

# Innovative Technology for Computer Professionals Computer

SEPTEMBER 2011

## SECURITY AND PRIVACY IN AN ONLINE WORLD

CYBERBULLYING, P. 93

SOFT BIOMETRICS, P. 106

HAS EVERYTHING BEEN INVENTED?, P. 112

<http://www.computer.org>



Computer SEPTEMBER 2011

Online Security and Privacy//Server Energy Proportionality//Cyberphysical Spaces

Volume 44 Number 9




**INTERNATIONAL  
CONFERENCE ON  
PARALLEL  
PROCESSING**

# **CALL FOR PAPERS**

- The 41<sup>st</sup> Annual Conference -

## **2012 International Conference on Parallel Processing (ICPP-2012)**

<http://www.icpp2012.org>

**Pittsburgh, PA, USA  
September 10-13, 2012**

**Sponsored by**

The International Association for Computers and Communications (IACC)

**In cooperation with**

The University of Pittsburgh, USA

### **Organizing & Program Committees**

#### **General Chair**

Wu-chun Feng, Virginia Tech, USA

#### **Program Chair**

Manish Parashar, Rutgers Univ., USA

#### **Program Vice-Chairs**

##### **Architecture**

Hiroshi Nakashima, Kyoto Univ., Japan

##### **Algorithms Design and Parallelism**

Srinivas Aluru, Iowa State Univ., USA

##### **Programming Models, Languages and Envs.**

Thierry Priol, INRIA, France

##### **Compilers and Runtime Systems**

Barbara Chapman, Univ. of Houston, USA  
Xipeng Shen, William & Mary, USA

##### **Networking and Communications**

Ron Brightwell, Sandia National Labs., USA  
Guihai Chen, Shanghai Jiaotong Univ., China

##### **Performance Modeling and Evaluation**

Krishna Kant, National Science Foundation, USA  
David Lowenthal, Univ. of Arizona, USA

##### **Applications**

David Abramson, Monash Univ., Australia

#### **Program Committee Members**

(Please see the conference web page)

#### **Workshops Co-Chairs**

Pavan Balaji, Argonne Nat'l Lab, USA  
Heshan Lin, Virginia Tech, USA

#### **Awards Co-Chairs**

Guang R. Gao, Univ. of Delaware, USA  
Yu-Chee Tseng, Nat. Chiao Tung Univ., Taiwan

#### **Publications Chair**

Feng Qin, Ohio State University, USA

#### **Publicity Co-Chairs**

Martin Schulz, Lawrence Livermore National Lab, USA  
Bary Rountree, Lawrence Livermore National Lab, USA

#### **Local Arrangements Chair**

Youtao Zhang, Univ. of Pittsburgh, USA

#### **International Liaison Co-Chairs**

Tarek S. Abdelrahmaner, U. Toronto, Canada  
Yves Robert, Ecole Normale Supérieure de Lyon, France  
Pen-Chung Yew, Academia Sinica, Taiwan

#### **Registration Chair**

Carrie Stein, Ohio State University, USA

#### **Steering Committee Chair**

Ten H. Lai, Ohio State University, USA

### **Scope**

The International Conference on Parallel Processing (ICPP) provides a forum for engineers and scientists in academia, industry and government to present their latest research findings in all aspects of parallel and distributed computing.

ICPP 2012 will be focused on 3 crosscutting themes: **(1) Disruptive Technologies (Multicore/Manycore, Accelerators, Clouds), (2) Data-Intensive Competing and (3) Green Computing.** The meeting will be organized around the following tracks:

- Architecture
- Algorithm Design and Parallelism
- Programming Models, Languages & Environments
- Networking and Communication
- Performance Modeling and Evaluation
- Compilers and Run-Time Systems
- Applications

### **Paper Submission**

Paper submissions should be formatted according to the IEEE standard double-column format with a font size 10 pt or larger. Each paper is strictly limited to 10 pages in length. Submissions should represent original, substantive research results. We will not accept any paper which, at the time of submission, is under review for or has already been published (or accepted) for publication in another conference or journal venue. See the conference website for electronic paper submission instructions.

### **Conference Timeline**

Paper Submission Deadline	<b>March 02, 2012</b>
Author Notification	June 01, 2012
Final Manuscript Due	July 06, 2012

**Workshops** with more narrowly focused scope will be held from September 10th to 13th. Proposals should be submitted to the Workshops Co-Chairs, Pavan Balaji ([balaji@mcs.anl.gov](mailto:balaji@mcs.anl.gov)) and Heshan Lin ([hlin2@vt.edu](mailto:hlin2@vt.edu)) by **November 1, 2011.**

**Proceedings** of the conference and workshops will be published in CD format and will be available at the conference.

**For Further Information** please contact:

Manish Parashar, Rutgers University, [parashar@rutgers.edu](mailto:parashar@rutgers.edu)

# Innovative Technology for Computer Professionals

# Computer

**Editor in Chief**

**Ron Vetter**  
University of North Carolina  
Wilmington  
[vetterr@uncw.edu](mailto:vetterr@uncw.edu)

**Associate Editor in Chief**

**Sumi Helal**  
University of Florida  
[helal@cise.ufl.edu](mailto:helal@cise.ufl.edu)

**Associate Editor in Chief, Research Features**

**Kathleen Swigger**  
University of North Texas  
[kathy@cs.unt.edu](mailto:kathy@cs.unt.edu)

**Associate Editor in Chief, Special Issues**

**Bill N. Schilit**  
Google  
[schilit@computer.org](mailto:schilit@computer.org)

**Computing Practices**

**Rohit Kapur**  
Synopsis  
[rohit.kapur@synopsys.com](mailto:rohit.kapur@synopsys.com)

**Perspectives**

**Bob Colwell**  
[bob.colwell@comcast.net](mailto:bob.colwell@comcast.net)

**Web/Multimedia Editor**

**Charles R. Severance**  
[csev@umich.edu](mailto:csev@umich.edu)

**2011 IEEE Computer Society President**

**Sorel Reisman**  
[s.reisman@computer.org](mailto:s.reisman@computer.org)

**Area Editors****Computer Architectures**

**Tom Conte**  
Georgia Tech  
**Steven K. Reinhardt**  
AMD

**Distributed Systems**

**Jean Bacon**  
University of Cambridge

**Graphics and Multimedia**

**Oliver Bimber**  
Johannes Kepler University Linz

**High-Performance Computing**

**Vladimir Getov**  
University of Westminster

**Information and Data Management**

**Naren Ramakrishnan**  
Virginia Tech

**Multimedia**

**Savitha Srinivasan**  
IBM Almaden Research Center

**Networking**

**Ahmed Helmy**  
University of Florida

**Security and Privacy**

**Rolf Oppliger**  
eSECURITY Technologies

**Software**

**Robert B. France**  
Colorado State University

**David M. Weiss**

Iowa State University

**Web Engineering**

**Simon Shim**  
San Jose State University

**Column Editors****Discovery Analytics**

**Naren Ramakrishnan**  
Virginia Tech

**Education**

**Ann E.K. Sobel**  
Miami University

**Entertainment Computing**

**Kelvin Sung**  
University of Washington, Bothell

**Green IT**

**Kirk W. Cameron**  
Virginia Tech

**Identity Sciences**

**Karl Ricanek**  
University of North Carolina, Wilmington

**In Development**

**Chris Huntley**  
Fairfield University

**Industry Perspective**

**Sumi Helal**  
University of Florida

**Invisible Computing**

**Albrecht Schmidt**  
University of Stuttgart

**The Known World**

**David A. Grier**  
George Washington University

**The Profession**

**Neville Holmes**  
University of Tasmania

**Security**

**Jeffrey M. Voas**  
NIST

**Social Computing**

**John Riedl**  
University of Minnesota  
**Software Technologies**  
**Mike Hinchey**  
Lero—the Irish Software Engineering Research Centre

**Advisory Panel**

**Carl K. Chang**  
Editor in Chief Emeritus  
Iowa State University  
**Hal Berghel**  
University of Nevada, Las Vegas  
**Doris L. Carver**  
Louisiana State University  
**Ralph Cavin**  
Semiconductor Research Corp.  
**Rick Mathieu**  
James Madison University  
**Naren Ramakrishnan**  
Virginia Tech  
**Theresa-Marie Rhyne**  
Consultant  
**Alf Weaver**  
University of Virginia

**Publications Board**

**David A. Grier** (chair),  
**Alain April**, **David Bader**,  
**Angela R. Burgess**, **Jim Cortada**, **Hakan Erdogmus**,  
**Frank E. Ferrante**, **Jean-Luc Gaudiot**, **Paolo Montuschi**,  
**Dorée Duncan Seligmann**,  
**Linda I. Shafer**, **Steve Tanimoto**,  
**George Thiruvathukal**

**Magazine Operations Committee**

**Dorée Duncan Seligmann** (chair), **Erik R. Altman**, **Isabel Beichl**, **Krishnendu Chakrabarty**, **Nigel Davies**, **Simon Liu**, **Dejan Milojičić**, **Michael Rabinovich**, **Forrest Shull**, **John R. Smith**, **Gabriel Taubin**, **Ron Vetter**, **John Viega**, **Fei-Yue Wang**, **Jeffrey R. Yost**

**Editorial Staff**

**Judith Prow**  
Managing Editor  
[jprow@computer.org](mailto:jprow@computer.org)  
**Chris Nelson**  
Senior Editor

**Contributing Editors**

**Camber Agrelius**  
**Lee Garber**  
**Bob Ward**

**Design and Production**

**Larry Bauer**  
**Design**  
**Olga D'Astoli**  
**Cover Design**  
**Kate Wojogbe**  
**Jennie Zhu**

**Administrative Staff**

**Products and Services Director**  
**Evan Butterfield**  
**Senior Manager, Editorial Services**  
**Lars Jentsch**

**Manager, Editorial Services**

**Jennifer Stout**  
**Senior Business Development Manager**  
**Sandy Brown**  
**Senior Advertising Coordinator**  
**Marian Anderson**

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

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

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

# Innovative Technology for Computer Professionals

# Computer

[www.computer.org/computer](http://www.computer.org/computer)

## CONTENTS

### COVER FEATURES

#### 21 **GUEST EDITOR'S INTRODUCTION** **Security and Privacy in an Online World**

**Rolf Oppliger**

Most of the perimeter-oriented security mechanisms we currently rely on no longer work in an online world where it's increasingly difficult if not impossible to define the perimeter and separate the trusted inside from the untrusted outside. The articles included in this special issue are intended to provide a comprehensive picture of the security challenges and privacy concerns that apply to this new environment.

#### 23 **Malicious and Spam Posts in Online Social Networks**

**Saeed Abu-Nimeh, Thomas M. Chen, and Omar Alzubi**

A large-scale study of more than half a million Facebook posts suggests that members of online social networks can expect a significant chance of encountering spam posts and a much lower but not negligible chance of coming across malicious links.

#### 29 **Security Vulnerabilities in the Same-Origin Policy: Implications and Alternatives**

**Hossein Saedian and Dan S. Broyles**

The same-origin policy overly restricts Web application development while creating an ever-growing list of security holes, reinforcing the argument that the SOP is not an appropriate security model.

#### 38 **Secure Collaborative Supply-Chain Management**

**Florian Kerschbaum, Axel Schröpfer, Antonio Zilli, Richard Pibernik, Octavian Catrina, Sebastiaan de Hoogh, Berry Schoenmakers, Stelvio Cimoto, and Ernesto Damiani**

The SecureSCM project demonstrates the practical applicability of secure multiparty computation to online business collaboration.

#### 44 **The Final Frontier: Confidentiality and Privacy in the Cloud**

**Francisco Rocha, Salvador Abreu, and Miguel Correia**

The boundary between the trusted inside and the untrusted outside blurs when a company adopts cloud computing. The organization's applications—and data—are no longer onsite, fundamentally changing the definition of a malicious insider.

#### 51 **Securing the Internet of Things**

**Rodrigo Roman, Pablo Najera, and Javier Lopez**

The Internet of Things offers a vision of extreme interconnection that will bring unprecedented convenience and economy, but ensuring its safe and ethical use will require novel approaches.

#### 60 **Sticky Policies: An Approach for Managing Privacy across Multiple Parties**

**Siani Pearson and Marco Casassa Mont**

The EnCoRe project has developed a technical solution for privacy management using machine-readable policies that is applicable in a broad range of domains.

### PERSPECTIVES

#### 69 **Trends in Server Energy Proportionality**

**Frederick Ryckbosch, Stijn Polfliet, and Lieven Eeckhout**

Server energy proportionality, as quantified by the proposed EP metric, has improved significantly, from 30-40 percent in 2007 to 50-80 percent today, but much more can be done to move systems closer to ideal.

### RESEARCH FEATURE

#### 73 **The Three Rs of Cyberphysical Spaces**

**Vivek Menon, Bharat Jayaraman, and Venu Govindaraju**

Integrating recognition with spatiotemporal reasoning enhances the overall performance of information retrieval.

For more information on computing topics, visit the Computer Society Digital Library at [www.computer.org/csdl](http://www.computer.org/csdl).



IEEE Computer Society: <http://computer.org>  
 Computer: <http://computer.org/computer>  
[computer@computer.org](mailto:computer@computer.org)  
 IEEE Computer Society Publications Office: +1 714 821 8380

Flagship Publication of the IEEE  
 Computer Society

September 2011, Volume 44, Number 9

## 6 The Known World

Leisure Science  
**David Alan Grier**

## 9 32 & 16 Years Ago

*Computer*, September 1979 and 1995  
**Neville Holmes**

## 11 Patent Law

Ten Things to Know When Your Patent Application Is Allowed  
**Brian M. Gaff and Catherine J. Toppin**

### NEWS

## 14 Technology News

IPv6: Any Closer to Adoption?  
**Neal Leavitt**

## 18 News Briefs

**Lee Garber**

### ERRATA

In "A Personal History of the IBM PC" (Aug. 2011, pp. 19-25), the publication in which the MITS Altair cover story appeared was incorrectly identified on page 19. The sentence should read as follows: "Widespread personal computing began with the publication of a cover story on the Micro Instrumentation Telemetry Systems (MITS) Altair in the January 1975 issue of *Popular Electronics*."

In "IBM PC Retrospective: There Was Enough Right to Make It Work" (Aug. 2011, pp. 26-33), the manufacturer of the 6502 was incorrectly identified on page 28. The sentence should read as follows: "The three choices were Motorola, Intel, and the MOS Technology 6502."

*Computer* regrets these errors.

### MEMBERSHIP NEWS

## 88 IEEE Computer Society Connection

## 91 Call and Calendar

### COLUMNS

## 93 Social Computing

Let's Gang Up on Cyberbullying  
**Henry Lieberman, Karthik Dinakar, and Birago Jones**

## 97 Green IT

Software Bloat and Wasted Joules: Is Modularity a Hurdle to Green Software?  
**Suparna Bhattacharya, K. Gopinath, Karthick Rajamani, and Manish Gupta**

## 102 In Development

Onshore Mobile App Development: Successes and Challenges  
**Christopher L. Huntley**

## 106 Identity Sciences

What Are Soft Biometrics and How Can They Be Used?  
**Karl Ricanek Jr. and Benjamin Barbour**

## 112 The Profession

Has Everything Been Invented?  
 On Software Development and the Future of Apps  
**Alessio Malizia and Kai A. Olsen**

### DEPARTMENTS

- 4 Elsewhere in the CS
- 37 Computer Society Information
- 80 Career Opportunities
- 87 Bookshelf



IEEE  
 computer society



Reuse Rights and Reprint Permissions: Educational or personal use of this material is permitted without fee, provided such use: 1) is not made for profit; 2) includes this notice and a full citation to the original work on the first page of the copy; and 3) does not imply IEEE endorsement of any third-party products or services. Authors and their companies are permitted to post the accepted version of their IEEE-copyrighted material on their own Web servers without permission, provided that the IEEE copyright notice and a full citation to the original work appear on the first screen of the posted copy. An accepted manuscript is a version which has been revised by the author to incorporate review suggestions, but not the published version with copyediting, proofreading and formatting added by IEEE. For more information, please go to: [http://www.ieee.org/publications\\_standards/publications/rights/paperversionpolicy.html](http://www.ieee.org/publications_standards/publications/rights/paperversionpolicy.html).

Permission to reprint/republish this material for commercial, advertising, or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to the IEEE Intellectual Property Rights Office, 445 Hoes Lane, Piscataway, NJ 08854-4141 or [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org). Copyright © 2011 IEEE. All rights reserved.

Abstracting and Library Use: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01925.

IEEE prohibits discrimination, harassment, and bullying. For more information, visit [www.ieee.org/web/aboutus/whatis/policies/p9-26.html](http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html).

## ELSEWHERE IN THE CS



# Computer Highlights Society Magazines

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

## Software

Trustworthiness is a crucial characteristic when it comes to evaluating any product, even more so for open source software. In “A Survey on Open Source Software Trustworthiness” in *Software*’s September/October issue, Chris Lewis and Jim Whitehead report the results of a survey conducted to identify the reasons and motivations that lead software companies to adopt or reject open source software. Using the study’s results, the authors then ranked the specific trust factors used when selecting an open source software component or product according to importance. The motivations and importance ranking of factors might be useful to both developers and users of open source software.

## Intelligent Systems

The July/August issue of *IS* focuses on social computing and cultural modeling. The issue includes articles that are the result of a special workshop on social computing and modeling held in Chiang Mai, Thailand, under the auspices of the National Electronics and Computer Technology Center in Bangkok and sponsored by the US Office of Naval Research Global. In “Social Computing and Cultural Modeling,” authors Rebecca Goolsby and Chandra Curtis present a spectrum of issues concerning social media, sociotechnical behavior, and modeling approaches to understanding the new global village.

## IEEE Computer Graphics AND APPLICATIONS

A new, robust generation of physics-based animation approaches maintains a standard of visual quality as high as that derived from data-driven synthesis. And even with

the holy grail of control principles that describe human motion still a mystery, the animation research community continues to forge its own path. Researchers have learned that they don’t need to solve the problem of biological control, nor is it necessary to throw out the advantages of animator control and motion capture. Instead, current research aims to find the best of all worlds, judiciously combining physics with human-motion examples, animator inputs, or both. A July/August special issue of *CG&A* brings together four examples of the innovations in this exploding area.

## Computing SCIENCE & ENGINEERING

High-fidelity climate models are the workhorses of modern climate change sciences. In the September/October issue of *CiSE*, “Climate Change Modeling: Computational Opportunities and Challenges” by Dali Wang, Wilfred Post, and Bruce Wilson focuses on several computational issues associated with climate change modeling, covering simulation methodologies, temporal and spatial modeling restrictions, and the role of high-end computing, as well as the importance of data-driven regional climate impact modeling.

## IEEE SECURITY & PRIVACY BUILDING CONFIDENCE, RELIABILITY, AND TRUST

The study of trust should be multidisciplinary. This primarily means including computing and information science on one hand, and psychology on the other. Although some research projects have already employed multidisciplinary approaches, they’ve rarely included all the necessary ingredients. Furthermore, the core of the trust phenomenon is often overlooked. Writing in the July/August issue of *S&P*, author Denis Trček of the University of Ljubljana says that researchers should develop methodologies that support a quantitative treatment by using qualitative assessments of trust. Read “Trust Management in the Pervasive Computing Era.”

## IEEE pervasive COMPUTING MOBILE AND UBIQUITOUS COMPUTING

The July-September issue of *PvC* covers the ever-evolving world of automotive pervasive computing. The



automotive telematics community has moved well beyond the problems of connecting vehicles to each other and to the infrastructure. The focus of inquiry is now turning to interactions between the vehicle and its occupants. This special issue includes articles by teams of authors from MIT, UC San Diego, Kassel University, and the University of South Florida, among others.

## IEEE Internet Computing

IC's July/August issue features an essay on identity authentication by Internet pioneer Vint Cerf. In "Secure Identities," Cerf discusses the US government's recent National Strategy for Trusted Identities in Cyberspace proposal to strengthen identity credentials while increasing protection of personally identifiable information. "As I read it," Cerf concludes, "the [NSTIC] program's intent is to establish both a technical basis and sustainable ecosystem for competing and interoperable identifier authentication products and services that achieve minimum security levels. Such mechanisms can strongly enable the healthy evolution and expansion of online products and services."

## IEEE micro

Exponential growth in cores is among the factors driving recent multicore and many-core processors with relatively large die sizes. *Micro's* July/August theme is "Big Chips." Guest editors Andrew Kahng (University of California, San Diego) and Vijayalakshmi Srinivasan (IBM T.J. Watson Research Center) present six articles that look at some of the tradeoffs inherent in big chips with respect to performance, power density, communications scalability, and packaging and cooling costs. The articles also explore new technologies such as 3D integration. Taken together, they provide a snapshot and sampling of chip design activity in this area.

## IEEE MultiMedia

In "Web-Scale Multimedia Analysis: Does Content Matter?," in *MultiMedia's* April-June issue, author Malcolm Slaney of Yahoo Research pits fast Fourier transforms against metadata to solve multimedia content problems and finds the metadata winning in three examples from music similarity, movie recommendations, and image tagging. Slaney calls himself a content person who would never ignore content analysis but sees it as "hard, perhaps even AI-complete." So he cautions readers, "Every object comes with a context, and those who ignore this signal harm science and their chance of success."

## IT Professional

TECHNOLOGY SOLUTIONS FOR THE ENTERPRISE

*IT Pro's* The July/August issue includes an article on the April 2011 intrusion into the PlayStation3 network that suggests scenarios based on co-opting the enormous computing power and gameplay information available in gaming networks. In "Ender Wiggin Played Mafia Wars Too," Phil LaPlante of Pennsylvania State University keys off a 1985 award-winning science fiction novel, *Ender's Game*, about a boy who thinks he's using a military training simulator in war games but is actually leading a space force in real combat. LaPlante asks, "Were the PS3 intruders actually trying to create some kind of botnet or harvest the by-products of an existing one?" He surveys some possibilities.

## IEEE Design & Test

of Computers

In "FPGA-Based Particle Recognition in the HADES Experiment" in *D&T's* July/August issue, researchers from the Royal Institute of Technology and Justus Liebig University Giessen describe a data acquisition and trigger system for the high-acceptance di-electronic spectrometer, which investigates hadron particles produced from collisions with accelerated beam particles. The system is based on a reconfigurable FPGA cluster and algorithm for efficiently filtering huge data quantities for rare events relevant to the experiment.

## IEEE Annals

of the History of Computing

Two articles in the *Annals* July-September issue give firsthand details about the Internet's origins and structures. As director of the US Department of Defense Advanced Research Projects Agency beginning in 1971, Stephen Lukasik authorized most of the expenditures for the Arpanet. In "Why the Arpanet Was Built," Lukasik sheds light on the basis and context for the DoD expenditures to create the first operational packet-switching computer network. In "Host Tables, Top-Level Domain Names, and the Original of Dot Com," Elizabeth (Jake) Feinler describes the organizations and people involved in maintaining the official Arpanet Host Table and in transitioning to the Domain Name System. Feinler was the principal investigator and later director of the Network Information Systems Center at SRI from 1972 to 1989.

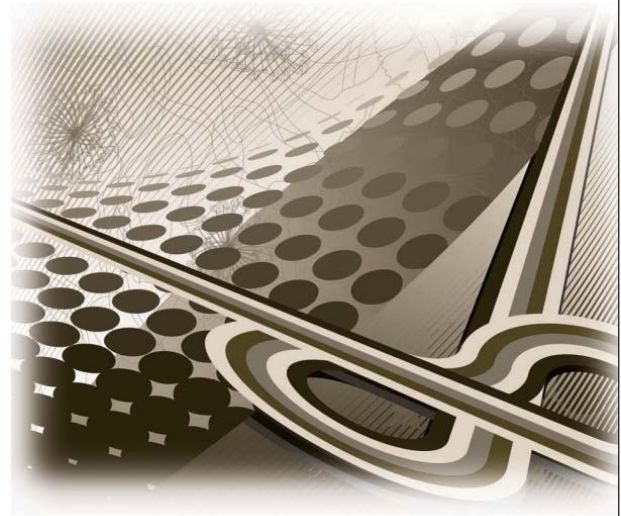
To access free articles and online extras from these publications, visit  
<http://computingnow.computer.org>

## THE KNOWN WORLD

# Leisure Science

**David Alan Grier**

*George Washington University*



**While it might seem otherwise to most individuals, we live in an age of leisure as much as we live in an age of information or industry or globalization.**

**F**or many years, I've followed a strict discipline of buffered output to protect myself from the undue influence of the past. I live in small quarters that lack all the accoutrements of storage: attic, cellar, basement, garage, toolshed, and burial ground.

To keep the evidence of gross materialism from accumulating in rooms, I keep a buffer, in the form of a larger cardboard box, in the closet. When I find something that's cluttering the house, I place it in the buffer. If I determine that I need the object, I retrieve it. But after a decent interval, I tape the buffer shut without looking at what remains inside. I then give it to one of those charitable agencies that's devoted to recycling wayward souls and the unwanted goods of the middle class.

The key is to avoid looking. Once you look, you risk bonding with an object. A pleasant memory will form, and you'll decide that nostalgia requires you to save something "for just a little longer." As you remove the object from the buffer, you'll find something else that evokes another

pleasant emotion, and you'll feel bound by some sense of fairness to save the second item. So the process will continue in an iterative fashion until the box is empty, and the room is cluttered. Buffered output solves this problem by putting a barrier between yourself and the fatal attraction of ideas outgrown.

I admit that buffered output has failed me on a few occasions when I unintentionally disposed of an irreplaceable household item. However, the embarrassment from such failures never lasts and pales in comparison with the benefits of a home free of clutter.

I've argued that industries and research fields could benefit from such a procedure, especially when they're attempting to undertake new areas of activity. It seems this would be especially useful at the current time, when we're looking at the "leisurification" of computer science.

## THE INDUSTRIALIZATION OF LEISURE

Leisurification—transferring ideas from the world of leisure to the world

of industry—is a concept I discovered many years ago when I moved into an office and found a collection of journals that hadn't been properly handled with the buffered output process. These journals were from a field called leisure science, a discipline in which researchers engage in the study of play.

In reading through the articles, I discovered that their focus was somewhat scattered, lacking a common theme or approach. This was apparent even when the articles addressed a specific topic, such as the operation of a household consisting of a couple in which one member is currently employed outside the home and one is retired, which the authors identified as a mixed-leisure household.

In studying the mixed-leisure household, the authors used every conceivable tool of social science to cover all possible combinations of partners. They treated the common scenario of the retired husband and the working wife from a sociological standpoint, but they looked at same-sex couples with economic models.



As I started to move these periodicals to the local recycling bin, I concluded that the contributors to this publication had undertaken a task that was fundamentally self-contradictory. They defined leisure as one concept and then were determined to prove that it was something else. They simply dropped an idea that no longer worked.

Traditionally, and here tradition dates only to the start of the factory age, leisure was defined as the activities that had nothing to do with production. While those activities ranged from hobbies to arts, sports, and ceremonies, all had “the common economic characteristic of being nonindustrial,” noted one early scholar of the field.

However, shortly after leisure was defined as “nonindustrial,” a variety of institutions rushed to create industrialized leisure, which consisted of traditional activities that were augmented by capital and dominated by systematic procedures. They created major league baseball, which was industrialized sport; sight-seeing tours, which were industrialized walks; and amusement parks, which were industrialized fantasy—the purest form of industrialized leisure.

Amusement parks were a direct application of industrial principles to leisure activities. The earliest of these parks were owned by electrical utilities or trolley lines that were looking for ways to generate new demand for their products.

These companies had “excess capacity at night, on weekends, and during the holidays, precisely the times when an amusement park did business,” explained David Nye, a scholar of these parks. The companies located the parks at the end of trolley lines, exploiting cheap land and low fares. They draped the new parks with lights and developed electrically driven rides as novelties. “The amusement park was a machine of illusions,” Nye argued, “the logical

counterpart of the new industrial factory.”

Originally, the engineering community generally kept its distance from the new forms of industrial leisure. In the first half of the 20th century, even though electrical rides presented substantial problems related to efficiency and safety, the precursors to IEEE, the AIEE, and the IRE published not a single article about electrical rides, although they did discuss the problems associated with lighting the buildings at world’s fairs. The “engineer responsible for planning the lighting for an exposition must approach his problem with the purpose of the exposition in

### Amusement parks were a direct application of industrial principles to leisure activities.

mind,” explained an engineer from General Electric. “Originality and appropriateness should be the outstanding characteristic of lighting installations of this class.”

### LEISURIFICATION OF INDUSTRY

A significant change occurred in industrialized leisure in the last three decades of the 20th century. The driving force behind this change was videogames. Originally, videogames represented just another step in the industrialization of leisure. They borrowed ideas from simulation, computer graphics, real-time computation, and other developments from computer research.

Unlike the engineers who worked on the early amusement parks, the designers of these new electronic games were pleased to discuss their accomplishments and describe the technological accomplishments behind their work. “The home TV game industry has just been through

a year of unprecedented growth,” explained IEEE member Ralph Baer, who discussed the underlying engineering in computer games. Baer compared the industry to the technology of telephones and TV and explained the nature of “microprocessor controlled TV games.”

The discussion of computer games quickly shifted from the games themselves to their applications beyond entertainment. Barely six months after Baer wrote his article, Steven Bristow, an Atari engineer who worked on the original *Pong* game, suggested the future of gaming technology. “What can be done on a TV game?” Bristow asked. “Literally anything can be done,” was his response. Bristow went on to suggest battle games, driving games, and simulation games. “Anything that is enjoyable and fun,” he concluded, “will be or has been simulated for fun and profit on the screen of a TV game.”

### GAMES AND GOVERNANCE

The game industry grew quickly during the 1980s and 1990s, and the technology was soon adapted for use in a variety of industrial applications.

Having concluded long ago that industry had determined that games were a flexible and powerful training tool, I was a little surprised when I recently received a report from the National Research Council on the value of serious games to governance and policy problems. “The technical and cultural boundaries between modeling, simulation, and games are increasingly blurring,” the report began, “providing broader access to capabilities in modeling and simulation and further credibility to game-based applications.”

After quickly scanning the report’s summary, I was about to conclude that it offered nothing new and consign it to my system of buffered output so that it might ultimately grace the bookseller’s table at some thrift sale. “Incomprehensible policy

## THE KNOWN WORLD

reports: 12 for \$1.00," the sign would read.

However, something made me pause. I wondered why the NRC was involved and started to think about the story behind the report. The council doesn't address technology problems unless someone brings it a question. Despite what many people think, the NRC isn't a government agency. As an independent organization that's part of the National Academies of Science, it's chartered to answer questions posed by units of the US government. Clearly, someone had asked the council about videogames, and the council had felt moved to respond.

I pulled the report from my box and flipped through its pages. The report didn't specify the underlying question or the organization that had asked it, but both were easy to identify. The report's major recommendation stated that the US Department of Defense should start using commercial videogame technology for some of its war games.

"There," I thought to myself, "problem solved." The DoD wanted to use this technology for its war games. Because it faced some resistance, a deputy secretary of something or another asked for an NRC report to handle the objections.

