer system that has a core layer and a cache layer bonded face to face. This system is thinned on the core side to expose the TSVs, and an aluminum layer is added to create wirebonding pads that connect to them. A die micrograph is shown in Figure 9.

The next system will have four layers including two core layers and two cache layers. It will be formed by bonding two core–cache pairs face to face, thinning both pairs on the cache side to expose the TSVs, adding a copper layer, and then bonding back to back. One core layer is thinned to expose TSVs and same aluminum wirebonding pads are added.

The final system will include all seven layers, including two core layers, two cache layers, and three DRAM layers. The wirebonding sites will be on the DRAM, which has much larger layers than the core and cache layers.

Figure 10 shows silicon measurements and interpolated data for the fabricated 2-layer system for different cluster configurations running a Dhrystone test. The default NTC cluster (slow/slow) configuration operates with four cores at 10 MHz and caches at 40 MHz, achieving 3,930 DMIPS/W. Based on FO4 delay scaling, 10 MHz is projected to translate to 45 MHz in 45 nm SOI CMOS. Latency critical threads can operate in boosted modes at 8× higher frequency. One-core and two-core boosted modes provide the same throughput, 100 DMIPS/cluster (estimated as 450 in 45 nm), while enabling a tradeoff between latency and power (Figure 10). The system

Figure 9. Die micrograph of two-layer system. Two layers are bonded face to face. The clusters can be seen through the backside of the core layer silicon. Wirebonding pads line the top and bottom edges. The system consists of 16 clusters of 4 cores (64 total cores).



bus operates at 160–320 MHz, which supplies an ample 2.23–4.46GB/s memory bandwidth. The latency of the bus ranges from 1 core cycle latency for 10 MHz cores to 6 cycles when cores are boosted to 80 MHz. As a point of comparison, the ARM Cortex-A9 in a 40 nm process is able to achieve 8000 DMIPS/W.[1] At peak system efficiency, Centip3De achieves 3930 DMIPS/W in a much older 130 nm process, and when scaled to 45 nm Centip3De achieves 18,500 DMIPS/W (Table 1).

The results in Figure 10 present many operating points to select based on workload or workload phase characteristics. When efficiency is desired, choosing the slow core and slow memory results in the most efficient design. For computationally intensive workloads, additional throughput can be obtained, at the expense of power, by using the fast core and slow memory configuration. For memory bound workloads, the core can remain slow and the bus speed can be increased to provide more memory bandwidth. For workloads or phases that require higher single-thread performance (to address serial portions of code), the number of cores in a cluster can be reduced and the core speeds increased.

When boosting clusters within a fixed TDP environment, it may require disabling or down-boosting other clusters to compensate for that cluster's increase in power consumption. A package with a 250 mW TDP can support all 16 clusters in four-core mode (configuration 16/0/0/0, with the number of clusters in each mode designated as 4C/3C/2C/1C). Up to five clusters can be boosted to three-core mode (11/5/0/0) while remaining within the budget. To boost a cluster to one-core mode within the TDP, however, would require disabling other clusters, resulting in system configuration 9/0/0/1. By boosting clusters, Centip3De is able to efficiently adapt to processing requirements. Figure 11 shows a range of system configurations under a fixed TDP of 250 mW. On the left are high-efficiency configurations, with more aggressively boosted configurations on the right, which provide single-threaded performance when needed. Overall, the Centip3De design offers programmers with a wide range of power/throughput/single-thread performance points at which to operate.

5. DISCUSSION

In designing Centip3De we learned several lessons. First, supporting 3D LVS and DRC checking was a challenge

Figure 10. Power analysis of the 64-core system. Power breakdowns for 4-, 3-, 2-, and 1-core modes. Each mode has a slow (NTC) core option or a fast (boosted) option where the frequency/voltage is increased. Each option provides a trade-off in the efficiency/throughput/single-thread performance space. Overall, Centip3De achieves a best energy efficiency of 3930 DMIPS/W.

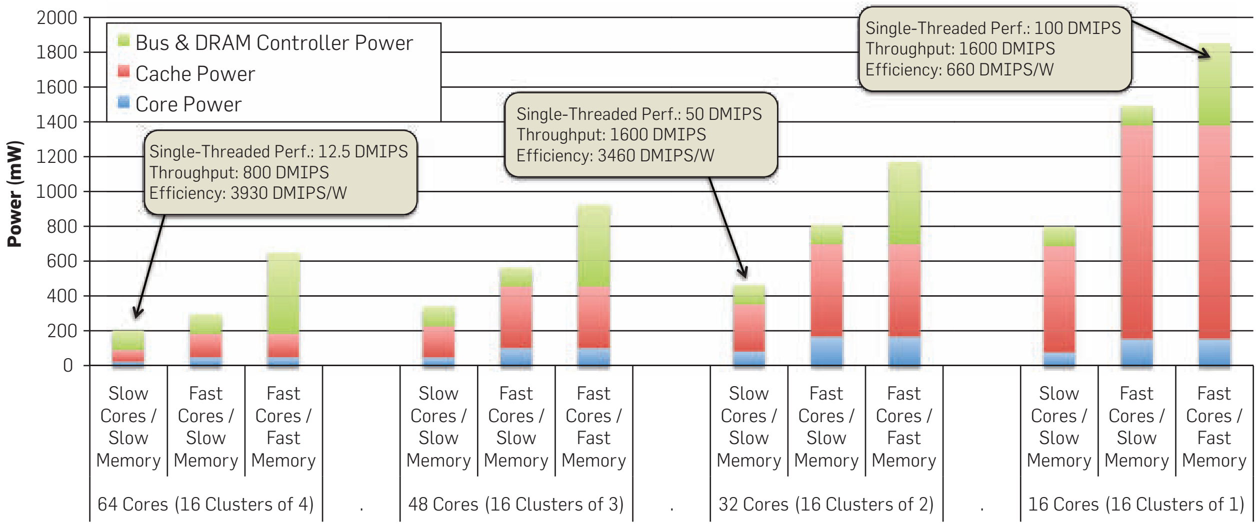
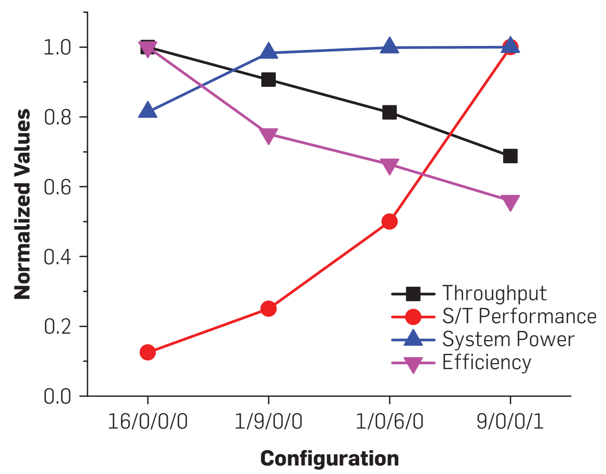


Table 1. Design and technology data. Connection numbers include only signals and do not include minimum density numbers for TSV or power/gnd connections. F2F connections are for a single F2F bonding interface.

Logic layer dimensions	2.66 × 5 mm
Technology	130 nm
Metal layers	5
Core layer devices	28.4 M
Cache layer devices	18.0 M
Core layer thickness	12 μm
F2F connection pitch	5 μm
F2F connections/cluster	1,591
Bus F2F connections	2,992
Total F2F connections	28,485
B2B connection pitch	5 μm
Total B2B connections (TSV)	3,024
DRAM connection pitch	25 μm
DRAM connections (TSV)	3,624

that required a large amount of design time to develop scripts for these operations. The good news is that in the two years since we first started Centip3De, EDA tools have made significant progress in supporting 3D integration. Second, when we designed Centip3De we were concerned with the amount of clock skew that would be introduced by TSVs, so we designed highly tunable clock delay generators as insurance against timing mismatch. However, measurements show that the skew through the TSVs was quite small. Spice simulations indicate that a significant amount of power (~30–60%) is being used in these delay generators, particularly in NTC mode (~60%). Unfortunately, we were unable to subtract the unnecessary power these delay generators consume, because they were not on their own supply rail. If we were able to reduce that power using less tunable delay generators, we expect the efficiency would be far better at NTC than

Figure 11. Range of system configurations under a 250 mW fixed TDP. The number of clusters in each mode is listed as 4-core/3-core/2-core/1-core. Each configuration emphasizes a different cluster mode, with the most energy-efficient configurations on the left, and the highest single-threaded performance on the right.



we observed, achieving ~9900 DMIPS/W in 130 nm and ~46,000 DMIPS/W when scaled to 45 nm.

6. CONCLUSION

This paper presented the design of the Centip3De system. Centip3De utilizes the synergy that exists between 3D integration and near-threshold computing to create a reconfigurable system that provides both energy-efficient operation and techniques to address single-thread performance bottlenecks. The original Centip3De design is a 7-layer 3D stacked design with 128-cores and 256MB of DRAM. Silicon results were presented showing a 2-layer, 64-core system in 130 nm technology which achieved an energy efficiency of 3930 DMIPS/W. At 46.4 M devices, Centip3De is one of the largest academic projects to date.

Acknowledgments

This work was funded and organized with the help of DARPA, Tezzaron, ARM, and the National Science Foundation.

References

1. ARM Cortex-A9. http://www.arm.com/products/processors/cortex-a/cortex-a9.php.
2. ARM Cortex-M3. http://www.arm.com/products/CPUs/ARM_Cortex-M3.html.
3. Dennard, R., Gaensslen, F., Rideout, V., Bassous, E., LeBlanc, A. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE J. Solid State Circ. 9*, 5 (Oct. 1974), 256–268,.
4. Dreslinkski, R.G., Zhai, B., Mudge, T., Blaauw, D., Sylvester, D. An energy efficient parallel architecture using near threshold operation. In *Proceedings of the Parallel Architecture and Compilation Techniques* (2007).
5. Fick, D., Dreslinski, R., Giridhar, B., Kim, G., Seo, S., Fojtik, M., Satpathy, S., Lee, Y., Kim, D., Liu, N., et al. Centip3de: A 3930 dmips/w configurable near-threshold 3d stacked system with 64 arm cortex-m3 cores. In *ISSCC 2012* (2012).
6. Hanson, S., et al. Ultralow-voltage, minimum-energy CMOS. *IBM J. Res. Dev. 50*, 4/5 (Jul/Sep 2006), 469–490.
7. Kim, T.H., Persaud, R., Kim, C. Silicon odometer: an on-chip reliability monitor for measuring frequency degradation of digital circuits. *IEEE J. Solid State Circ. 43*, 4 (Apr. 2008), 874–880.
8. Knickerbocker, J.U., et al. Three-dimensional silicon integration. *IBM J. Res. Dev. 52*, 6 (Nov. 2008), 553–569.
9. Loh, G. 3d-stacked memory architectures for multi-core processors. In *ACM SIGARCH Computer Architecture News* (2008), volume 36. IEEE Computer Society, 453–464.
10. Qazi, M., Stawiasz, K., Chang, L., Chandrakasan, A. A 512 kb 8t sram macro operating down to 0.57 μv with an ac-coupled sense amplifier and embedded data-retention-voltage sensor in 45 nm soi cmos. *IEEE J. Solid State Circ. 46*, 1 (Jan. 2011), 85–96.
11. Rusu, S., et al. A 45 nm 8-Core Enterprise Xeon Processor. In *Proceedings of ISSCC* (Feb. 2009).
12. Soeleman, H., Roy, K. Ultra-low power digital subthreshold logic circuits. In *Proceedings of the 1999 International Symposium on Low Power Electronics and Design* (1999), 94–96.
13. Swanson, R., Meindl, J. Ion-implanted complementary mos transistors in low-voltage circuits. *IEEE J. Solid State Circ. 7*, 2 (Apr. 1972), 146–153.
14. Tezzaron Semiconductor FaStack® Technology. http://tezzaron.com/technology/FaStack.htm.
15 Tezzaron Semiconductor Octopus DRAM. http://www.tezzaron.com/memory/Octopus.html.
16. Vittoz, E., Fellrath, J. Cmos analog integrated circuits based on weak inversion operations. *IEEE J. Solid State Circ. 12*, 3 (Jun. 1977), 224–231.
17. Wang, A., Chandrakasan, A. A 180 mv fft processor using subthreshold circuit techniques. In *IEEE International Solid-State Circuits Conference 2004, ISSCC 2004. Digest of Technical Papers* (15–19 Feb. 2004), volume 1, 292–529,.
18. Zhai, B., Blaauw, D., Sylvester, D., Hanson, S. A sub-200 mv 6t sram in 0.13 μm cmos. In *IEEE International Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers* (Feb. 2007), 332–606.
19. Zhai, B., Nazhandali, L., Olson, J., Reeves, A., Minuth, M., Helfand, R., Pant, S., Blaauw, D., Austin, T. A 2.60pj/inst subthreshold sensor processor for optimal energy efficiency. In *IEEE VLSI Technology and Circuits* (2006).