I looked at a few more pages and noted that the report urged the DoD to stay ahead of all potential threats. It claimed that one of the 2008 presidential candidates had appeared weak because he admitted that he'd never used e-mail. It was a trivial point, but it added a little sting to the report's conclusion.

I was about to make a third attempt to dispose of the report, but an irrational attachment had started to form in my mind. There was something deeper in the report, a story about the nature of leisurification. Certainly, any organization that could build multibillion-dollar fighter planes could find a way to spend a couple hundred thousand dollars to buy

some commercial gaming equipment.

That something else in the report was the asymmetry of the leisure process. We find it easy to apply industrial methods to leisure. We find it harder to admit that our leisure informs our production processes.

### THE LEISURIFICATION PROCESS

The great benefit of computer games is that they provide a new way of developing technical and nontechnical skills. "As gamers play together in groups large and small," argued the NRC report, "they gain specific new ideas in how to communicate,

**Computer games are moving from the world of leisure to the world of production in a way that will likely have a substantial social impact.**

organize, and act in ad hoc collaborative environments, a skill that will be in increasing demand in a global transient workplace." While this educational process is one of the great advantages of transferring ideas from the world of leisure to the world of industry, it also represents one of the most serious threats to leisurification.

The conventional alternative to learning by experience is learning through schools. In addition to conveying knowledge and intellectual skills to students, schools also socialize them, giving them the abilities to work within established society and social institutions.

We find it difficult to shake the idea that individuals without social skills are prepared to work in modern society. The NRC report's authors had tried to address this concern. "For many years," they explained, "the popular perception of digital games has been that they are solitary experiences enjoyed by

those who shun social interaction, such as introverted teenage males." The writers countered this claim by arguing that "These are mistaken assumptions, as digital games have evolved to be highly social environments that appeal to a much broader demographic."

It's difficult to offer a strong defense of the claim that your colleagues are indeed able to engage with the larger society when your report is filled with the kinds of details that are familiar only to people who spend a great deal of time playing games. The report discussed aspects of *Spore*, *World of Warcraft*, and *Grand Theft Auto*. It talked about the soundtracks of various games and explained some of the commercial products that are advertised by careful placement in games. These were ideas that the report wasn't quite ready to abandon, an old vision of games that needed to be expelled from the house so that a new one could take its place.

**C**omputer games indeed represent the mature technologies of computer science. They're moving from the world of leisure to the world of production in a way that will likely have a substantial social impact. Some will likely be good. Some might not be as welcome. However, to see that process clearly, we have to put aside an old vision of games. We might not have had to abandon outmoded ideas when technology industrialized leisure, but we clearly have to do so when it leisurizes industry. **□**

*David Alan Grier, an associate professor of international science and technology policy at George Washington University, is the author of the upcoming book, The Company We Keep.*

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



## 32 &amp; 16 YEARS AGO

**SEPTEMBER 1979**

**CURRICULA** (p. 3) “While we continue to assist educators in carrying out the intent of the committee’s 1976 Model Curriculum at the undergraduate level, we have set as a goal the designing of three other curricula reports at the graduate, community college, and pre-college (elementary and secondary) levels. The first draft of the curricula materials guidelines for software engineering at the master’s level has passed through an extensive study by the Executive Committee of the society.”

**SCHOOLS** (p. 4) “Elementary and secondary schools have now acquired many microcomputers and significant terminal access; however, there is a great need to bring to this level of education the benefits of better educational software maintenance and delivery systems. This might now be provided by advanced downloading techniques from educational networks to microcomputer systems, for example. We hope that computer communications and pre-college professionals will join hands in this effort.”

**NETWORK PROTOCOLS** (p. 8) “One of the primary reasons for the rapid growth of networking is the greater understanding of how to best organize, design, and implement the protocol necessary to support efficient network operation.”

“However, all recent protocol work has been moving in the direction of a hierarchical multilayered structure, with the implementation details of each layer transparent to all other layers in the hierarchy.”

**INTERFACING** (p. 12) “A number of characteristics define a complete interface. They include electrical, physical, and functional characteristics, as well as procedures needed to facilitate transfer of control information and data across the interface. There are a number of protocols that can be involved in different applications and modes of operation; therefore, ISO and ANSI are defining a basic architecture that identifies the interface levels so that they may be independently treated.”

**PROTOCOL FORMALISM** (p. 20) “A great deal of confusion surrounds the words ‘specification’ and ‘verification’ as they apply to computer communication protocols. Hence our first goal will be the definition of these concepts in the context of a layered model of protocols. Next, an overview of various approaches emphasizes the use of more formal specification and verification techniques. We conclude by reviewing some recent applications of these techniques, and suggesting some directions for future work.”

**INTERHOST PROTOCOL** (p. 29) “When computer communication development first entered the present era (with the beginning of construction of the Arpanet in 1968), researchers unaccountably used the word ‘protocol’

to mean a set of communication procedures and conventions. Thus, we use the term ‘host-to-host protocol’ for the subject of this paper.”



**RESOURCE SHARING** (p. 47) “There are two major classes of protocols in a general-purpose computer network: communications protocols and resource sharing protocols. Communications protocols, often referred to as lower-level protocols, are primarily concerned with the reliable transfer of data, while resource sharing protocols, often referred to as higher-level protocols, are primarily concerned with performing remote operations. This article discusses two basic classes of resource sharing protocols: terminal and file transfer job protocols.”

**NETWORK ARCHITECTURES** (p. 58) “General-purpose networking mechanisms must serve a wide range of applications, be very adaptable to changes in these application requirements, and integrate new hardware and software components as necessary. The structure, interfaces, and capabilities that comprise these networking mechanisms define the architecture of the computer network.”

“This article presents some of the current conceptual and implementation developments in computer network architectures. The material is presented in the order that a designer of a network architecture might follow in pursuing that design.”

**SEMICONDUCTOR TECHNOLOGY** (p. 92) “The impact of semiconductor technology on computer systems has been felt mainly due to two elements—memory and microprocessors. The most widespread influence has been exerted by semiconductor random access memory, which has already displaced magnetic core as the dominant main memory technology. Future trends in memory technology will affect computer system organizations. Emerging charge-coupled device and magnetic bubble memory technologies will shape the structure of the memory hierarchy in mass storage systems.”

**PROGRAMMING** (p. 122) “The problem is that computer scientists want to make people think and express themselves in a way that is easy to translate into machine code. The much harder problem of understanding how people really think and express themselves, and translating this into a machine language, has been dropped by the computer scientists.”

## 32 &amp; 16 YEARS AGO

**SEPTEMBER 1995**

**PATENTS** (p. 6) “Today, the US patent system (as well as patent systems in other countries) is arguably used in a manner not anticipated by the framers of the Constitution. Companies use the patent system as a method of competing in the global economy. The real question is, should the patent system be used in this manner?”

**THE INFO AGE** (p. 8) “What is the Info Age? Everyone uses the term, but nobody defines it. So here goes. In this and subsequent essays, I formulate the principles of the Info Age and use the computer industry as proof of its existence. In this series, I answer some burning questions: What is the Info Age, how does it differ from the Machine Age, where is it leading, and what does it have to do with computers?”

**IMAGE RETRIEVAL** (p. 18) “Images are being generated at an ever-increasing rate by sources such as defense and civilian satellites, military reconnaissance and surveillance flights, fingerprinting and mug-shot-capturing devices, scientific experiments, biomedical imaging, and home entertainment systems. For example, NASA’s Earth Observing System will generate about 1 terabyte of image data per day when fully operational. A content-based image retrieval (CBIR) system is required to effectively and efficiently use information from these image repositories. Such a system helps users (even those unfamiliar with the database) retrieve relevant images based on their contents. ...”

**IMAGE CONTENT** (p. 23) “One of the guiding principles used by QBIC [Query By Image Content] is to let computers do what they do best—quantifiable measurement—and let humans do what they do best—attaching semantic meaning. QBIC can find ‘fish-shaped objects,’ since shape is a measurable property that can be extracted. However, since fish occur in many shapes, the only fish that will be found will have a shape close to the drawn shape. This is not the same as the much harder semantical query of finding all the pictures of fish in a pictorial database.”

**IMAGE RETRIEVAL** (pp. 40-41) “... One goal of the Chabot project is to integrate image analysis techniques into the retrieval system so that image requests do not depend solely on stored textual information. As a first step, we have implemented a simple method for color analysis, which we describe in this article. By using both color and textual information for the images, we can locate pictures of red flowers (like anemones and azaleas) and Lake Tahoe sunsets.”

**CAPTIONED IMAGES** (p. 49) “The interaction of textual and photographic information in an integrated text/image database environment is being explored at the Center

of Excellence for Document Analysis and Recognition (CEDAR), SUNY, Buffalo. Specifically, our research group has developed an automatic indexing system for captioned pictures of people; the indexing information and other textual information is subsequently used in a content-based image retrieval system. Our approach presents an alternative to traditional face identification systems; it goes beyond a superficial combination of existing text-based and image-based approaches to information retrieval. ...”

**SHAPE RETRIEVAL** (p. 57) “An object’s shape is typically described through an image or drawing. Large collections of object drawings or images, called shape databases, already exist or are being created in several application areas. Selecting or retrieving a subset of shapes or images that satisfy certain specified constraints is a central problem in shape database management. This article addresses the problem of similar-shape retrieval, where shapes or images in a shape database that satisfy the specified shape-similarity constraints with respect to the query shape or image must be retrieved from the database. ...”

**TECHNOLOGY ASSESSMENT** (p. 82) “Who needs technology assessment? Apparently not the US Congress, since it has just closed down its Office of Technology Assessment (OTA), effective October 1, 1995, after only 20 years of existence. Others, including this author, are not so sure. I believe the nation will sorely miss an important experiment in incorporating a better understanding of technology into government policy making.”

**COMPLEX SYSTEMS** (pp. 85-86) “We thus define the engineering of complex computer systems as all activities pertinent to specifying, designing, prototyping, building, testing, operating, maintaining, and evolving complex computer systems. While in the past, relatively noncomplex ‘traditional’ systems sufficed for most computer control applications, the new and emerging demands of applications and the evolution of computer architectures and networks now essentially *force* systems to be complex, given our current understanding of how to engineer these systems. ...”

*PDFs of the articles and departments from Computer’s September 1979 and 1995 issues are available through the IEEE Computer Society’s website: [www.computer.org/computer](http://www.computer.org/computer).*

**Editor: Neville Holmes; [neville.holmes@utas.edu.au](mailto:neville.holmes@utas.edu.au)**

## PATENT LAW

# Ten Things to Know When Your Patent Application Is Allowed

**Brian M. Gaff**, *Edwards Angell Palmer & Dodge LLP*  
**Catherine J. Toppin**, *General Electric Corp.*



The third in a series of three articles provides basic points to keep in mind once you've successfully filed for a patent.

In the second article in this three-part series (*Computer*, Aug. 2011, pp. 13-15), we listed 10 things to keep in mind while your patent application was pending before the US Patent Office. If you've ultimately convinced the patent examiner that your invention is patentable, the USPTO will send you a Notice of Allowance with instructions to pay the official issue fee (currently \$1,510) within three months. The USPTO will officially issue your patent shortly after it receives the issue fee.

Congratulations—you've earned your patent! But before you celebrate, keep the following 10 points in mind to facilitate issuance of your patent.

## IDS SUBMISSIONS

Filing an Information Disclosure Statement (IDS) is the way to tell the patent examiner about "prior art." In general, prior art is published material that existed before your invention and is relevant to it. Your patent attorney typically files one or more IDSs while your application is pending to list the prior art for the examiner to evaluate while investigating your invention's patentability.

The USPTO requires that you fully disclose the relevant prior art of which you're aware in an IDS. Failure to do so could result in your patent being declared "unenforceable"—that is, worthless—at a later date. It's very important that you tell your patent attorney about all relevant prior art you are aware of and work with the attorney to ensure that the proper IDSs are filed.

## CONTINUING APPLICATIONS

It's common for an inventor who continues research and development work while an application is being prosecuted—that is, pending—to make additional discoveries that are related to the application. One way to apply for a patent on these discoveries is to file a continuing application. This is a new application, but as the name implies, it's related to the first application. You can usually file a continuing application quickly, but you must do so before your first patent issues. Consequently, consider whether to file any continuing application well before the deadline for paying the issue fee.

## CHECKING PATENT TERM ADJUSTMENT

As described in the second article in this series, patent term adjustment (PTA) is the way the USPTO compensates for delays in prosecution. If your application experienced delays, the USPTO will calculate how much time to add to the life, or term, of your patent and send you a notice detailing the additional time. Carefully examine this notice for its accuracy and, if you disagree with the amount of time listed, promptly file your objection with the USPTO.

## OWNERSHIP AND TITLE

Before paying the issue fee, you should verify that your application's listed ownership, or chain of title, is correct. In the US, inventors own their patents by default. However, inventors who work for a company likely have an obligation to assign their patent applications and resulting patents to their employer. This allows the owner to add its name on the first page of the patent.

It's important that you work with your patent attorney to ensure that the proper owner is identified. An



## PATENT LAW

inventor who isn't the owner needs to sign an assignment document that is sent to and recorded at the USPTO.

### CERTIFICATES OF CORRECTION

When the USPTO officially issues your patent, it will send you an impressive-looking "ribbon copy." This is a hard copy of the patent that's bound in heavy paper with a seal (the ribbon) on the front cover.

Read through the entire ribbon copy, paying special attention to the claims at the end. Typographical errors can occur, and it's important to identify each one. If an error is significant enough to require correction, file a Request for a Certificate of Correction with the USPTO. There is no charge for filing this request if the error is the USPTO's fault.

### MAINTENANCE FEES

The USPTO requires that a patent owner pay a series of three maintenance fees during the patent's term. The fees are due 3.5, 7.5, and 11.5 years after the patent is issued and are currently \$980, \$2,480, and \$4,110, respectively. Failure to pay any one of these fees on time will result in the patent being declared "abandoned." An abandoned patent cannot be enforced against an infringer.

There are procedures for paying an overdue maintenance fee and restoring an abandoned patent, but they're only applicable in limited circumstances. Those procedures won't work if an affirmative decision was made not to pay a maintenance fee. Give very careful consideration to any decision about nonpayment of a maintenance fee.

### REISSUE

Sometimes an inventor realizes that an issued patent includes errors that are so significant that a certificate of correction can't remedy the situation. This can occur, for example, if the drawings or specification contain inaccurate information, or

if the claims don't cover what they should have covered. The USPTO has a reissue procedure to handle this. To use this procedure, the patent owner surrenders the original patent and applies for a new one that incorporates the necessary corrections. The USPTO will examine the reissue application and, if satisfied with it, will reissue the corrected patent under a new patent number that has an "RE" prefix.

Reissue applications filed within two years of the issuance of the original patent are entitled to include new claims that are broader than those in the original patent. This can be important if you discover that a

**Anyone—not just the inventor or patent owner—can ask the USPTO to reexamine an issued patent at any time.**

competitor has a product that could infringe your patent if the claims were slightly different. Make it a point to talk to your patent attorney at least once before the two years are up to determine if you can file a reissue application with new claims that could cover a competitor and further protect your invention.

### REEXAMINATION

Anyone—not just the inventor or patent owner—can ask the USPTO to reexamine an issued patent at any time. This is typically done because someone believes that prior art, especially prior art that the patent examiner didn't see during prosecution, is significant and could render the issued patent invalid. The inventor or patent owner could request reexamination to verify that the prior art will not invalidate the patent. A competitor could request reexamination in an attempt to invalidate the patent.

Regardless of who asks for a reexamination, if the USPTO grants the request, a new round of prosecution begins, typically with a different patent examiner. This prosecution, however, usually proceeds faster than the original prosecution. Although you're permitted to amend your issued claims during the reexamination, it's advisable to avoid doing so, or at least keep one claim unchanged. The USPTO will issue a reexamination certificate at the conclusion of the reexamination that identifies any changes made to the claims, including whether any of the originally issued claims were canceled.

### FOREIGN FILING

Filing patent applications in foreign countries is a complex topic. US patent attorneys typically engage and manage "patent agents" in other countries to handle overseas filings and communications with foreign patent offices. If you've filed applications outside the US that are related to your newly issued patent, let the patent offices in those other countries know about the issuance of your US patent. This can carry significant weight in some foreign patent offices and may facilitate the issuance of foreign patents.

### PATENT MARKING

Once you have an issued US patent, the law entitles you to mark the corresponding patent number on items your patent covers. This marking provides notice to others that an item is protected by a patent. This can be important if patent infringement litigation results because the accused infringer will be deemed to have knowledge of your patent simply because its number was marked on covered items. This can increase the amount of money awarded to a patent owner when someone is found to have infringed a patent.


Over the past few years, there's been an increase in the number of

lawsuits alleging “false marking,” where the number of an expired patent continued to be marked on items. Recent changes in the law, however, have largely eliminated these suits, so risks associated with marking are generally minimal at this time.

### THE FUTURE

Currently, several changes to US patent law are under consideration. One significant change is the proposed move away from the current “first to invent” system to a “first to file” system. The current system allows the applicant to prove that, under certain conditions, he or she was the first to make an invention even though someone else filed a patent application beforehand. The proposed system would automatically deem the first person to file the application to be the inventor. Until the law changes, however, keeping

all of the records that describe your inventions is essential because they could be important when defending your patents.

**A**fter enduring the long journey from your invention to your issued patent, it may be tempting to “forget” about the patent and move on to other things. Don’t do that. Check on the patent from time to time. Consider what your competitors are doing. Think about reissue applications. Consider licensing opportunities. Ensure that your maintenance fees are being paid on time. Above all, keep innovating and thinking about new applications you can file that reflect your hard work and perseverance. 

*Brian M. Gaff, a partner at the Edwards Angell Palmer & Dodge LLP law firm, is a senior member of IEEE. Contact him at [bgaff@eapdlaw.com](mailto:bgaff@eapdlaw.com).*

*Catherine J. Toppin, a patent counsel at the Global Patent Operation of General Electric Corp., is a member of IEEE. Contact her at [catherine.toppin@ge.com](mailto:catherine.toppin@ge.com).*

The content of this article is intended to provide accurate and authoritative information with regard to the subject matter covered. It is offered with the understanding that neither IEEE nor the IEEE Computer Society is engaged in rendering legal, accounting, or other professional services or advice. If legal advice or other expert assistance is required, the services of a competent professional person should be sought.

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



Enroll now.

### ON THIS BATTLEFIELD, EDUCATION IS YOUR BEST DEFENSE.

Cyber attacks are being waged all over the world, creating an unprecedented demand for trained professionals to protect our country’s data assets and develop cybersecurity policies. Help meet the demand with a bachelor’s or master’s degree in cybersecurity. Whether you plan to work for Cyber Command taking down cyber terrorists or for private industry battling hackers, UMUC can help you make it possible.

- Designated as a National Center of Academic Excellence in Information Assurance Education by the NSA and DHS
- BS and MS in cybersecurity and MS in cybersecurity policy available
- Programs offered entirely online
- Interest-free monthly payment plan available, plus financial aid for those who qualify

**CYBERSECURITY**

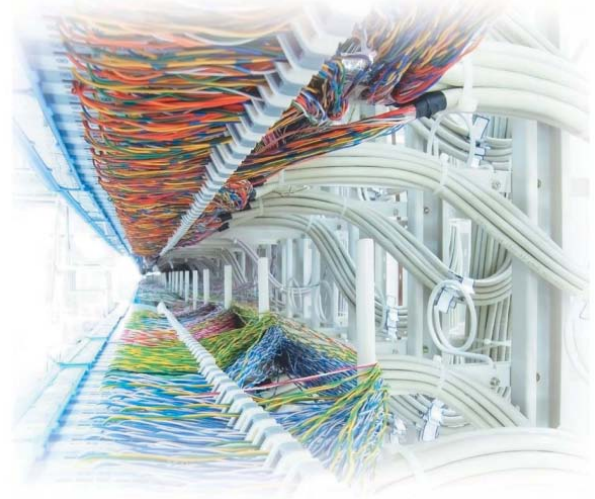
800-888-UMUC • [umuc.edu/cyberwarrior](http://umuc.edu/cyberwarrior)

 **UMUC**  
University of Maryland University College  
Copyright © 2011 University of Maryland University College

## TECHNOLOGY NEWS

# IPv6: Any Closer to Adoption?

Neal Leavitt



**For years, Internet engineers have said adopting IPv6 is important because the number of available IPv4 addresses is rapidly decreasing. However, IPv6 adoption is still minimal.**

**F**or years, Internet engineers have talked about the importance of adopting IPv6, the latest version of the Internet's primary communications protocol.

However, 16 years after the Internet Engineering Task Force (IETF) adopted IPv6, more than 99 percent of the Internet is still based on the older IPv4.

With about 2 billion Internet users worldwide—many utilizing multiple connected devices—and billions more possibly going online in the future, the IPv4 address supply will soon run out.

The Asia-Pacific Network Information Centre (APNIC), one of five regional Internet registries (RIRs) that allocate IP addresses to members in their geographic area, has almost exhausted its IPv4 addresses.

"It's just a matter of time before the remaining registries exhaust their address space, too," said Yahoo IPv6 evangelist Jason Fesler.

IPv6 provides many more Internet addresses than IPv4. Thus, proponents of the new protocol warn that the Internet could experience higher operation costs, less innovation, and more network complexity if IPv6

usage doesn't increase substantially over the next few years.

However, that hasn't occurred.

Network operators have been reluctant to switch to IPv6 for economic more than technical reasons, said Tom Coffeen, director of global network architecture for Limelight Networks, a content delivery network operator.

The lack of content available over IPv6 networks and the dearth of IPv6 clients have also made immediate adoption less appealing.

"Adoption has been seen as a risk-management initiative with little potential for a compelling return on investment," Coffeen said. "However, the recent exhaustion of IPv4 addresses should change that calculus for most operators."

## RUNNING OUT OF IPv4 ADDRESSES

IP versions 0 through 3 were development versions of the Internet Protocol used between 1977 and 1979.

In September 1981, the IETF released IPv4, which has 32-bit addresses and enables about 4.3 billion Internet addresses.

APNIC chief scientist Geoff Huston estimated the projected IPv4-address exhaustion date will be 12 February 2012 for European networks; 25 July 2013 for African networks; 17 December 2013 for the US, Canada, and some Caribbean islands; and 9 April 2014 for Latin America and other parts of the Caribbean.

The rapidly increasing adoption of smartphones that connect to the Internet has accelerated this process, noted Alain Fiocco, senior director of architecture and marketing for Cisco Systems and head of the company's IPv6 program.

This has implications for business continuity and e-commerce, according to Danny McPherson, chief security officer with VeriSign, which provides Internet infrastructure services and operates two of the Internet's 13 root name servers.

Businesses that want to expand their networks and otherwise use more IP addresses, as well as ISPs that want to serve additional customers, will require the additional addresses that IPv6 provides, said John Curran, president and CEO of the American Registry for Internet Numbers (ARIN), an RIR.



## IPv6 BRIEFING

In 1994, the Internet Engineering Task Force initiated development of the IPv6 suite of protocols, which were designed to replace IPv4. The IETF published the IPv6 standard in 1995.

Unlike IPv4, which has 32-bit addresses, IPv6 has 128-bit addresses. Thus, the new protocol increases the number of available IP addresses to  $2^{128}$  (about  $3.4 \times 10^{38}$ ) from IPv4's  $2^{32}$  (about 4.3 billion).

IPv6 also offers other benefits. For example, the protocol specifies a new, simplified packet format designed to minimize header processing by routers.

In addition, support for the IP Security standard is mandatory in IPv6 but optional in IPv4. Another advantage is that IPv6 hosts can autoconfigure when connected to an IPv6 network.

And the protocol's large address space enables multiple levels of hierarchy and greater flexibility in addressing and routing.

This could be the case particularly for major ISPs in fast-growing economies, noted Syracuse University professor of information studies Milton L. Mueller.

Until IPv6 takes off, the prices of IPv4 addresses on the secondary market could skyrocket, said Shawn Morris, manager of IP development at NTT America.

### IPv6 ADOPTION

Yahoo's Fesler estimated that only about 0.2 percent of Internet addresses are IPv6-based.

Nonetheless, most major backbone networks—such as those belonging to Amazon, Comcast, and Verizon—and some key router makers—like Billion Electric, Cisco, D-Link, Juniper Networks, and ZyXEL Communication—have deployed IPv6.

In fact, most enterprise and ISP network equipment sold during the past few years is IPv6 compatible, noted Leo Vegoda, number resources manager for the Internet Corporation for Assigned Names and Numbers' Internet Assigned Numbers Authority.

VeriSign's McPherson said his company has seen a fourfold IPv6 traffic increase over its infrastructure—to 0.9 percent of the total—in the past year.

"While it may seem like a small amount, 0.9 percent of [our average daily] 60 billion [Domain Name System (DNS)] queries is pretty significant," he noted.

ARIN's Curran added that the demand for IPv6 addresses from both ISPs and big companies jumped 50 percent from 2009 to 2010 and has continued rising this year.

### WORLD IPv6 DAY

On 8 June 2011, nearly 400 organizations—including Akamai Technologies, Facebook, Google, Limelight Networks, and Yahoo—participated in a 24-hour global IPv6 trial.

The goal was to determine how well IPv6 would run on a large scale over an entire day. According to Curran, most end users didn't experience problems. "That's what we were hoping for," he said. "At the same time, it was a good learning experience."

NTT America's Morris said his company's network had no trouble handling the 80 percent increase in IPv6 traffic.

Most of the participating organizations used dual stacking, with their networks running both IPv4 and IPv6. That way, a computer that couldn't connect via IPv6 could do so via IPv4.

### STANDING IN THE WAY

IPv6 adoption faces several noteworthy challenges.

For example, said Arbor Networks president Rob Malan, "The little things will be the problem, such as figuring out why a customer's DNS doesn't work with IPv6, having trained people that can configure firewall policies, and troubleshooting IPv6 routing."

"The additional complexity for network operations teams is also significant," he added.

Measuring IPv6 adoption is difficult. No single agency or standards group has comprehensive statistics about how much Internet traffic is based on IPv6 or IPv4. Explained Yahoo's Fesler, "There's

no single point in the Internet to do measurements."

There are also significant differences in the two protocols' underlying technology. Enabling IPv6 is thus more demanding than simply flicking a switch, noted APNIC's Huston. "There is a required investment in technology, operational process, and skill sets for providers," he said.

Added Fesler, problems could occur if a user has a firewall that doesn't understand and tries to block IPv6 traffic. However, he noted, the number of people this could affect is small and steadily shrinking.

Also, he pointed out, improvements in OSs and Web browsers are quickly making this a nonissue.

### No IPv6 backward compatibility with IPv4

The headers of IPv4 and IPv6 packets are significantly different. For this and other reasons, the two protocols don't interoperate.

Thus, to serve both types of networks, service providers will need to run dual stacks.

Said Ed Moyle, senior analyst for market research firm Security Curve, "The infrastructures will have to exist side-by-side for the next few years. Meanwhile," he added, "users running only IPv4 won't be able to reach parts of the IPv6 Internet as it grows."

## TECHNOLOGY NEWS

## ADOPTING IPv6

**A**s adoption of IPv6 takes off, users with older devices and other hardware that support just IPv4 might not be able to reach destinations supported by IPv6-only networks.

If IPv6 isn't adopted widely, the lack of IPv4 addresses will close the Internet to start-ups, explained Chief Scientist Geoff Huston of the Asia-Pacific Network Information Centre. APNIC is one of five regional Internet registries that allocate IP addresses to members in their geographic area.

"That is going to allow incumbents to dictate the terms and conditions of competition," Huston said.

The only alternative to IPv6 that supports continued network growth is to have many devices within a network share an external IP address.

To do this, organizations must use network-address-translation equipment. NAT boxes translate the private address that a device has within an organization into a public address for use on the Internet.

However, many Internet experts say NAT isn't a good solution to the IPv4 address shortage.

They say this approach adds complexity to and can reduce the performance of enterprise networks. Purists say NAT equipment breaks the Internet's end-to-end nature, keeping users from communicating directly with one another without intermediate devices altering their packets.

"NAT more or less constrains user applications to talk only to servers and not directly to other user devices," said Matt Levine, director of engineering for Akamai Technologies, which operates a content-delivery network.

NAT disrupts the direct, point-to-point connections that make popular real-time applications like streaming possible.

Also, the technology deployed on a large scale is expensive to operate, noted Yahoo IPv6 evangelist Jason Fesler.

Users have to wait for each IPv6 connection to time out before the system tries making an IPv4 connection. This causes slow webpage loading.

A goal of World IPv6 Day was to gauge how big an issue this could be. "The meltdown predicted by the pessimists didn't occur," said Cisco's Fiocco.

**Y**ahoo's Fesler predicted that IPv6 adoption will increase by 30 to 45 percent during the next three years.

APNIC running out of IPv4 addresses will probably drive further IPv6 adoption in the Asia-Pacific region, which could encourage more implementation globally, said Security Curve's Moyle.

However, noted Syracuse University's Mueller, "Unless radical new applications are developed that take advantage of IPv6's greater address space, not much will change. These new apps will probably have to wait until there is more adoption."

And, he added, "Anyone who migrates to IPv6 still must run IPv4 to maintain compatibility with those who don't migrate. This means that expanding networks will still need new IPv4 addresses." **■**

*Neal Leavitt is president of Leavitt Communications ([www.leavcom.com](http://www.leavcom.com)), a Fallbrook, California-based international marketing communications company with affiliate offices in Brazil, China, France, India, and the UK. He writes frequently on technology topics and can be reached at [neal@leavcom.com](mailto:neal@leavcom.com).*

**Editor: Lee Garber, Computer;**  
[l.garber@computer.org](mailto:l.garber@computer.org)

## Security concerns

The migration from IPv4 to IPv6 will present security challenges.

Low IPv6 demand has kept security companies from developing many features for the technology, said Lawrence Orans, research director with market-research firm Gartner Inc.

Over years of heavy use, experts have found and fixed numerous problems with IPv4. IPv6 is 16 years old but hasn't been widely implemented. It thus might still have security issues to deal with, according to Fesler.

Organizations that run dual-stack IPv4-IPv6 architectures will face complexity that could yield significant security problems. For example, firewall users will have to create separate sets of rules for both types of traffic.

While transitioning from IPv4 to IPv6, organizations are using various approaches—including tunneling and multiprotocol label switching—

to translate from one protocol to the other.

According to VeriSign's McPherson, each translation process potentially creates a vulnerability. For example, when users tunnel IPv6 traffic to IPv4 networks, they utilize a virtual private network. However, McPherson said, a VPN to a network beyond the originator's control could result in either security exposure or unauthorized data access.

## IPv6 brokenness

*IPv6 brokenness* occurs in tunneled or dual-stack deployments when the system tries to use unreliable or faulty IPv6 connections rather than properly functioning IPv4 connectivity.