Ronald G. Dreslinski, David Fick, Bharan Giridhar, Gyouho Kim, Sangwon Seo, Matthew Fojtik, Sudhir Satpathy, Yoonmyung Lee, Daeyeon Kim, Nurrachman Liu, Michael Wieckowski, Gregory Chen, Dennis Sylvester, David Blaauw, and Trevor Mudge (rdreslin, dfick, bharan, coolkgh, swseo, mfojtik, sudhirks, sori, daeyeonk, rachliu, wieckows, grgkchen, dennis, blaauw, tnm}@eecs.umich.edu), University of Michigan, Ann Arbor, MI.

Baylor University
Department of Computer Science
Lecturer of Computer Science

The Department of Computer Science seeks a dedicated teacher and program advocate for a lecturer position beginning August, 2014. The ideal candidate will have a master's degree or Ph.D. in Computer Science or a related area, a commitment to undergraduate education, effective communication and organization skills, and industry/academic experience in game development, especially with graphics and/or engine development. For position details and application information please visit: http://www.baylor.edu/hr/index.php?id=81302

Baylor, the world's largest Baptist university, holds a Carnegie classification as a "high-research" institution. 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 deepening its distinctive Christian mission as described in Pro Futuris (http://www.baylor.edu/profuturis/).

Baylor is a Baptist university affiliated with the Baptist General Convention of Texas. As an AA/EEO employer, Baylor encourages minorities, women, veterans, and persons with disabilities to apply.

Baylor University
Department of Computer Science
Assistant, Associate or Full Professor of Computer Science

The Department of Computer Science seeks a productive scholar and dedicated teacher for a tenured or tenure-track position beginning August, 2014. The ideal candidate will hold a terminal degree in Computer Science or a closely related field, demonstrate scholarly capability and an established and active independent research agenda in one of several core areas of interest, including, but not limited to, game design and development, software engineering, computational biology, informatics, machine learning and large-scale data mining. A successful candidate will also exhibit a passion for teaching and mentoring at the graduate and undergraduate level. For position details and application information please visit: http://www.baylor.edu/hr/index.php?id=81302

Baylor, the world's largest Baptist university, holds a Carnegie classification as a "high-research" institution. 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 deepening its distinctive Christian mission as described in *Pro Futuris* (http://www.baylor.edu/profuturis/).

Baylor is a Baptist university affiliated with the Baptist General Convention of Texas. As an AA/EEO employer, Baylor encourages minorities, women, veterans, and persons with disabilities to apply.

Boston College
Assistant Professor, Computer Science

The Computer Science Department of Boston College invites applications for a tenure-track Assistant Professorship beginning September, 2014. Applications from all areas of Computer Science will be considered. Applicants should have a Ph.D. in Computer Science or related discipline, a strong research record, and a commitment to undergraduate teaching.

We will begin reviewing applications on December 1, 2013, and will continue considering applications until the position is filled. Additional information about the department and the position is available at www.cs.bc.edu. **Submit applications online at apply.interfolio.com/22805**.

Bradley University
Computer Science & Information Systems Department
Tenure-track Assistant Professor

The Computer Science and Information Systems Department at **Bradley University** invites applications for a **tenure-track assistant professor** position starting August 2014. The position requires a PhD in Computer Science or a closely related field to be completed prior to start date.

Please **visit** http://www.bradley.edu/offices/business/humanresources/opportunities/faculty/position.dot?id=1a178f81-8ed8-4ccc-b892-d150c-b48eb36 **for full position description and application process**.

California State University, Fullerton
Department of Computer Science
Assistant Professor

The Department of Computer Science invites applications for **three tenure-track positions** at the **Assistant Professor** level starting fall 2014. For a complete description of the department, the position, desired specialization and other qualifications, please visit http://diversity.fullerton.edu/.

Carleton College
Department of Computer Science
Assistant Professor

Carleton College invites applications for a tenure-track position in computer science, in any area of specialization, preferably at the rank of **assistant professor**, beginning September 1, 2014. A Ph.D. or its imminent completion is required.

Carleton is a highly selective liberal arts college with outstanding, enthusiastic students. We seek an equally enthusiastic computer scientist committed to excellence in teaching, curriculum design, ongoing research, and undergraduate research advising. We are particularly interested in applicants who will strengthen the departmental commitment to students from underrepresented groups. Any area of specialization is welcome. The appointee would be expected to teach courses at a range of levels, from the introductory level to advanced undergraduate electives.

Carleton's Department of Computer Science (cs.carleton.edu) includes six faculty and an experienced full-time system administrator, and typically graduates between 25 and 40 majors annually.

To apply, please visit the Carleton College website at jobs.carleton.edu and complete the online application, including a cover letter, CV, brief statements describing teaching philosophy and research agenda, a graduate transcript, and contact information for three letters of reference (at least one of which should address your teaching). Questions about the position or the application process may be directed to csjobs@carleton.edu.

Applications completed by December 15, 2013, including reference letters, will receive full consideration.

Carleton College does not discriminate on the basis of race, color, creed, ethnicity, religion, sex, national origin, marital status, veteran status, actual or perceived sexual orientation, gender identity and expression, status with regard to public assistance, disability, or age in providing employment or access to its educational facilities and activities. We are committed to developing our faculty to better reflect the diversity of our student body and American society. Women and members of minority groups are strongly encouraged to apply.

Carnegie Mellon University
Tepper School of Business
Junior Faculty Tenure-Track Position in Business Analytics

The Tepper School of Business at Carnegie Mellon University seeks candidates for a tenure-track faculty position at the Assistant Professor level in Business Analytics. The ideal candidate will be a main player in the school's analytical approach to business, a long-standing differentiator of the Tepper business education. The ideal candidate will be substantially involved in new research and educational initiatives across disciplines that concern the use of data for decision making. Applicants are expected to have a doctoral degree at the time of appointment (Fall 2014) in Business, Computer Science, Information Systems, Operations Research, Statistics, or another related field. Applicants are also expected to have a strong commitment to and proven record of research excellence involving cutting-edge analytical methods in business. The Tepper School and Carnegie Mellon University have a strong culture of collaboration across disciplines. This open environment provides unique opportunities for highly innovative work. We expect a "data for decision making" focus to have a large and synergistic overlap with other research areas such as marketing, service and supply-chain management, healthcare management, finance, etc.

Applicants should submit a vita, research statement, up to five publications or working papers, and three recommendation letters to: bapost@andrew.cmu.edu. Material may also be sent via the Postal Service to Ms. Rosanne Christy, Faculty Search Coordinator for Business Analytics, Carnegie Mellon University, Tepper School of Business, Posner 233, 5000 Forbes Avenue, Pittsburgh, PA 15213. If you have any questions about the application, please contact Ms. Christy at 412-268-1320. **To ensure full consideration, applications must be received by December 1, 2013.**

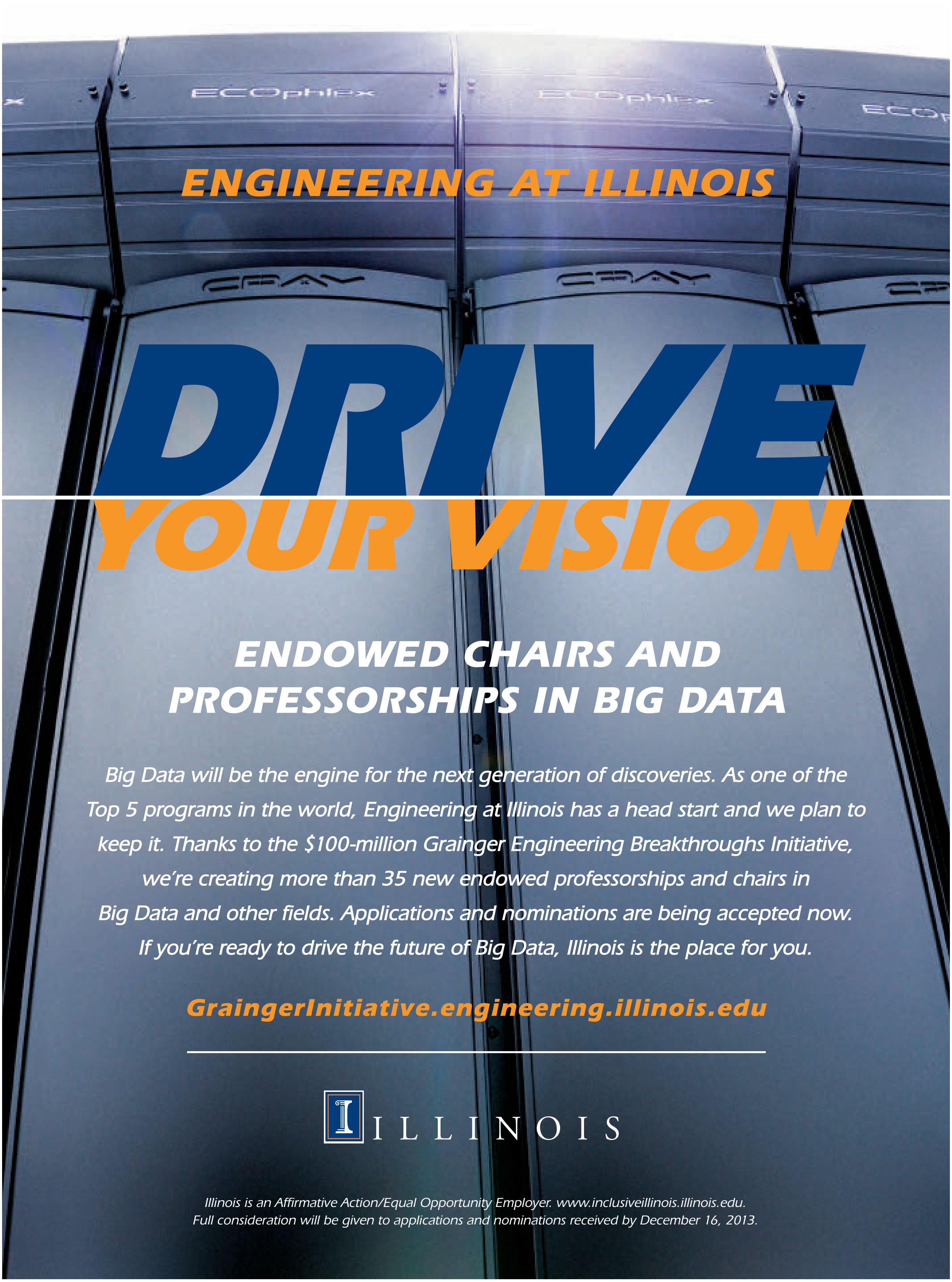
Carnegie Mellon is an equal opportunity, affirmative action employer with particular interest in identifying women and minority applicants for faculty positions.

Clemson University
School of Computing
Director, School of Computing

The Clemson University School of Computing invites applications and nominations for the Director of the School of Computing. The individual who holds this administrative position will serve as the executive officer of the school and report to the Dean of the College of Engineering and Science. Candidates are expected to merit the rank of Professor with tenure and have an internationally recognized record of scholarship and leadership, with demonstrated administrative experience. For the appropriate candidate, an endowed chair may be considered. The school seeks a Director excited to provide leadership with shared governance in areas including research and funding, faculty and student recruiting, promotion and tenure, and curriculum management. Additional information about the School of Computing and the Director position is available at http://www.clemson.edu/ces/computing/director-position.html. Informal enquiries may be directed to the Search Committee Chair, Jason O. Hallstrom, at jasonoh@clemson.edu.

The School of Computing comprises three academic units representing a broad cross-section of computing and its applications: (1) the Division of Computer Science, (2) the Division of Visual Computing, and (3) the Division of Human-Centered Computing. The school includes 31 tenured/tenure-track faculty, 6 lecturers, 419 undergraduates, and 215 graduate students. Competitive funding for FY12 was over $5 million.

Clemson University is the land-grant institution of South Carolina, enrolling approximately 15,000 undergraduates and 4,000 graduate students. A faculty of 1,400 and staff of 3,000 support 84 undergraduate degree offerings, 73 master's degree programs, and 40 Ph.D. programs. Research and economic development activities are enhanced by public-private partnerships at three innovation campuses, and six research and education centers located throughout South Carolina.


ENGINEERING AT ILLINOIS
DRIVE
YOUR VISION
ENDOWED CHAIRS AND PROFESSORSHIPS IN BIG DATA
Big Data will be the engine for the next generation of discoveries. As one of the Top 5 programs in the world, Engineering at Illinois has a head start and we plan to keep it. Thanks to the $100-million Grainger Engineering Breakthroughs Initiative, we're creating more than 35 new endowed professorships and chairs in Big Data and other fields. Applications and nominations are being accepted now. If you're ready to drive the future of Big Data, Illinois is the place for you.
GraingerInitiative.engineering.illinois.edu
ILLINOIS
Illinois is an Affirmative Action/Equal Opportunity Employer. www.inclusiveillinois.illinois.edu.
Full consideration will be given to applications and nominations received by December 16, 2013.

Today, Clemson University is ranked 25th among national public universities by *U.S. News & World Report*. Clemson University is described by students and faculty as an inclusive, student-centered community characterized by high academic standards, a culture of collaboration, school spirit, and a competitive drive to excel.

Applicants must have an earned doctorate in computer science, computer engineering, or a closely related field. Applicants should submit a current vita and a minimum of five references with full contact information. (References will be contacted only after receiving follow-up approval from the applicant/nominee.) Electronic submissions (PDF files) to directorsearch@clemson.edu are preferred, but applications and nominations can also be mailed to Director Search, 212 McAdams Hall, Clemson University, Clemson, SC 29634, USA. Application materials must be received by December 15, 2013 to receive full consideration, though the search will remain open until the position is filled.

Clemson University is an Affirmative Action/Equal Opportunity employer and does not discriminate against any individual or group of individuals on the basis of age, color, disability, gender, national origin, race, religion, sexual orientation, veteran status or genetic information.

Colorado State University

Department of Computer Science

Tenure-Track Faculty in Computer Science

Colorado State University is accepting applications for a tenure-track assistant professor in Computer Science, beginning fall 2014. Only candidates in bioinformatics/computational biology will be considered. This position is part of a university-wide effort to recruit additional faculty in biosciences and bioinformatics. Applications must be received by December 16, 2013. Submit materials at http://cns.natsci.colostate.edu/employment/Compsci/. CSU is an EO/EA/AA employer. Colorado State University conducts background checks on the final candidates.

Dartmouth College

The Geisel School of Medicine at Dartmouth College

Biomedical Informaticist

TENURE-TRACK FACULTY POSITION IN BIOMEDICAL INFORMATICS

The **Collaboratory for Healthcare and Biomedical Informatics** at the **Geisel School of Medicine at Dartmouth** seeks an outstanding individual who is pursuing a vigorous, independent research program focused on innovative informatics methods and technologies that will advance human health, medical care or biomedical sciences.

Applicants must have earned a PhD and/or MD degree, have formal advanced training in biomedical informatics, computer science, or a related field, and have a successful track record of peer-reviewed publications. Evidence of extramural funding is essential. The position carries a faculty appointment in the Geisel School of Medicine at the rank of Assistant or Associate Professor, commensurate with experience.

Applicants should submit a PDF file of their *curriculum vitae* and 2-3 page description of their research interests and future research plans, as well as have 3 letters of reference sent directly to the search committee, by December 1, 2013. Review of applications will begin immediately. All materials should be sent to Informatics.Faculty.Search@Dartmouth.edu, and addressed to:

Amar Das, MD, PhD
Director, Collaboratory for Healthcare and Biomedical Informatics
Geisel School of Medicine at Dartmouth, HB 7751
Lebanon, NH 03766, USA

Geisel School of Medicine is an Affirmative Action/Equal Opportunity Employer. We will extend equal opportunity to all individuals without regard for gender, race, religion, color, national origin, sexual orientation, age, disability, handicap or veteran status.

Florida State University

Department of Computer Science

Tenure-Track Assistant Professor Positions

The Department of Computer Science at the Florida State University invites applications for multiple tenure-track Assistant Professor Positions to begin August 2014. Positions are 9-mo, full-time, tenure-track, and benefits eligible. Outstanding applicants with research and teaching strengths in Big Data and Cyber Security are particularly

Faculty Positions in Computer Science

Ecole polytechnique fédérale de Lausanne

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

The School of Computer and Communication Sciences at EPFL invites applications for faculty positions in computer and communication sciences. We are seeking candidates for positions at the tenure-track assistant professor or tenured senior professor levels.

Successful candidates will develop an independent and creative research program, participate in both undergraduate and graduate teaching, and supervise PhD students.

Candidates from all areas of computer science will be considered, but preference will be given to candidates in the fields of cyber-physical risk management, data analytics, machine learning, security and privacy.

EPFL, in Lausanne, Switzerland, is a dynamic science and engineering university ranked among the top research institutions in Europe. It offers a rich international environment with English as its common language, as well as competitive salaries, significant start-up resources, and outstanding research infrastructure.

To apply, please follow the application procedure at
https://academicjobsonline.org/ajo/jobs/3050.

The following documents are requested in PDF format: curriculum vitae including publication list, brief statements of research and teaching interests, names and addresses (including e-mail) of 3 references for junior positions and 6 for senior positions. Screening will start on **December 1, 2013**. Further questions can be addressed to :

Prof. Willy Zwaenepoel
Chairman of the recruiting committee
School of Computer and Communication Sciences
EPFL
CH-1015 Lausanne

recruiting.ic@epfl.ch

For additional information on EPFL, please consult:
http://www.epfl.ch or http://ic.epfl.ch

EPFL is an equal opportunity employer.

encouraged to apply. Outstanding applicants specializing in other emerging research areas are also welcome to apply. Applicants should hold a PhD in Computer Science or closely related field, and have excellent research and teaching accomplishments or potential. The department offers degrees at the BS, MS, and PhD levels. The department is an NSA Center of Academic Excellence in Information Assurance Education (CAE/IAE) and Research (CAE-R).

FSU is classified as a Carnegie Research I university. Its primary role is to serve as a center for advanced graduate and professional studies while emphasizing research and providing excellence in undergraduate education. Further information can be found at:

http://www.cs.fsu.edu

Screening will begin January 1, 2014 and will continue until the position is filled. Please apply online with curriculum vitae, statements of teaching and research philosophy, and the names of five references, at

http://www.cs.fsu.edu/positions/apply.html

Questions can be e-mailed to Prof. Mike Burmester, Faculty Search Committee Chair, recruitment@cs.fsu.edu or to Prof. Robert van Engelen, Department Chair, chair@cs.fsu.edu.

The Florida State University is a Public Records Agency and an Equal Opportunity/Access/Affirmative Action employer, committed to diversity in hiring.

Georgia Regents University

Assistant (or Associate) Professor of Computer Science

Tenure track position in Computer Science especially programming languages, applied theory, algorithms, and operating systems.

PhD in Computer science or related field, demonstrated capability or promise for teaching & research.

Apply URL: http://www.higheredjobs.com/details.cfm?JobCode=175799939

Inquiries about the position should be directed to Todd Schultz, tschultz@gru.edu or 706-667-4534.

Lehigh University

Assistant or Associate Professor

Applications are invited for a tenure-track position at the **Assistant or Associate Professor** level in the Computer Science and Engineering Department (http://www.cse.lehigh.edu) of Lehigh University to start in August 2014. Outstanding candidates in all areas of computer science will be considered, with priority given to candidates with a research focus in either Data Mining or Cybersecurity, both defined broadly.

The successful applicant will hold a Ph.D. in Computer Science, Computer Engineering, or a closely related field. The candidate must demonstrate a strong commitment to quality undergraduate and graduate education, and the potential to develop and conduct a high-impact research program with external support. Applicants should have an interest in teaching core courses in computer science as well as courses in their research area. The successful applicant will also be expected to contribute to interdisciplinary research programs.

The faculty of the Computer Science and Engineering department maintains an outstanding international reputation in a variety of research areas, and includes ACM and IEEE fellows as well as six NSF CAREER award winners. The department currently has 14 full-time faculty members and more than 250 undergraduate and 70 graduate students, more than half of whom are doctoral students. We offer B.A., B.S., M.S., and Ph.D. degrees in Computer Science and jointly oversee B.S., M.S., and Ph.D. degree programs in Computer Engineering with the department of Electrical and Computer Engineering. We also offer a B.S. in Computer Science and Business with the College of Business and Economics.

Lehigh University is a private, highly selective institution that is consistently ranked among the top 40 national research universities by U.S. News & World Report. Located in Bethlehem, Pennsylvania, Lehigh is 80 miles west of New York City and 50 miles north of Philadelphia, providing an accessible and convenient location that offers an appealing mix of urban and rural lifestyles. Lehigh Valley International Airport, just six miles from campus, provides easy access to the world. Lehigh Valley cities and towns are regularly listed as among the best places to live in the country.

Applications can be submitted online at https://academicjobsonline.org/ajo/jobs/3051.

THE UNIVERSITY OF TENNESSEE, KNOXVILLE

The Department of Electrical Engineering and Computer Science (EECS) at The University of Tennessee (UT) is seeking candidates for multiple faculty positions in computer science, computer engineering, and software engineering, which are further specified as follows. Successful candidates will be expected to teach at both undergraduate and graduate levels, to establish a vigorous funded research program, and to have a willingness to collaborate with other faculty in research.

EECS is housed in a new $37.5 million teaching and research facility completed in 2012. The department currently has an enrollment of more than 550 undergraduate and 220 graduate students, with a faculty of 40, and research expenditures that exceed $13 million per year. EECS has strong research partnerships with organizations such as the nearby Oak Ridge National Laboratory.

ERICSSON-HARLAN D. MILLS CHAIR OF SOFTWARE ENGINEERING

Applications and nominations are invited for the Ericsson-Harlan D. Mills Chair of Software Engineering. Associate Professors and recently promoted Full Professors are strongly encouraged to apply, although consideration will be given to all outstanding candidates. Applicants should have demonstrated an outstanding independent research and teaching program in software engineering or in a related software-encompassing field, and should be capable of working with other faculty to further strengthen our existing research. Applicants should have an earned Ph.D. in Computer Science or equivalent and be eligible for appointment at the rank of full Professor.

COMPUTER ENGINEERING AND COMPUTER SCIENCE

Applications are sought for two tenure-track junior-level faculty positions in Computer Science and in Computer Engineering. Applicants should have an earned Ph.D. in Computer Engineering, Computer Science, or a related field. Individuals with expertise in computer architecture, high performance computer systems, parallel and distributed computing, performance modeling and simulation, computing theory, cloud computing, data analytics, digital systems, VLSI, computational accelerators, dependable and secure systems, software engineering, intelligent systems and robotics are encouraged to apply, although qualified applicants from all areas of computer science and computer engineering will be considered.

The University of Tennessee welcomes and honors people of all races, genders, creeds, cultures, and sexual orientations, and values intellectual curiosity, pursuit of knowledge, and academic freedom and integrity. Interested candidates should apply through the departmental web site at http://www.eecs.utk.edu and submit a letter of application, a curriculum vitae, a statement of research and teaching interests, and contact information for three references. Review of applications will begin on December 15, 2013, and continue until the positions are filled.

The University of Tennessee is an EEO/AA/Title VI/Title IX/Section 504/ADA/ADEA institution in the provision of its education and employment programs and services.