STAY  
CONNECTED

## TWITTER

| @ComputerSociety  
| @ComputingNow

## FACEBOOK

| [facebook.com/IEEE.ComputerSociety](https://facebook.com/IEEE.ComputerSociety)  
| [facebook.com/ComputingNow](https://facebook.com/ComputingNow)

## LINKEDIN

| IEEE Computer Society  
| Computing Now

# Now, manage both your UPS and your energy proactively.



**Energy usage and energy cost reporting:**  
Save energy and money by tracking energy usage and costs over time.



**CO<sub>2</sub> emissions monitoring:**  
Reduce environmental impact through increased understanding of CO<sub>2</sub> emissions.



**Risk assessment:**  
Identify and proactively manage threats to availability (e.g., aging batteries).



## Only APC Smart-UPS saves money and energy without sacrificing availability.

Today's more sophisticated server and networking technologies require higher availability. That means you need more sophisticated power protection to keep your business up and running at all times. But that's not all. In today's economy, your UPS must safeguard both your uptime *and* your bottom line. Only APC by Schneider Electric™ helps you meet both of these pressing needs. Specifically, the APC Smart-UPS™ family now boasts models with advanced management capabilities, including the ability to manage your energy in server rooms, retail stores, branch offices, network closets, and other distributed environments.

### Intelligent UPS management software

PowerChute™ Business Edition, which comes standard with Smart-UPS 5 kVA and below, enables energy usage and energy cost reporting so you can save energy and money by tracking energy usage and costs over time; CO<sub>2</sub> emissions monitoring to reduce environmental impact through increased understanding; and risk assessment reporting so you can identify and proactively manage threats to availability (e.g., aging batteries).

### Best-in-class UPS

Our intelligent, interactive, energy-saving APC Smart-UPS represents the combination of more than 25 years of Legendary Reliability™ with the latest in UPS technology including an easy-to-read, interactive, alphanumeric LCD display to keep you informed of important status, configuration, and diagnostic information, a unique battery life expectancy predictor, and energy-saving design features, like a patent-pending "green" mode. Now, more than ever, every cost matters and performance is critical. That's why you should insist on the more intelligent, more intuitive APC Smart-UPS.

## Why Smart-UPS is a smarter solution



### Intuitive alphanumeric display

Get detailed UPS and power quality information at a glance – including status, about, and diagnostic log menus in up to five languages.



### Configurable interface

Set up and control key UPS parameters and functions using the intuitive navigation keys. On rack/tower convertible models, the display rotates 90 degrees for easy viewing.



### Energy savings

A patent-pending "green" mode achieves online efficiencies greater than 97 percent, reducing heat loss and utility costs.



Download White Paper #24, "Effect of UPS on System Availability," and register to **WIN APC Smart-UPS 1500VA rack/tower LCD 120 V**, a \$779 value!

Visit [www.apc.com/promo](http://www.apc.com/promo) Key Code **f769v** Call 888-289-APCC x6298

# APC™

by Schneider Electric



NEWS BRIEFS

**Start-up Unveils Energy-Efficient Transistors**

A start-up has developed a transistor technology that promises to reduce processors' energy consumption by at least 50 percent while maintaining performance levels.

Reducing power consumption is a major goal of chip makers, particularly for processors used in mobile devices such as laptops and smartphones. The less power the devices' chips use, the longer their batteries last.

SuVolta says its PowerShrink technology could be used in multiple products, including microprocessors, static RAM chips, and SoCs, all of which are important parts of mobile systems.

The company hasn't released many details of its technology. However, it notes that the approach reduces power leakage in transistors and can be built with existing fabrication technology. That is a huge benefit for chipmakers, who don't want to spend large sums of money on new fabrication equipment.

Today's integrated circuits often have millions of transistors, which turn on or off at each transistor's

*threshold voltage.* The threshold voltage can vary greatly across transistors within an IC and among different ICs. This variation reduces performance and increases power consumption because higher voltages must be used to ensure that all transistors are fully on or off.

PowerShrink uses a deeply depleted channel transistor. This DDC approach includes a channel structure that decreases the threshold-voltage variation by 50 percent, compared to conventional transistors, thereby reducing leakage and maintaining performance.

SuVolta says PowerShrink will work with small and large feature sizes.

The company expressed hope that chip makers see PowerShrink as a way to compete with Intel's recently announced Tri-Gate technology. Tri-Gate transistors employ a third gate stacked on top of two vertical gates providing three times the surface area for electrons to travel. Intel says this reduces leakage and consumes less power.

SuVolta plans to license its technology, scheduled to be in production next year, to chip makers and already has a large buyer in



Fujitsu Semiconductor. Companies such as ARM, Broadcom, and Cypress Semiconductor have expressed interest in PowerShrink.

**McAfee: Massive Cyberattack Targeted US and Others**

Security vendor McAfee says it has identified a major multiyear cyberspying campaign against the US and numerous other countries, as well as about 70 corporations and organizations.

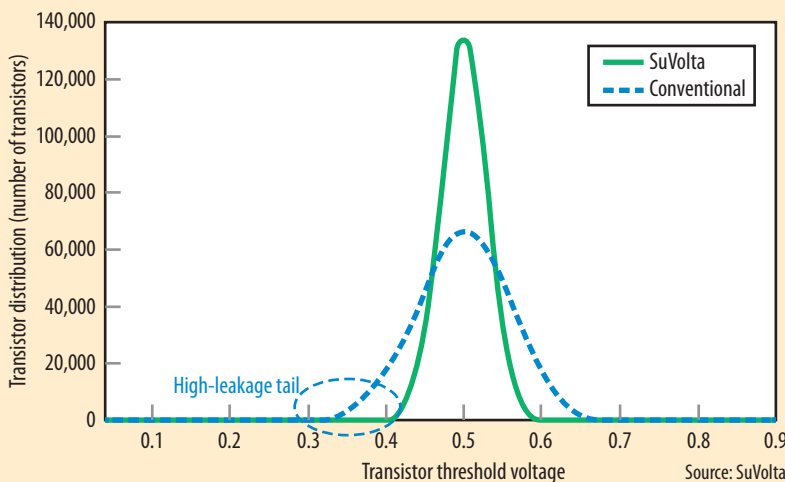
The massive cyberespionage effort—which McAfee calls Operation Shady RAT (remote access tool)—has been ongoing for at least five years and appears to be state sponsored. The firm, which Intel recently acquired, found 19 intrusions lasting more than a year and five lasting more than two years.

In a report on the subject, McAfee said that many of the victims don't even know that they've been attacked, even though they might have lost confidential information such as government secrets and intellectual property.

According to the report, the intrusions are much more extensive and dangerous than the recent high-profile attacks by hacker groups such as Anonymous and Lulzsec.

While investigating intrusions at several US defense contractors, McAfee said it discovered the extent of the cyberespionage campaign by accessing one of the hundreds of command-and-control servers the hackers were operating. The company then examined the server's logs, which contained information about the attacks from mid-2006.

McAfee said it had been aware of the server since 2009 but only recently discovered that the attackers had erred by configuring the server to



SuVolta's PowerShrink technology promises to save energy by reducing the range of threshold voltages that turn the various transistors on a chip on and off. The wide variance in threshold voltages on traditional chips can reduce performance and increase energy consumption.

generate logs that identified every IP address they controlled.

The security vendor is working with US government agencies to shut down the server. The firm has also helped several victimized companies investigate the intrusions. However, other organizations have reportedly refused help, denying they had been successfully attacked.

Some security companies downplay the McAfee report's importance, saying such incidents aren't new.

In describing the attacks, the report said, "The compromises themselves were standard procedure for these types of targeted intrusions: A spear-phishing e-mail containing an exploit is sent to an individual with the right level of access at the company; and the exploit, when opened on an unpatched system, will trigger a download of the implant malware."

McAfee didn't go into more detail. However, spear phishing typically uses e-mail to attack a specific organization and access confidential data.

As is the case with general phishing attacks, spear-phishing e-mail appears to come from a trusted source, often someone in a position of authority within the recipient's organization who might request confidential information.

Frequently, the message asks victims to log into a fake but realistic-looking webpage that requests their username and password or that asks them to click on a link that subsequently and surreptitiously downloads malware.

According to the McAfee report, the attacks it discovered have targeted government agencies, the United Nations, nonprofit and other organizations, and companies in the US, South Korea, Taiwan, Canada, Japan, the UK, Switzerland, South Korea, Vietnam, and Indonesia.

Among those reportedly attacked were the Associated Press, the International Olympic Committee, natural-gas companies, technology

## IBM DEVELOPS SEARCH ENGINE FOR PEOPLE WHO CAN'T READ

**M**any people, particularly in the developing world, can't read. This makes it impossible for them to use many Internet technologies that could benefit them.

With this in mind, IBM has developed a speech-based search engine that the illiterate can work with.

Scientists at IBM Research-India have created the Spoken Web system, which utilizes telephone numbers in place of URLs. Users can input the numbers into their phones to listen to spoken Web-based information or share information such as crop prices.

The researchers recently developed a search engine that employs speech recognition to determine what a user is looking for and to find its location on the Spoken Web.

The system can generate many results. However, users can't scan the results to choose the ones they like, as they can with a traditional search engine. Instead, the IBM engine announces how many results were found and recommends ways to filter the list. This process continues until there are no more than five results, which the system then reads to the user.

IBM said it tested the technology with 40 farmers in India, who found it easy to use. The researchers are continuing to work on their speech-recognition software.

firms, and US defense contractors. McAfee didn't name most of the victimized corporations, saying it didn't want to alarm shareholders or customers.

McAfee's analysis indicated a single group of attackers conducted all of the assaults. Because the five-year campaign targeted Taiwan and Olympics-related organizations, the Center for Strategic and International Studies—a US-based public-policy think tank—suggests China is responsible for the intrusions. China, which denies the allegations, hosted the 2008 Summer Olympic Games.

### First Smart Grid Standards Approved

A group representing companies and government agencies has approved the first six smart-grid-interoperability standards, addressing areas such as device interoperability for easier information exchange and requirements for upgrading smart electric meters.

The Smart Grid Interoperability Panel, which the US National Institute of Standards and Technology (NIST) formed in 2009, recently adopted the specifications as part of its Catalog of Standards.

The SGIP is using the work of several standards organizations in

developing specifications designed to enable manufacturers and developers to build smart-grid-related systems that work together.

Smart-grid technology promises to use intelligent networking and automation to enable a two-way flow and analysis of information between electricity suppliers and consumers. Suppliers would use the data to efficiently and effectively control the delivery of electricity to consumers.

Last year, NIST released the first version of its Framework and Roadmap for Smart Grid Interoperability Standards, which identified 75 existing specifications. A second version, currently in draft form, names 83 proposed standards that the SGIP is studying.

The SGIP doesn't have regulatory power, but proponents hope the panel's broad public and private membership and the requirement that at least 75 percent of its members approve standards will encourage adoption.

The new standards, approved by more than 90 percent of the members, include:

- the Internet Protocol Suite for the Smart Grid, which would help devices exchange information;



## NEWS BRIEFS



Qualcomm's mirasol display uses the same approach that produces vivid colors on some butterflies and birds. The screen not only shows images that are visible in bright sunlight, it also reduces energy usage.

- a data model for the online exchange of information between energy suppliers and customers;
- standards for electric-vehicle plugs;
- approaches for communication between plug-in vehicles and the grid;
- requirements for upgrading household smart meters; and
- guidelines for assessing standards for wireless smart-grid communications.

### New Display Uses 'Natural' Approach to Generate Colors

Qualcomm has designed an energy-efficient display that produces colors with an approach like that used by the wings of some butterflies and birds. This enables the mirasol display to show images that are easy to see even in bright light.

The screen is an example of the increasingly popular *biomimetics* approach, in which technology uses techniques found in nature to solve problems.

The mirasol display, a microelectromechanical system that utilizes a technique called *interferometric modulation* technology, consists of two electrically conductive plates. One is a thin-film stack on a glass substrate; the other is a reflective membrane. An air gap separates the plates.

When light hits the display, it reflects off both plates. Depending on the height of the air gap, the light reflecting off one plate will be out of phase with that reflecting off the other. The resulting constructive and destructive interference creates colors.

The interference between the light bouncing off different surfaces produces the iridescent colors seen on birds such as peacocks and on the wings of some butterflies.

The mirasol display controls the color it displays by running electricity through the plates and thereby adjusting the size of the gap between them.

Many displays use a backlight to boost visibility. The mirasol display doesn't need this because it works with reflected light.

By eliminating the backlight, the mirasol display uses less energy than conventional screens. In addition, the screen will hold its image until signaled to change without refreshing. This avoids the need to expend energy refreshing the image constantly, as LCDs must do.

The display's reflected-light images remain bright even in sunlight. The backlight used in many other types of displays, on the other hand, tends to wash out in strong ambient light.

Qualcomm says its technology could be employed in many hardware applications, from mobile phones to flat-panel monitors.

### Antennas Use Plasma to Make Wireless Networks Faster

A new type of antenna uses plasma to focus radio waves and enable ultrafast wireless networks.

UK-based Plasma Antennas has designed a plasma silicon antenna (PSiAN), which can be built with the same well-established, cost-efficient techniques used to make silicon chips.

Traditional directional antennas that transmit high-frequency radio waves require expensive materials

or precise manufacturing techniques.

A PSiAN consists of thousands of diodes on a silicon chip. Each activated diode generates a cloud of electrons, also known as *plasma*. If the cloud is dense enough, it reflects high-frequency radio waves.

Users could turn on select diodes to change the reflecting area's shape, which could then focus the waves in one direction and steer them as desired, thereby speeding up transmissions. Typical antennas send signals in all directions, causing dispersion that reduces performance.

PSiAN is a solid-state antenna small enough to fit within a mobile phone. Some potential users may tend to favor this type of antenna over gas plasma antennas, which are larger and have moving parts.

Proponents say PSiAN's ability to focus radio waves would be ideal for use with Wireless Gigabit technology, which provides data rates up to 7 Gbits per second—fast enough to quickly download video—at a range of 10 meters.

WiGig operates in the unlicensed 60-GHz frequency band. Signals in this range disperse quickly unless they are focused.

Using PSiAN with technologies such as WiGig could make it easier for users to download video and other data-intensive content to their smartphones.

PSiAN could also be used in small radar systems for cars, which could improve driving in low visibility.

Some experts note that plasma antennas like PSiAN might not be useful for some purposes because they operate at high frequencies. The signals couldn't penetrate walls and thus aren't ideal for some indoor uses, they explain.

Plasma Antennas says its product could be ready for commercial use within two years. **■**

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



## GUEST EDITOR'S INTRODUCTION



# Security and Privacy in an Online World

Rolf Oppliger, *eSECURITY Technologies*

**Because it's increasingly difficult if not impossible to define the perimeter that separates the trusted inside from the untrusted outside, many security and privacy mechanisms no longer work in an online world.**

**D**ue to the amazing advances in information and communications technologies, we're heading for an online world in which convenience goods have unprecedented computing power and are permanently connected to the Internet or stored in the cloud.

The Internet is everywhere, and now people are talking about the Internet of Things (IoT). Look at your own belongings; it's likely that you carry around at least one or possibly several handheld devices such as smartphones that are permanently connected to the Internet. Each device has computing power that was sufficient for navigating a rocket to the moon 40 years ago. Now we use that power to download and play songs and movies, access social media such as Facebook or Twitter, run e-mail or messenger software, or access any of the other myriad apps people have created recently.

## A NEW APPROACH TO SECURITY AND PRIVACY

The online world not only changes our way of living, but also the way we approach security and privacy. In particular, most security mechanisms we rely on in our daily lives are perimeter-oriented, meaning that they basically

protect the edges of a particular domain. However, many of these mechanisms no longer work in an online world, mainly because it's increasingly difficult if not impossible to define the perimeter and separate the trusted inside from the untrusted outside.

A mobile user typically requires remote access from a handheld device even if it's located outside the corporate network. In some cases, the user employs a device to access the network, a situation the community refers to as BYOD (bring your own device). BYOD provides some unique challenges for a company's chief information security officer. The bottom line is that perimeters need to be permeable to some extent, and many entities must exist on either side of a perimeter. Deperimeterization occurs naturally, and this not only poses new security challenges but also raises privacy concerns.

## IN THIS ISSUE

There's a large gap between the general knowledge and wisdom pertaining to computer and information security, which focuses on perimeter protection, and what's really needed in practice. That is where the content in this special issue comes into play. Comprising six contributions, the content addresses some of the security and privacy challenges that apply to the online world.

In "Malicious and Spam Posts in Online Social Networks," Saeed Abu-Nimeh, Thomas M. Chen, and Omar Alzubi report on an empirical analysis of Facebook posts that leads to the conclusions that an overwhelmingly large fraction of posts is spam, and only a much smaller fraction is malicious. This is good news, and it contradicts the publications that elaborate on the intrinsic dangerousness of social media in general, and social media posts in particular.

## GUEST EDITOR'S INTRODUCTION

In "Security Vulnerabilities in the Same-Origin Policy: Implications and Alternatives," Hossein Saiedian and Dan S. Broyles assess the effectiveness of the same-origin policy (SOP) that is at the core of many Internet and Web security technologies in use today. Their assessment is not particularly encouraging, but they also propose ways for the SOP to evolve in the future.

"Secure Collaborative Supply-Chain Management" by Florian Kerschbaum and coauthors demonstrates the practical applicability of secure multiparty computation to business collaboration. In particular, the authors report the key findings of the European research project SecureSCM, which applies secure computation protocols in the supply-chain management realm.

In "The Final Frontier: Confidentiality and Privacy in the Cloud," Francisco Rocha, Salvador Abreu, and Miguel Correia focus on the confidentiality and privacy challenges of cloud computing, assuming, for example, that the cloud operator might have malicious employees operating as insiders. They also offer proposals to address these issues.

"Securing the Internet of Things" by Rodrigo Roman, Pablo Najera, and Javier Lopez provides an overview of the security challenges and protection mechanisms related to the IoT. As our world is heading in this direction, properly understanding and addressing the challenges and coming

up with appropriate protection mechanisms is key for the IoT's future deployment.

In their article titled "Sticky Policies: An Approach for Managing Privacy across Multiple Parties," Siani Pearson and Marco Casassa Mont not only address the practically relevant question of how to handle privacy management across multiple parties, they also propose a solution.

**T**he contributions selected for inclusion in this special issue are intended to provide a comprehensive picture of some of the most important topics related to security and privacy in the online world. We hope that some readers will become interested in directing their research activities to resolving the problems concerning these topics. *Computer* is strongly committed to providing additional coverage related to ongoing developments in the area of online security and privacy in future issues. **□**

*Rolf Oppliger, the founder and owner of eSECURITY Technologies, is an adjunct professor of computer science at the University of Zurich. Contact him at [rolf.oppliger@esecurity.ch](mailto:rolf.oppliger@esecurity.ch).*



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

# Innovative Technology for Computer Professionals

## Computer

### Welcomes Your Contribution

*Computer*  
magazine  
looks ahead  
to future  
technologies



- **Computer**, the flagship publication of the IEEE Computer Society, publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.
- Articles selected for publication in **Computer** are edited to enhance readability for the nearly 100,000 computing professionals who receive this monthly magazine.
- Readers depend on **Computer** to provide current, unbiased, thoroughly researched information on the newest directions in computing technology.

**To submit a manuscript for peer review, see *Computer's* author guidelines:**

**[www.computer.org/computer/author.htm](http://www.computer.org/computer/author.htm)**

## COVER FEATURE



# Malicious and Spam Posts in Online Social Networks

Saeed Abu-Nimeh, *Damballa Inc.*

Thomas M. Chen and Omar Alzubi, *Swansea University, Wales*

**A large-scale study of more than half a million Facebook posts suggests that members of online social networks can expect a significant chance of encountering spam posts and a much lower but not negligible chance of coming across malicious links.**

**T**he popularity of online social networks has been growing exponentially. Launched in February 2004, Facebook—the world’s largest social network—had 250 million active users by July 2009 but doubled that number within just one year. According to Facebook’s own statistics ([www.facebook.com/press/info.php?statistics](http://www.facebook.com/press/info.php?statistics)), the average user has 130 friends and creates 90 pieces of content—news stories, blog posts, notes, photos, hyperlinks, and so on—monthly. The total user population spends 700 billion minutes on Facebook and shares more than 30 billion pieces of content each month.

The most obvious threat to users in social networks is loss of privacy. In July 2010, a security researcher revealed that the account details of more than 100 million Facebook accounts were publicly accessible through search engines.<sup>1</sup> In addition to loss of privacy, social network users face spam and various malicious threats including social engineering, identity theft, browser exploits, and malware.

Hackers target online social networks for several reasons:

- such networks contain large target populations;
- there is an abundance of personal information to steal or exploit;

- joining is fairly easy;
- users tend to have a high level of trust in one another and for objects (messages, links, photos, applications) within the networks;
- the variety of shared Web content, including hyperlinks and applications, exposes users to a range of potential attack vectors; and
- social graphs are highly interconnected, offering the potential for viral dissemination of malware and other attacks—a property of human social networks made famous by the “six degrees of separation” postulated by social psychologist Stanley Milgram.<sup>2</sup>

Despite the popularity and widely recognized security risks of online social networks, there have been very few large-scale investigations of the real extent of malicious threats. The “Related Work” sidebar summarizes the general findings of these efforts.

To assess the prevalence of malicious and spam posts in Facebook, we analyzed more than half a million posts with the help of Defensio, a Facebook application that protects users from such content as well as filters profanity and blocks URL categories. Our analysis revealed that a significant fraction of Facebook posts is spam and a much smaller fraction is malicious.

## FACEBOOK ARCHITECTURE

Typical for online social networks, Facebook is designed to allow a community of users to easily share information, messages, links, photos, and videos. After filling in a profile page, users can choose various levels of information access for different visitors. In addition, users can establish connections to designate “friends” or join groups. They can also send messages to one another through the Social



## COVER FEATURE

## RELATED WORK

**M**ost work on the risks of online social networks has focused on privacy concerns, but numerous researchers have looked at security threats.

Lynn Greiner noted that many attacks exploit the implicit trust between users in a social network, which makes people more likely to click on fake links or fall for social engineering schemes.<sup>1</sup>

Weimin Luo and colleagues surveyed numerous general threats to social networks and identified various attacker motivations and attack vectors—namely, spam, applications, malware, Web vulnerabilities, browser plug-ins, and social engineering.<sup>2</sup>

A team led by Tom Jagatic showed that it is easy to use data from social networks to hone phishing attacks.<sup>3</sup> They discovered that the success rate of phishing increases dramatically when e-mail appears to come from friends. Supporting this point, Garrett Brown and colleagues<sup>4</sup> found that, although Facebook itself does not reveal users' e-mail addresses, most such addresses can be obtained through public databases linked to the network. Furthermore, on most publicly accessible Facebook profiles, contextual information is available that hackers could exploit to generate context-aware spam. These researchers discovered that even a fraction of users with closed (private) profiles is vulnerable to such spam.

Several studies have considered the security risks related to Facebook applications.

Andrew Besmer and coauthors pointed out that Facebook app users are initially asked for permission to allow access to their profile data, but even if they do not consent, an app can still request such data on behalf of a friend who installed it.<sup>5</sup>

Constantinos Patsakis, Alexandros Asthenidis, and Abraham Chatzidimitriou carried out a case study with a malicious app on Facebook.<sup>6</sup> The app ostensibly was a slide show of dog pictures, but it also collected information about users' systems including IP address, browser version, operating system version, and open ports. Although the app only profiled users, it could have collected friend lists and sent messages to them, or executed arbitrary code.

Elias Athanasopoulos and colleagues examined ways to turn a social network into a botnet, demonstrating a proof-of-concept malicious Facebook app.<sup>7</sup> When a user activated the application, it displayed an image but also embedded hidden frames with inline images hosted at a designated target. Each time the user clicked within the app, it fetched the inline images without the target's awareness. The experiment suggests that an adversary taking full advantage of popular social utilities could generate a high volume of distributed denial-of-service traffic toward a target.

## References

1. L. Greiner, "Hacking Social Networks," *netWorker*, Mar. 2009, pp. 9-11.
2. W. Luo et al., "An Analysis of Security in Social Networks," *Proc. 8th IEEE Int'l Conf. Dependable, Autonomic, and Secure Computing (DASC 09)*, IEEE CS Press, 2009, pp. 648-651.
3. T.N. Jagatic et al., "Social Phishing," *Comm. ACM*, Oct. 2007, pp. 94-100.
4. G. Brown et al., "Social Networks and Context-Aware Spam," *Proc. 2008 ACM Conf. Computer Supported Cooperative Work (CSCW 08)*, ACM Press, 2008, pp. 403-412.
5. A. Besmer et al., "Social Applications: Exploring a More Secure Framework," *Proc. 5th Symp. Usable Privacy and Security (SOUPS 09)*, ACM Press, 2009, article no. 2.
6. C. Patsakis, A. Asthenidis, and A. Chatzidimitriou, "Social Networks as an Attack Platform: Facebook Case Study," *Proc. 2009 8th Int'l Conf. Networks (ICN 09)*, IEEE CS Press, 2009, pp. 245-247.
7. E. Athanasopoulos et al., "Antisocial Networks: Turning a Social Network into a Botnet," *Proc. 11th Int'l Conf. Information Security (ISC 08)*, LNCS 5222, Springer, 2008, pp. 146-160.

Inbox system, which functions as a closed e-mail service.

The unique features in Facebook include the Wall and News Feed. The Wall serves as a virtual bulletin board for people to post notes, comments, or other feedback about a person or group. Friends can write on one another's wall, and groups have walls for their members to communicate. News Feed aggregates and streams information about friends' activities.

The Facebook Platform was started in 2007 to let software developers create applications (in PHP or Java) that run in the Facebook environment. Facebook currently includes more than 550,000 active apps. These are actually installed on the developer's server, not on the Facebook server. Facebook calls an app when a user requests the application URL. The app communicates with Facebook using the Facebook API (application programming interface) or Facebook Query Language (FQL, similar to SQL). It returns content to Facebook formatted by the Facebook Markup Language (FBML, similar to HTML), which Facebook in turn presents to the user's Web browser.

Apps can interact and integrate with core Facebook services. For example, apps can access a user's friends list—say, to send invitations—or post to a user's news feed. One important built-in Facebook application is Links, which manages a user's link collection. Users can share links to interesting objects, and these links also appear on users' profile pages and in their news feeds. Recently, Facebook went further to offer Like buttons for any website. If a Facebook user clicks a Like button, the system adds a link to that website to the user's activity stream, which friends can see in their news feed.

## SECURITY THREATS

With their rising popularity, Facebook and other online social networks will become even more attractive targets than before.

Social engineering is an obvious attack vector because of the implicit trust most users have in the social network environment. A hacker can use a compromised account to send malware-infected messages to the account holder's friends, many of whom will accept the message at face value. Another social engineering attack lures users to a phishing site designed to look like Facebook and with a similar URL, and tries to trick them into submitting personal data.

Facebook users are also attractive targets for spam, including fake Facebook invitations, news stories, or Like messages. Fake e-mail becomes even more effective if an attacker can steal personal information from users' accounts.

Malware is another growing problem. A well-known example is the Koobface worm and its many variants, which have targeted Facebook as well as other online social networks such as Myspace, Twitter, and hi5 for more than two years. Koobface (an anagram of Facebook) spreads through messages to the friends of users who have an infected system. The message includes a video link that purportedly directs recipients to download an update to Flash Player but instead downloads the worm. Part of the worm payload is a Trojan horse that joins the computer to a peer-to-peer botnet.

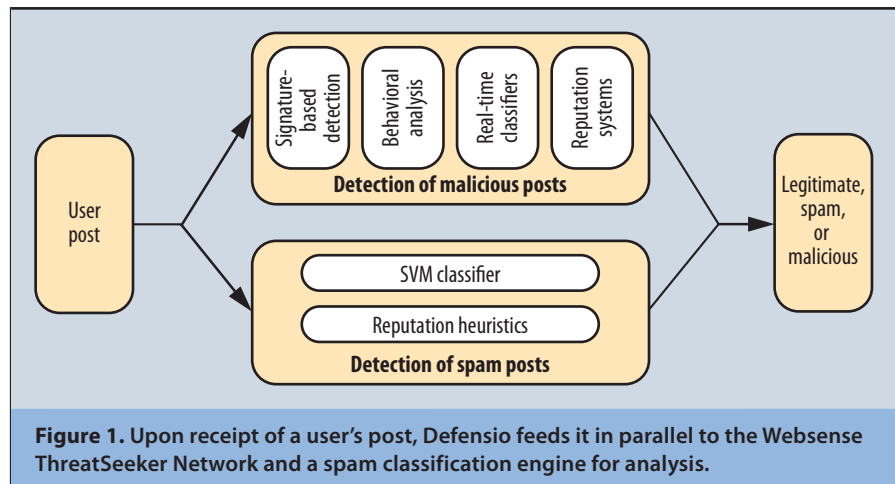
Malicious links are widely used for attacks, often taking users to a phishing site or drive-by download. Links can be shared in numerous ways in Facebook—for example, through messages, comments on a wall, shared news feed items, or the built-in Links application. A clickjacking worm has exploited the Like feature to spread such links: users receive messages with various subject lines that entice them to click a link; the link leads to a blank page with a hidden inline frame that publishes the initial message on their Facebook page, giving the appearance that they like the malicious link.

Facebook is quick to respond to suspicious or malicious links discovered by users or security companies and reportedly shares phishing and blacklist data with companies such as McAfee, MarkMonitor, and Microsoft. It also claims automated systems proactively detect and flag accounts with anomalous activity like sending many messages in a short time or messages with known bad links.

## DEFENSIO OVERVIEW

Defensio ([www.facebook.com/apps/application.php?id=177000755670](http://www.facebook.com/apps/application.php?id=177000755670)) is a Facebook application from Websense that monitors posts in a user's profile and determines whether they are legitimate, spam, or malicious (malware). In our study, we used the app to analyze only those posts that contained URLs. Although malicious links are clearly not the only threat to Facebook users, they are helpful in understanding the network's overall risk exposure.

To write a Facebook application, a developer registers with Facebook to access the Facebook API, which enables the app to read/write data from/to Facebook. In addition, Facebook provides an authentication mechanism that lets apps access the general information in users' profiles. It does not provide apps with access to users' private information; further, most apps require users' consent to access their data. When the user installs the app and allows it to



**Figure 1.** Upon receipt of a user's post, Defensio feeds it in parallel to the Websense ThreatSeeker Network and a spam classification engine for analysis.

access his data, the app registers the user with Defensio and provides him with a Defensio key with the Defensio API.