Review of applications will begin December 1, 2013 and will continue until the position is filled. Lehigh University is an Equal Opportunity/Affirmative Action Employer and is the recipient of an NSF ADVANCE Institutional Transformation award for promoting the careers of women in academic science and engineering. Lehigh provides comprehensive benefits including domestic partner benefits (see also http://www.lehigh.edu/worklifebalance/). Questions concerning this search may be sent to faculty-search@cse.lehigh.edu.

National University of Singapore
Department of Computer Science
Multiple Tenure-Track Faculty Positions

The Department of Computer Science, National University of Singapore (NUS), has openings for several tenure-track faculty positions. Our main focus is on candidates at the Assistant Professor level with research interests in the following areas:

- ► Cyber-physical systems
- ► Big data analytics
- ► Security
- ► Sensor data modelling and learning

These areas are to be viewed in a broad sense, and we are particularly interested in candidates whose research interests cut across these and related areas. We seek candidates demonstrating excellent research potential and a strong commitment to teaching. We will also seriously consider exceptional candidates in other areas of computer science. Further, we will consider candidates at senior ranks (Associate and Full Professor) who have an outstanding record of research accomplishments.

We are an internationally top-ranked department with low teaching loads, excellent facilities, and intensive external collaborations. Significant funding opportunities abound for strong candidates. The research of the faculty covers all the major areas of computer science and is well represented at prestigious international conferences and journals. The department has a thriving PhD programme and it actively strives to attract the best students from the region and beyond. More information can be found at http://www.comp.nus.edu.sg/.

NUS offers highly competitive salaries and generous benefits, while Singapore provides a vibrant international environment with world-class health care, excellent infrastructure, a warm climate and low taxes.

Interested candidates are invited to send, via electronic submission, the following materials to the Chair of the CS Search Committee, Prof. P.S. Thiagarajan, at csrec@comp.nus.edu.sg

- ► Cover letter
- ► Curriculum Vitae
- ► A teaching statement
- ► A research statement
- ► Contact information for at least three references

Applications will be reviewed as they are received and will continue until the positions are filled. However, to ensure maximal consideration applicants are encouraged to submit their materials by December 15, 2013.

New Mexico State University
Computer Science Department
Assistant Professor

The Computer Science Department at New Mexico State University invites applications for a tenure-track position at the Assistant Professor level, with appointment starting in the Fall 2014 semester. We are seeking strong candidates with research expertise that can effectively complement the research foci of the department; we are particularly interested in expertise in the areas of High Performance Computing (HPC) and/or Big Data. Applications from women, members of traditionally under-represented groups, and other individuals interested in contributing to the diversity and excellence of the academic community are strongly encouraged.

For the full position announcement, please visit

http://www.cs.nmsu.edu/~epontell/position2013.html

APPLY FOR THIS JOB

Contact Person: CS Faculty Search Chair
Email Address: cssearch@cs.nmsu.edu
Phone: 575-646-6239
Fax: 575-646-1002
Apply URL: http://cssearch@cs.nmsu.edu

THE CHINESE UNIVERSITY OF HONG KONG

Applications are invited for:-

Faculty of Engineering
Professors / Associate Professors / Assistant Professors
(Ref. 1314/032(255)/2)

The Faculty of Engineering invites applications for several faculty posts at Professor / Associate Professor / Assistant Professor levels with prospect for substantiation in the interdisciplinary area of 'Big Data Analytics', which is a new strategic research initiative supported by the University's Focused Innovations Scheme and will complement current/planned strengths in different Departments under the Faculty. To lead the big data research initiative, senior academics in this area are particularly welcome to apply.

Currently, the Faculty is seeking candidates in the following areas:

- Theoretical, mathematical and algorithmic aspects in large data analytics;
- Large scale software systems and architecture in large data analytics;
- Application areas in large data analytics (including information analytics, network/Web analytics, financial analytics, or bio/medical analytics, etc.).

Applicants should have (i) a PhD degree; and (ii) a strong scholarly record demonstrating potential for teaching and research excellence. The appointees will be expected to (a) teach at both undergraduate and postgraduate levels; (b) develop a significant independent research programme with external funding; and (c) supervise postgraduate students. Appointments will normally be made on contract basis for three years initially, which, subject to performance and mutual agreement, may lead to longer-term appointment or substantiation later. Applications will be accepted until the posts are filled. Further information about the Faculty is available at *http://www.erg.cuhk.edu.hk*.

Salary and Fringe Benefits
Salary will be highly competitive, commensurate with qualifications and experience. The University offers a comprehensive fringe benefit package, including medical care, plus a contract-end gratuity for appointments of two years or longer, and housing benefits for eligible appointees. Further information about the University and the general terms of service for appointments is available at *http://www.per.cuhk.edu.hk*. The terms mentioned herein are for reference only and are subject to revision by the University.

Application Procedure
Please send full resume, copies of academic credentials, a publication list with abstracts of selected published papers, details of courses taught and evaluation results (if any), a research plan, a teaching statement, together with names of three to five referees, to the Dean, Faculty of Engineering by e-mail to *recruit-bda@erg.cuhk.edu.hk*. For enquiries, please contact Professor John C.S. Lui, the leader of this strategic initiative (e-mail: *cslui@cse.cuhk.edu.hk*). Applicants are requested to clearly indicate that they are applying for the posts under 'Big Data Analytics Initiative'. The Personal Information Collection Statement will be provided upon request. Please quote the reference number and mark 'Application – Confidential' on cover.

Computer Science

Multiple faculty positions are available at Cornell's Department of Computer Science, based in Ithaca, New York.

Candidates are invited to apply at all levels including tenured, tenure-track, or lecturer, and from all areas of computer science and related fields.

Tenured and tenure track faculty must hold the equivalent of a Ph.D.; applicants for the position must have demonstrated an ability to conduct outstanding research. Lecturers must hold the equivalent of a Master's degree, with a Ph.D. preferred.

To ensure full consideration, applications should be received by December 1, 2013, but will be accepted until all positions are filled.

Applicants should submit a curriculum vita, brief statements of research and teaching interests, and arrange to have at least three reference letters submitted at:
https://academicjobsonline.org/ajo/jobs/2813

Diversity and inclusion have been and continue to be a part of our heritage. Cornell University is a recognized EEO/AA employer and educator.

New York University/Courant Institute of Mathematical Sciences
Department of Computer Science
Faculty

The department expects to have several regular faculty positions beginning in September 2014 and invites candidates at all levels. We will consider outstanding candidates in any area of computer science.

Faculty members are expected to be outstanding scholars and to participate in teaching at all levels from undergraduate to doctoral. New appointees will be offered competitive salaries and startup packages, with affordable housing within a short walking distance of the department. New York University is located in Greenwich Village, one of the most attractive residential areas of Manhattan.

The department has 35 regular faculty members and several clinical, research, adjunct, and visiting faculty members. The department's current research interests include algorithms, cryptography and theory; computational biology; distributed computing and networking; graphics, vision and multimedia; machine learning; natural language processing; scientific computing; and verification and programming languages.

Collaborative research with industry is facilitated by geographic proximity to computer science activities at AT&T, Google, IBM, Bell Labs, NEC, and Siemens.

Please apply at https://cs.nyu.edu/webapps/facapp/register

To guarantee full consideration, applications should be submitted no later than December 1, 2013; however, this is not a hard deadline, as all candidates will be considered to the full extent feasible, until all positions are filled. Visiting positions may also be available.

New York University is an equal opportunity/affirmative action employer.

North Dakota State University
Department of Computer Science
Assistant Professor

Department of Computer Science, North Dakota State University seeks to fill a tenure-track Assistant Professor position starting Fall of 2014. PhD required. NDSU offers degrees in Computer Science and Software Engineering. Research and teaching excellence is expected. Normal teaching loads are 3 courses per year. Salary is competitive.

The Department has 20 Faculty, approximately 200 graduate (PhD and MS) and 300 BS/BA students. NDSU is a Carnegie research extensive class institution.

Fargo is a clean, growing metropolitan area of 250,000 that consistently ranks very high in national quality-of-life surveys. We have low levels of crime and pollution, excellent schools, short commutes and proximity to the Minnesota lakes. The community has a symphony, an opera, a domed stadium, a community theater, three universities and many other opportunities for advancement, recreation and entertainment.

Minimum Qualifications:
Ph.D in Computer Science or a closely related area. Experience or other evidence of potential for excellence in research, undergraduate and graduate teaching and service. Effective oral and written communication skills.

Qualifications preferred but not required:
Evidence of potential for excellence in research in software engineering, software systems, or closely related areas. Capability for teaching a variety of courses in computer science.

To Apply: see https://jobs.ndsu.edu/

For additional information see http://cs.ndsu.edu/positions.htm

Review of applications begins December 1, 2013. Applications will be accepted until position is filled.

North Dakota State University is an Equal Opportunity/AA employer.

This position is exempt from North Dakota Veterans' Preference requirements.

Ohio State University
Computer Science and Engineering Department
Multiple Tenured or Tenure-track Appointments
Assistant, Associate or Full Professor level

The **Computer Science and Engineering Department** at the Ohio State University seeks candi-

---

Computer and Information Sciences
TEMPLE UNIVERSITY
Senior Faculty

Applications are invited for senior faculty positions for Associate or Full Professor in the Department of Computer and Information Sciences (CIS) at Temple. These positions are open to all CIS areas. Applicants are expected to have an outstanding track record.

The CIS Dept. has two undergraduate degree programs in Computer Science (CS) and in Information Science and Technology (IS&T), master's programs in CIS and IS&T, and a PhD program in CIS. The CIS Dept. has undergone considerable growth in research in the past few years, during which research funding, publication rates, and the number of Ph.D. students has significantly increased. CIS is scheduled to move to a new building next year.

Please submit applications with all requested information online at **http://academicjobsonline.org#3244**. For latest updates, including openings in other ranks, check **http://www.temple.edu/cis** or send email to committee chair Dr. Eugene Kwatny at **gkwatny@temple.edu**. Review of candidates will begin immediately and will continue until the positions are filled.

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

JOIN THE INNOVATION.

Qatar Computing Research Institute seeks talented scientists and software engineers to join our team and conduct world-class applied research focused on tackling large-scale computing challenges.

We offer unique opportunities for a strong career spanning academic and applied research in the areas of Arabic language technologies including natural language processing, information retrieval and machine translation, distributed systems, data analytics, cyber security, social computing and computational science and engineering.

Scientist applicants must hold (or will hold at the time of hiring) a PhD degree, and should have a compelling track record of accomplishments and publications, strong academic excellence, effective communication and collaboration skills.

Software engineer applicants must hold a degree in computer science, computer engineering or related field; MSc or PhD degree is a plus.

We also welcome applications for post doctoral researcher positions.

As a **national research institute** and proud member of Qatar Foundation, our research program offers a collaborative, multidisciplinary team environment endowed with a comprehensive support infrastructure.

Successful candidates will be offered a highly competitive compensation package including an attractive tax-free salary and additional benefits such as furnished accommodation, excellent medical insurance, generous annual paid leave, and more.

For full details about our vacancies and how to apply online please visit
http://www.qcri.qa/join-us/
For queries, please email
QFJobs@qf.org.qa

/QCRI.QA @QatarComputing QatarComputing QatarComputing www.qcri.qa dates for **multiple tenured or tenure-track appointments at the assistant, associate or full professor level**. The specific searches being conducted this year include:

- ► One targeted position, open rank, in the area of cybersecurity, broadly defined. We are specifically interested in applicants with research interests in network security, physical layer/information theoretic security, cyberphysical systems, secure systems, data privacy, cryptography, or programming language security.
- ► One position (open rank) open to all areas of computer science and engineering.
- ► In addition, data analytics is slated for significant growth and we anticipate open positions in data mining and cloud computing.

The department is committed to enhancing faculty diversity; women, minorities, and individuals with disabilities are especially encouraged to apply.

Applicants should hold or be completing a Ph.D. in CSE or a closely related field, have a commitment to and demonstrated record of excellence in research, and a commitment to excellence in teaching.

To apply, please submit your application via the online database. The link can be found at:

https://www.cse.ohio-state.edu/cgi-bin/portal/fsearch/apply.cgi.

Review of applications will begin in December and will continue until the positions are filled.

The Ohio State University is an Equal Opportunity/Affirmative Action Employer.