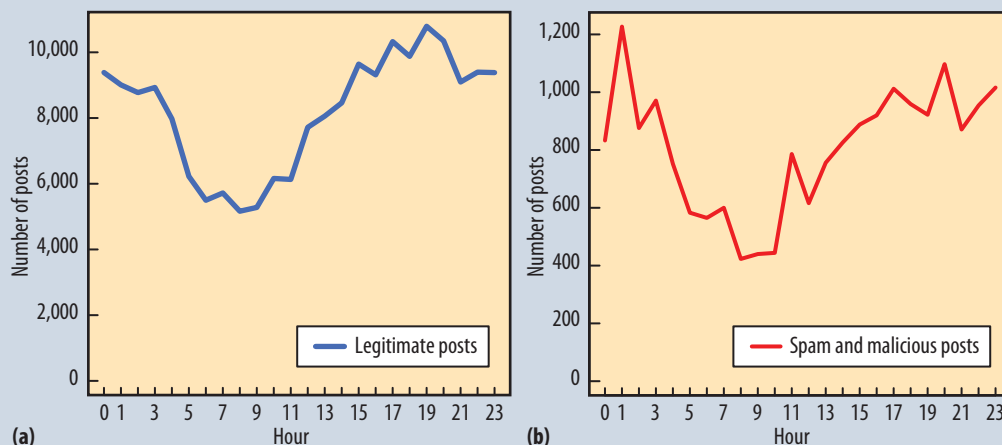
The app starts monitoring posts in the user's profile. It adds these to a *stream queue* and sends them in batches to Defensio for classification. The app associates each post with a Defensio user key to keep track of the recipient. After determining the status of the post, Defensio sends it to the *pending queue* to await user action. Users can request Defensio to immediately delete posts it classifies as spam or malicious, or they can request a notification e-mail and manually delete them.

Upon receipt of a user's post, Defensio feeds it in parallel to the ThreatSeeker Network ([www.websense.com/content/ThreatSeeker.aspx](http://www.websense.com/content/ThreatSeeker.aspx)), Websense's proprietary system for detecting malicious URLs, and a spam classification engine,<sup>3</sup> as Figure 1 shows. ThreatSeeker analyzes URLs in posts using a combination of signature-based detection, behavioral analysis of Web components, real-time content classifiers, and reputation systems, and accordingly flags those it determines to be malicious. The spam classification engine extracts the text from the post and runs it through a support vector machine (SVM) classifier, which assigns a score to the text. In parallel, the engine uses reputation heuristic rules to assign a score to the sender's identity. The engine then calculates a weighted average of the SVM and reputation scores and, based on this value, categorizes the post as either spam or ham (legitimate).

## RESULTS AND EVALUATION

Using Defensio data logs, we collected all Facebook posts containing a URL during a 21-day period, 22 June to 12 July 2010. These 502,624 posts were submitted by more than 25,000 users from 19 different countries. Each post had a timestamp indicating the date and time it was posted. In addition, Defensio had classified every post as legitimate, spam, or malicious. Our goal was not to evaluate the accuracy of Defensio's detection scheme but to

**COVER FEATURE**



**Figure 2.** Facebook posts containing URLs per hour over 21 days: (a) legitimate; (b) spam and malicious.

survey the temporal and network-level properties of those posts containing URLs that Defensio had determined to be malicious or spam.

**Table 1. Network properties of Facebook posts.**

Posts	Unique hosts	IP addresses	IP blocks	ASNs	Hosting countries
All	11,352	6,552	2,931	1,588	78
Legitimate	10,393	6,256	2,828	1,541	74
Spam	1,049	507	362	243	37
Malicious	156	127	104	74	24

**Temporal properties**

Approximately 215,999 of the Facebook posts contained URLs, averaging 10,286 each day. Thus, approximately two out of five posts contained a URL. The vast majority of these posts, 91 percent, were legitimate. A significant portion of posts, 8.7 percent, were spam: 18,693 total posts, averaging 890 per day. Only 0.3 percent of posts—644, an average of 31 per day—were malicious.

Figure 2 shows the number of Facebook posts at each hour totaled over all 21 days. The volume of legitimate posts rose steadily during the day to peak between 18:00 and 19:00 PST, perhaps because people socialize the most

**Table 2. Top 10 domains and their frequency.**

All posts		Legitimate posts		Spam posts		Malicious posts	
Host	Frequency	Host	Frequency	Host	Frequency	Host	Frequency
apps.facebook.com	115,560	apps.facebook.com	102,464	apps.facebook.com	13,094	nobrain.dk	103
facebook.com	42,010	facebook.com	41,749	facebook.com	223	mcdonaldsexposed.info	63
youtube.com	9,952	youtube.com	9,781	youtube.com	166	facebook.com	38
foursquare.com	706	foursquare.com	673	myspace.com	116	giveaway-madness.com	36
reddit.com	477	reddit.com	476	open.spotify.com	69	clicklikebro.info	21
p.ly	288	p.ly	288	foursquare.com	33	chkths.info	20
myspace.com	242	feedproxy.google.com	211	runkeeper.com	28	video.mcdonalds-revealed.com	17
hotmail.com	222	causes.com	208	ahmad.ly	20	truth.mcdonalds-revealed.com	14
flickr.com	213	hotmail.com	206	flickr.com	19	www.mjacksonsalive.com	12
feedproxy.google.com	211	maximumpc.com	203	hotmail.com	16	thecoolapps.com	11



**Table 3. Top 10 ASNs and their frequencies.**

All posts		Legitimate posts		Spam posts		Malicious posts	
ASN	Frequency	ASN	Frequency	ASN	Frequency	ASN	Frequency
AS32934	148,036	AS32934	135,731	AS32934	12,272	AS30736	104
AS15169	10,575	AS15169	10,383	AS15169	181	AS25653	63
AS14618	1,683	AS14618	1,633	AS33739	101	AS26347	53
AS33070	927	AS33070	870	AS43650	74	AS23522	36
AS21844	924	AS21844	844	AS26496	52	AS26496	36
AS26496	726	AS26496	638	AS33070	52	AS32934	33
AS36351	638	AS36351	623	AS10913	50	AS21844	32
AS20940	398	AS20940	398	AS14618	49	AS30058	20
AS8075	395	AS8075	371	AS21844	48	AS27458	15
AS3561	349	AS3561	340	AS36024	26	AS15169	11

after work. The pattern of spam and malicious posts looks roughly similar, rising during the day and early evening and then dipping in the early morning, but it also exhibits more irregularities. There was a sharp peak between 00:00 and 01:00 PST and a smaller peak between 19:00 and 20:00 PST. These irregularities might have occurred because spam and malicious posts are mostly planted by automated means, with the peaks representing bursts of activity by these programs.

### Network properties

For each URL extracted from the data, we resolved the IP address, autonomous system number (ASN), IP block, and country hosting the URL. To obtain the IP address-to-country mappings, we relied on MaxMind's geolocation database (<http://geolite.maxmind.com/download/geoip/database>), which uses regional Internet registries' whois information.

Table 1 summarizes the number of unique hosts, IP addresses, IP blocks, and ASNs, as well as the number of hosting countries. The values largely correspond to the proportionate volume of each category of URL. An unexpected revelation was that spam and malicious URLs were a small fraction of the total volume but were hosted in a disproportionately large number of countries, suggesting that malicious activities are geographically widespread.

Table 2 lists the top 10 domains and their frequencies (total number of appearances). Because the URLs were predominantly legitimate, the most frequently appearing domains in all posts were similar to those in legitimate posts. Most URLs in spam posts were hosted in the Facebook domain. In contrast, malicious links were hosted in various unusual domains.

Table 3 lists the top 10 ASNs and their frequencies. Because an ASN covers several IP addresses and IP blocks, we do not summarize the top IP addresses and IP blocks for

**Table 4. Top 10 ASNs not hosting any legitimate content.**

ASN	AS name	Country
AS6581	BKCNET "SIA" IZZI	Latvia
AS20597	ELTEL-AS ELTEL.NET	Russia
AS42560	BA-GLOBALNET-AS	Bosnia and Herzegovina
AS43134	COMPLIFE-AS	Moldova
AS19194	JOVITA Sentris Network LLC	US
AS34104	GLOBAL-AS Iletisim Hizmetleri	Turkey
AS25751	VCLK Valueclick Inc.	US
AS29650	HOSTING365-AS	Ireland
AS30361	SWIFTWILL2	US
AS50144	LALIB-AS	Portugal

each URL category. Note that some ASNs hosted only spam or malicious content and no legitimate content. Table 4 lists the top 10 ASNs that only hosted malicious or spam content. AS6851, which tops the list, has been heavily linked with malicious activities, especially the Koobface worm.

Table 5 summarizes the top 10 hosting countries and their frequencies. Most of the URLs were hosted in the US, Denmark, Norway, the UK, and Canada. Checking these countries against the locations of the Defensio application users revealed that the majority of users were from the US, followed by Norway, Germany, the UK, and Canada, which explains why these countries top the list. All the URLs hosted in three countries—Latvia, Morocco, and Paraguay—appeared in malicious or spam posts.


**O**nline social networks are a convenient way to keep informed about activities, share messages and multimedia with family and friends, and meet new people with similar interests. At the same time, they expose users to numerous security threats.

## COVER FEATURE

Table 5. Top 10 hosting countries and their frequencies.

All posts		Legitimate posts		Spam posts		Malicious posts	
Country	Frequency	Country	Frequency	Country	Frequency	Country	Frequency
US	187,340	US	172,590	US	14,344	US	406
Germany	1,215	Germany	1,175	Luxembourg	79	Denmark	110
Norway	989	Norway	975	Germany	33	Malaysia	14
UK	952	UK	933	Malaysia	28	France	13
Canada	395	Canada	377	Netherlands	28	Netherlands	10
Netherlands	320	Namibia	282	Canada	18	Germany	7
France	297	France	266	France	18	UK	5
Israel	239	Israel	235	UK	14	Turkey	5
Spain	221	Spain	220	Norway	14	Latvia	4
Denmark	208	Italy	187	Hong Kong	10	Austria	3

Our study suggests that Facebook users can expect a significant chance (9 percent) of encountering spam posts and a much lower but not negligible chance (0.3 percent) of coming across malicious links. The study also found the domains in spam posts to be mostly commonplace while those in malicious links tend to be unusual. Not unexpectedly, links in spam and malicious posts appear to be primarily hosted in some of the most technologically advanced western countries, but a few smaller countries, such as Latvia, host a substantial proportion of malicious content and little or no legitimate content.

Malicious links are an important risk indicator but do not portray the entire threat landscape. Much more research is needed to gain a better grasp of the true extent and nature of security threats in online social networks, especially social engineering and malware. 

### Acknowledgment

The authors thank Websense Inc. for providing access to the Defensio data.

### References

1. N. Bilton, "Researcher Releases Facebook Profile Data," *The New York Times*, 28 July 2010; <http://bits.blogs.nytimes.com/2010/07/28/100-million-facebook-ids-compiled-online>.
2. S. Milgram, "The Small-World Problem," *Psychology Today*, May 1967, pp. 61-67.
3. S. Abu-Nimeh and T. Chen, "Proliferation and Detection of Blog Spam," *IEEE Security & Privacy*, Sept./Oct. 2010, pp. 42-47.

**Saeed Abu-Nimeh** is a senior researcher at Damballa Inc., a network security company based in Atlanta, Georgia. His research interests include Web security, spam and phishing detection, and machine learning. Abu-Nimeh received a PhD in computer science from Southern Methodist University. Contact him at [sabunimeh@damballa.com](mailto:sabunimeh@damballa.com).

**Thomas M. Chen** is a professor of networking in the School of Engineering at Swansea University, Wales. His research interests include Web filtering, Web classification, network traffic classification, smart grid security, privacy, cybercrime, and malware. Chen received a PhD in electrical engineering from the University of California, Berkeley. He is a senior member of IEEE. Contact him at [t.m.chen@swansea.ac.uk](mailto:t.m.chen@swansea.ac.uk).

**Omar Alzubi** is a PhD student in the School of Engineering at Swansea University. His research interests include machine learning for detecting Web threats, social network spam, and elliptic curve cryptography. Alzubi received an MS in computer and network security from the New York Institute of Technology. Contact him at [omar\\_zubi@hotmail.com](mailto:omar_zubi@hotmail.com).



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



## COVER FEATURE



# Security Vulnerabilities in the Same-Origin Policy: Implications and Alternatives

Hossein Saiedian, *University of Kansas*

Dan S. Broyles, *Sprint Nextel*

**The same-origin policy, a fundamental security mechanism within Web browsers, overly restricts Web application development while creating an ever-growing list of security holes, reinforcing the argument that the SOP is not an appropriate security model.**

One of the first security measures that Internet browsers incorporated was the same-origin policy. As early as Netscape Navigator 2.0, SOP prohibited data sharing between origins—any unique host (such as a website), port, or application protocol. So, for example, SOP prevents one site's documents from accessing the document contents or properties from other sites. Thus, the SOP makes it possible for users to visit untrusted websites without allowing them to manipulate data and sessions on trusted sites.

If you browse `http://example.com/index.htm`, the SOP in the browser would accept or reject script and data accesses from the following sources:

- `http://example.com/about.htm` (port 80): accept
- `https://example.com/doc.html` (port 443): reject
- `http://google.com/search.php` (port 80): reject
- `http://dev.example.com/more.htm` (port 80): reject

By default, the SOP does not allow subdomains such as `dev.example.com` to interact with the primary domain.

However, by using the `<document.domain="example.com">` script in various subdomains, the SOP permits data sharing between the pages of `example.com` and `dev.example.com`. Doing so can cause problems, however; the script would let pages from a subdomain such as `userpages.example.com` access and alter pages from another subdomain, such as `payments.example.com`.

The SOP incorrectly assumes that all directory paths within the URL belong to the same source. For example, URLs `www.example.com/~john` and `www.example.com/~mary` have the same origin, even though they belong to different users and therefore should not trust each other. Another problem with the SOP is that it prevents developers from delivering dynamic multisource data. As the Internet and Web technology have progressed, the SOP has not evolved to keep up with the security needs of a more complex system, allowing malicious users to circumvent and exploit it.

In principle, the SOP restriction is a good security measure because it aims to protect data integrity and confidentiality. However, it has not kept up with changes in Web technology. The first Web browsers were not designed with security in mind, so developers added the SOP mechanism later to meet some basic security needs.

With the advent of JavaScript, Ajax, Web services, and mashups, clever programmers and hackers have found creative ways to subvert the SOP. Any SOP exploitation can expose a Web application to attack from malicious code, even if that exploitation comes from a well-intentioned developer. In addition, those who correct security flaws must account for the Web's unique environment, such as




## COVER FEATURE

statelessness and code mobility.<sup>1</sup> To further complicate matters, SOP rules and implementations differ between resources, DOM objects, XMLHttpRequests, cookies, Flash, Java, JavaScript, ActiveX, Silverlight, plug-ins, and browsers.

Inexperienced Web programmers who do not know that certain objects and actions—such as form submissions and tags like `<script>` and `<img>`—are not subject to SOP might copy JavaScript from other websites without understanding the security implications. Much like the early Web browser itself, such developers focus on functionality first and security last.

SOP weaknesses have led to attacks such as cross-site request forgery (CSRF), cross-site scripting (XSS), and Web cache poisoning. Attempts to fix these exploits have had only limited success; they tend to patch individual exploits without actually correcting the underlying security problems. In other words, the SOP is not the correct



**Inexperienced Web programmers who do not know that certain objects and actions are not subject to the SOP might copy JavaScript from other websites without understanding the security implications.**

security mechanism and requires redesign to meet the access-control requirements of Web-based assets. The Web security community is still debating how best to implement such a major undertaking. However, it seems clear that the current SOP lacks two basic access-control principles: the separation of privilege and least privilege.

Professional Web developers know about these deficiencies and the many effective mitigation techniques available, but many websites are built by nonprofessional developers with limited experience.

### NEED FOR DATA IN WEB APPLICATIONS

Internet activity is moving away from traditional searching and navigating toward an interactive and application-like activity in which browsers deliver dynamic, customized content. Users can enter their own content on Web forums, and social networking sites and mashups incorporate content from many users and third-party sites.

Jim Mischel has noted that the Web browser is the platform of the future, but in its current state, the SOP makes it difficult to share remote data and exposes too many vulnerabilities.<sup>2</sup> JavaScript and Ajax make modern feature-rich websites possible, bringing applications directly to users and improving efficiency and per-

formance. However, the SOP makes it difficult for Web applications from one source to obtain and display data from another. Developers use two common and powerful techniques to circumvent the SOP and obtain data from other domains; the first uses an Ajax proxy, and the second uses JavaScript object notation with padding (JSONP) script tag injection.

### Ajax proxy

XMLHttpRequest objects, the cornerstone of Ajax technology, make dynamic Web applications possible. The SOP restricts XMLHttpRequest calls much like it does any other script running in a browser, allowing such requests only between applications and servers from the same source.

Imagine a Web application that displays current stock price information hosted by a remote webserver. If the user enters a URL such as `www.getyourstocks.com/current.php?ticker=msft&format=json`, the remote webserver will return the current stock price in the following format:

```
{
    "ticker": "msft",
    "current": "24.5",
    "lastclose": "24.0",
    "pctchange": "2.1",
    "30dayavg": "23.45"
}
```

JavaScript makes it easy to format and display such return data on the webpage, but the SOP forbids the developer from making a request to `http://getyourstocks.com` from within his webpage. However, he can set up an application proxy server on his webserver, ask it to obtain the data from the other server, and deliver it through the server to the user. Since the page makes an XMLHttpRequest to the webserver proxy, which has the same origin as the Web application, the SOP allows it. A proxy server is functional, but slow and inefficient. It would be far better if the Web application could query the remote server directly.<sup>2</sup>

### JSONP script tag injection

Another approach to getting this outside data into the Web application is to place the call to `getyourstocks.com` inside a JavaScript function. In this example, if `getyourstocks.com` supports JSONP, then the programmer could add a JavaScript function on the page called, for example, `showCurrent`, that displays the data once it returns from `getyourstocks.com`:

```
function showCurrent(data){
    // display the contents of the data
}
```

Then all the application needs is a `<script>` tag that makes the request to the remote server:

```
<script type="text/javascript"
src="http://www.getyourstocks.com/current.
php?ticker=msft&format=json&callback=show
current">
</script>
```

The remote server will return the requested data so that it calls the callback function, `showCurrent`:

```
showCurrent({
  "ticker": "msft",
  "current": "24.5",
  "lastclose": "24.0",
  "pctchange": "2.1",
  "30dayavg": "23.45"
});
```

The webpage will execute the returned JavaScript as if it were native to the page. However, if a hacker were to alter the `getyourstocks.com` site so that it returns malicious JavaScript instead of stock quote data, the browser running this Web app will execute the malicious code. By circumventing SOP, the developer has introduced a security hole.

So how do developers obtain data for a Web application and still maintain security? They need a better security policy. Basic security logic suggests that third-party entities should not have the same access rights as trusted entities, so a policy that lets application developers determine access rights for each object might solve many SOP issues.

## HACKER EXPLOITS

Hackers use various exploits to take advantage of SOP deficiencies. There are many variations of these attacks, and hackers create new exploits all the time. Here, the purpose is not to enumerate every SOP vulnerability and attack, but to outline the fundamental flaws in the SOP so that readers can better analyze the proposed solutions.

### Cross-site request forgery

Security experts identified the earliest CSRF as a confused deputy attack. In this type of attack, hackers lure a victim to a malicious website to submit a form that points to a trusted target site in which the victim might have an active session. The trusted webserver receives and processes the form submission request, which looks identical to a legitimate request from the trusted website.

CSRF includes any malicious webpage with scripts that make unauthorized requests to trusted sites, hoping to take advantage of users who have an active session with the trusted site. For example, a user visits her bank's web-

site regularly and has an active session open with the bank when she browses to a malicious website containing code that calls the bank's webserver. The call requests a transfer to another account, robbing the unwitting user. The browser caches the user's active session information within itself, so the malicious request to the bank's server looks exactly like a legitimate user request.

Today, banks employ various measures to thwart such attacks, but other sites do not, especially those of small- and medium-size companies whose Web developers do not see the need for security measures covering SOP exploits. Using the Referrer header could be effective against CSRF, but Web applications frequently block this header because of privacy concerns, so an application that enforces it will exclude many users. Applications also strip Referrer headers from all HTTPS requests. Still worse, hackers can modify the Referrer header, making it unreliable.<sup>3</sup>

**The SOP is not the correct security mechanism and requires redesign to meet the access-control requirements of Web-based assets.**

Adam Barth and his colleagues recommended augmenting browser policy to use an Origin header as opposed to the Referrer header to provide CSRF and click-jacking protection.<sup>4</sup> The Mozilla security model currently proposes this fix (<https://wiki.mozilla.org/Security/Origin>).

### Cross-site scripting

There are two basic ways attackers implement XSS. The first method, considered nonpersistent, introduces malicious scripts in GET or POST requests that show up in pages returned by the server. For example, an attacker sends an e-mail containing a specially crafted hyperlink to a trusted website; the URL string includes malicious instructions reflected in the page returned by the webserver. The attack takes place when the user clicks on the hyperlink.

The second attack, considered persistent, occurs when the attacker injects malicious scripts into GET or POST actions that the server stores and then dynamically presents to the victim. For example, a malicious user logs into a forum and adds comments containing malicious JavaScript to a discussion thread as follows.<sup>5</sup>

```
<html><head><title>Paul's Blog</title>
</head><body>

  <h1>Scavenger Hunt!</h1>

  <hr>

  <h2>Paul: I will award the student
  bringing me the following items:</h2>
```

## COVER FEATURE

```

<ul>
<li>Yellow #2 pencil</li>
<li>Secretary's middle name</li>
<li>Number of ceiling tiles in our lab
</li>
</ul>
<hr>
<h4>Comments</h4>
Karthick: What will we get?
<script>
//malicious script that modifies the
above list
</script>
<hr>
</body></html>

```

The forum database stores this code as part of the thread. The code executes anytime a user browses that page.

**The most vulnerable websites are those of small- to medium-size companies or institutions, which often do not understand the threat.**

With the famous Samy MySpace worm, a user (Samy) exploited an XSS vulnerability in the MySpace profile form submission page and attached code to his profile. The injected code included a CSRF attack wherein anyone viewing his profile page would automatically add the same code to their profile page, make Samy their hero, and request Samy as a friend.<sup>6</sup> Another XSS attack technique is to embed an invalid image object and use the “onerror=” event to redirect the user to a webpage the attacker chooses. Any code injected into a webpage receives full rights as if it were part of the original page. The code can read and write other page elements, cookies, and browser history.

Proper form validation and input sanitization can prevent XSS attacks. The Samy attack infected a million user accounts in the first day, and the damage to MySpace's reputation is incalculable. The estimated cost to repair the damage from two smaller XSS attacks, Code Red and Slammer, was \$2.6 billion and \$1 billion, respectively.

In DOMXSS, a more recent version of an XSS attack, webpage code running on the client browser uses DOM objects related to the URL string and other environment variables that users can influence. If applications do not properly validate these objects, an attacker can use them to introduce malicious code into the page. This attack targets

Flash and other embedded programmable objects that have access to user-manipulated DOM objects and environment variables. In DOMXSS attacks, the client-side code embeds the malicious code into the webpage; in traditional XSS attacks, the server embeds the malicious code.

### Dynamic pharming

Dynamic pharming employs Domain Name System (DNS) hijacking to deliver a Web document with malicious JavaScript code, and then in a separate <iframe>, the attacker uses DNS rebinding techniques to load the authentic website the user expected.<sup>7</sup> In this way, the user is actually interacting with a trusted website while, behind the page, an attacker can monitor transactions and steal session cookies and passwords. Since the malicious page and the <iframe> appear to have the same origin, SOP allows the malicious page to interact with the legitimate website.

There are several ways to hijack the DNS name. For example, an attacker can set up a wireless hotspot in an airport. When unsuspecting users connect to this “free” hotspot, the attacker's router can intercept their DNS requests and redirect these requests to a site of the attacker's choosing. Victims might never suspect the attack since the browser indicates the domain that they trust and expect.

### PROTECTION CONSIDERATIONS FOR DEVELOPERS

Even if a website has little traffic and contains no confidential information, the fact that someone is interested enough to visit the site makes it a potential target for attack. As Figure 1 shows, the most vulnerable websites are those of small- to medium-size companies or institutions, which often do not understand the threat. If well-mannered users can find the site, so can hackers.

It is thus important for developers of even small sites to harden their sites against attacks. No strategy can guarantee protection against SOP attacks, but many make it more difficult for such attacks to succeed. When attackers exploit SOP vulnerabilities, they can steal passwords and cookies, log keystrokes, and alter information. Having good protection against CSRF attacks buys you nothing if a hacker can hijack user sessions. Therefore, it makes sense to prioritize security measures by protecting against XSS first, CSRF second.

### Scanning tools and services

Scanning services and vulnerability-checking tools find common flaws, such as not filtering user input.<sup>8</sup> Many popular vulnerability-checking tools are free, and they test for basic SQL injection and XSS vulnerabilities, system configuration problems, default passwords, and so on. However, they fail to detect CSRF or DOMXSS



attacks, which do not exactly fit the automated scanning template; moreover, such services might not be updated regularly enough to detect the newest exploits. For a fee, professional scanning services can examine Web code and simulate traditional and nontraditional attacks in a safe environment.

### Confine untrusted domain data to their own <iframe>

When dealing with untrusted content, open it up in an <iframe> with its own JavaScript execution context and its own DOM elements. Using a different domain lets you leverage the browser's SOP to isolate code and elements on the main page from malicious code or elements on the <iframe>. This disassociation is beneficial but not always possible. Mashups, for example, depend on the interaction of data and components from multiple sources.

### Avoid eval() and dynamically generated code

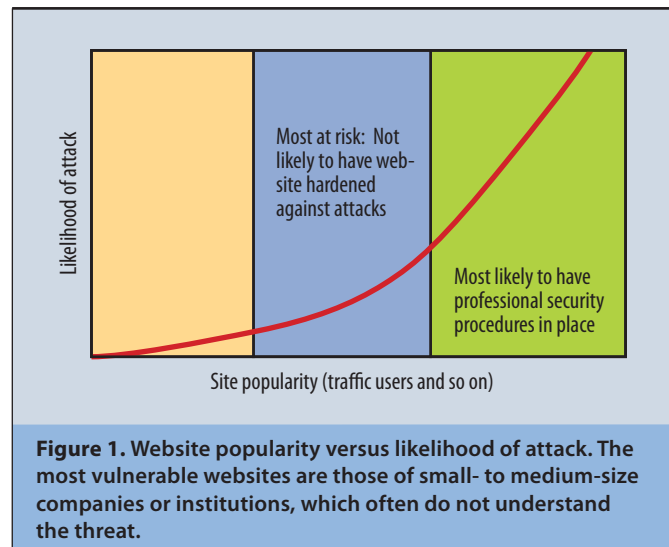
The eval() function lets the browser execute any string as JavaScript code. Web applications that do not properly validate input data risk executing malicious code. Avoid dynamically generated code unless absolutely necessary. JSON strings are meant to be a relatively safe subset of JavaScript that lets data safely pass through a Web application's eval() function. However, attackers might attempt to pass malformed JSON strings to your application, so use regular expressions or parsejson() to check for non-JSON strings.

### HTML validation and escape of untrusted data

Web servers must validate all input, including URLs, query strings, and post input. Sites that host blogs, forums, comments, reviews, and social networks let users contribute their own content, including HTML code and rich data. As previously noted, malicious users exploit SOP by uploading their own JavaScript routines. In the Samy MySpace worm example, MySpace in fact did filter the profile page submissions, but Samy circumvented the filters by breaking up the filtered words over multiple lines. Better filtering would have frustrated the attack. Server-side validation can remove potentially malicious tags and scripts from untrusted user-supplied content and reduce the threat of XSS attacks. Client-side validation is not reliable for security purposes. Use one of the freely available security-focused encoding libraries to help validate untrusted data.

### Use the HttpOnly cookie attribute

The HttpOnly cookie attribute is a cookie security control option that, if set, prevents JavaScript from accessing or modifying a cookie, making it more difficult for an attacker to steal or abuse a session.<sup>9</sup> Other cookie security parameters are also useful. The path attribute



restricts cookie access to a specific path in the URL, which makes it a little more secure than SOP. When set, the secure flag instructs the browser to only grant access to cookies from an HTTPS request. And setting the expires attribute to a date in the past makes the browser delete the cookie immediately instead of waiting until it closes.

### Use cryptographic tokens or captchas for high-risk GET/POST requests

Jeremiah Grossman<sup>6</sup> and Thomas Schreiber<sup>10</sup> both advocated including a cryptographic token to all links and forms that modify server-side data to provide strong protection against CSRF attacks. Presenting the user with a link or form that includes an unpredictable element specific to that action, user, and session disrupts these attacks. To implement this token, you can use a hidden input form element with a value from a keyed cryptographic hash like HMAC\_sha1(Action\_Name + Secret, SessionID). Before the server executes the request, it generates the same code or hash and compares it against the user-submitted one; if the two do not match, the server aborts the action.

Captcha and other challenge-response mechanisms are also effective, but they affect the user experience, so use them with care. Cryptographic synchronizer tokens are not visible to users and require no additional user steps.

### Avoid third-party code

Instead of pointing to a JavaScript or image file on a remote server, copy the file to your webserver and reference it from there. Do not trust third-party ads. If a webpage must contain ads, make sure they come from a reputable company with an excellent security record.

Any scripts injected into the website, including those for ads, have complete access to all webpage content.

## COVER FEATURE

Third-party scripts lower the website's security boundary. If a hacker alters that script, that and every other site using the same script will put confidential information at risk. Multiple websites using scripts from a handful of entities creates the potential for a single point of compromise. If it is absolutely necessary to use third-party scripts, check them for vulnerabilities at [secunia.com](http://secunia.com).

### User precautionary reminders

If a Web application involves sensitive user information, then its users will probably appreciate security reminders. Use e-mail or website notifications encouraging users to log out immediately after using the Web application; turn off JavaScript or use white-list plug-ins like No-Script; use a different browser to access secure or sensitive websites

**Basic security logic suggests that third-party entities should not have the same access privileges as trusted entities.**

than the one being used to browse the Internet freely, especially with tabbed browsing; and do not allow browsers to remember usernames and passwords.

### Mozilla Content Security Policy

One important security measure recently implemented in Mozilla Firefox 4 is the Content Security Policy (CSP). Aimed at mitigating XSS and click-jacking attacks, the CSP employs a set of directives that define the security policy for all types of webpage content on the webpage (<https://wiki.mozilla.org/Security/CSP/Specification>). The Web developer or administrator specifies a list of hosts or URIs that can supply each content type. Additionally, the CSP restricts common attack vectors in the client browser, denying inline `<script>` tags, calls to `eval()`, and other methods of creating code from strings.

### NEW BROWSER SECURITY MODELS

The preceding techniques mitigate but do not solve the root problem—the lack of an appropriate security model. Many proposed new models aim to alleviate browser security, but two that stand out as innovative and noteworthy also complement each other.

### Cryptographic server identity (locked SOP)

Chris Karlof and colleagues introduced a method that enforces access control not only via a website's host, port, and application protocol, but also through the webserver's cryptographic identity: a browser only grants access if the server's public Secure Sockets Layer (SSL) key matches the key from the locked Web objects.<sup>7</sup> This is crucial to protect

against pharming attacks, which manipulate DNS records and return the attacker's IP address with the target site's name. Victims are unaware that they are under attack since the URL in their browser shows the expected host name. Using the webserver's cryptographic identity, the browser would detect and deny any server whose cryptographic identity does not match the website's SSL public keys.