Purdue University

Tenure-Track/Tenured Faculty Positions

The Department of Computer Science at Purdue University is entering a phase of sustained expansion. Applications for tenure-track and tenured positions at the Assistant, Associate and Full Professor levels beginning August 2014 are being solicited. Outstanding candidates in all areas will be considered.

The Department of Computer Science offers a stimulating and nurturing academic environment with active research programs in all areas of our discipline. Information about the department and a description of open positions are available at http://www.cs.purdue.edu.

Applicants should hold a PhD in Computer Science, or related discipline, be committed to excellence in teaching, and have demonstrated excellence in research. Successful candidates will be expected to teach courses in computer science, conduct research in their field of expertise, and participate in other department and university activities. Salary and benefits are competitive. Applicants are strongly encouraged to apply online at https://hiring.science.purdue.edu. Alternatively, hardcopy 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 in fall 2013, and will continue until positions are filled. A background check will be required for employment. Purdue University is an Equal Opportunity/Equal Access/Affirmative Action employer committed to achieving a diverse workforce.

San Diego State University

Department of Computer Science

Assistant Professor of Computer Science

The Department of Computer Science at SDSU seeks an Assistant Professor with a PhD in Computer Science or a closely related field and with expertise in networks and large data infrastructure and analysis. For more details and application procedures, please see http://cs.sdsu.edu/position . **SDSU is an equal opportunity/Title IX employer**.

Shanghai Jiao Tong University (SJTU)

Department of Computer Science and Engineering (CSE)

Assistant Professor

The **Department of Computer Science and Engineering (CSE)** at **Shanghai Jiao Tong University (SJTU)** is seeking to fill several tenure-track positions in computer science at the rank of Assistant Professor and above, starting January 2014 and September 2014.

Shanghai Jiao Tong University is one of the oldest and most prestigious universities in China, and CSE is premier in computer science research and education. Candidates for these positions are sought for the well-recognized computer science program (ACM class) at Zhiyuan College of SJTU which provides an outstanding undergraduate education to a select group of 30 research-oriented students. Over the last ten years, students from the ACM class have won five

Colorado School of Mines

Department of Electrical Engineering & Computer Science

Assistant Professor(s)

Colorado School of Mines invites applications for two (2) Assistant Professor positions in Computer Science. Research specializations of interest are HPC/parallel computing, robotics, real-time/mobile/embedded systems, and privacy/security. The successful candidate will teach undergraduate and graduate courses as well as develop an externally funded research program that includes graduate students and results in publications in conferences and journals. University/College and professional service is expected. Required is a Ph.D. in Computer Science, Computer Engineering, Electrical Engineering, or Software Engineering. Applicants must demonstrate the potential for success in teaching, scholarship and service. Superb interpersonal and communication skills are required.

For the complete job announcement and directions on how to apply, visit: http://inside.mines.edu/HR-Academic-Faculty

Mines is an EEO/AA employer.

Millersville University

DEPARTMENT OF COMPUTER SCIENCE ASSISTANT PROFESSOR

Full-time, tenure-track beginning August 2014. **Required:** Ph.D. in Computer Science or closely related field. Recent record of post-secondary teaching and scholarly experience in the Computer Science related area. Ability and willingness to teach a wide range of Computer Science courses, documented by one of the following: transcript, teaching philosophy, letter of reference, peer review of teaching, student evaluations or teaching awards. Commitment to excellence in teaching at the undergraduate level. Commitment to working with students from diverse backgrounds. Excellent oral and written communication skills. Successful interview and teaching demonstration.

Millersville University, highly regarded for its instructional quality and strong commitment to diversity, is a student-centered institution and one of 14 institutions of the PA State System of Higher Education. The campus is located in historic Lancaster County within a short drive of Baltimore, Philadelphia, Washington, New York and the Atlantic Ocean beaches. Millersville enrolls approximately 9,000 undergraduate and graduate students and has over 900 full-time employees, and is nationally ranked as one of the top regional public institutions of higher learning by *U.S. News & World Report*. The Computer Science Department offers an ABET accredited B.S. degree in Computer Science. There are seven full-time faculty and about 200 majors. The teaching load is 24 semester hours per year, usually three 4-credit courses with two preparations per semester. The department has two teaching labs and two research labs.

Additional information about the program can be found at **http://cs.millersville.edu/**. Send questions to **search@cs.millersville.edu**. **To apply:** Go to **https://jobs.millersville.edu** and create a faculty application. **The following application materials are required:** A letter of application addressing how your qualifications match the requirements. A curriculum vitae. One-page statement of teaching philosophy and one-page statement of research philosophy. Names, address, email addresses, and phone numbers of three references, who will be asked to upload letters of reference directly to website. At least two letters must address teaching effectiveness. Copies of undergraduate and graduate transcripts. Official transcript will be required at time of appointment. Full consideration will be given to applications received by **January 13, 2014**.

An EO/AA institution. www.millersville.edu gold medals in the ACM International Collegiate Programming Contest.

Professor John Hopcroft, 1986 Turing Award recipient, is chairing the committee on curriculum development and faculty recruiting. Since December 2011, he has spent two months a year teaching at Zhiyuan College. In May 2012, he was appointed Special Counselor to President Jie Zhang.

An internationally competitive package for salary and benefits will be offered. Strong candidates in all areas will be considered with special consideration given to systems and networking, architecture, machine learning, theory, and security. In addition to the teaching duties at Zhiyuan College's ACM class, faculty members are required to teach graduate level courses, to supervise Ph.D. students, and to conduct research in the CSE. The overall teaching load is one course per semester.

SJTU makes a great effort to provide opportunities for the development of young faculty, including a startup research grant. There are a number of sources for additional research funding. The positions are provided in strong cooperation with Microsoft Research Asia (MSRA) with opportunities for research collaborations. Candidates are encouraged to apply to the Thousand Talents Program for extra funding and benefit support. Our equal opportunity and affirmative action program seek minorities, women, and non-Chinese scientists.

The criteria for promotion will be professional reputation as judged by international experts in the candidate's field and excellence in teaching.

Applications, including vita and the names of three references, should be sent to Professor John Hopcroft (jeh@cs.cornell.edu) and to Bing Li (binglisjtu@sjtu.edu.cn).

The application deadline is January 31, 2014 for positions starting in September 2014. Applications for starting earlier will be reviewed immediately.

The State University of New York at Buffalo

Department of Computer Science and Engineering

Multiple Faculty Positions at Full, Associate, and Assistant Professor Levels

Junior and Senior Positions in All Core Areas

We invite outstanding candidates to apply for several tenured or tenure-track faculty positions at all ranks. We welcome qualified applicants in all core areas of Computer science and Engineering, especially Robotics, Software Systems, Hardware Systems, Big Data Analytics, Cloud computing, and Theory/Algorithms.

Junior Position in Cyber-Physical Systems and Analytics

We invite outstanding candidates to apply for an opening at the assistant professor level to work closely with the newly established Institute of Sustainable Transportation and Logistics (ISTL). Research areas of interest include, but are not limited to: cyber-physical systems; connected and automated vehicles; data-acquisition and analysis; and end-to-end workflow process management.

The Department of Computer Science and Engineering has 34 tenured and tenure-track faculty and four teaching faculty members. Eight members of our faculty are NSF CAREER award recipients. Our faculty members are actively involved in successful interdisciplinary programs and centers devoted to biometrics, bioinformatics, biomedical computing, cognitive science, document analysis and recognition, high performance computing, information assurance and cyber security, and computational & Data science and engineering.

Candidates are required to have a Ph.D. in Computer Science/Engineering or related field by August 2014, and should demonstrate potential for excellence in research, teaching and mentoring.

Applications should be submitted by December 31, 2013, electronically via http://www.ubjobs.buffalo.edu/.

The State University of New York at Buffalo is an Equal Opportunity Employer/Recruiter.

Texas A&M University

Department of Computer Science and Engineering

Department Head

The Dwight Look College of Engineering at Texas A&M University invites nominations and applications for the position of head of the Department of Computer Science and Engineering.

MSOE

MILWAUKEE SCHOOL OF ENGINEERING

SOFTWARE ENGINEERING FACULTY

The Milwaukee School of Engineering invites applications for a full-time faculty position in our Software Engineering program beginning in the fall of 2014. Rank will depend on qualifications and experience of the candidate. Applicants must have an earned doctorate degree in Software Engineering, Computer Engineering, Computer Science or closely related field, as well as relevant experience in software engineering practice.

The successful candidate must be able to contribute in several areas of software engineering process and practice while providing leadership in one of the following: computer security, networks, software architecture and design, or software requirements.

MSOE expects and rewards a strong primary commitment to excellence in teaching at the undergraduate level. Continued professional development is also expected.

Our ABET-accredited undergraduate software engineering program had its first graduates in Spring 2002. Founded in 1903, MSOE is a private, application-oriented university with programs in engineering, business, and nursing. MSOE's 15+ acre campus is located in downtown Milwaukee, in close proximity to the Theatre District and Lake Michigan.

Please visit our website at: http://www.msoe.edu/hr/ for additional information including requirements and the application process.

MSOE IS AN EQUAL OPPORTUNITY/AFFIRMATIVE ACTION EMPLOYER

Faculty Search

ShanghaiTech University

School of Information Science and Technology

The School of Information Science and Technology (SIST) in the new ShanghaiTech University invites applications to fill multiple tenure-track and tenured positions. Candidates should have an exceptional academic record or strong potential in frontier areas of information sciences.

ShanghaiTech is founded as a world-class research university for training future scientists, entrepreneurs, and technology leaders. Besides keeping a strong research profile, successful candidates should also contribute to undergraduate and graduate education within SIST.

Compensation and Benefits:
Salary and startup fund are highly competitive, commensurate with academic experience and accomplishment. ShanghaiTech also offers a comprehensive benefit package which includes housing. All regular faculty members are hired within ShanghaiTech's new tenure-track system commensurate with international practice and standards.

Academic Disciplines:
We seek candidates in all cutting edge areas of information science and technology. Our recruitment focus includes, but is not limited to: computer architecture and technologies, nano-scale electronics, high speed and RF circuits, intelligent and integrated signal processing systems, computational foundations, big data, data mining, visualization, computer vision, bio-computing, smart energy/power devices and systems, next-generation networking, as well as inter-disciplinary areas involving information science and technology.

Qualifications:
- Well developed research plans and demonstrated record/potentials;
- Ph.D. (Electrical Engineering, Computer Engineering, Computer Science, or related field);
- A minimum relevant research experience of 4 years.

Applications:
Submit (all in English) *a cover letter, a 2-page research plan, a CV including copies of 3 most significant publications, and names of three referees* to: sist@shanghaitech.edu.cn.

Deadline: December 31st, 2013 (or until positions are filled). We have 10 positions for this round of faculty recruitment.

For more information, visit http://www.shanghaitech.edu.cn.

In response to the national demand for more qualified engineers, the Look College has embarked on “25 by 25,” an ambitious enrollment growth initiative to more than double the college’s current enrollment to 25,000 engineering students by the year 2025. The Department of Computer Science and Engineering will be an integral part of this planned growth.

We seek a dynamic, innovative candidate who can lead the growth in students, faculty, and staff that is envisioned. We are looking for an innovative thinker with a strategic vision for guiding the department to a higher level of excellence that can communicate this vision to a constituency that includes academia, government, industry, and alumni. Candidates should possess proven leadership and administrative skills, and an established reputation as a scholar consistent with an appointment to the rank of professor of computer science and engineering with tenure. Candidates must have a Ph.D. in computer science, computer engineering, or a related field.

The Department of Computer Science and Engineering has 36 tenured and tenure-track faculty members and three full-time lecturers. The department currently has faculty with a number of national distinctions, including National Academy of Engineering membership, ACM and IEEE Fellows, and ACM Distinguished Scientists and Engineers. The department has a strong and vibrant research program, with 47 percent of the faculty receiving NSF CAREER/NYI/PYI awards. The department offers bachelor’s, master’s, and Ph.D. degrees in computer science and - jointly with the Department of Electrical and Computer Engineering - in computer engineering, to roughly 275 graduate and 830 undergraduate students. In recent years, the department has built a strong national reputation based on the quality of its faculty and programs; among public institutions, its graduate computer engineering program ranks 12th and its graduate computer science 27th in the most recent U.S. News & World Report rankings. More information about the department is available at http://www.cse.tamu.edu.