This proposal includes two policies: weak and strong locked SOPs. In traditional SSL server connections, the browser warns the user if the SSL certificate is unsigned or if it has any errors, but users tend to ignore the warning. SSL warnings can indicate a DNS spoofing or man-in-the-middle attack. Using the weak locked SOP, the browser would only allow a locked Web object to access another locked Web object if the standard SOP would have allowed it and if the object's certificate had no errors or warnings.

For the strong locked SOP, the browser tags locked Web objects with the public key of the webserver at the other end of the SSL connection. A browser implementing the strong policy would only allow access between locked Web objects if the standard SOP would have allowed it and if their tags match.

In a dynamic pharming attack, the attacker controls the main page, and the `<iframe>` contains the genuine trusted website, but only the true website server could produce the cryptographic credentials necessary to verify its identity. Therefore, the strong locked SOP would prevent the attacker's page from accessing anything on the genuine page. By authenticating the server in this way, the strong locked SOP prevents dynamic pharming attacks. The locked SOPs, however, do nothing to thwart XSS or CSRF attacks.

### Escudo Web protection

Basic security logic suggests that third-party entities should not have the same access privileges as trusted entities. If various sections within webpages included access-control mechanisms, a programmer could wall off untrusted scripts and content from accessing or changing trusted code or sensitive information.

Escudo is a new Web browser protection model that uses mandatory access control to wall off content from various sources and levels of trustworthiness.<sup>7</sup> By enforcing access rules similar to those found in some file systems, it seeks to enforce the separation of privilege and the principle of least privilege, the lack of which contributes heavily to XSS and CSRF attacks. With Escudo, Web developers identify the principles and objects in the code along with their levels of trustworthiness, and the Web browser implements those access decisions.

Unlike the CSP, which lists allowable sources of each content type for the whole page, Escudo is more granular; it identifies access rights to specific sections and elements

of the page, regardless of the content source. Developers assign all the elements of each webpage to a protection ring based on the trustworthiness of those elements and their protection requirements. The developer is free to apply as many rings as is necessary to protect the application's security.

Escudo lets developers define the meaning of a particular ring number, but ring level 0 is the most privileged. Principals can access elements with equal or lesser privilege, so a principal in ring level 2 can only perform operations on elements in ring levels 2 and higher. To assign ring levels, the developer encapsulates various page elements inside a div tag with a new attribute called ring. In addition to the ring boundaries, Escudo also incorporates access-control lists (ACLs), which let developers specify the minimum privilege level to read, write, or use a particular element.

The following code example defines a set of rings and access-control assignments:

```
<div ring=2 r=1 w=0>
...
<div ring=3 r=2 w=0>
...
</div>
</div>
```

In this code segment, the outer ring is level 2. The ACL assignments require that a principal must have a ring level of 1 to read the element ( $r = 1$ ), and ring level 0 to modify it ( $w = 0$ ). The combination of ring levels and ACLs gives Escudo a high degree of access granularity and lets developers employ the principle of least privilege in various parts of their applications.

For nested rings, inner rings must have a lower privilege level than outer rings, or else the Escudo security policy in the browser ignores them. This prevents untrusted sources from injecting code with a higher privilege level. Furthermore, div tags can include markup randomization attributes such as nonces to prevent injected code from splitting the div tag and creating a new div region with elevated privileges. Properly configured, the Escudo-enabled browser assigns untrusted principals to the least-privileged ring, where they cannot access or alter the rest of the page.

As the following example shows, Escudo rings separate untrusted page elements from trusted elements, thereby preventing a malicious user from altering the content:

```
<html><head><title>Paul's Blog</title>
</head><body>

<div ring=2 r=0 w=0 x=0
nonce=23409750497590487>
```

```
<h1>Scavenger Hunt!</h1>
<hr>
<h2>Paul: I will award the student
bringing me the following items:</h2>
<ul>
<li>Yellow #2 pencil</li>
<li>Secretary's middle name</li>
<li>Number of ceiling tiles in our
lab</li>
</ul>
</div nonce=23409750497590487>
<hr>
<h4>Comments</h4>
<div ring=3 r=1 w=1 x=1
nonce=23409750497590487>
Karthick: What will we get?
<div ring=0 r=0 w=0 x=0>
<script>
//malicious script to modify the
above list
</script>
</div>
</div nonce=23409750497590487>
<hr>
</body></html>
```

The attempt to embed a ring level of 0 will also fail, because it resides within ring level 3.

Using attributes in the HTTP header, Escudo rings can also protect cookies, browser API code, and browser history. Escudo is backward compatible; Web browsers that do not support the mechanism will simply ignore the Escudo attributes in the div tag and implement SOP as always.

**W**eb developers need better control and security mechanisms. Data and code from untrusted sources should not have the same privileges as the trusted programmer's code. Looking forward, the CSP should continue to improve; future development may include protection attributes similar to Escudo to allow fine-grained access control. Escudo itself could benefit by incorporating a valid-only flag, similar to the locked SOP, instructing the browser to ignore items within the div tag unless the SSL certificate(s) for the locked items in the tag are valid.

As the Internet evolves, the Web browser's fundamental security mechanism also must evolve. If the future Web



## COVER FEATURE

browser is to be an effective user interface for experiencing the Internet, privacy, trust, and security will be among its most important qualities. ■

## References

1. A. Rubin and D. Geer, "A Survey of Web Security," *Computer*, Sept. 1998, pp. 34-41.
2. J. Mischel, "Browser Applications and the Same Origin Policy," *informIT*, 6 Aug. 2010; [www.informit.com/guides/content.aspx?g=dotnet&seqNum=809](http://www.informit.com/guides/content.aspx?g=dotnet&seqNum=809).
3. J. Grossman, "Cross-Site Request Forgery: The Sleeping Giant," *WhiteHat Security*, July 2007; [www.whitehatsec.com/home/assets/WPCSRF072307.pdf](http://www.whitehatsec.com/home/assets/WPCSRF072307.pdf).
4. A. Barth, C. Jackson, and J.C. Mitchell, "Robust Defenses for Cross-Site Request Forgery," *Proc. 15th ACM Conf. Computer and Communications Security (CCS 08)*, ACM Press, 2008, pp. 75-87.
5. K. Jayaraman et al., "ESCUDO: A Fine-grained Protection Model for Web Browsers," *Proc. IEEE 30th Int'l Conf. Distributed Computing Systems (ICDCS 10)*, IEEE CS Press, 2010, pp. 231-240.
6. J. Grossman, "Cross-Site Scripting Worms and Viruses: The Impending Threat and the Best Defense," *WhiteHat Security*, Apr. 2006; <http://net-security.org/dl/articles/WHXSSThreats.pdf>.
7. C. Karlof et al., "Dynamic Pharming Attacks and Locked Same-Origin Policies for Web Browsers," *Proc. 14th ACM Conf. Computer and Communications Security (CCS 07)*, ACM Press, 2007, pp. 58-71.
8. M. Curphey and R. Arawo, "Web Application Security Assessment Tools," *IEEE Security & Privacy*, July/Aug. 2006, pp. 32-41.
9. S. Crites, F. Hsu, and H. Chen, "OMash: Enabling Secure Web Mashups via Object Abstractions," *Proc. 15th ACM Conf. Computer and Communications Security (CCS 08)*, ACM Press, 2008, pp. 99-107.
10. T. Schreiber, "Session Riding: A Widespread Vulnerability in Today's Web Applications," *SecureNet GmbH*, Dec. 2004; [www.securenet.de/papers/Session\\_Riding.pdf](http://www.securenet.de/papers/Session_Riding.pdf).

*Hossein Saiedian is a professor of software engineering in the Department of Electrical Engineering and Computer Science at the University of Kansas, where he also is a member of the Information and Telecommunication Technology Center. His research focuses on software engineering, particularly technical and managerial models for quality software development. Saiedian received a PhD in computer science from Kansas State University. He is a senior member of IEEE. Contact him at [saiedian@ku.edu](mailto:saiedian@ku.edu).*

*Dan S. Broyles is an engineer in the technology development organization at Sprint Nextel, where he provides spectrum analysis applications and automation tools for internal engineering teams. He received an MS in information technology from the University of Kansas. Contact him at [daniel.s.broyles@sprint.com](mailto:daniel.s.broyles@sprint.com).*



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

“All writers are vain,  
selfish and lazy.”

—George Orwell, “Why I Write” (1947)

(except ours!)



The world-renowned IEEE Computer Society Press is currently seeking authors. The CS Press publishes, promotes, and distributes a wide variety of authoritative computer science and engineering texts. It offers authors the prestige of the IEEE Computer Society imprint, combined with the worldwide sales and marketing power of our partner, the scientific and technical publisher Wiley & Sons.

For more information contact Kate Guillemette, Product Development Editor, at [kguillemette@computer.org](mailto:kguillemette@computer.org).

IEEE  

**CS Press**  
[www.computer.org/cspress](http://www.computer.org/cspress)

# IEEE computer society

**PURPOSE:** The IEEE Computer Society is the world's largest association of computing professionals and is the leading provider of technical information in the field.

**MEMBERSHIP:** Members receive the monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

**COMPUTER SOCIETY WEBSITE:** [www.computer.org](http://www.computer.org)

**OMBUDSMAN:** To check membership status or report a change of address, call the IEEE Member Services toll-free number, +1 800 678 4333 (US) or +1 732 981 0060 (international). Direct all other Computer Society-related questions—magazine delivery or unresolved complaints—to [help@computer.org](mailto:help@computer.org).

**CHAPTERS:** Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

**AVAILABLE INFORMATION:** To obtain more information on any of the following, contact Customer Service at +1 714 821 8380 or +1 800 272 6657:

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

## PUBLICATIONS AND ACTIVITIES

**Computer:** The flagship publication of the IEEE Computer Society, *Computer*, publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.

**Periodicals:** The society publishes 13 magazines, 18 transactions, and one letters. Refer to membership application or request information as noted above.

**Conference Proceedings & Books:** Conference Publishing Services publishes more than 175 titles every year. CS Press publishes books in partnership with John Wiley & Sons.

**Standards Working Groups:** More than 150 groups produce IEEE standards used throughout the world.

**Technical Committees:** TCs provide professional interaction in more than 45 technical areas and directly influence computer engineering conferences and publications.

**Conferences/Education:** The society holds about 200 conferences each year and sponsors many educational activities, including computing science accreditation.

**Certifications:** The society offers two software developer credentials. For more information, visit [www.computer.org/certification](http://www.computer.org/certification).

## NEXT BOARD MEETING

13–14 Nov., New Brunswick, NJ, USA

## EXECUTIVE COMMITTEE

**President:** Sorel Reisman\*

**President-Elect:** John W. Walz\*

**Past President:** James D. Isaak\*

**VP, Standards Activities:** Roger U. Fujii†

**Secretary:** Jon Rokne (2nd VP)\*

**VP, Educational Activities:** Elizabeth L. Burd\*

**VP, Member & Geographic Activities:** Rangachar Kasturi†

**VP, Publications:** David Alan Grier (1st VP)\*

**VP, Professional Activities:** Paul K. Joannou\*

**VP, Technical & Conference Activities:** Paul R. Croll†

**Treasurer:** James W. Moore, CSDP\*

**2011–2012 IEEE Division VIII Director:** Susan K. (Kathy) Land, CSDP†

**2010–2011 IEEE Division V Director:** Michael R. Williams†

**2011 IEEE Division Director V Director-Elect:** James W. Moore, CSDP\*

\*voting member of the Board of Governors †nonvoting member of the Board of Governors

## BOARD OF GOVERNORS

**Term Expiring 2011:** Elisa Bertino, Jose Castillo-Velázquez, George V. Cybenko, Ann DeMarle, David S. Ebert, Hironori Kasahara, Steven L. Tanimoto

**Term Expiring 2012:** Elizabeth L. Burd, Thomas M. Conte, Frank E. Ferrante, Jean-Luc Gaudiot, Paul K. Joannou, Luis Kun, James W. Moore

**Term Expiring 2013:** Pierre Bourque, Dennis J. Frailey, Atsuhiro Goto, André Ivanov, Dejan S. Milošević, Jane Chu Prey, Charlene (Chuck) Walrad

## EXECUTIVE STAFF

**Executive Director:** Angela R. Burgess

**Associate Executive Director; Director, Governance:** Anne Marie Kelly

**Director, Finance & Accounting:** John Miller

**Director, Information Technology & Services:** Ray Kahn

**Director, Membership Development:** Violet S. Doan

**Director, Products & Services:** Evan Butterfield

## COMPUTER SOCIETY OFFICES

**Washington, D.C.:** 2001 L St., Ste. 700, Washington, D.C. 20036-4928

**Phone:** +1 202 371 0101 • **Fax:** +1 202 728 9614

**Email:** [hq.ofc@computer.org](mailto:hq.ofc@computer.org)

**Los Alamitos:** 10662 Los Vaqueros Circle, Los Alamitos, CA 90720-1314

**Phone:** +1 714 821 8380

**Email:** [help@computer.org](mailto:help@computer.org)

## MEMBERSHIP & PUBLICATION ORDERS

**Phone:** +1 800 272 6657 • **Fax:** +1 714 821 4641 • **Email:** [help@computer.org](mailto:help@computer.org)

**Asia/Pacific:** Watanabe Building, 1-4-2 Minami-Aoyama, Minato-ku, Tokyo 107-0062, Japan

**Phone:** +81 3 3408 3118 • **Fax:** +81 3 3408 3553

**Email:** [tokyo.ofc@computer.org](mailto:tokyo.ofc@computer.org)

## IEEE OFFICERS

**President:** Moshe Kam

**President-Elect:** Gordon W. Day

**Past President:** Pedro A. Ray

**Secretary:** Roger D. Pollard

**Treasurer:** Harold L. Flescher

**President, Standards Association Board of Governors:** Steven M. Mills

**VP, Educational Activities:** Tariq S. Durrani

**VP, Membership & Geographic Activities:** Howard E. Michel

**VP, Publication Services & Products:** David A. Hodges

**VP, Technical Activities:** Donna L. Hudson

**IEEE Division V Director:** Michael R. Williams

**IEEE Division VIII Director:** Susan K. (Kathy) Land, CSDP

**President, IEEE-USA:** Ronald G. Jensen



revised 23 August 2011

## COVER FEATURE



# Secure Collaborative Supply-Chain Management

**Florian Kerschbaum and Axel Schröpfer**  
SAP Research Karlsruhe, Germany

**Antonio Zilli**  
University of Salento, Italy

**Richard Pibernik**  
EBS Business School, Germany

**Octavian Catrina**  
University of Mannheim, Germany

**Sebastiaan de Hoogh and Berry Schoenmakers**  
Eindhoven University of Technology, Netherlands

**Stelvio Cimato and Ernesto Damiani**  
Università degli studi di Milano, Italy

**The SecureSCM project demonstrates the practical applicability of secure multiparty computation to online business collaboration. A prototype supply-chain management system protects the confidentiality of private data while rapidly adapting to changing business needs.**

**T**he Internet's ubiquity and the advent of cloud computing enable new forms of collaboration, while services such as Facebook and Twitter let people communicate and interact in novel ways. These trends impact businesses as well as consumers, but can companies safely operate in this new world?

It is well known that effective supply-chain collaboration reduces inefficiencies such as excess costs, inadequate service levels, and environmental pollution. Nevertheless, companies often do not collaborate due to a lack of trust: they fear that data they share could be used outside its intended purpose. For example, business partners need to know one another's production costs to optimally schedule warehousing and production, but they could also use such information to manipulate future price negotiations.

Researchers have long tried to reconcile these two conflicting goals of information sharing and protecting

confidentiality. Cryptographers developed a theoretical solution more than 25 years ago: *secure multiparty computation*. According to this concept, multiple parties compute a function on their joint confidential input. Each party learns the result but nothing else about the other parties' inputs. The "Secure Multiparty Computation" sidebar describes this subfield of cryptography in more detail and provides an example of its use.

The adoption of secure multiparty computation for real-world business problems still faces four key obstacles. First, researchers must analyze observed inefficiencies in the supply chain and derive a formula that accurately addresses them. Second, researchers must develop a secure protocol to implement this computation. Designing such protocols is a challenging task, as their efficiency is low and performance optimizations often require manual security verification. Third, researchers must carry out practical experiments to evaluate the solution's computational performance. Finally, there must be a risk assessment to determine how adoption will impact the economic environment.

## CASE STUDY: AEROENGINE SUPPLY CHAIN

The European research project SecureSCM ([www.securescm.org](http://www.securescm.org)) has completed all four of these steps in a case study using an aeroengine parts manufacturer's real supply chain.



In the aerospace industry, coordination among supply-chain actors is critical to success. During program execution, many parts flow from one stage of the supply chain to the next, and at each stage these parts are assembled into a new component and continue their flow. This movement of products forms a pyramidal supply chain: at the apex is the *program leader firm*; below and directly connected to it are three or more *prime partners*, each in charge of a main section of the aircraft such as the engine, airframe, and avionics.

Prime partners manage numerous suppliers that provide the most important component modules. For example, in the engine section of the supply chain, the prime partner decouples the engine into the low-pressure turbine, the high-pressure turbine, the combustion chamber, the gearbox, and so on, and assigns the detailed design and manufacturing of these modules to different suppliers. As in the previous stage, the module suppliers require their own parts and raw materials. Given this complex network of firms and relationships, synchronizing product flow is essential to efficiently managing stock.

Downstream companies leading the supply chain are linked by long-term partnerships because designing and assembling engines and other complete aircraft parts require both highly specialized skills and knowledge and considerable research investment. In contrast, upstream firms that produce nonaerospace-specific products exploit short-term business opportunities.

The SecureSCM case study focused on the supply chain of the shroud nozzle, which is part of the engine's low-pressure turbine. Figure 1 shows the supply chain's structure, which includes the shroud nozzle manufacturer, an external manufacturer that is an outsourcing partner, and five component suppliers. The final customer is the engine manufacturer.

We interviewed the various actors in the supply chain, who provided several reasons for protecting the confidentiality of their data. The manufacturer wanted to protect short-term market demand and its capacity by limiting the external manufacturer's awareness of its business power. The suppliers sought to protect their capacity and maintain low costs by not causing the manufacturer

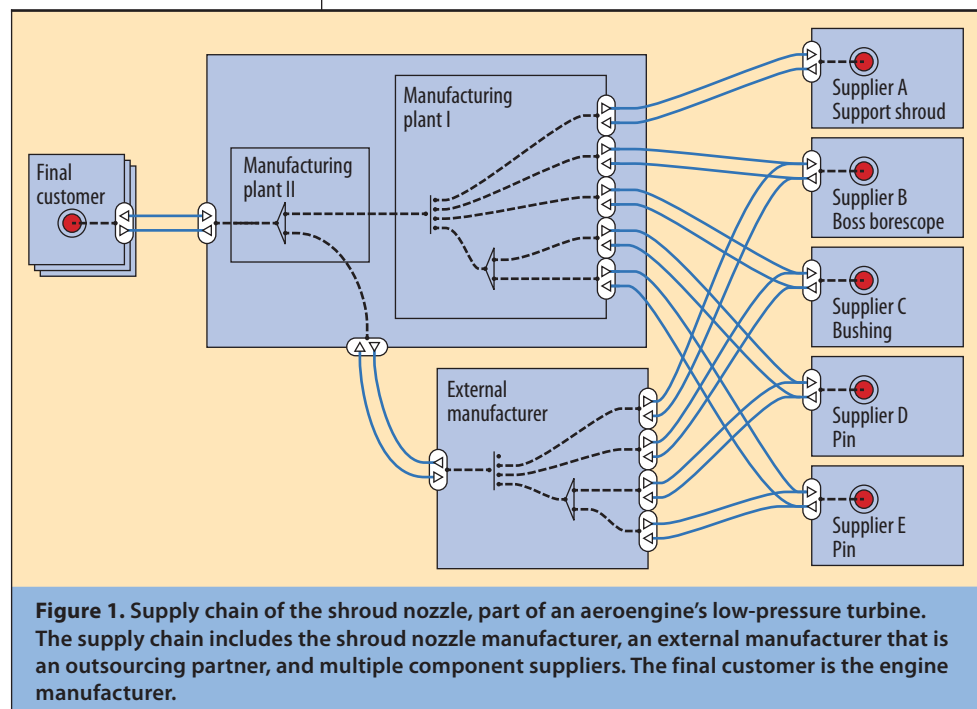
## SECURE MULTIPARTY COMPUTATION

**S**ecure multiparty computation is a cryptographic technique that allows several parties to compute any function without any party having to disclose its input to another. Each party's input remains private to that party, but the result can be made available to all, or only to a subset, of the other parties.

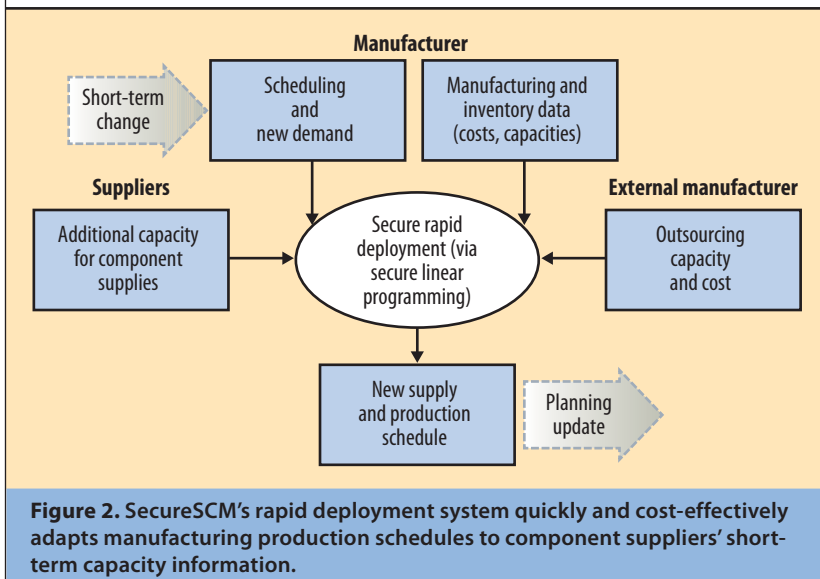
Consider the following simple example. Alice, Bob, and Charlie each have a number  $x_A$ ,  $x_B$ , and  $x_C$ , respectively, as private input and want to compute  $x = x_A + x_B + x_C$ . However, they do not want to disclose their private data to one another. Alice chooses a random number  $r$  and privately sends  $r + x_A$  to Bob. Bob adds his input and privately sends  $r + x_A + x_B$  to Charlie. Charlie does the same with his input and sends  $r + x_A + x_B + x_C$  back to Alice. Alice recalls  $r$ , subtracts it from the received value  $r + x_A + x_B + x_C$ , and announces  $x = x_A + x_B + x_C$ . Observing the messages exchanged between Alice, Bob, and Charlie, it is easy to see that none of them learns the input of the other parties—for example, Alice's random choice  $r$  blinds the value she sends to Bob, and Alice does not see the message (including  $r + x_A + x_B$ ) Bob sends to Charlie.

Cryptography research has proven that such a protocol exists for any efficiently computable function for any finite number of parties. It is important to note that secure computation does not rely on a trusted third party to perform computations and ensure data privacy, but is based on decentralized computation implemented through cryptographic protocols.

to look to other potential suppliers. They also wanted to conceal their production, warehousing, and shipping costs and their inventory status from the manufacturers to protect their business and production strategy. Moreover, and somewhat orthogonal to trust, all of the parties desired to protect their confidential data to meet their partners'



## COVER FEATURE



expectations: a firm unable to protect its confidential data is not a good partner.

### PROBLEM ANALYSIS

In standard operating mode, the shroud nozzle manufacturer procures four components from different suppliers that are delivered to a raw materials and component warehouse. Depending on the production schedule, the manufacturer transports these components to a facility where they are preassembled. It then ships the preassembled components to a different warehouse, from which they are delivered to a different manufacturing plant for final assembly, testing, and quality control.

In general, this part of the supply chain is characterized by a stable production schedule that the program leader determines well in advance. Sometimes, however, the program leader must, on short notice, order spare parts, which can lead to an upsurge in demand that the current supply chain cannot easily handle. The manufacturer has no means to quickly change production schedules and obtain extra parts from its standard suppliers. To cover this additional demand and satisfy its customers' service-level requirements, the manufacturer employs, on a contract basis, an external manufacturer that procures the necessary components, assembles the shroud nozzle, and delivers it to the manufacturer.

The outsourcing partner charges a premium price that exceeds the standard supply-chain cost by an order of magnitude. The manufacturer would therefore like to implement a system that enables it to quickly reschedule production to accommodate short-term demand peaks. However, the manufacturer has no information about the suppliers' capacity and thus cannot quickly place orders with them. One reason for this is that the suppliers are not willing to share detailed real-time information about their

capacity utilization and ability to provide additional components.

To remedy this problem, SecureSCM proposed a rapid deployment system, depicted in Figure 2, that utilizes the suppliers' short-term capacity data to quickly adapt manufacturing production schedules. The objective is to minimize total cost—including supply, manufacturing, inventory, and backordering costs—while respecting all relevant capacity constraints and fulfilling overall customer demand.

We formulated this optimization problem using *linear programming*, which minimizes an objective function subject to linear equality and inequality constraints. The objective function is the sum of all supply-chain costs; the equality

constraints model the flow of goods, and the inequality constraints implement the capacity constraints. To protect sensitive and private data, the protocol relies on secure multiparty computation.

### SECURE MULTIPARTY COMPUTATION PROTOCOL

Secure multiparty computation uses generic cryptographic protocols to protect data privacy and deliver correct outputs in the presence of an adversary—which can be a single entity or a coalition of parties—that controls communications. There are two basic models: in the *passive adversary* model, semi-honest parties reveal some secret data but otherwise continue to follow the protocol; in the *active adversary* model, parties deviate from the protocol to carry out malicious attacks.

Generic protocols for multiparty computation are roughly based on either *homomorphic encryption* or *linear secret sharing*. Both schemes represent a computable function as a Boolean or arithmetic circuit over a finite field or ring; linear functions of secret values are locally computed, while multiplication requires interaction between parties. With homomorphic encryption, the parties encrypt their secret inputs. With secret sharing, the parties randomly split their secret inputs into shares and give one share to each of the other parties. If enough parties combine their shares, they can reconstruct the secret; otherwise, they learn nothing about it. We used protocols based on linear secret sharing in the case study because they are more efficient than those based on homomorphic encryption.

The performance of generic protocols is insufficient for secure linear programming and similar large-scale applications. Our first task, therefore, was to develop a collection of efficient protocols based on linear secret sharing for operations on primitive data types and arrays.<sup>1,2</sup>

Current protocols provide all basic operations with binary, signed integer, and signed fixed-point data types as well as secret indexing for vectors and matrices. Floating-point protocols are still too complex for large-scale applications. We constructed the protocols using a small set of building blocks based on the same security model and cryptographic techniques, thus simplifying analysis and application development.

The protocols' effectiveness is primarily determined by the amount of exchanged data (communication complexity) and the number of sequential interactions (round complexity). Our pragmatic approach focused on achieving maximum efficiency while meeting rigorous but realistic security requirements. We obtained important performance and scalability improvements by precomputing secret random values and optimizing the building blocks for parallel computation of large batches of operations. In particular, we reduced communication complexity by encoding data in small fields that match the size of the data types; this is possible for protocols based on secret sharing because privacy does not rely on computational assumptions (as it does in the case of homomorphic encryption), and share conversions between different fields are efficient. Furthermore, families of building blocks that exploit different tradeoffs allow the application developer to select the variant that best suits the algorithm and execution environment.

Research on solving linear programs has produced two types of iterative algorithms: *interior-point* methods have polynomial complexity, but *simplex* techniques can require exponentially many iterations in the worst case. However, this is rarely an issue, and simplex algorithms are more suitable for secure computation due to their simpler iterations.


Previous secure linear programming protocols<sup>3</sup> solved only linear programs for which the null solution is feasible—that is, it does not violate any constraint. Because this condition is generally not satisfied, we developed a two-phase simplex protocol for solving any linear program. The inputs are shared-secret values representing the constraints (capacities) and the objective function (costs), while the outputs are the shared-secret final simplex data and a public value indicating the termination condition: optimal, unbounded, or infeasible. The termination condition proved useful in the case study for debugging the supply-chain model. The simplex data structure is protected throughout the computation through secret sharing as well as *secret indexing*, a cryptographic method that hides the indexes of changes in the simplex data. To avoid exponentially many iterations, the protocol also reveals the number of iterations.

The type of simplex algorithm a secure multiparty computation protocol uses has an important impact on performance and scalability. The algorithm must take into account the complexity of the secure operations: arithmetic,

comparison, and indexing. After evaluating different variants, we obtained the most efficient protocol with a simplex algorithm using fixed-point arithmetic.<sup>4</sup> This algorithm reduces communication complexity by using more compact data types, more efficient data structures, and fewer secure comparisons (one of the main performance bottlenecks). It is thus more suitable than other algorithms, including some variants that are superior for nonsecret data.

## IMPLEMENTATION

Implementing secure multiparty computation protocols is difficult: the programmer must manage complicated cryptographic routines and handle many auxiliary tasks, such as communication. To address this challenge, we designed L1, a small programming language that makes it easier to rapidly implement and evaluate such protocols during a project cycle.<sup>5</sup>



**The type of simplex algorithm a secure multiparty computation protocol uses has an important impact on performance and scalability.**

L1 is a procedural language similar to C or Java that also supports parallel execution. It offers primitives for cryptographic operations, such as homomorphic encryption and secret sharing, as well as for synchronous and asynchronous communication. All communication occurs over secure and authenticated channels using the Secure Sockets Layer protocol. The programmer can combine different secure computation protocols and even partially reveal intermediate results, if deemed safe.

An L1 program represents the code run by a specific player, instead of the function implemented by all players, but also incorporates *player-specific code*, such that the same program can be used for all parties. Consider the following example:

```

1  send(1, "share_" + id(), share);
2  1: {
3      for (int32 i=1; i<=players; i++)
4          shares[i]=readInt("share_"+i);
5      output(reconstruct(shares));
6  }
```

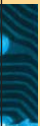
Line 1 sends all parties' shares to the first party. The player-specific code in lines 2-6 tells the first party to receive all shares and to reconstruct and output the secret.



## COVER FEATURE

After implementing a protocol in L1, we evaluate it using a testbed. Because protocol performance depends on the computer system as well as the network characteristics, we simulate two scenarios. In one scenario, the servers are connected via a local area network (LAN). This could be the case if the parties are cohosts in a common data center. In the other, more common, scenario, the servers are connected via a wide area network (WAN). In this case, we reduce the bandwidth and increase network latency.