Texas A&M, a land-grant, sea-grant, and space-grant institution, is one of the six largest universities in the United States and has more than 50,000 students. Today, the Dwight Look College of Engineering is one of the largest and best endowed in the nation, and it ranks among the top institutions in every significant national poll, including sixth for graduate programs and eighth for undergraduate programs among public institutions by U.S. News & World Report. The Look College has long enjoyed national leadership status in engineering education, and currently has more than 11,000 engineering students in 13 departments. Approximately 25 percent of the engineering students are graduate students.

Letters of application should include
(1) a full curriculum vitae,
(2) a two-page statement summarizing the candidate’s vision and goals for the department and leadership philosophy, and
(3) the names and addresses of at least five references.

Applications will be accepted until the position is filled; screening will begin immediately. **Nominations or applications should be sent to** headsearch@cse.tamu.edu.

The members of Texas A&M Engineering are all Affirmative Action/Equal Employment Opportunity Employers. It is the policy of these members in all aspects of operations each persons shall be considered solely on the basis of qualifications, without regard to race, color, sex, religion, national origin, age, disabilities or veteran status.

University of California, Santa Barbara
Department of Electrical and Computer Engineering
Tenure-track Faculty Position in Computer Engineering

The Electrical and Computer Engineering Department at the UC, Santa Barbara invites applications for a tenure-track faculty position in the area of Computer Engineering with a start date of Fall quarter, 2014. For more details please visit – http://www.ece.ucsb.edu/employment/. An EO/AA employer.

University of Central Florida
Computer Science
Tenure-Track Faculty Positions

The Computer Science Division in the College of Engineering and Computer Science at UCF is looking for exceptional tenure-track faculty at all levels, especially at senior levels. We are interested in hiring candidates with an established re-

---

udjobs
Employer of Choice

FACULTY POSITION IN ELECTRICAL AND COMPUTER ENGINEERING

The Department of Electrical and Computer Engineering at the University of Delaware invites nominations and applications for a tenure-track or tenured faculty position in electrical and computer engineering with a focus on cybersecurity or software engineering. We also invite candidates that complement the department's traditional strengths in (1) Computer Engineering, (2) Signal Processing, Communications & Controls, (3) Nanoelectronics, Electromagnetics, and Photonics, and (4) Biomedical Engineering.

The University of Delaware is located in northern Delaware, which is in close proximity to a banking industry hub, and to the US Army’s Communications-Electronics Research, Development and Engineering Center (CERDEC) and supporting businesses. Cybersecurity is a key area of focus for both industries and the University of Delaware has established research and education partnerships with JP Morgan Chase, CERDEC, and several defense businesses that can provide faculty members with research funding, data, and access to personnel working in cybersecurity. ECE initiatives are further supported by over 40,000 square feet of departmental facilities, including a state-of-the art 7,000 sq ft clean room for nano-fabrication, and fueled with over $10M/year in research expenditures.

Applicants should have a Ph.D. in electrical and computer engineering, computer science, or a closely related field. Candidates should have a demonstrated research record in an area of electrical engineering or computer engineering commensurate with their level of experience, have the potential to successfully secure competitive research awards from federal, state, and industry, have a passion and ability to teach core undergraduate courses and integrate research with teaching at the graduate level. Candidates are expected to supervise Masters and Ph.D. students as well as mentor undergraduate research.

Review of applications will begin immediately and will continue until the position is filled. We desire to fill the position by September 1, 2014. The University of Delaware is an equal opportunity employer.

For additional information regarding the position listed below and all open positions please visit the UDJOBS website at: www.udel.edu/udjobs Job ID 101217

Dare to be first.
UNIVERSITY OF DELAWARE

The UNIVERSITY OF DELAWARE is an Equal Opportunity Employer

Computer Science

Multiple senior faculty positions in computer science are available at Cornell's new Cornell NYC Tech campus in New York City. Faculty hired in these positions will be in the Department of Computer Science, which spans the Ithaca and New York City campuses, but their teaching and research will be based in New York City. We will consider only candidates at the Associate and Full Professor level, but will consider candidates from all areas of computer science and related fields.

Candidates must hold the equivalent of a Ph.D., must have demonstrated an ability to conduct outstanding research, and should also have a demonstrated interest in technology commercialization, entrepreneurship and social impact mission of the campus. In addition, experience in international programs and/or pre-college (K-12) education is of interest. This search may include Cornell faculty positions that are part of the Jacobs Technion-Cornell Innovation Institute. To ensure full consideration, applications should be received by December 1, 2013, but will be accepted until all positions are filled. Candidates should submit a curriculum vita, brief statements of research and teaching interests on-line at
https://academicjobsonline.org/ajo/jobs/2814

Diversity and inclusion have been and continue to be a part of our heritage. Cornell University is a recognized EEO/AA employer and educator.

cord of high-impact research and publication in: *big data, human-computer interaction, computer security, mobile computing, software engineering*, and *theory of computing*, though we will also consider remarkable candidates in other areas. We offer a competitive salary and start up package with generous benefits.

All applicants must have a Ph.D. from an accredited institution in an area appropriate to Computer Science and a strong commitment to the academic process, including teaching, scholarly publications, and sponsored research. Successful candidates will have a record of high-quality publications and be recognized for their expertise and the impact of their research.

Computer Science at UCF has a rapidly-growing educational and research program with nearly $4.6 million in research contracts and expenditures annually and over 179 graduate students, including 113 Ph.D. students. Computer Science has strong areas of research in Computer Vision, Machine Learning, Virtual and Mixed Reality, and HCI. More information about the Computer Science Division can be found at http://www.cs.ucf.edu/.

Research sponsors include NSF, NIH, NASA, DOT, DARPA, ONR, and other agencies of the DOD. Industry sponsors include AMD, Boeing, Canon, Electronic Arts, General Dynamics, Harris, Hitachi, Intel, Lockheed Martin, Oracle, SAIC, Symantec, Toyota USA, and Walt Disney World, as well as local startups.

UCF has about 60,000 students and is the nation's second largest university. Located in Orlando, UCF is at the center of the I-4 High Tech Corridor. The corridor has an excellent industrial base that includes: software, defense, space, simulation and training, and a world-renowned entertainment industry. Adjacent to UCF is a thriving research park that hosts more than 100 high-technology companies and the Institute for Simulation and Training. The Central Florida area is designated by the State of Florida as the Center of Excellence in Modeling and Simulation. UCF also has an accredited medical school, which opened in 2009. Great weather, easy access to the seashore, one of the largest convention centers in the nation, and one of the world's best airports are just a few features that make Orlando an ideal location. UCF is an Equal Opportunity/Affirmative Action employer. Women and minorities are particularly encouraged to apply.

To submit an application, please go to: http://www.cs.ucf.edu/facsearch

University of Central Missouri
Department of Mathematics and Computer Science
Assistant Professor of Computer Science - Tenure Track

The Department of Mathematics and Computer Science at the University of Central Missouri is accepting applications for a tenure-track position in Computer Science at the rank of Assistant Professor. Ph.D. in Computer Science is required. All areas in computer science will be considered with preference given to candidates with expertise in computer security, software engineering or big data analytics. For more information about the position and the application process, visit http://www.ucmo.edu/math-cs/openings.cfm.

University of Chicago
Department of Computer Science
Associate Professor

The Department of Computer Science at the **University of Chicago** invites applications from exceptionally qualified candidates in the areas of theory of computing, and systems for faculty positions at the rank of **Associate Professor**.

Systems is a broad, synergistic collection of research areas spanning systems and networking, programming languages and software engineering, software and hardware architecture, data-intensive computing and databases, graphics and visualization, and systems biology. Particular areas of focus include formal definition design and implementation of programming languages, data-intensive computing systems and algorithms, large scale distributed and collaborative systems, heterogeneous computer architectures, reliable computing systems, and self-tuning systems.

Theory of Computing is striving to understand and explore the most basic and fundamental principles underlying computation and related disciplines. While mathematical in its core, it also has very strong connections with machine learning, economics, bioinformatics and natural language processing, to name just a few. We encourage applications from researchers in any area whose work contains a significant theoretical component.

The University of Chicago has the highest standards for scholarship and faculty quality, is dedicated to fundamental research, and encourages collaboration across disciplines. We encourage strong connections with researchers across the campus in such areas as mathematics, natural language processing, bioinformatics, logic, molecular engineering, and machine learning, to mention just a few. Applicants must have a doctoral degree in Computer Science or a related field such as Mathematics or Statistics. Applicants are expected to have established an outstanding research program and will be expected to contribute to the department's undergraduate and graduate teaching programs.

The Department of Computer Science (cs.uchicago.edu) is the hub of a large, diverse computing community of two hundred researchers focused on advancing foundations of computing and driving its most advanced applications. Long distinguished in theoretical computer science and artificial intelligence, the Department is now building strong systems and machine learning groups. The larger community in these areas at the University of Chicago includes the Department of Statistics, the Computation Institute, the Toyota Technological Institute, and Argonne's Mathematics and Computer Science Division.

The Chicago metropolitan area provides a diverse and exciting environment. The local economy is vigorous, with international stature in banking, trade, commerce, manufacturing, and transportation, while the cultural scene includes diverse cultures, vibrant theater, world-renowned symphony, opera, jazz, and blues. The University is located in Hyde Park, a Chicago neighborhood on the Lake Michigan shore just a few minutes from downtown.

A cover letter, curriculum vitae including a list of publications, a statement describing past and current research accomplishments and outlining future research plans, a description of teaching philosophy and a reference contact list consisting of three people are required.

Review of complete applications will begin January 15, 2014, and will continue until all available positions are filled.

All applicants must apply through the University's Academic Jobs website.

To apply for the position of Associate Professor-Theory, go to: http://tinyurl.com/kkldt2f

To apply for the position of Associate Professor-Systems, go to: http://tinyurl.com/ll8h8a3

The University of Chicago is an Affirmative Action / Equal Opportunity Employer.

University of Chicago
Department of Computer Science
Lecturer

The Department of Computer Science at the **University of Chicago** invites applications for the position of **Lecturer**. Subject to the availability of funding, this would be a three year position with the possibility of renewal. This position involves teaching in the fall, winter and spring quarters. The successful candidate will have exceptional competence in teaching and superior academic credentials, will carry responsibility for teaching computer science courses and laboratories. Applicants must have a Ph.D. in Computer Science or a related field at time of appointment and have experience teaching Computer Science at the College level.

The Chicago metropolitan area provides a diverse and exciting environment. The local economy is vigorous, with international stature in banking, trade, commerce, manufacturing, and transportation, while the cultural scene includes diverse cultures, vibrant theater, world-renowned symphony, opera, jazz and blues. The University is located in Hyde Park, a Chicago neighborhood on the Lake Michigan shore just a few minutes from downtown.

Applicants must apply on line at the University of Chicago Academic Careers website at http://tinyurl.com/m3z99rw.

Applicants must upload a curriculum vitae with a list of publications and a one page teaching statement. In addition, three reference letters will be required. Review of complete applications, including reference letters, will begin January 15, 2014, and continue until the position is filled.

The University of Chicago is an Affirmative Action / Equal Opportunity Employer.

University of Colorado, Boulder

Department of Computer Science

Assistant Professor

The Department of Computer Science at the University of Colorado Boulder seeks outstanding candidates for a tenure-track position in human-centered computing (HCC). The opening is targeted at the level of Assistant Professor, although outstanding senior candidates at higher ranks may be considered.

We seek candidates with promising research records in the areas of human-computer interaction, computer-supported cooperative work, social computing, ubiquitous computing, and information visualization. Candidates should have an orientation to computer science as their primary teaching home, though an interdisciplinary research program is welcomed and has been a hallmark of HCC research at CU.

Applications must be submitted on-line at http://www.jobsatcu.com/postings/71832

The University of Colorado is an Equal Opportunity/Affirmative Action employer.