For our case study, we connected eight servers representing the various supply-chain actors. Each server had a dual-core 2.6-GHz AMD Opteron 885 processor and 4 Gbytes of RAM. For brevity, we report here only the results of our best protocol variant. For the LAN scenario, with a bandwidth of 1,000 Mbits per second and a latency of 1 millisecond, the protocol runtime was 41 minutes and 4 seconds. For the WAN scenario, with a bandwidth of 10 Mbps and a latency of 20 ms, the runtime was 1 hour, 1 minute, and 38 seconds—a reasonable performance penalty for business purposes.



**Secure multiparty computation can facilitate other forms of collaboration besides supply-chain management.**

## RISK ASSESSMENT

Secure multiparty computation prevents supply-chain partners from learning one another's input, but it does not prevent them from altering their own. Partners who cannot reconcile their objectives with the common good might be tempted to distort the optimization to yield results more favorable to themselves. It is therefore necessary to assess the risk of any conflicts of interest.

Supply-chain actors are characterized by strategic interdependence—the payoff (typically in revenue) for any action depends on what the other actors do. All actors attempt to maximize their own payoff vis-à-vis others' response strategies, and equilibrium is attained when no actors would obtain a higher payoff by unilaterally deviating from their course.

In such a scenario, it is possible to compute a risk profile for each actor based on several factors that could contribute to selfish behavior:

- *fairness*—the actor's perception of the supply chain's fairness, modeled as the distance between its fairness value (the average of the contributions that each actor gives to all possible coalition configurations) and the actor's actual profit obtained from the chain;
- *payoff*—the actor's perception of the potential outcome of a unilateral action; and


- *context*—the actor's role and position within the supply chain.

We translate these three factors into a probability value and then compute a risk profile for each actor in the supply chain by multiplying the probability of a unilateral attack by its impact. It is then an easy matter to set up an incentive scheme that defines the way benefits and penalties devolve to the actors, motivates collaborative behavior, and punishes disruptive behavior.

Researchers can use a prototype supply-chain risk simulator that emerged from the SecureSCM project<sup>6</sup> to simulate a particular attack in different scenarios and predict the supply chain's response ([www.mathworks.com/matlabcentral](http://www.mathworks.com/matlabcentral)).

**S**ecureSCM demonstrates the practical applicability of secure multiparty computation to online business collaboration. Using part of an aeroengine manufacturer's supply chain, we designed and implemented a collaborative planning system that protects the confidentiality of private data while rapidly adapting to changing business needs. To our best knowledge, this is the first system to run secure multiparty protocols on such a large problem instance.

While data confidentiality is a clear prerequisite of online business collaboration, it is not sufficient. Because partners must put their trust in electronic systems, user acceptance is a critical issue. Secure multiparty computation raises the risk of selfish behavior, and all parties must feel confident that no one can game the system. Nevertheless, the companies we worked with in the case study have expressed an interest in ultimately adopting the system, which we continue to improve.

Secure multiparty computation can facilitate other forms of collaboration besides supply-chain management. Whenever parties working together are reluctant to exchange data due to confidentiality or privacy concerns, this approach can be beneficial. For example, law-enforcement agencies have successfully used secure multiparty computation in criminal investigations.<sup>7</sup> 

## References

1. O. Catrina and S. de Hoogh, "Improved Primitives for Secure Multiparty Integer Computation," *Proc. 7th Conf. Security and Cryptography for Networks (SCN 10)*, LNCS 6280, Springer, 2010, pp. 182-199.
2. O. Catrina and A. Saxena, "Secure Computation with Fixed-Point Numbers," *Proc. 14th Int'l Conf. Financial Cryptography and Data Security (FC 10)*, LNCS 6052, Springer, 2010, pp. 35-50.
3. T. Toft, "Solving Linear Programs Using Multiparty Computation," *Proc. 13th Int'l Conf. Financial Cryptography and Data Security (FC 09)*, LNCS 5629, Springer, 2009, pp. 90-107.

4. O. Catrina and S. de Hoogh, "Secure Multiparty Linear Programming Using Fixed-Point Arithmetic," *Proc. 15th European Symp. Research in Computer Security (ESORICS 10)*, LNCS 6345, Springer, 2010, pp. 134-150.
5. A. Schröpfer, F. Kerschbaum, and G. Müller, "L1—An Intermediate Language for Mixed-Protocol Secure Computation," *Proc. 35th Ann. IEEE Computer Software and Applications Conf. (COMPSAC 11)*, IEEE CS Press, 2011; <http://leprint.iacr.org/2010/578.pdf>.
6. M. Anisetti et al., "Using Incentive Schemes to Alleviate Supply Chain Risks," *Proc. Int'l Conf. Management of Emergent Digital EcoSystems (MEDES 10)*, ACM Press, 2010, pp. 221-228.
7. F. Kerschbaum and A. Schaad, "Privacy-Preserving Social Network Analysis for Criminal Investigations," *Proc. 7th ACM Workshop on Privacy in the Electronic Society (WPES 08)*, ACM Press, 2008, pp. 9-14.

**Florian Kerschbaum** is a research architect at SAP Research Karlsruhe, Germany. His research interests include security and privacy in distributed applications and applied cryptography. Kerschbaum received a PhD in computer science from Karlsruhe Institute of Technology. Contact him at [florian.kerschbaum@sap.com](mailto:florian.kerschbaum@sap.com).

**Axel Schröpfer** is a PhD student at SAP Research Karlsruhe. His research interests include applied cryptography and programming languages. Schröpfer received an MSc in computer science from Karlsruhe University of Applied Science. Contact him at [axel.schroepfer@sap.com](mailto:axel.schroepfer@sap.com).

**Antonio Zilli** is a research fellow in the Center for Business Innovation at the University of Salento, Italy. His research focuses on secure knowledge-management practices and aerospace value-network systems. Zilli received a degree in physics from the University of Lecce, Italy. Contact him at [zilli@unisalento.it](mailto:zilli@unisalento.it).

**Richard Pibernik** is a professor of supply-chain management at EBS Business School in Wiesbaden, Germany, and an adjunct professor at the Zaragoza Logistics Center, Spain. His research interests include supply-chain collaboration, demand fulfillment, and supply-chain risk management. Pibernik received a PhD in business administration from Goethe University, Frankfurt, Germany. Contact him at [richard.pibernik@ebs.edu](mailto:richard.pibernik@ebs.edu).

**Octavian Catrina** is a senior research associate in the Department of Mathematics and Computer Science at the University of Mannheim, Germany. His research interests include cryptographic protocols, secure multiparty computation, and security of distributed systems. Catrina received a PhD in telecommunications from Politehnica University Bucharest, Romania. Contact him at [octavian.catrina@uni-mannheim.de](mailto:octavian.catrina@uni-mannheim.de).

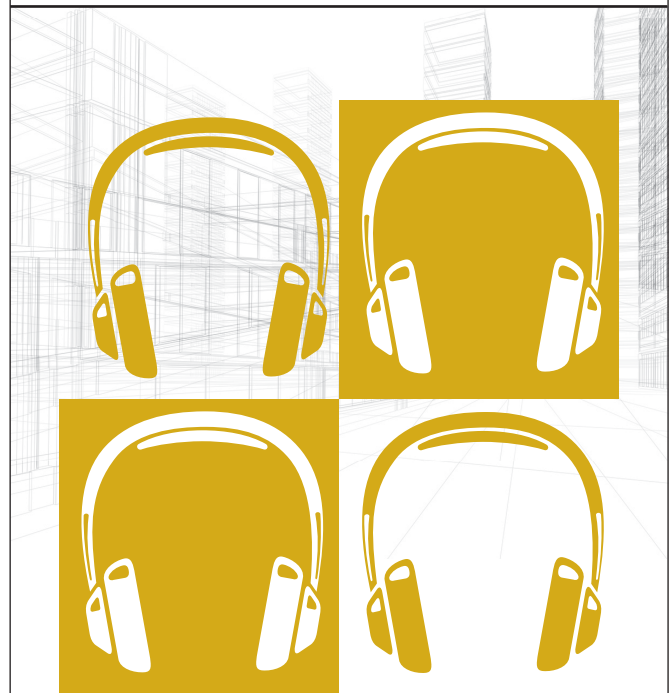
**Sebastiaan de Hoogh** is a PhD student at Eindhoven University of Technology, Netherlands. His research interests include applied cryptography and secure multiparty computation. De Hoogh received an MSc in mathematics from Eindhoven University of Technology. Contact him at [s.j.a.d.hoogh@tue.nl](mailto:s.j.a.d.hoogh@tue.nl).

**Berry Schoenmakers** is an associate professor of cryptography at Eindhoven University of Technology and an advisor on cryptography at Philips Research Labs. His research focuses on privacy-protecting protocols as applied to electronic voting, electronic payment systems, and secure multiparty computation, and on implementing systems based on such protocols. Contact him at [berry@win.tue.nl](mailto:berry@win.tue.nl).

**Stelvio Cimato** is an assistant professor of computer technology at Università degli Studi di Milano, Italy. His research interests include cryptography, network security, and Web applications. Cimato received a PhD in computer science from Università di Bologna, Italy. Contact him at [stelvio.cimato@unimi.it](mailto:stelvio.cimato@unimi.it).

**Ernesto Damiani** is a professor and heads the School of Computer Science at the Università degli Studi di Milano. His research interests include knowledge extraction and processing, system and network security, and software process engineering. Damiani received a PhD in computer science from the Università degli Studi di Milano. He is a senior member of IEEE and an ACM Distinguished Scientist. Contact him at [ernesto.damiani@unimi.it](mailto:ernesto.damiani@unimi.it).

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



## LISTEN TO GRADY BOOCH "On Architecture"

podcast available at

**cn** <http://computingnow.computer.org>

## COVER FEATURE



# The Final Frontier: Confidentiality and Privacy in the Cloud

**Francisco Rocha**, *University of Lisbon, Carnegie Mellon University*

**Salvador Abreu**, *University of Évora, Portugal*

**Miguel Correia**, *Technical University of Lisbon, Portugal*

**The boundary between the trusted inside and the untrusted outside blurs when a company adopts cloud computing. The organization's applications—and data—are no longer onsite, fundamentally changing the definition of a malicious insider.**

**M**any companies have embraced the benefits of cloud computing because of its pay-per-use cost model and the elasticity of resources that it provides. But from a data confidentiality and privacy viewpoint, moving a company's IT systems to a public cloud poses some challenges. System protection is often based on perimeter security, but in the cloud, the company's systems run on the cloud provider's hardware and coexist with software from both the provider and other cloud users. Simply put, the cloud blurs the formerly clear separation between the trusted inside and the untrusted outside.

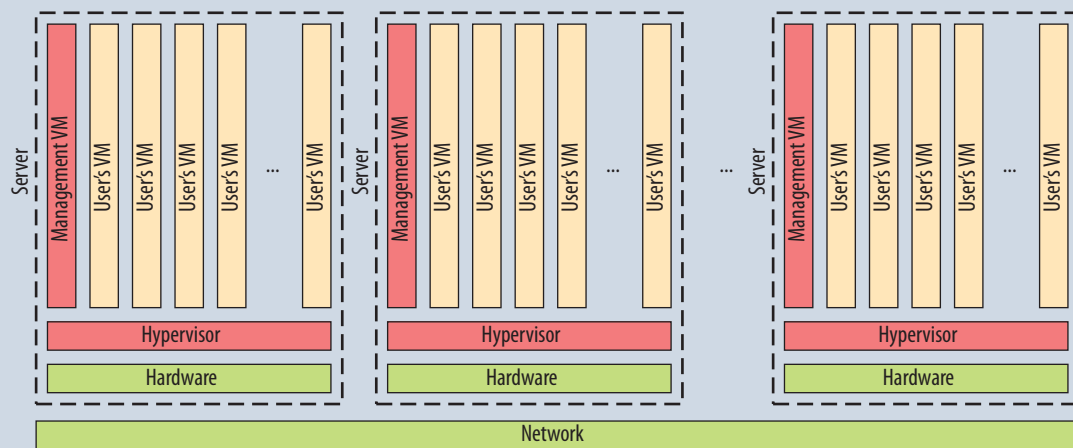
Although researchers have identified numerous security threats to the cloud,<sup>1</sup> malicious insiders still represent a significant concern. Security threats take on new dimensions in this new environment, as cloud operators and system administrators are unseen, unknown, and not onsite. Confidential data such as passwords, cryptographic keys, or files are just a few commands away from access by a malicious or incompetent system administrator.

Cloud providers are well aware of these concerns, as demonstrated in a recent roundtable including senior staff representing the sector's major companies.<sup>2</sup> One participant stated that his company has "very strict procedures in place for when our employees are allowed to access the machines the customer data resides on. We keep track of every action they take on those machines, and we log all that information for later audits so that we can ensure that all employees are behaving consistently with our privacy policy." Another participant added, "We have zero tolerance for insiders abusing that trust."

Although these policies are important—even essential—they fall short of solving the problem. Preventing physical access is not effective against remote attacks, and monitoring or auditing only detects an attack after it happens, which is usually too late. Interestingly, a participant in that same roundtable replied to a question about security and trust in the cloud by stating that "there are some things that will never go into [the cloud], for example, our SAP back end."

So how can an organization store confidential and private data in the cloud in a way that prevents its disclosure by malicious attacks inside the cloud? To the best of our knowledge, no single current approach solves this problem. Instead, some options include providing isolated environments<sup>3</sup> via trusted computing and the Trusted Platform Module (TPM).<sup>4,5</sup> DepSky prevents data disclosures but only in storage clouds via replication.<sup>6</sup>





**Figure 1. Infrastructure as a service.** This cloud provider's infrastructure illustrates three of possibly tens of thousands of cloud servers, each running virtual machines from one or more users.

A proposed solution based on the TPM offers protection against malicious insider threats in infrastructure-as-a-service (IaaS) clouds.

### INFRASTRUCTURE AS A SERVICE

The National Institute of Standards and Technology identifies three classifications for cloud services: infrastructure as a service, platform as a service, and software as a service.<sup>7</sup> We focus here on IaaS because it provides the user with the highest level of control over the infrastructure.

The main enabler of IaaS is native, or type I, virtualization.<sup>8</sup> This technology first appeared in the late 1960s in IBM mainframes such as the System/360. In native virtualization, the hardware—typically a server—runs a layer of software called a *hypervisor* or *VM monitor* that supports the execution of several virtual machines (VMs) with their own operating systems and software instances. The hypervisor essentially provides each VM with both an interface that is almost indistinguishable from the hardware and isolation from other VMs. For almost a decade, hypervisors were simply a software layer, but their efficiency today relies on hardware support from mainstream processors.

IaaS cloud services such as Amazon EC2, Rackspace, and Eucalyptus provide an interface for instantiating VMs from system images. Users can, for instance, request 50 VMs by running an image of a certain version of Linux and the Tomcat application server. Resources are elastic in the sense that the cloud can instantiate or delete VMs dynamically. The cloud provides APIs to help with load balancing among application servers.

Figure 1 offers a simplified representation of an IaaS infrastructure that shows three of possibly tens of thousands of cloud servers, each running VMs from one or more users. The servers contain a special VM that supports management operations such as launching a VM, deleting

a VM, taking a snapshot of a VM's memory, migrating a VM to a different server, monitoring VM performance, and backing up VM files.

### INSECURITY

A recent report about the Community Emergency Response Team's insider threat database found more than 550 insider attacks, including cases of sabotage, fraud, and intellectual property theft.<sup>9</sup> The report did not mention cloud computing specifically, but the threat in this context is clear from more recent cases, such as that of a disgruntled ex-employee of cloud storage provider CyberLynk. His attack deleted one entire season of a TV show of one of the company's clients ([www.courthousenews.com/2011/03/31/35406.htm](http://www.courthousenews.com/2011/03/31/35406.htm)).

Because IaaS gives cloud users more control over infrastructure, cloud providers necessarily have less control over it. In fact, providers own just a limited layer of software—the hypervisor and the management VM. Although this should limit what a malicious insider can do with a user's VMs and data, it is not enough. A recent study found three devastating yet simple-to-execute malicious insider attacks in IaaS clouds,<sup>10</sup> all of which assume the cloud's administrators have log-in access to the management VM and can run two operations: taking a snapshot of the VM's memory and mounting logical disk volumes. In this study, the hypervisor was Xen and the management VM was Xen's domain 0 with Linux (Xen's default management VM). The attacks happened as follows:

- *Extraction of cleartext passwords from a VM's memory.* Systems often store passwords in memory as clear-text, so the first attack required only three steps. First, the attacker ran Xen's `xm dump-core` command in the console to obtain a memory snapshot and store it in a file. Then, the attacker ran the `strings` command

## COVER FEATURE

Table 1. Trusted Platform Module components.

Component class	Function provided/data stored
Functional units	Random numbers Cryptographic hashes Message authentication codes (HMAC) RSA key-pair generation RSA encryption and decryption
Nonvolatile memory	Endorsement key Storage root key Owner authorization secret key
Volatile memory	RSA key pairs Platform configuration registers Key handles Authorization session handles

to extract strings from that file. Finally, he launched a dictionary attack with these strings to get both log-in passwords and the Apache Web server pass-phrases used to protect private keys.

- *Extraction of private keys from a VM's memory.* Systems typically store private keys of asymmetric cryptographic algorithms such as RSA in memory. At first glance, these keys are numbers, so they might seem indistinguishable from other data. However, they are usually stored in standard formats, such as PKCS#1, whose byte patterns are identifiable. The second attack consisted simply of running the `xm dump-core` command to obtain a memory snapshot and the `rsakeyfind` tool to search for certain byte patterns and extract the keys.
- *Extraction of files from the disk.* Xen typically uses Linux's logical volume manager (LVM), which attackers can use to obtain a copy of a VM's files. The third attack consisted of running `lvcreate` to create a new volume with a copy of all the VM files, `kpartx` to add a partition map to the new volume, `vgscan` to obtain the new volume's name, `vgchange` to activate the volume, and `mount` to make it accessible as a normal file system.

A cloud administrator's access to the management VM makes these attacks possible. Although a malicious cloud administrator can perform these attacks, another cloud user or a malicious insider in the user's infrastructure cannot exploit it.

## TRUSTED COMPUTING

To the best of our knowledge, the term *trusted computing* first appeared in the early 2000s to designate the work of what later became the Trusted Computing Group (TCG; [www.trustedcomputinggroup.org](http://www.trustedcomputinggroup.org)). The key idea was that security must be based on a few mechanisms implemented in hardware, such as cryptographic key storage. Hardware-based security is far from new, but the TCG made

important contributions by creating a standard hardware component, the Trusted Platform Module,<sup>11</sup> and defining a set of core functions based on this component.

## Trusted Platform Module

The TPM chip is currently found on the motherboard of many commercial PCs (its presence can be checked in the BIOS settings). It is supposed to be tamperproof, providing a set of functions that the software in the PC can call with the assistance of a device driver and a library for the programming language used.

The TPM essentially contains functional units and memory. As Table 1 indicates, the TPM's functional units are related to cryptography, including random number generation, key generation, hash functions, and RSA encryption.

Nonvolatile memory stores two important public/private key pairs whose private parts never leave the TPM—the endorsement key (EK) and storage root key—as well as a password called the owner authorization secret key. The EK uniquely identifies a TPM, whereas the storage root key encrypts keys to be stored outside the TPM; memory available inside the TPM is limited. Volatile memory stores several keys and handles, as well as a minimum of 16 platform configuration registers (PCRs) that store measurements, that is, cryptographic hashes of code modules.

## Root of trust for measurement

Systems typically use the TPM and PCRs to provide a root of trust for measurement (RTM). The objective is to give reliable measurements for assessing whether the system is in a trusted state—that is, to verify its integrity. These measurements are cryptographic hashes of certain code modules, such as the master boot record (MBR) or the operating system kernel. Cryptographic hash functions such as SHA-1 or SHA-256 are collision resistant, meaning it is computationally infeasible to find a different input that provides the same output/hash. Therefore, if a certain code module has a hash `h1`, it is impossible to substitute it with different code with the same hash `h1`.

When a system such as a server turns on, the PCRs are set to zero. In the boot process, several modules run in sequence, each one starting the next one—first the BIOS, then the MBR, the kernel, and so forth. To create the root of trust, each module calculates and stores in one PCR the hash of the next module. The BIOS provides a static RTM (SRTM) in the sense that it is trusted to provide the TPM with the correct hash of the first module it executes. This process creates a set of hashes in the TPM that the component can provide to challengers—processes in other computers charged with verifying whether the system is in a trusted state, meaning the system is running a certain version of the MBR characterized by having a certain hash, a certain version of the kernel, and so on.

As presented, this process has a serious vulnerability: after booting a configuration (MBR, kernel) that challengers do not consider trusted, the system modifies some of the PCRs in the TPM to hashes that the challengers trust; this would trick the challengers into believing that the configuration is the one the hashes represent, when this is not the case. To avoid this, the TPM does not have an operation to write a value into a PCR, only to *extend* a PCR. So, instead of storing the hash provided by whatever calls the TPM in a PCR, the extension operation stores a hash of the PCR's previous value concatenated with the input hash. Due to the collision resistance property, it is impossible to extend a PCR so that its state becomes a trusted hash. This means that the TPM design itself avoids this vulnerability.

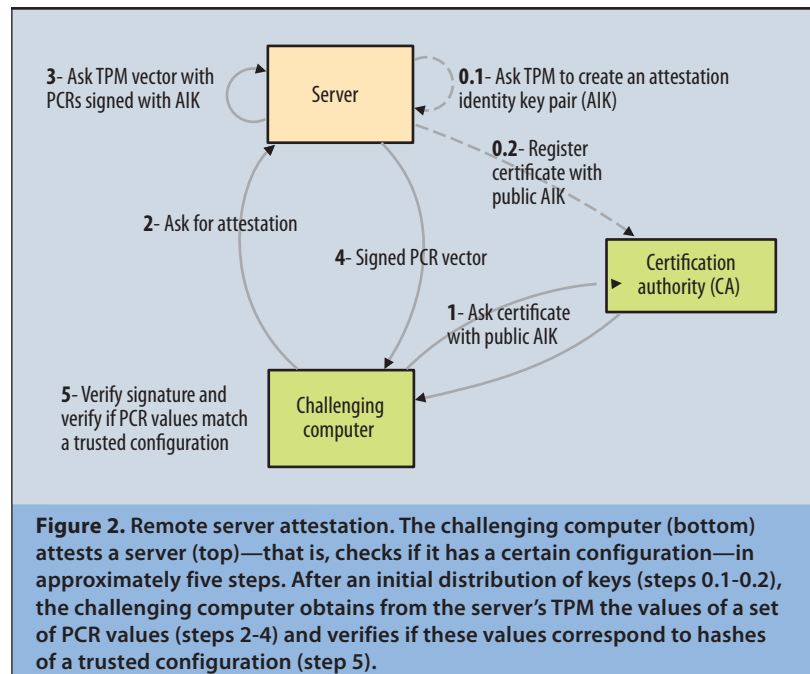
The SRTM requires the BIOS to be trusted, which could be problematic because it is possible to change BIOS content. The notion of dynamic RTM (DRTM), which was made possible by AMD's Secure Virtual Machine (SVM) and Intel's Trusted Execution Technology extensions to the x86 architecture, removes this limitation.<sup>4</sup> The main difference between SRTM and DRTM is that the latter enables the system to start protected code at any time, not just at boot time. For this to be possible, the extensions provide instructions to put the CPU in a clean state, akin to a restart, but from which it is possible to return to normal operation. This clean state represents a new root of trust.

### Remote attestation

One use for an RTM is remote attestation, which allows challengers to verify a remote system's integrity simply by asking for the values of some of its PCRs.

Figure 2 illustrates the remote attestation process for a challenging computer (bottom) attesting a server (top). Before the process starts, the server requires its TPM to create an attestation identity key (AIK) pair (step 0.1), obtains the public key, and then registers this key in a certification authority (CA) that issues a signed certificate (0.2). The remote attestation process officially starts with the challenging computer asking the CA for the certificate (1) and then asking the server for attestation—that is, for the values of a set of PCRs (2). The server obtains these PCR values signed with the AIK (3) from the TPM and sends them to the challenging computer (4). Finally, the challenging computer verifies the signature, which only the TPM can make, and determines whether the PCR values correspond to a trusted configuration (5).

However, remote attestation suffers from a major problem: although a challenger might consider a software



module as trusted, this does not mean that it is trustworthy. The module might be plagued with buffer-overflow and command-injection vulnerabilities that would let an attacker subvert its operation. The number of vulnerabilities in software is believed to be proportional to its size, so reducing the attested code size is an important goal. More precisely, attestation should be done only on security-critical software, the Trusted Computing Base.<sup>4</sup> This is a problem for SRTMs in particular because the complete operating system kernel must be trusted for trust to be put in modules loaded later in the system. With a DRTM, it is possible to attest smaller modules.

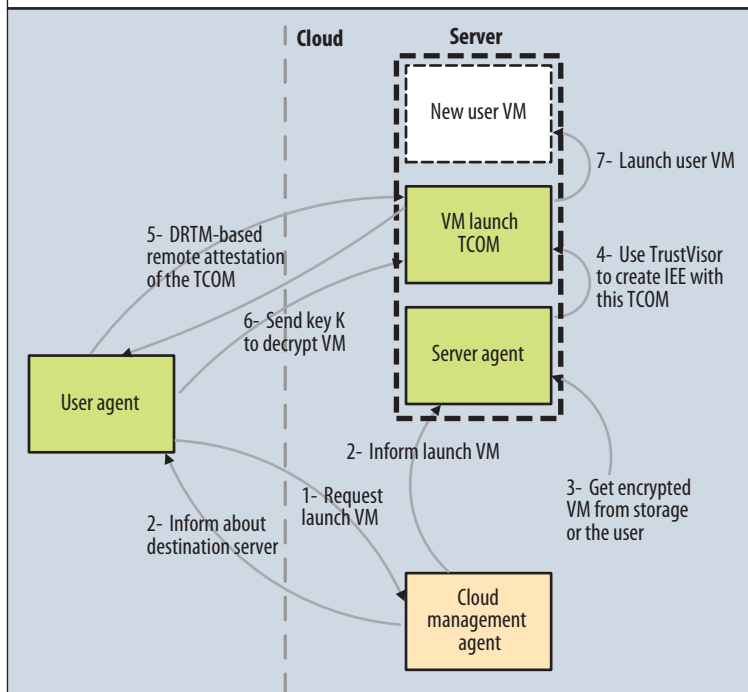
### PROTECTING CONFIDENTIALITY AND PRIVACY

To protect data confidentiality and privacy, the cloud must prevent certain attacks and give users the ability to assess whether the necessary mechanisms are in place, instead of simply trusting the cloud provider. This latter requirement might seem excessive, but the potential problem here is a malicious insider who might return arbitrarily incorrect information on the provider's behalf, so a solution can't be based on trusting the same provider.

User VMs reside in the cloud in three places: in servers (as in Figure 1), in the network during deployment and migration, and backed up on disks. To prevent data disclosure, users should limit what someone can do with the VM on a server and force it to be encrypted on the network and when backed up. Therefore, servers must run a trusted virtualization environment (TVE), comprising a hypervisor and a management VM that will not provide certain operations to administrators (such as snapshots and volume



## COVER FEATURE



**Figure 3. VM launch operation.** The user agent contacts the cloud management agent to identify which server to use for launch.

mount, due to the aforementioned vulnerabilities) and will support only trusted versions of others (launch, migrate, and backup VMs). Referring to one TVE is a simplification; cloud users can trust several TVE configurations with access to these TVEs' hashes and measurements. To attain the second aspect of this proposed solution, users can check these hashes with attestation. This process requires a public-key infrastructure (PKI) to provide certificates with public AIKs; the CA in Figure 2 is part of this PKI. Managing and distributing the measurements might require a third party, which could be the PKI provider or a different organization.

During server boot, the TPM's PCRs store the measurements of several loaded modules. Users can ensure that the hypervisor and management VM are loaded in a particular server via these hashes. This process is especially important to ensure that certain operations are not provided, such as memory snapshots, volume mount, and untrusted VM migration. It is not enough to provide a trusted VM migration operation:<sup>5</sup> a malicious administrator could let the user's VM launch in a server with a TVE and allow the attestation to finish, but then relocate the VM to a server that does not contain a TVE, just a hypervisor such as Xen with memory snapshots available. This administrator could then perform attacks.

A DRTM-based solution can help prove to cloud users that the modules responsible for launching, migrating, or destroying their VMs are trustworthy; we call these *trusted cloud operation modules* (TCOMs). To the best of

our knowledge, only one efficient implementation exists for DRTM-based attestation—TrustVisor, which is based on the idea of two-level integrity measurement.<sup>4</sup> TrustVisor uses both the hardware TPM and a software  $\mu$ TPM (which provides similar functionality) to extend TCOM measurement into the  $\mu$ TPM during runtime, when the module is needed. This removes the TPM from attestations done during runtime, which is important because TPM operations are slow (signatures take approximately 500 ms in some versions). Furthermore, TrustVisor isolates TCOM memory during its execution, thereby creating an isolated execution environment (IEE).

## VM OPERATIONS

The most critical VM operations—launching, migrating, backing up, and destroying—involve interaction among four components: the server agent, the TCOM of each operation in the server, the user agent, and the cloud management agent. The user agent is naturally trusted; the TCOM and server agent are trusted because they are subject to remote attestation. A malicious insider can control the cloud management agent, so it is not trusted.

### Launching a VM

As Figure 3 shows, to launch the VM, the user agent requests it by contacting the cloud management agent. This agent then selects a server on which to launch the VM (the scheduling operation) and informs the user agent and the server agent about this decision. The server agent obtains the VM either from storage in the cloud or from the user agent directly.

The server agent uses TrustVisor to create an IEE in which to start the VM launch TCOM. Then, the user agent does a DRTM-based attestation of this module and uses a simple protocol to send it key  $K$  to decrypt the VM. In this protocol, the TCOM sends the user agent its public key  $PK$ , the user agent encrypts  $K$  with  $PK$  and sends it, and the TCOM uses the corresponding private key to obtain  $K$ . Encryption prevents anyone from observing the VM content in the network.

Finally, the TCOM launches the VM. The module's execution in an IEE prevents the administrator from obtaining the key from the management VM memory itself.

### VM migration

VM migration is slightly more complex. In addition to performing a DRTM-based attestation of the VM's migration of TCOM in the origin server, the user agent also must attest the TCOM in the destination server. This ensures that the VM is moving into a TVE and that the entire process runs under the control of trustworthy modules.

Prior to migration, the TCOM must encrypt the VM in the origin server so that it is unreadable while it transfers through the network. This involves sending a session key similar to what happens in the launching process. Notice that the process starts with the cloud management agent and that, unlike VM migration in current clouds, this solution requires the user agent to actively participate in the beginning of the operation (to do the first attestation) and therefore be aware of the migration process.

### VM backup

Backing up a VM requires steps similar to launching a VM: after the initial signaling from the cloud infrastructure, the user agent attests the module, the agent and the module establish a session key, and the module encrypts the VM. The process finishes with the module sending the encrypted VM to its destination and deleting the key.

### VM termination


To securely terminate a VM, the hypervisor must clean the memory region and disk space used during VM execution. This is paramount to avoid leaking confidential data to the next VM that uses that memory region. If information leaks here, an attacker could use a special-purpose VM randomly launched in public cloud infrastructures to find the confidential data left forgotten in RAM. Attestation must be used again to ensure that this module is trustworthy.

Because the AIK is simply an RSA key pair, the cloud must prove to the user that a real TPM created this pair, not a malicious insider. For this purpose, the TPM must provide both the public AIK, in a certificate signed by the TPM's private EK, and the certificate with the public EK, which states that a certain TPM vendor created it and has a valid certificate chain to a root CA, such as Verisign.

**T**he solution we present here has one main problem: the gap between a measurement (a hash) and a complex software module's functionality. Checking that a hash belongs to a list of trusted hashes is trivial, but actually trusting that a hash represents a trustworthy complex module is quite different. This is particularly true for vulnerabilities in hypervisors or anomalies in virtualization that can allow a malicious user VM to attack another one.<sup>12</sup> Clearly, work remains to be done related to trustworthy hypervisors, management VMs, and the operation modules we propose.

Another research topic is the management of such a solution in a production environment. Different companies will develop different software modules, and various evaluation organizations will evaluate and certify the modules, measurements, and cloud providers. All these organizations must cooperate effectively under the pressure of

a short time to market. An additional issue is updating the measurements and revoking those that correspond to modules that eventually become untrusted.

Although we focused here on reinforcing the cloud infrastructure with trusted computing mechanisms, an entirely different approach distributes trust among several cloud providers. The idea is that the user does not trust the cloud provider and its administrators but instead trusts that there is no collusion among malicious insiders of more than a certain number of clouds. Currently, this idea has been applied in the context of cloud computing by a single system, DepSky,<sup>6</sup> whose purpose is to guarantee the confidentiality, integrity, and availability of data stored in the cloud. To do this, DepSky scatters data in storage clouds from four different providers, using Byzantine quorum system algorithms to assure data integrity and availability, even in the presence of data losses or corruptions in some of the clouds. This solution allows the implementation of secure storage clouds, but it is unclear how it can be extended to support replicated VMs. 

### Acknowledgments

This work was partially supported by the EC through project TLOUDS (FP7/2007-2013, ICT-257243) and by the FCT through the PIDDAC Program funds (INESC-ID multiannual funding), RC-Clouds (PCT/EIA-EIA/115211/2009), and REGENESYS (PTDC/EIAEIA/100581/2008). We thank the TLOUDS partners and members of the Navigators group for many inspiring discussions on topics covered in this article.

### References

1. Cloud Security Alliance, *Top Threats to Cloud Computing*, vol. 1, Mar. 2010; <https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf>.
2. E. Grosse et al., "Cloud Computing Roundtable," *IEEE Security & Privacy*, Nov./Dec. 2010, pp. 17-23.
3. D.G. Murray, G. Milos, and S. Hand, "Improving Xen Security through Disaggregation," *Proc. 4th ACM SIGPLAN/SIGOPS Int'l Conf. Virtual Execution Environments (VEE 08)*, ACM Press, 2008, pp. 151-160.
4. J.M. McCune et al., "TrustVisor: Efficient TCB Reduction and Attestation," *Proc. IEEE Symp. Security and Privacy (SSP 10)*, IEEE CS Press, 2010, pp. 143-158.
5. N. Santos, K.P. Gummedi, and R. Rodrigues, "Towards Trusted Cloud Computing," *Proc. 1st Workshop Hot Topics in Cloud Computing (HotCloud 09)*, Usenix, 2009; [www.usenix.org/event/hotcloud09/tech/full\\_papers/santos.pdf](http://www.usenix.org/event/hotcloud09/tech/full_papers/santos.pdf).
6. A. Bessani et al., "DepSky: Dependable and Secure Storage in a Cloud-of-Clouds," *Proc. European Conf. Computer Systems (EuroSys 11)*, ACM Press, 2011, pp. 31-46.
7. P. Mell and T. Grance, *The NIST Definition of Cloud Computing, Recommendation of the Nat'l Inst. Standards and Technology*, 2009; [http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145\\_cloud-definition.pdf](http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_cloud-definition.pdf).
8. M. Rosenblum and T. Garfinkel, "Virtual Machine Monitors: Current Technology and Future Trends," *Computer*, May 2005, pp. 39-47.

## COVER FEATURE

9. M. Hanley et al., "An Analysis of Technical Observations in Insider Theft of Intellectual Property Cases," tech. report CMU/SEI-2011-TN-006, Software Eng. Inst., Carnegie Mellon Univ., 2011.
10. F. Rocha and M. Correia, "Lucy in the Sky without Diamonds: Stealing Confidential Data in the Cloud," *Proc. 41st Int'l Conf. Dependable Systems and Networks Workshops (DSN 11)*, IEEE CS Press, 2011, pp. 129-134.
11. Trusted Computing Group, TPM Main Specification, v1.2, rev. 103, 2007; [www.trustedcomputinggroup.org/resources/tpm\\_main\\_specification](http://www.trustedcomputinggroup.org/resources/tpm_main_specification).
12. T. Ristenpart et al., "Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds," *Proc. 16th ACM Conf. Computer and Comm. Security (CCS 09)*, ACM Press, 2009, pp. 199-212.

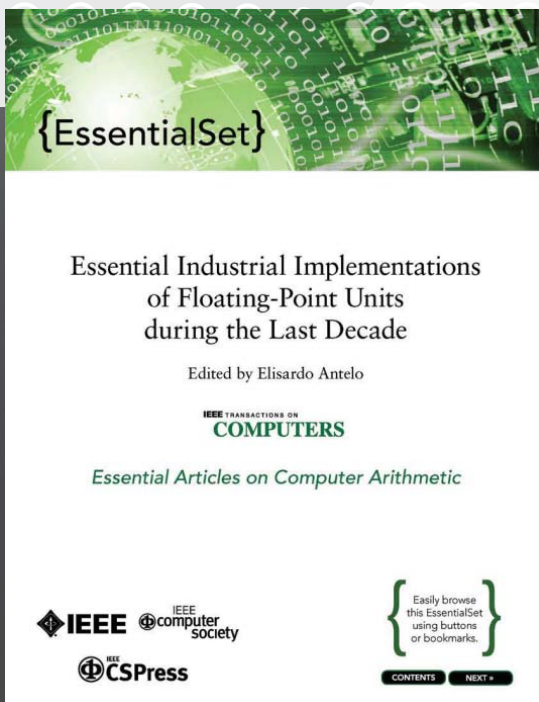
**Francisco Rocha**, formerly affiliated with the University of Lisbon, Portugal, and Carnegie Mellon University, is an information security consultant whose research interests include application and network security, trusted and trustworthy computing, cloud computing security, access control, and security architecture and design. Rocha received an MSc in information technology and informa-

tion security from Carnegie Mellon University, an MS from the University of Lisbon, and a licenciatura degree from the Polytechnic Institute of Beja, Portugal. Contact him at [rocha.francisco@gmail.com](mailto:rocha.francisco@gmail.com).

**Salvador Abreu** is an associate professor in the Department of Computer Science at the University of Évora, Portugal. His research interests include programming language design and implementation, and parallel and distributed computing models. Abreu received a PhD in computer science from the Universidade Nova de Lisboa. Contact him at [spa@di.uevora.pt](mailto:spa@di.uevora.pt) or via [www.di.uevora.pt/~spa](http://www.di.uevora.pt/~spa).

**Miguel Correia** is an associate professor in the Instituto Superior Técnico at the Technical University of Lisbon. He is also a researcher in the Distributed Systems Group at INESC-ID. His research interests include security, intrusion tolerance, distributed systems, and cloud computing. Contact him at [miguel.p.correia@ist.utl.pt](mailto:miguel.p.correia@ist.utl.pt) or via <http://homepages.gsd.inesc-id.pt/~mpc>.

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



NEW from  CS Press

## ESSENTIAL INDUSTRIAL IMPLEMENTATIONS OF FLOATING-POINT UNITS DURING THE LAST DECADE: VOLUMES 1 & 2

Edited by Elisardo Antelo


Surveys the industrial design of floating-point units during the last decade. This EssentialSet is broken into two volumes, sold separately.

PDF edition • \$15 each (\$9 members) • 103 & 79 pp.

Order Online:  
[computer.org/store](http://computer.org/store)



## COVER FEATURE



# Securing the Internet of Things

Rodrigo Roman, Pablo Najera, and Javier Lopez, *University of Malaga, Spain*

**In the Internet of Things vision, every physical object has a virtual component that can produce and consume services. Such extreme interconnection will bring unprecedented convenience and economy, but it will also require novel approaches to ensure its safe and ethical use.**

In the Internet of Things (IoT), everything real becomes virtual, which means that each person and thing has a locatable, addressable, and readable counterpart on the Internet. These virtual entities can produce and consume services and collaborate toward a common goal. The user's phone knows about his physical and mental state through a network of devices that surround his body, so it can act on his behalf. The embedded system in a swimming pool can share its state with other virtual entities. With these characteristics, the IoT promises to extend “anywhere, anyhow, anytime” computing to “anything, anyone, any service.”

Several significant obstacles remain to fulfill the IoT vision, chief among them security. The Internet and its users are already under continual attack, and a growing economy—replete with business models that undermine the Internet's ethical use—is fully focused on exploiting the current version's foundational weaknesses. This does not bode well for the IoT, which incorporates many constrained devices. Indeed, realizing the IoT vision is likely to spark novel and ingenious malicious models. The challenge is to prevent the growth of such models or at least to mitigate their impact.

Meeting this challenge requires understanding the characteristics of things and the technologies that empower the IoT. Mobile applications are already intensifying users' interaction with the environment, and researchers have made considerable progress in developing sensory devices to provide myriad dimensions of information to enrich the user experience.

However, without strong security foundations, attacks and malfunctions in the IoT will outweigh any of its benefits. Traditional protection mechanisms—lightweight cryptography, secure protocols, and privacy assurance—are not enough. Rather, researchers must discover the full extent of specific and often novel obstacles. They must analyze current security protocols and mechanisms and decide if such approaches are worth integrating into the IoT as is or if adaptations or entirely new designs will better accomplish security goals.

The proper legal and technical framework is also essential. To establish it, analysts must thoroughly understand the risks associated with various IoT scenarios, such as air travel, which has many interrelated elements, including safety, privacy, and economy.<sup>1</sup> Only then is it possible to justify the cost of developing security and privacy mechanisms.

All these requirements underline some critical first steps in successfully implementing IoT security measures: understand the IoT conceptually, evaluate Internet security's current state, and explore how to move from solutions that meet current requirements and constraints to those that can reasonably assure a secure IoT.

## INFRASTRUCTURE SEEDS

The “Objects in a Superconnected World” sidebar describes some of the characteristics of the things in the

## COVER FEATURE

## OBJECTS IN A SUPERCONNECTED WORLD

Since the IoT's inception, governments and other organizations have tried to capture its essence in words, some more successfully than others.<sup>1</sup> In a nutshell, the IoT is a worldwide network of interconnected objects. Each object that surrounds a person, from books and cars to appliances and food, has a virtual avatar that behaves as an active entity. In this context, all IoT objects have five main characteristics:

**Existence.** Things, such as a car, exist in the physical world, but specific technologies, such as an embedded communication device, enable the existence of a thing's virtual persona.

**Sense of self.** All things have, either implicitly or explicitly, an identity that describes them, such as car, Porsche, or license plate number. Objects can process information, make decisions, and behave autonomously.

**Connectivity.** Things can initiate communication with other entities. As a result, both an element in their surroundings and a remote entity can locate and access them.

**Interactivity.** Things can interoperate and collaborate with a wide range of heterogeneous entities, whether human or machine, real or virtual. As such they produce and consume a wide variety of services.

**Dynamicity.** Things can interact with other things at any time, any place, and in any way. They can enter and leave the network at will, need not be limited to a single physical location, and can use a range of interface types.

An optional sixth characteristic is environmental awareness. Sensors might enable a thing to perceive physical and virtual data about its environment, such as water radiation or network overhead. This characteristic is optional because not all things will exhibit it, such as an object enhanced with a lower-end radio frequency identification (RFID) tag.

The combination of various technologies has enabled objects to exhibit these characteristics, allowing them to become virtual beings. Energy-efficient microcontrollers act as brains, imbuing objects with embedded intelligence. Sensor technology provides

objects with sensory receptors, and RFID provides a way for them to distinguish one another, much like people recognize a face. Finally, low-energy wireless technology, such as specified in IEEE 802.15.4, supplies the virtual counterparts of voice and hearing.

Multiple applications already use these and other technologies, such as machine-to-machine communication, virtual worlds, and robotics. To be a virtual being, an IoT object needs only enough technology to realize its role and complete its mission. A tire can simply provide information about itself and its state, but a car will be much more technologically complex because it must be aware of its surroundings as well as its own state.

RFID in pharmaceutical environments, location-aware mobile applications, and smart metering systems are all essentially "intranets of things," in which objects are isolated from those in other scenarios and domains. IoT applications will have greater scope and flexibility, being able to interact not only with objects in other scenarios and domains but also with real and virtual entities.

Figure A shows an application involving a smart meter with current capabilities—an intranet-of-things scenario—and a smart meter as part of the IoT.

Another example is weather stations, which will send anonymous queries to pedestrians' personal networks to create a city temperature and humidity map that business owners can integrate with other data to decide the best place to build an ice cream shop. Likewise, virtual supermarket goods will interact with a clerk to notify him of a strawberry yogurt shortage; with a potentially allergic shopper to provide her with ingredient information; and with third-party applications, such as event planners, to reveal if the shopper's friends like strawberry yogurt.<sup>2</sup>

At present, only partial IoT instances exist, mainly those that provide information services through centralized systems and interfaces. These IoT application forerunners hint at the possible benefits of a full-blown IoT and can serve as foundational elements for building this new virtual world. New companies are providing a centralized interface to access raw sensor data worldwide. Such data can help launch an IoT application. The personal network paradigm is another example of a partial instance. Local entities, such as wearable objects, interact indirectly with external services, such as a fitness monitor, through a central server such as a smartphone. In this way, users and their objects interact with their environment, deciding what data they want to share and with whom. These and similar applications are a start, but to attain the IoT's full benefits, work must continue.

## References

1. H. Sundmaeker et al., eds., "Vision and Challenges for Realizing the Internet of Things," *IoT European Research Cluster*, Mar. 2010; [www.internet-of-things-research.eu](http://www.internet-of-things-research.eu).
2. E. Fleisch, "What Is the Internet of Things? An Economic Perspective," white paper, WP-BIZAPP-053, AutoID Labs, Jan. 2010; [www.autoidlabs.org](http://www.autoidlabs.org).

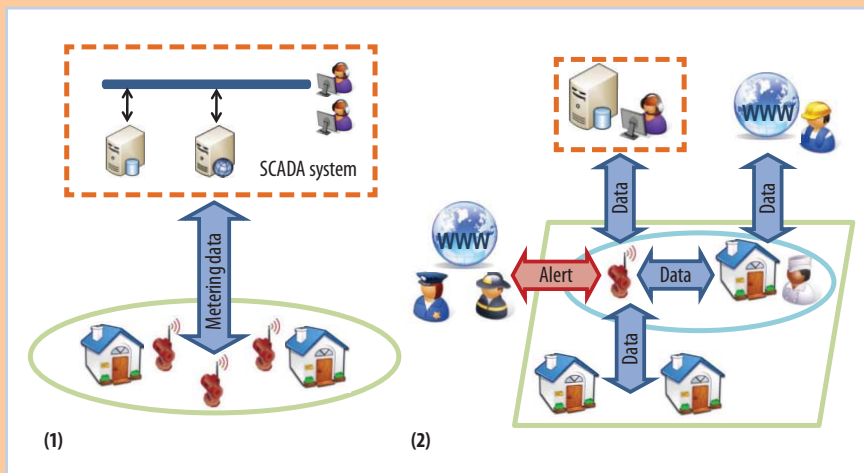



Figure A. A smart meter application in two scenarios. (1) In the intranet-of-things scenario, the meter interacts only with the supervisory control and data acquisition (SCADA) system. (2) In the IoT scenario, the meter interacts with the SCADA system, household members, other houses, and emergency personnel.

envisioned IoT and some existing applications that are arguably partial IoT instances. The path to the IoT is not a single step; rather it is the gradual incorporation of IoT applications into the real world, which involves giving objects virtual personas and thinking outside the box. For example, researchers could enhance fishing vessels with sensors and communication systems that offer shared services about the state of the sea and other facets. Objects that belong to a virtual world can be made aware of objects outside that world—including the services that other objects and entities provide. Sensors that monitor agricultural fields can access weather reports for the area and adapt the irrigation systems accordingly. Developers can also decrease system dependence on a centralized architecture, creating autonomous applications. Mobile phones without Internet connection in a disaster area can collaborate to propagate the location of a sensor-enabled water source.



**Traditional public-key infrastructures will almost certainly not scale to accommodate the IoT's amalgam of contexts and devices.**

With this staggered approach, society might be able to enjoy IoT's benefits, while analysts and researchers tackle the infrastructure complexities. The problems will be both technical and semantic, requiring interoperable mechanisms that can connect entities as well as help them understand each other. Distributed services must be reliable, not only offering availability but also adapting themselves in case of malfunction. A governance model must scale to billions of devices all over the world. Within these metachallenges are issues such as bootstrapping, mobility, scalability, data processing, standardization, and billing.

### **COPING WITH OLD AND NEW THREATS**

Not surprisingly, even a staggered approach to developing the IoT presents a daunting task for security. What protection measures are possible as billions of intelligent things cooperate with other real and virtual entities in random and unpredictable ways? The IoT's highly distributed nature and use of fragile technologies, such as limited-function embedded devices in public areas, create weak links that malicious entities can exploit. Easily accessible objects in unprotected zones, such as city streets, are vulnerable to physical harm. Like compromising botnets, some objects would try to hinder services from the inside. Additional threats include the existence of a domino effect between intertwined services and user profiling through data collection and other methods.

To avoid these threats, the IoT must have strong security foundations built on a holistic view of security for all IoT elements at all stages—from object identification to service provision, from data acquisition to infrastructure governance, all security mechanisms must consider each object's life cycle and services from the very beginning of that object's existence.

### **Protocol and network security**

Heterogeneity greatly affects the degree of infrastructure protection. Highly constrained devices that use low-bandwidth standards, such as IEEE 802.15.4, must open a secure communication channel with more powerful devices—for example, sensor nodes scattered in a smart city would communicate with smartphones or PDAs. Securing this channel requires optimal cryptography algorithms and adequate key-management systems, as well as security protocols that connect all these devices through the Internet. Although it is not clear how many resources will be available to such constrained devices once the IoT truly takes off, it makes sense to optimize security as much as possible to improve future service provision.

In a bottom-up approach, cryptography is the cornerstone for network infrastructure protection. Although standards such as AES might work for some IoT devices, others, such as passive RFID tags, might be extremely constrained. Cryptographic mechanisms must be smaller and faster but with little or no reduction in security level. Mechanisms could include symmetric algorithms, hash functions, and random number generators.

In this approach, if cryptography is the brick, the mortar is key-management infrastructures that establish keying material, for example, shared secret keys. Making this mortar requires associating previously unrelated and sometimes highly constrained objects in an extremely dynamic environment. Manual configuration works only in small and personal environments, and traditional public-key infrastructures will almost certainly not scale to accommodate the IoT's amalgam of contexts and devices. There is also the issue of rekeying devices to keep information flow safe in the long run.

Further up the network infrastructure are the communication layers. Clearly, the IoT must extensively use Internet standards for communication and service provision. Still, some devices, such as sensors that check the state of runway lights, will lack the resources to implement the Internet security mechanisms that normally protect these kinds of interactions. Therefore, security protocols require some forward-looking adaptation. Subtle differences between IoT and Internet protocols might lead to gaps in end-to-end security. Thus, adaptations should not only fulfill the IoT's performance requirements but also provide the protocol's original security properties in the context of the Internet architecture.<sup>2</sup>




## COVER FEATURE

**Data and privacy**

Privacy is one of the most sensitive subjects in any discussion of IoT protection. The data availability explosion has created Big Brother-like entities that profile and track users without their consent. The IoT's anywhere, anything, anytime nature could easily turn such practices into a dystopia. Users would have access to an unprecedented number of personalized services, all of which would generate considerable data, and the environment itself would be able to acquire information about users automatically.

Although a dystopia is the worst-case scenario, the IoT could certainly exacerbate a range of undesirable situations. Facebook accounts already affect a user's employability and personal interactions. Imagine exponentially more such exposure opportunities.

**Privacy by design.** One viable solution is privacy by design, in which users would have the tools they need to manage their own data. The solution is not too far from current reality. Whenever users produce a data fragment,



**As developers create a worldwide object network, they must build an infrastructure that allows mutual object authentication.**

they can already use dynamic consent tools that permit certain services to access as little or as much of that data as desired. Taking that idea a step further, a user in Central Park could provide a location-based service with the information that he's in New York City, but not that he's in a specific park.

**Transparency.** Transparency is also essential, since users should know which entities are managing their data and how and when those entities are using it. Stakeholders such as service providers must be part of this equation, which might make take-it-or-leave-it license agreements obsolete. Businesses will adjust their services according to the amount of personal data the user provides.

**Data management.** A huge issue is deciding who manages the secrets. Technically, cryptographic mechanisms and protocols protect data throughout the service's life cycle, but some entities might lack the resources to manage such mechanisms. In other words, one data management policy will not fit all situations. Consequently, there must be policies on how to manage various kinds of data as well as some policy-enforcement mechanism. Developing and enforcing such data management policies is not trivial. It requires interpreting, translating, and optimally reconciling a series of rules, each of which might be in a different language. And any policies must align with legislation on data protection, which itself could change.

**Identity management**

In the IoT, identity management requires considering a staggering variety of identity and relationship types, all of which must follow four object identity principles:

- An object's identity is not the same as the identity of its underlying mechanisms. The x-ray machine in the radiology department might have an IP address, but it should also have its own identity to distinguish it from other machines.
- An object can have one core identity and several temporary identities that change according to its role. A hospital is always a hospital, but it can temporarily be more significant as a conference locale or a shelter.
- An object can identify itself using its identity or its specific features. A food's virtual identity is defined by its ingredients and quantity.
- Objects know the identity of their owners. The device that controls a user's glucose level should know how that information fits in that user's overall health.

Objects can also be in groups, which some mechanism must manage. A house could have several appliances that only certain residents and visitors can use at particular times. The refrigerator could lock itself after midnight to any resident or visiting teenagers, but remain open for the adults.

Proving identity is an important part of identity management. As developers create a worldwide object network, they must build an infrastructure that allows mutual object authentication. There must be a balance between centralized management and a distributed, hierarchical approach.

Mechanisms for anonymization and the creation of pseudonyms are also important building blocks. Because the IoT deals with multiple contexts, an entity is not likely to reveal its identity all the time. In a vehicular network, for example, a police car can reveal its identity to cars and staff at the police station, but keep its identity hidden during undercover work unless it is interacting with other police cars.

As these examples illustrate, identity management in the IoT offers both challenges and the opportunity to improve security. A promising approach is to combine diverse authentication methods for humans and machines. With this combination, a user could open an office door using bioidentification (such as a fingerprint) or an object within a personal network, such as a passport, identity card, or smartphone. Combining authentication methods can prevent any loss of overall system security. Such combinations typically take the form of what I am + what I know or what I have + what I know.

Authorization is also an identity management concern. Authentication and authorization share open research issues, such as finding a balance between centralized and

distributed systems to answer the question of who's in charge of defining and publishing roles. However, specific topics, such as delegation, fall mainly under the authorization umbrella. An IoT element can delegate certain tasks to other objects for a limited time. For example, an object in the user's personal network, such as his phone, can check on his behalf to see if his suitcase contains all the needed clothes for an upcoming conference.

Granularity is another authorization issue. The services that an object provides would depend on the number of credentials presented. For example, a classroom could provide anyone who asks with the name of the course being taught, but it would release the syllabus of that course only to students with authorization certificates from the dean.

### Trust and governance

Trust is essential to implement the IoT. In this context, trust is more than the mechanisms that reduce the uncertainty of objects as they interact, although such mechanisms are important in helping objects choose

**Although governance offers stability, support for political decisions, and a fair enforcement mechanism, it can easily become excessive.**

an adequate partner for their needs. In the IoT, such mechanisms must be able to define trust in a dynamic, collaborative environment and understand what it means to provide trust throughout an interaction.

But trust also encompasses how users feel while interacting in the IoT. Feelings of helplessness and being under some unknown external control can greatly undermine the deployment of IoT-based applications and services. There must be support for controlling the state of the virtual world. Users must be able to control their own services, and they must have tools that accurately describe all their interactions so that they can form an accurate mental map of their virtual surroundings.

Governance will help strengthen trust in the IoT. A common framework for security policies will support interoperability and ensure security's continuity. Defining adequate enforcement mechanisms will go a long way toward simplifying data protection.

A governance framework can also help reduce liability. If someone can attribute a malicious action to a particular user or agent, it will be possible to punish that user or the agent's owner. But governance is a double-edged sword. On the one hand, it offers stability, support for political decisions, and a fair enforcement mechanism. On the other hand, it can easily become excessive, fostering an environment in which people are continuously monitored and

controlled. If the current Internet's partially solved governance problem is any indication, it will take the combined efforts of several research communities to address the challenges of a governance framework when countless stakeholders and billions of objects join the mix.

### Fault tolerance

Clearly, the IoT will be more susceptible to attack than the current Internet, since billions more devices will be producing and consuming services. Highly constrained devices will be the most vulnerable, and malicious entities will seek to control at least some devices either directly or indirectly. In this context, fault tolerance is indispensable to assure service reliability, but any solution must be specialized and lightweight to account for the number of constrained and easily accessible IoT devices.

Achieving fault tolerance in the IoT will require three cooperative efforts. The first is to make all objects secure by default. Aside from designing secure protocols and mechanisms, researchers must work on improving software implementation quality, since it might not be feasible to provide a software patch for billions of devices.

The second effort is to give all IoT objects the ability to know the state of the network and its services. This system would need to give feedback to many other elements; for example, a watchdog system could acquire data as part of supplying qualitative and quantitative security metrics. An important task in this second effort is to build an accountability system that will help monitor state.

Finally, objects should be able to defend themselves against network failures and attacks. All protocols should incorporate mechanisms that respond to abnormal situations and let the object gracefully degrade its service. Objects should be able to use intrusion-detection systems and other defensive mechanisms to ward off attackers.

Once an attack affects their services, IoT elements should be able to act quickly to recover from any damage. Such elements can use feedback from other mechanisms and IoT entities to map the location of unsafe zones, where an attack has caused service outages, as well as trusted zones—areas with no service outages. Such information can be the basis for implementing various recovery services, such as having objects access trusted zone services first. Mechanisms could also inform human operators of any damaged zone and then perform maintenance operations. This infrastructure self-management is a key IoT tenet.

### WORK IN PROGRESS

Researchers, governments, and industries are committed to developing and standardizing identity and security mechanisms for IoT building blocks. Table 1 lists some of the more mature efforts, excluding work-in-progress standards and government recommendations such as EU

## COVER FEATURE

Table 1. Standards for IoT technologies.

Standard	Purpose	Security	URL
ISO/IEC 14443	Architecture for contactless proximity cards	Information flow protection (AES)	<a href="http://www.iso.org/iso/identification_cards.html">www.iso.org/iso/identification_cards.html</a>
IEC 62591 (WirelessHART)	Protocol for industrial wireless sensor networks	Encryption, authentication, key management	<a href="http://www.hartcomm.org">www.hartcomm.org</a>
GS1 keys	Identification system	Unique identifier definition	<a href="http://www.gs1.org/gsmp/kc/epcglobal">www.gs1.org/gsmp/kc/epcglobal</a>
unicode	Hardware-agnostic identifier	Unique identifier definition	<a href="http://www.uidcenter.org">www.uidcenter.org</a>

Table 2. IETF standards that might be implemented in the IoT.

Standard	Purpose	URL
6LowPAN	IP connectivity	<a href="http://datatracker.ietf.org/wg/6lowpan">http://datatracker.ietf.org/wg/6lowpan</a>
ROLL	IP connectivity	<a href="http://datatracker.ietf.org/wg/roll">http://datatracker.ietf.org/wg/roll</a>
CoRE	Lightweight REST Web service architecture	<a href="http://datatracker.ietf.org/wg/core">http://datatracker.ietf.org/wg/core</a>
CoAP	Generic Web protocol definition	<a href="http://datatracker.ietf.org/wg/core">http://datatracker.ietf.org/wg/core</a>

recommendation C(2009) 3200.

Although these standards and mechanisms are good first steps, much additional work is required to build a robust and secure IoT. Again, a holistic view is vital: it is important to protect the IoT's building blocks, but its features create new requirements that are equally significant.

The design of specific security IoT mechanisms is still in its infancy, but recent developments are encouraging and could provide some degree of protection to existing IoT applications, such as the different instances of the IBM Smarter Planet initiative ([www.ibm.com/smarterplanet/us/en/?ca=v\\_smarterplanet](http://www.ibm.com/smarterplanet/us/en/?ca=v_smarterplanet)) or ventures such as Pachube (<https://pachube.com>) and Arrayent (<http://arrayent.com>).

### Cryptography and protocols

Researchers are already making strides toward developing better cryptographic algorithms and modes for IoT devices. The ISO/IEC 29192 standards aim to provide lightweight cryptography for constrained devices, including block and stream ciphers and asymmetric mechanisms. As of August 2011, these standards were still under development, but some algorithms are available. Sony's CLEFIA is a novel block cipher that supports 128-bit keys ([www.sony.net/Products/cryptography/clefi/about/index.html](http://www.sony.net/Products/cryptography/clefi/about/index.html)). The eSTREAM project ([www.ecrypt.eu.org/stream](http://www.ecrypt.eu.org/stream)) studied the robustness of stream ciphers such as Salsa20/12 and Trivium, which are extremely useful in embedded systems.

Research on lightweight dedicated hash functions has just started. The winner of the SHA-3 competition—scheduled to finish in late 2012—should lay the foundation for more work on a new class of hash functions for long-term security. The competition's goal is to develop a new cryptographic hash algorithm that converts a variable-length

message into a short message digest. The digest will be part of generating digital signatures, message authentication codes, and many other security applications in the information infrastructure (<http://csrc.nist.gov/groups/ST/hash/sha-3/index.html>).

It is already possible to construct lightweight hash functions based on lightweight block ciphers. As an alternative to these lightweight algorithms, existing optimizations in operational modes can make data processing more efficient. Both AES-CCM and AES-GCM offer data integrity and confidentiality. Another optimization is algorithm management in a cross-layer architecture, where various security mechanisms share one algorithm.<sup>3</sup>

Other communities, such as the Internet Engineering Task Force, aim to implement Internet standards in the IoT. Table 2 lists these standards and their purpose.