University of Illinois at Urbana-Champaign

Department of Electrical and Computer Engineering (ECE)

Faculty Positions

The Department of Electrical and Computer Engineering (ECE) at the University of Illinois at Urbana-Champaign invites applications for **faculty positions** at all areas and levels in Computing, broadly defined, with particular emphasis on big data, including complex data analysis and decision making; scalable hardware and software systems; parallel, high-performance, and energy-efficient computing; reliable and secure computing; bioinformatics and systems biology; and networking and distributed computing, among other areas. From the transistor and the first computer implementation based on von Neumann's architecture to the Blue Waters petascale computer – the fastest computer on any university campus, ECE Illinois faculty have always been at t he forefront of computing research and innovation. Applications are encouraged from candidates whose research programs specialize in traditional, nontraditional, and interdisciplinary areas of electrical and computer engineering. The department is engaged in exciting new and expanding programs for research, education, and professional development, with strong ties to industry.

Qualified senior candidates may also be considered for tenured full Professor positions as part of the Grainger Engineering Breakthroughs Initiative (http://graingerinitiative.engineering.illinois.edu), which is backed by a $100-million gift from the Grainger Foundation to support research in big data and bioengineering, broadly defined. In addition, the University of Illinois is home to Blue Waters - one of the most powerful supercomputers in the world, supported by the National Science Foundation and developed and operated by the University of Illinois' National Center for Supercomputing Applications. Qualified candidates may be hired as Blue Waters Professors who will be provided substantial allocations on and expedited access to the supercomputer. To be considered as a Blue Waters Professor, candidates need to mention Blue Waters as one of their preferred research areas in their online application, and include a reference to Blue Waters in their cover letter.

Please visit http://jobs.illinois.edu to view the complete position announcement and application instructions. Full consideration will be given to applications received by December 15, 2013.

Illinois is an AA-EOE. www.inclusiveillinois.illinois.edu

University of Kansas

Department of Electrical Engineering and Computer Science

Electrical Engineering/Computer Engineering / Computer Science Faculty

The University of Kansas (KU) Department of Electrical Engineering and Computer Science (EECS) seeks outstanding individuals for five tenure track positions in the disciplines of electrical engineering, computer engineering, or computer science. Successful candidates are expected to contribute to the development of academic and research programs and to contribute to the research community. Applicants must have an earned doctorate or equivalent in electrical engineering, computer science, computer engineering, or related fields at the time of appointment. The following positions are open at the assistant professor (tenure-track) level, with the possibility of higher rank for the exceptional applicant, with experience in:

Electronic/Photonic Devices & Sensors -- semiconductor materials, devices and sensors; nanotechnology and MEMS; high power electronics; meta-materials, plasmonics, and spintronics; optoelectronics, nanophotonics, and ultrafast photonics; or microwave, millimeter-wave, and THz devices

High Speed Digital Systems – advanced computer architectures; hybrid hardware/software systems; embedded systems; digital logic synthesis and analysis for VLSI and FPGAs; multi-core and system-on-a-chip architectures; or high assurance computer architectures

Bioinformatics – machine learning, data mining, statistical learning, distributed databases, or big data analytics focused on computational life sciences problems

High Performance Computing (HPC) – parallel computing employing multi-core, GPU, or special-purpose architectures; parallel and distributed algorithms; or data-intensive computing

Image Processing/Visualization – image and video processing, analysis, and vision; scientific visualization; or analysis of algorithms, data structures, or graph theory.

Exceptional applicants in other closely related areas to the above topics may be considered.

EECS leads the KU School of Engineering with 36 faculty members and a research volume of over $8 million per year. The EECS department offers undergraduate and graduate degrees in electrical engineering, computer engineering, computer science, interdisciplinary computing, and information technology. The department has approximately 500 undergraduate and 250 graduate students. EECS faculty engage in collaborative research projects across multiple areas of interest within the department and the University. There are many interdisciplinary opportunities involving schools at KU and the KU Medical Center.

The KU School of Engineering is rapidly expanding and plans to add thirty new faculty lines in the next 5 years with expanded research and teaching facilities. The University of Kansas is focused on four key campus-wide strategic initiatives: (1) Sustaining the Planet, Powering the World; (2) Promoting Well-Being, Finding Cures; (3) Building Communities, Expanding Opportunities; and (4) Harnessing Information, Multiplying Knowledge. For more information, see http://www.provost.ku.edu/planning/themes/. Successful candidates should describe in their application materials how their work addresses one or more of KU's strategic initiatives.

Applications should be submitted at http://employment.ku.edu, **select "Search Faculty Jobs"**. Applications should include a letter of application, curriculum vita, a statement of research interests and plans, a statement of teaching interests and plans, and contact information for three references. Applications will be reviewed beginning December 1, 2013 and will be accepted until the position is filled. The appointment will be effective as negotiated. Questions can be sent to: EECS_Search@eecs.ku.edu. EOE M/F/D/V.

University of Michigan, Ann Arbor

Department of Electrical Engineering and Computer Science

Computer Science and Engineering Division

Faculty Positions

Applications and nominations are solicited for multiple faculty positions in the Computer Science and Engineering (CSE) Division. Highly qualified candidates from all areas of computer science and computer engineering will be considered for positions at all levels and are encouraged to apply. Particularly encouraged are applicants with research interests in the following areas.

1. Software Systems, including distributed systems and networking.
2. Computer Vision, including research applied to robotics and transportation problems.
3. Medical Computing, including machine learning, big data, probabilistic reasoning, and visualization approaches to medicine and, more broadly, healthcare.
4. Theoretical Computer Science, in which we are particularly interested in candidates for an endowed professorship at the Associate or Full Professor level.

Qualifications include an outstanding academic record, a doctorate or equivalent in computer science or computer engineering or a discipline relevant to the above areas, and a strong commitment to teaching and research. **Applications must be received by January 1, 2014**.

To apply, please complete the form at:
http://www.eecs.umich.edu/eecs/etc/jobs/csejobs.html

Electronic applications are strongly preferred, but you may alternatively send resume, teaching statement, research statement and names of three references to:

Professor Satinder Singh Baveja, Chair, CSE Faculty Search
Department of Electrical Engineering and Computer Science
University of Michigan
2260 Hayward Street
Ann Arbor, MI 48109-2121

The University of Michigan is a Non-Discriminatory/Affirmative Action Employer with an Active Dual-Career Assistance Program. The college is especially interested in candidates who can contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

University of North Florida
School of Computing
Tenure-track Assistant Professor

The School of Computing at the **University of North Florida** invites applications for a **tenure-track assistant professor** position beginning August 2014. Interested applicants must have an earned Ph.D. in computer science or closely related discipline. Preferred research areas include software engineering, traditional computer science, and big data. A commitment to excellence in teaching undergraduate and graduate courses in computing and guiding graduate students thesis work is expected plus demonstrated ability to establish/maintain a productive research program. Applicants must complete an online application at www.unfjobs.org to include: a letter of interest, a curriculum vitae, a list of three references with contact information, a teaching statement, and a vision for research. Please direct questions to Dr. Bob Roggio at broggio@unf.edu. Initial review of applications will begin on October 1, 2013.

University of North Texas
Department of Computer Science and Engineering
Assistant Professor and Lecturers

The Department of Computer Science and Engineering at the University of North Texas (UNT) is seeking candidates for one tenure track faculty position at the Assistant Professor level and two non-tenure track Lecturers beginning August 15, 2014.

The tenure track faculty position will expand our existing strengths in Natural Language Processing and closely related areas. This tenure track faculty position will expand our existing strengths in Natural Language Processing and closely related areas. Outstanding candidates with significant publication records and research funding in other areas will also be considered. Candidates should have demonstrated the potential to excel in research in one or more of the areas and in teaching. A Ph.D. in Computer Science or a closely related field is required at the time of appointment.

The lecturer positions will focus on introductory courses spanning all programs in Computer Science and Engineering and advanced undergraduate courses in Computer Engineering. A Ph.D. in Computer Science, Computer Engineering, or a closely related field is required at the time of appointment.

The Computer Science and Engineering department is home to 790 Bachelors students, 150 Masters students and 82 Ph.D. students. Additional information about the department is available at the department's website at: www.cse.unt.edu.

Application Procedure:
All applicants must apply online to: https://facultyjobs.unt.edu. Submit nominations and questions regarding the tenure track position to Dr. Yan Huang (huangyan@unt.edu), and regarding the Lecturer positions to Dr. Armin R. Mikler (mikler@cse.unt.edu).

Application Deadline:
The committee will begin its review of applications on November 1, 2013 and continue to accept and review applications until the positions are closed.

The University:
A student-focused public research university, UNT is the nation's 25th largest public university and the largest, most comprehensive in the Dallas-Fort Worth area. It is dedicated to providing an excellent educational experience to its 36,000 students while powering the North Texas region through innovative education and research. UNT's ultimate mission is to give a green light to greatness by helping its students, region, state and nation excel.

The University of North Texas is an AA/ADA/EOE committed to diversity in its educational programs.

University of Oregon
Department of Computer and Information Science
Faculty Position
Assistant Professor

The Department of Computer and Information Science (CIS) seeks applications for a tenure track faculty position at the rank of Assistant Professor, beginning Fall 2014. The University of Oregon is an AAU research university located in Eugene, two hours south of Portland, and within one hour's drive of both the Pacific Ocean and the snow-capped Cascade Mountains.

The department seeks candidates in one of our focus areas – parallel computing, distributed systems, and data sciences. We will also consider candidates in theoretical computer science who can collaborate with faculty members in our focus areas or in security. Applicants must have a Ph.D. in computer science or closely related field, a demonstrated record of excellence in research, and a strong commitment to teaching. A successful candidate will be expected to conduct a vigorous research program and to teach at both the undergraduate and graduate levels.

We offer a stimulating, friendly environment for collaborative research both within the department, which expects to grow substantially in the next few years, and with other departments on campus. The CIS Department is part of the College of Arts and Sciences and is housed within the Lorry Lokey Science Complex. The department offers B.S., M.S. and Ph.D. degrees. More information about the department, its programs and faculty can be found at http://www.cs.uoregon.edu.

Applications will be accepted electronically through the department's web site. Application information can be found at http://www.cs.uoregon.edu/Employment/ . Review of applications will begin January 15, 2014, and continue until the position is filled. Please address any questions to faculty.search@cs.uoregon.edu.

The University of Oregon is an equal opportunity/affirmative action institution committed to cultural diversity and is compliant with the Americans with Disabilities Act. We are committed to creating a more inclusive and diverse institution and seek candidates with demonstrated potential to contribute positively to its diverse community.

University of São Paulo
Institute of Mathematics and Statistics
Department of Computer Science
Tenure-Track Positions
Assistant Professor

The Institute of Mathematics and Statistics of the University of São Paulo (IME-USP) invites applications for faculty positions at the Assistant Professor level. The Department is accepting applications in all areas of Computer Science.

We expect candidates with strong potential for research and teaching ability. The candidates should have a PhD in Computer Science or a related area. The selected candidate will be responsible for developing research and for teaching to the programs of the department (undergraduate and graduate courses). Deadlines and documents required for the application are specified at www.ime.usp.br/dcc/faculty_position The documents and selection interview may be either in Portuguese or English.

The University of São Paulo - USP is one of the most prestigious educational institutions in South America. It is the best ranked Brazilian university. The Department of Computer Science of the IME-USP is responsible for the BSc, MSc and PhD degrees in Computer Science, offering some of the most competitive courses in Brazil. The Department traditionally has open positions every year and potential candidates should contact us by email at any time.

More information:

http://www.ime.usp.br/dcc

Contact: mac@ime.usp.br

University of South Carolina Upstate

Assistant Professor of Informatics

Assistant Professor of Informatics needed 8/16/14 to teach information management and information technology. For complete job description, requirements and online application submission process go to www.uscupstate.edu/jobs. (Requisition #006662)

Contact: Dr. Ron Fulbright, rfulbright@uscupstate.edu; 864-503-5683; 800 University Way, Spartanburg, SC 29303

The USC Upstate is an equal opportunity institution.

University of Texas at Austin

Department of Computer Science

Tenured/Tenure-Track Faculty Positions

The Department of Computer Science of the University of Texas at Austin invites applications for tenure-track positions at all levels. Outstanding candidates in all areas of Computer Science will be considered, particularly in Formal Methods, Big Data, and Robotics. All tenured and tenure-track positions require a Ph.D. or equivalent degree in computer science or a related area at the time of employment.

Successful candidates are expected to pursue an active research program, to teach both graduate and undergraduate courses, and to supervise graduate students. The department is ranked among the top ten computer science departments in the country.

It has 44 tenured and tenure-track faculty members across all areas of computer science. Many of these faculty participate in interdisciplinary programs and centers in the University, including the Texas Advanced Computing Center (TACC), and those in Computational and Applied Mathematics, Computational Biology, and Neuroscience.

Austin, the capital of Texas, is a center for high-technology industry, including companies such as IBM, Dell, Freescale Semiconductor, Advanced Micro Devices, National Instruments, AT&T, Intel and Samsung. For more information please see the department web page: http://www.cs.utexas.edu/

The department prefers to receive applications online, beginning September 3, 2013. To submit yours, please visit http://www.cs.utexas.edu/faculty/recruiting.

Applicants for an assistant professor position must have at least three (3) referees send letters of reference directly to the address provided. Applicants for a tenured position (associate or full professor) must have at least six (6) referees send letters of reference directly.

Inquiries about your application may be directed to faculty-search@cs.utexas.edu. For full consideration of your application, please apply by January 31, 2014. Complete applications (including all reference letters) will begin being reviewed on December 16th. Women and minority candidates are especially encouraged to apply. The University of Texas is an Equal Opportunity Employer.

University of Wisconsin-Platteville

Department of Computer Science

Assistant Professor of Computer Science and/or Software Engineering

The Department of Computer Science and Software Engineering at the **University of Wisconsin-Platteville** invites applications for two tenure-track faculty positions at the assistant professor level.

Candidates must have a Ph.D. in Software Engineering or Computer Science or a closely related field. Applicants must have a strong commitment to undergraduate education and teaching excellence.

Position starting August 19, 2014. See http://www.uwplatt.edu/pers/faculty.htm for complete announcement details and application instructions. AA/EEO employer.

University of Wisconsin School of Medicine & Public Health

Department of Biostatistics & Medical Informatics

Morgridge Institute for Research

Assistant (tenure-track) or Associate/Full (tenured) Professor

Department of Biostatistics & Medical Informatics at the University of Wisconsin School of Medicine & Public Health, and Morgridge Institute for Research, seek Assistant (tenure-track) or Associate/Full (tenured) Professor in biostatistics or bioinformatics. Innovative position combines research in methodology with applications in virology; includes teaching, graduate student training. PhD in Biostatistics, Statistics, Bioinformatics, Computational Biology, Biomedical Informatics, Computer Sciences, or related area. Please see:

http://www.biostat.wisc.edu/About/Jobs/BMI-MIR-Ad-2013-final.pdf .

AA/EOE

University of Wisconsin-Stevens Point

Department of Computing and New Media Technologies

Assistant/Associate Professor

The fast growing and dynamic **Department of Computing and New Media Technologies** at the **University of Wisconsin-Stevens Point** invites applications for 3 tenure-line faculty positions at the rank of **assistant/associate professor** with appointments starting August 2014. The positions focus on: (1) **Software Development** - OO programming, algorithms and data structures, software engineering – analysis, design and testing of applications, networking, information assurance and cloud computing; (2) **Web Development** – server-side and client-side technologies, OO programming, rich Internet application technologies; and (3) **Multimedia** - multimedia design and development, digital video and audio production, aesthetics and human computer interaction (HCI), game development, 3D modeling. Check position descriptions and electronic application procedure at: http://www.uwsp.edu/equity/Pages/jobVacancies.aspx.

Applicant screening will begin on January 2, 2014 and continue until the positions are filled.

U.S. Air Force Academy

Department of Computer Science

Coleman-Richardson Chair of Computer Science

The United States Air Force Academy Department of Computer Science is accepting applications for the Coleman-Richardson Endowed Chair for the 2014-15 academic year. This appointment is designed to bring in distinguished professionals to enhance the cadet educational experience. Interested candidates should contact the Search Committee Chair, Dr. Barry Fagin, at barry.fagin@usafa.edu or 719-333-7377. Detailed information is available at http://www.usafa.edu/df/dfcs/dfcs/dfcsjobs.cfm. US citizenship is required.

U.S. Naval Academy

Computer Science Department

Distinguished Visiting Professor

The U.S. Naval Academy's Computer Science Department invites applications for one or more Distinguished Visiting Professors. The visiting professor is expected to have a strong reputation and technical expertise in Computer Science, Information Technology, or a closely related field. A Ph.D. or commensurate experience is required.

The start date of this position is flexible, and we have provisions for either full-time financial support or supplemental support for a professor on sabbatical. The position duration could vary from half a year to multiple years. The specific salary will be commensurate with experience and the expected contributions to the department. Such contributions may vary and might include teaching required or elective courses, developing new courses, leading research seminars, organizing relevant symposia, collaborating with faculty, and/or mentoring student research.

The Computer Science Department offers majors in Computer Science and Information Technology, and contributes to a new major in Cyber Operations. We currently have 100 CS majors, 70 IT majors and a faculty of 18. In addition to courses for CS, IT, and Cyber Operations majors, we also teach a course on cyber security to the entire freshman class.

The department is housed in a state of the art building overlooking the scenic Severn River. Our spaces provide outstanding office, laboratory, and research facilities for both students and faculty, including specialized labs for information assurance, networking, and robotics, as well as three micro-computing labs and two high performance computing labs. Further information is available at http://www.usna.edu/cs/ and http://www.usna.edu/.

Applicants should send a cover letter, curriculum vitae, and a list of three professional references. The cover letter should address in detail the expected contributions to the department, as described above. All materials should be sent to cssearch@usna.edu.

The United States Naval Academy does not discriminate in employment on the basis of race, color, religion, sex (including pregnancy and gender identity), national origin, political affiliation, sexual orientation, marital status, disability, genetic information, age, membership in an employee organization, retaliation, parental status, military service, or other non-merit factor.

Washington University in St. Louis
Department of Computer Science & Engineering
Faculty Positions

The Department of Computer Science & Engineering at Washington University seeks outstanding tenure-track faculty in all areas of computer science and engineering.

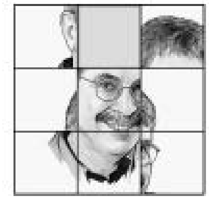
The department has been growing and plans to continue expanding its faculty size significantly in the coming years. We seek multiple talented and highly motivated individuals who will build transformative research programs, both through work in the core disciplines of computer science and computer engineering and through interdisciplinary collaborations with researchers in areas such as biomedicine, engineering, and the sciences. Successful candidates must show exceptional promise for research leadership and a strong commitment to high-quality teaching. Candidates will be expected to publish their research in peer-reviewed conferences and journals, to teach, and to participate in department and university service.

Our faculty is engaged in a broad range of research activities. Some key strategic themes of our research include: tight integration of computation with the human and physical worlds, the extraction of knowledge from massive data sets; and the design of safe, secure, and scalable computing systems. The impact of our work is magnified through interdisciplinary collaborations throughout the School of Engineering; with colleagues in the sciences, arts, and humanities; and with our world-renowned School of Medicine. Our doctoral graduates go on to leadership positions in both academia and industry. The department values both fundamental and applied research and has a strong tradition of successful technology transfer, and our faculty is known for its collegiality and for providing a supportive environment for new arrivals.

Washington University is a private university with 6,500 full-time undergraduates and 6,500 graduate students. It is nationally known for the exceptional quality of its student body.

We anticipate appointments at the rank of Assistant Professor; however, exceptionally qualified applicants may be considered for appointments at the Associate or Full Professor level. Applicants should hold a doctorate in Computer Science, Computer Engineering, or a closely related field. Qualified applicants should submit a complete application (cover letter, curriculum vitae, research statement, teaching statement, and contact information for at least three references) through AcademicJobsOnline at https://academicjobsonline.org/ajo/jobs/3095. Other communications may be directed to Prof. Roch Guérin, Department of Computer Science and Engineering, Campus Box 1045, Washington University, One Brookings Drive, St. Louis, MO 63130-4899.

Applications received by January 1, 2014 will receive full consideration. Washington University is an AA/EOE and is strongly committed to enhancing the diversity of its faculty; minority and women applicants are especially encouraged. Employment eligibility verification will be required upon employment.



DOI:10.1145/2524713.2524726

Peter Winkler

Puzzled
Coin Flipping

Each of these puzzles involves coin flipping. Simple stuff, right? Not necessarily…though solutions will indeed be provided in next month's column.

1. You have just joined the Coin Flippers of America, and, naturally, the amount of your annual dues will be determined by chance. First you must select a head-tail sequence of length five. A coin is then flipped (by a certified CFA official) until that sequence is achieved with five consecutive flips. Your dues is the total number of flips, in U.S. dollars; for example, if you choose HHHHH as your sequence, and it happens to take 36 flips before you get a run of five heads, your annual CFA dues will be $36. What sequence should you pick? HHHHH? HTHTH? HHHTT? Does it even matter?

2. Now you have entered your first CFA tournament, facing your first opponent. Each of you will pick a different head-tail sequence of length five, and a coin will be flipped repeatedly. Whoever's sequence arises first is the winner. You have the option of choosing first. What sequence should you pick?

3. Following the tournament (which you win), you are offered a side bet. You pay $1 and flip a coin 100 times; if you get exactly 50 heads, you win $20 (minus your dollar). If you lose, you are out only the $1. Even so, should you take the bet?

PHOTOGRAPHS BY JACOB HAMBLIN, ALEX KALMBACH

Readers are encouraged to submit prospective puzzles for future columns to puzzled@cacm.acm.org.

Peter Winkler (puzzled@cacm.acm.org) is William Morrill Professor of Mathematics and Computer Science at Dartmouth College, Hanover, NH.

10th World Congress on Services (SERVICES 2014)

Services Congress

June 27—July 2, 2014, Anchorage, Alaska, USA

Federation of 5 Service-Centric Conferences from Different Angles

(http://www.ServicesCongress.org)

7th International Conference on Cloud Computing (CLOUD 2014)

Cloud Computing

Cloud Computing is becoming a scalable services delivery and consumption platform in the field of Services Computing. The technical foundations of Cloud Computing include Service-Oriented Architecture and Virtualizations. Major topics cover Infrastructure Cloud, Software Cloud, Application Cloud, Social Cloud, & Business Cloud. Visit http://thecloudcomputing.org.

21th International Conference on Web Services (ICWS 2014)

ICWS Web Services

ICWS 2014 will feature **web-based** services modeling, design, development, publishing, discovery, composition, testing, QoS assurance, adaptation, and delivery technologies and standards. Visit http://icws.org.

11th International Conference on Services Computing (SCC 2014)

SCC Services Computing

SCC 2014 will focus on **services innovation lifecycle** e.g., enterprise modeling, business consulting, solution creation, services orchestration, optimization, management, and BPM. Visit confer-

3rd International Conference on Mobile Services (MS 2014)

Mobile Services

MS 2014 will feature **Wearable technology and applications**. Topics cover but not limited to all innovative aspects of wearable devices, programming models, integration, and domain specific solutions.

Visit themobileservices.org/2014.

International Congress on Big Data (BigData 2014)

Big Data

BigData 2014 aims to explore various aspects of Big Data including modeling, storage architecture (NOSQL), enterprise transformation, text mining, social networks, applied analytics and various applications, and Big Data As A Service. Visit http://ieeebigdata.org/2014.

Conference proceedings have been EI indexed. Extended versions of invited ICWS/SCC/CLOUD/MS/BigData papers will be published in IEEE Transactions on Services Computing (TSC, SCI & EI indexed), International Journal of Web Services Research (JWSR, SCI & EI indexed), International Journal of Business Process Integration and Management (IJBPIM), IT Pro (SCI & EI Indexed), and the worldwide Services Computing community's Open Access journals (IJCC, IJBD, and IJSC).

IEEE

IBM

IBM Research

hp

Services Computing

SAP

Submission Deadlines

ICWS 2014: 1/15/2014
CLOUD 2014: 1/15/2014
SCC 2014: 1/31/2014
MS 2014: 1/31/2014
BigData 2014: 2/10/2014
SERVICES 2014: 2/10/2014

Any questions, contact the organizing committee at services.ieeecs@gmail.com

HUAWEI

IEEE TRANSACTIONS ON SERVICES COMPUTING