Although researchers have met some interim implementation goals, various constraints make it difficult to fully achieve security through Internet standards.<sup>2</sup> Developers can tweak the IPsec protocol to provide network-layer security between Internet hosts and constrained devices,<sup>4</sup> but the remaining challenges are formidable. It will be no small task to deal with the coexistence of strong link-layer security and IPsec, for example, or the negotiation of keying material. Preshared keys are usable with previously known devices, and public-key cryptography is useful when the constrained object behaves as a client,<sup>5</sup> but the negotiation of dynamic keys between previously unknown entities is still an open problem.

### Identity and ownership

In certain IoT contexts, single-sign-on (SSO) mechanisms can be useful, since users need to authenticate only once to interact with various devices. However, traditional



Web 2.0 SSO (openID and Shibboleth, for example) were not designed to fulfill certain IoT requirements,<sup>6</sup> such as giving the user control over the choice of identity provider. Other mechanisms force users to employ a particular protocol, which can be problematic in a heterogeneous environment. Another issue is the lack of support for directional identities, in which objects broadcast their identities.

These issues strongly imply the need to adapt existing SSO mechanisms or create new ones that better fit the IoT. Although some approaches address this need through a hybrid architecture that combines all mechanisms through specially crafted middleware,<sup>6</sup> this topic still needs research.

One approach to verifying device ownership and owner identity is digital shadowing,<sup>7</sup> in which a user projects his virtual identity onto logical nodes. Digital shadows are based on the notion that a user's objects act on his behalf but do not store his identity, only a virtual identity that contains information about his attributes and the objects' sessions and interactions with the architecture. Therefore, digital shadows only implicitly indicate their owner's identity.

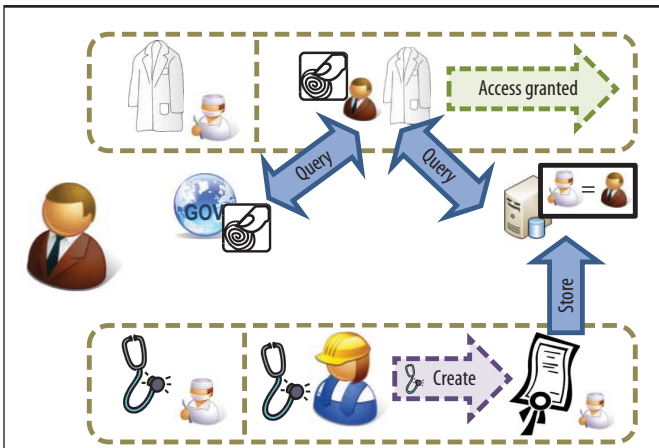
Figure 1 illustrates how digital shadowing might work with an electronic stethoscope and white coat. The doctor's fingerprints prompt a query to the government database, while his coat provides the digital shadow to the hospital database, which checks the doctor's role. Both authentication aspects (what I am + what I have) enable the doctor to enter a certain hospital area. The stethoscope records and stores the patient's heartbeats, signs the data on the doctor's behalf, and stores the data in the hospital database. The stethoscope can also check for any heartbeat anomalies by accessing other systems inside or outside the hospital.

The coat and fingerprint authentication scenario in Figure 1 might also benefit from revocable access delegation,<sup>8</sup> in which an RFID tag (the logical node) returns a valid ID (the virtual identity) only if the tag's owner has authorized the reader. These tags are essentially part of the user's digital shadow because they provide no user information (only a number), but any reader with explicit user authorization will know that they belong to that user. Because a tag's ID is not easy to link to its owner and the user can revoke authorization at any time, the digital shadow approach also accounts for privacy.

### Privacy protection

Various approaches are in development to protect the personal information of IoT users. The delegation mechanism is one privacy preservation proposal. An unauthorized RFID reader will retrieve only a random value, so it will not be able to track the user.

However, limited user access is not the only protection scenario. In some cases, users will want to provide information without revealing too much about themselves.



**Figure 1.** Instances of digital shadowing for a doctor. The doctor's white coat and electronic stethoscope store his virtual identity and act on his behalf. (Top) The coat and the doctor's fingerprints are elements of an authentication method. (Bottom) As the doctor uses the stethoscope, it not only records and stores the patient's heartbeats, but also signs the data on the doctor's behalf and stores it in the hospital database.

Some solutions in this context let the user find others who best match his preferences, without actually revealing such preferences to everyone. Other schemes let users maintain their location privacy even when making location-dependent queries.<sup>9</sup> For example, a user could try to locate someone nearby who likes Beethoven, without explicitly providing his own location and music preferences.

An interesting idea is the privacy coach,<sup>10</sup> in which an RFID reader in a mobile phone scans the tags embedded in some object, such as a loyalty card, and downloads the companion privacy policy. If the object's privacy policy does not match the user's preferences, the user can choose not to use the object. Conversely, whenever an RFID reader tries to read the mobile phone's signal, the phone can check the reader's privacy policy and ask for user consent. Finally, the privacy coach can protect the user's private physical space, such as a house, by scanning for malicious items or undesirable entities, such as objects left to monitor the house without the user's permission.<sup>11</sup>

**T**he IoT is already more than a concept. By complying with security requirements, it can fully bloom into a paradigm that will improve many aspects of daily life. Open problems remain in a number of areas, such as cryptographic mechanisms, network protocols, data and identity management, user privacy, self-management, and trusted architectures. Future research must also carefully consider the balance of governance and legal frameworks with innovation. Governance can

## COVER FEATURE

sometimes hinder innovation, but innovation in turn can inadvertently ignore human rights. The right balance will ensure stable progress toward realizing and securing the IoT as envisioned, and the benefits to humanity will be well worth the effort. **□**

### Acknowledgments

This work was partially supported by the European Union under the 7th Framework Programme for R&D (FP7) through the NESSOS (IST-256980) project and by the Spanish Ministry of Science and Innovation through the ARES (CSD2007-00004) and SPRINT (TIN2009-09237) projects. The latter is cofinanced by the European Regional Development Fund.

### References

1. B. Daskala, ed., *Flying 2.0—Enabling Automated Air Travel by Identifying and Addressing the Challenges of IoT & RFID Technology*, European Network and Information Security Agency, 2010; [www.enisa.europa.eu/media/press-releases/flying-2.0-study-of-internet-of-things-rfid-in-air-travel](http://www.enisa.europa.eu/media/press-releases/flying-2.0-study-of-internet-of-things-rfid-in-air-travel).
2. O. Garcia-Morchon et al., "Security Considerations in the IP-Based Internet of Things," IETF, Mar. 2011; <http://tools.ietf.org/html/draft-garcia-core-security>.
3. R. Roman, J. Lopez, and P. Najera, "A Cross-layer Approach for Integrating Security Mechanisms in Sensor Networks Architectures," *Wireless Comm. and Mobile Computing*, vol. 11, no. 2, 2011, pp. 267-276.
4. S. Raza, T. Voigt, and U. Roedig, "6LoWPAN Extension for IPsec," *Proc. Workshop Interconnecting Smart Objects with the Internet*, Internet Architecture Board, Mar. 2011; [www.iab.org/about/workshops/smartobjects](http://www.iab.org/about/workshops/smartobjects).
5. R. Roman et al., "Key Management Systems for Sensor Networks in the Context of the Internet of Things," *Computers & Electrical Eng.*, Mar. 2011, pp. 147-159.
6. H. Akram and M. Hoffmann, "Support for Identity Management in Ambient Environments—The Hydra Approach," *Proc. IEEE Int'l Conf. Advances in Human-Oriented and Personalized Mechanisms, Technologies, and Services (I-CENTRIC 08)*, IEEE CS Press, 2008, pp. 371-377.
7. A. Sarma and J. Girão, "Identities in the Future Internet of Things," *Wireless Personal Comm.*, Mar. 2009, pp. 353-363.
8. E. Rekleitis, P. Rizomiliotis, and S. Gritzalis, "A Holistic Approach to RFID Security and Privacy," *Proc. 1st Int'l Workshop Security of the Internet of Things (SecIoT 10)*, Network Information and Computer Security Laboratory, 2010; [www.nics.uma.es/seciot10/files/pdf/rekleitis\\_seciot10\\_paper.pdf](http://www.nics.uma.es/seciot10/files/pdf/rekleitis_seciot10_paper.pdf).
9. J. Sen, "Privacy Preservation Technologies in Internet of Things," *Proc. Int'l Conf. Emerging Trends in Mathematics, Technology, and Management*, 2011; <http://arxiv.org/ftp/arxiv/papers/1012/1012.2177.pdf>.
10. G. Broenink et al., "The Privacy Coach: Supporting Customer Privacy in the Internet of Things," *Proc. Workshop What Can the Internet of Things Do for the Citizen? (CIOT 2010)*, Radboud Univ., May 2010; <http://ldare.ubn.ru.nl/bitstream/2066/83839/1/83839.pdf>.
11. S. Radomirovic, "Towards a Model for Security and Privacy in the Internet of Things," *Proc. 1st Int'l Workshop Security of the Internet of Things (SecIoT 10)*, Network Information and Computer Security Laboratory, 2010; [www.nics.uma.es/seciot10/files/pdf/radomirovic\\_seciot10\\_paper.pdf](http://www.nics.uma.es/seciot10/files/pdf/radomirovic_seciot10_paper.pdf).

**Rodrigo Roman** is a researcher at the University of Malaga, Spain. His research interests include Internet of Things security, sensor network security, and security architectures. Roman received a PhD in computer science from the University of Malaga. He is a member of IEEE. Contact him at [roman@lcc.uma.es](mailto:roman@lcc.uma.es).

**Pablo Najera** is a doctoral candidate in computer engineering at the University of Malaga. His research interests include personal area network security, RFID security, and integration of security technologies. Najera received an MS in computer science engineering from the University of Malaga. Contact him at [najera@lcc.uma.es](mailto:najera@lcc.uma.es).

**Javier Lopez** is a full professor in the Department of Computer Science at the University of Malaga and head of the Network, Information, and Computer Security Laboratory. His research interests include security services, the protection of critical information infrastructures, and computer communications security. Lopez received a PhD in computer science from the University of Malaga. He is a member of IEEE and the ACM. Contact him at [jl@lcc.uma.es](mailto:jl@lcc.uma.es).

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

## Call for Articles

### IEEE Pervasive Computing

seeks accessible, useful papers on the latest peer-reviewed developments in pervasive, mobile, and ubiquitous computing. Topics include hardware technology, software infrastructure, real-world sensing and interaction, human-computer interaction, and systems considerations, including deployment, scalability, security, and privacy.


#### Author guidelines:

[www.computer.org/mc/pervasive/author.htm](http://www.computer.org/mc/pervasive/author.htm)

#### Further details:

[pervasive@computer.org](mailto:pervasive@computer.org)  
[www.computer.org/pervasive](http://www.computer.org/pervasive)






## Focus on Your Job Search


**IEEE Computer Society Jobs** helps you easily find a new job in IT, software development, computer engineering, research, programming, architecture, cloud computing, consulting, databases, and many other computer-related areas.

**New feature:** Find jobs recommending or requiring the IEEE CS CSDA or CSDP certifications!

Visit [www.computer.org/jobs](http://www.computer.org/jobs) to search technical job openings, plus internships, from employers worldwide.

<http://www.computer.org/jobs>

IEEE  computer society | **JOBS**



The IEEE Computer Society is a partner in the AIP Career Network, a collection of online job sites for scientists, engineers, and computing professionals. Other partners include Physics Today, the American Association of Physicists in Medicine (AAPM), American Association of Physics Teachers (AAPT), American Physical Society (APS), AVS Science and Technology, and the Society of Physics Students (SPS) and Sigma Pi Sigma.



## COVER FEATURE



# Sticky Policies: An Approach for Managing Privacy across Multiple Parties

Siani Pearson and Marco Casassa Mont, *HP Labs Bristol*

**Machine-readable policies can stick to data to define allowed usage and obligations as it travels across multiple parties, enabling users to improve control over their personal information. The EnCoRe project has developed such a technical solution for privacy management that is suitable for use in a broad range of domains.**

**C**urrent mechanisms for ensuring privacy protection across organizational boundaries rely on legal and business frameworks, including contracts and service-level agreements. Technical mechanisms complement such approaches by supporting enforcement and auditing of the organizational obligations they outline.

*Personally identifiable information* (PII), also referred to as personal data or personal information, is data that can be traced to a particular individual—for example, a name, address, phone number, Social Security number, national identity number, credit card number, e-mail address, password, or date of birth. Because of its sensitive nature, greater care must be taken in the handling of the subset of PII that includes financial or medical data.

In commercial contexts, meeting customers' expectations regarding privacy requires the protection and careful use of PII. For corporations, privacy includes the application of laws, policies, standards, and processes for managing an individual's PII. Privacy management identi-

fies the ways in which organizations and individuals can control the collection, usage, and sharing of personal data, including sensitive information.

Privacy management and compliance with regulatory requirements for data protection can help organizations foster trust with their customers. In a given context, there may be many different privacy-related regulatory requirements, including sector-specific laws, national legislative requirements, and transborder dataflow restrictions.<sup>1</sup> Although assessing requirements in a given situation can be complex, an established set of principles forms the basis of most privacy legislation worldwide.<sup>2</sup>

To provide mechanisms for online privacy management, substantial research has been conducted related to

- anonymization technologies;
- enforcement of privacy-enhanced access-control policies, as in the Prime and PrimeLife EU projects ([www.primelife.eu](http://www.primelife.eu));
- policy life-cycle management;
- satisfying global regulations relating to data protection, including tools for governance, risk, and compliance (GRC);
- modeling privacy regulations; and
- modeling organizational privacy policies down to the operational level.<sup>3</sup>

However, major issues remain outstanding, including how to provide more control to end users, how to gather and manage end users' consent (and subsequent revocations),

and how to make privacy management effective when information is transmitted across parties.

An approach based on *sticky policies*—conditions and constraints attached to data that describe how it should be treated—enables compliance with and enforcement of current requirements such as the US Health Insurance Portability and Accountability Act (HIPAA) of 1996 along with future needs emerging from the adoption of new technologies and models, including the storage and processing of sensitive data in the cloud.

### INFORMATION FLOW

In some scenarios, a user's confidential information flows across organizational boundaries. For example, a healthcare system could disclose personal data and preferences to a general practitioner via an online service provider (SP); the system also might need to share this information with hospital specialists, pharmaceutical companies, and other third parties involved in the healthcare supply chain. A similar situation might apply for a travel agency that needs to share data with various SPs such as hotel reservation brokers and car rental agencies.

More generally, these kinds of scenarios will be increasingly common in a cloud computing environment, where users interact with front-end SPs that will need to share part of the information with other SPs to supply the required services.

In all these situations, users must reveal personal and even sensitive information to receive a service, but they want to control how that information is used. They can directly control how their data should be processed, handled, and shared by explicitly expressing their preferences and data-handling policies. These choices must be respected all along the service provision chain, including allowing the user to update them. Achieving this objective requires propagating the user choices to all the SPs and deploying several mechanisms to ensure that the policies are respected. Moreover, the user can be actively involved in the selection of multiple, interchangeable services that will track and audit policy fulfillment.


### CHARACTERISTICS OF STICKY POLICIES

Depending on the degree of a policy's stickiness, the data might be encrypted, with access to the content allowed only upon the satisfaction of these policies. Specifically, the policies govern the use of associated data, and could specify the following:

- proposed use of the data—for example, for research, transaction processing, and so on;
- use of the data only within a given set of platforms with certain security characteristics, a given network, or a subset of the enterprise;

- specific obligations and prohibitions such as allowed third parties, people, or processes;
- blacklists; notification of disclosure; and deletion or minimization of data after a certain time; and
- a list of trusted authorities (TAs) that will provide assurance and accountability in the process of granting access to the protected data, potentially the result of a negotiation process.

Figure 1 illustrates the mechanisms for handling sticky policies. Our approach uses cryptographic mechanisms to strongly associate policies with the data. There can be different degrees of stickiness, but we adopt a strong binding as it provides better accountability. The data is encrypted and only accessible upon the acceptance and satisfaction of constraints and duties the policies impose.



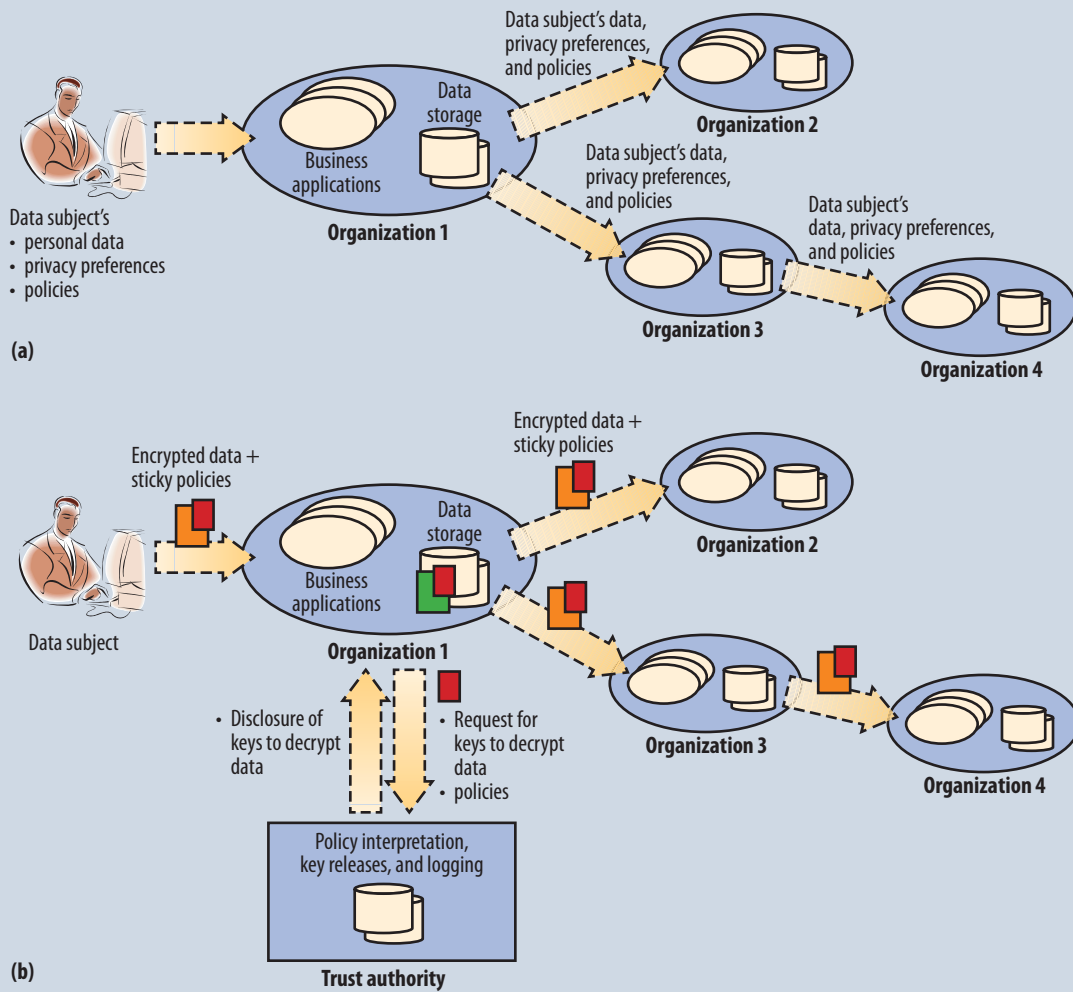
**Users can directly control how their data should be processed, handled, and shared by explicitly expressing their preferences and data handling policies.**

TAs provide assurance by keeping track of promises the involved parties make to access data, along with controlling access to such data. The TAs' role may be integrated with other functionality, such as being a consumer organization, a certification authority (CA), or a well-known organization. The TA's role also can be performed by a client-side software component or service that is under the control of end users or other parties, or it can be achieved using distributed components or a peer-to-peer mechanism.

The deployment of such a system is reasonably straightforward, as it does not require change from existing trusted third parties except for dealing with additional policy condition checks or from storage providers if they are used to store the data and an authenticated reference is passed around instead of the data. However, SPs would either need to manage packaged sticky policies or use an application to do this locally. This includes additional interactions with the TA and release of statements certifying their willingness to fulfill the policies. Hence, this technique is likely to be most suitable for service provision environments in which the increased trust and protection would justify the additional expense. Alternatively, business partners of goodwill enterprises that are trying to employ best practices might encourage its use.

Sticky policies are passed between organizations to capture obligations and other constraints that the receive-

COVER FEATURE



**Figure 1. High-level scenario and related management of sticky policies. (a) High-level scenario involving data disclosures across organizations. (b) Overview of sticky policy approach.**

ing parties must meet to access and use the associated personal data. For example, if the system passes a health-care record from a hospital to a research institution and then to a research team, the information might be in a form in which certain attributes such as medical results or personal information such as name and address are encrypted, with an associated sticky policy describing how parts of this could be used. For example, a patient wants this information to be released only to research teams, requests that it be deleted after three years, and asks to be notified every time the medical information is passed on. These constraints can be expressed in several ways, including using a simple XML format.

Sticky policies can help enable accountable management and disclosure of confidential data across boundaries. In the approach shown in Figure 1, personal, private, or confidential information is associated with machine-readable policies in a way that can't be compromised. The system

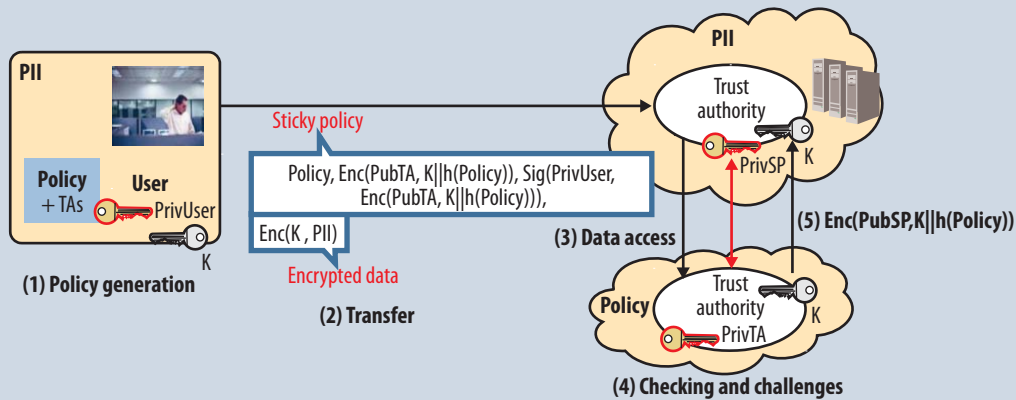
processes the information in a way that adheres to these constraints. As it replicates the data or fulfills the service provision request, mechanisms will be in place to ensure that the customer's preferences are respected all along the chain. Specifically, TAs need to retrieve keys to decrypt data and log all promises made by the requestors. This information can be used for forensic analysis if there are policy violations.

Figure 2 shows the basic mechanisms underpinning the management of sticky policies, which can be achieved using various cryptographic techniques, including public-key infrastructure-based and other approaches.

Our solution includes the following aspects:

- To more easily interpret and enforce policies, organizations impose a framework that defines their preferences and policies. In one approach to achieving





**Figure 2.** Core mechanisms underpinning the management of sticky policies. (1) Creation of sticky policies at the user side. (2) Sending sticky policies and data to the service provider. (3) Sending sticky policies to the agreed trust authority to get access to data. (4) Service provider interacts with trust authority to satisfy sticky policy constraints. (5) Getting cryptographic keys for use in accessing the data.

this, SPs publish a “manifesto” containing the list of supported (macro) policies and TAs and defining how these policies relate to access control and obligation behaviors the organization supports.

- A user (customer) can interact with an SP to select the granularity (ranging from coarse-grained to fine-grained) of applying policies to items or specific subsets of personal data to be disclosed and customize related preferences such as notification, period of time after deletion, set of agreed purposes, and the list of parties not to interact with.
- The user selects a subset of TAs that are to be trusted.
- Based on these selections, a client-side component supports the creation of sticky policies and their association to data—the bundling of policies, preferences, data, and TAs. In other words, the client-side component manages the packaging of data along with selected parameterized policies and TAs.
- Rather than passing the encrypted data directly to the SP, the user can select the option to refer to PII secured by a third party—a storage provider that stores the encrypted data.
- The system sends the encrypted data along with sticky policies to the SP.
- To gain access to the data, the SP needs to interact with one of the selected TAs (based on availability). During this interaction, the SP must assert its willingness to fulfill the customized sticky policies. Alternatively, depending on the policy requirements, the TA might be able to check this independently of such signed statements—for example, with reference to externally maintained blacklists or reputation management systems, or by verifying system properties using mechanisms such as trusted attestation or remote software verification. This creates an audit

trail available to the user and TA afterward in case of policy violations or misbehavior.

- The SP allows a predefined period of time for connection with the TA. The solution supports swapping data between TAs based on needs.
- Only after satisfying all these requirements and checking additional contextual information will the TA decide to release the keys for decrypting data.
- The TA will be able to decrypt and access the data regardless of whether it was directly disclosed or if only a reference to it was provided. In the latter case, the SP would need to fetch the data.

We envision the deployment within organizations of privacy-management components that complement identity and access management solutions, as tested in the context of the EnCoRe collaborative project ([www.encore-project.info](http://www.encore-project.info)). Specifically, these components will complement organizations’ middleware solutions, in the space of identity and access management, to provide privacy-aware access control, obligation management, data tracking, processing of sticky policies, and interactions with TAs. The role of these TAs is not just to release keys but also to provide accountability by means of logging and auditing, and subsequently supporting forensic analysis.

## CREATING STICKY POLICIES

The original sticky policy paradigm specified that privacy preferences should flow with personal data to make sure that they can always be enforced.<sup>4</sup> Subsequent research suggested a method for creating strong stickiness of policies to data.<sup>5</sup>

In a common central approach, customers allow SPs to have access to specific data based on agreed policies in interactions with interchangeable independent third par-

## COVER FEATURE

ties (the TAs). The access to data can be as fine-grained as necessary, based on policy definitions, underlying encryption mechanisms (supporting the stickiness of policies to the data), and a related key-management approach that specifically encrypts data based on the policy. A TA mediates access to data, checking for compliance to policies to release decryption keys, so that checking for compliance requires more than having the SP assert its willingness to do so. This provides users with fine-grained control over access and usage of their data, even in public cloud models.

Various techniques using different underlying encryption mechanisms can provide sticky-policy protection of data. In each case, the system can extend the selected technique to cover the propagation of data along the service provision chain. The process is analogous to user-to-SP protocols, in which the first SP can add policy constraints to form a superset of previous policy constraints. Multiple mechanisms currently used to exchange information can refine and deploy the proposed techniques, including Web technologies and protocols such as http/s, SOAP, and so on; document formatting and protection techniques such as Adobe and DRM; and various messaging tools including e-mail and instant messaging.

**The access to data can be as fine-grained as necessary, based on policy definitions, underlying encryption mechanisms, and a related key-management approach that specifically encrypts data based on the policy.**

These protocols apply not only to human users but also more broadly to machine-to-machine or service-to-SP interactions.

### Using public-key encryption techniques

When using public-key encryption, we assume that all the stakeholders have certified public or private key pairs from trusted CAs. An approach that enhances integrity binds policies to data by encrypting the data under a symmetric key that a sender and receiver conditionally share based on fulfillment of policies, and sticking the data to the policy using public-key enveloping techniques similar to the Public-Key Cryptography Standard (PKCS) 7. Figure 2 shows an example of this process, in which the labeled stages are as follows:

1. The sender generates the policy, together with a symmetric key  $K$  used to encrypt the data (for efficiency, a symmetric key is used rather than an asymmetric key). If desired, this process can be generalized to allow encrypting different attributes separately—that is, using

different symmetric keys generated at this stage—revealing only part of the information when an attribute is decrypted.

2. The sender generates a message to the SP. One part of the message is the data encrypted with  $K$ . The other part is a sticky policy, in which  $K$ , appended to the policy's hash, is encrypted with the TA's public key and then is signed using the user's private key. This makes it possible to verify the policy's source and integrity and binds  $K$  to the data and the policy. The system sends the resultant sticky policy together with the encrypted data to the SP.
3. The SP generates a message to the TA, which involves passing on just the sticky policy and encrypted shared keys.
4. The TA checks policies, potentially including challenges to the SP. The SP might need to provide signed statements about its policies.
5. If all checks are fulfilled, the TA releases the shared key. This generates a message from the TA to the SP, which involves encrypting  $K$  appended to the policy's hash with the SP's public key. The SP can get access to  $K$  to check the policy's integrity and then decrypt the PII.<sup>6</sup>

### Using identifier-based encryption

An identifier-based encryption (IBE) cryptographic schema can use any kind of string as a public encryption key, including a name, role, terms, or conditions.<sup>7</sup> The generation of the corresponding IBE decryption key can be postponed. A TA can generate this decryption key on the fly, under specific circumstances.

While it is conceptually similar to the PKI approach, we adapt IBE by mapping a sticky policy to an IBE encryption key. The TA's role is expanded to check the integrity and trustworthiness of the requestor's credentials and its IT environment before releasing the decryption key. It also logs and audits disclosures of confidential data.<sup>8</sup>

### VARIATIONS ON ENCRYPTION

We can potentially use any encryption mechanism to associate policies with data. For example, Voltage and Navajos provide format-preserving encryption and search-enabled encryption, respectively. If the operation involves indexing, it would still be possible to search and index encrypted attributes.

An alternative solution permits binding of privacy preferences to data and conveying the individual's consent as well.<sup>9</sup> However, this solution does not avoid the unauthorized use of data.

This approach can be adapted to support multiple verification and control functions. Instead of having individual certificates, each entity could be provided with a key component, called a "share." An option such as Shamir's threshold-based secret-sharing scheme<sup>10</sup> could be used

to require  $l$  of  $m$  shares for the cloud service provider to recover  $K$  and decrypt the PII, while still providing some redundancy among TAs. Secret-sharing schemes form a particular group of multiparty key establishment protocols that enable distribution of control or trust in critical activities. The central idea of such a  $(l, m)$  threshold scheme is that a key (in our case, the key used to encrypt the data) would be divided into  $m$  pieces (the shares), such that any  $l$  of them can be used to reconstruct the whole original key but using any number of shares less than  $l$  will not help to reconstruct the key.

Trusted computing group integrity-checking mechanisms can verify that the receiver's platform is trusted, its software state is conformant with the disclosure policies, and it correctly implements defined privacy-management mechanisms.

Furthermore, there are several variations on these approaches in terms of policy definition, the degrees of stickiness, and the fine-grained nature of the encryption that occurs. The mechanisms are independent of the particular representation used for the policies.

In addition, the protocols themselves can be amended. In the PKI approach, for example, the user can bind the policy to the data within a signing operation rather than within the encryption. Other options include using the signcryption algorithm specified in ISO/IEC 29150.2 ([www.iso.org/iso/iso\\_catalogue/catalogue\\_ics/catalogue\\_detail\\_ics.htm?ics1=35&ics2=040&ics3=&csnumber=45173](http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?ics1=35&ics2=040&ics3=&csnumber=45173)), performing a single operation and separately encrypting the data (or reference to the data). An alternative is to encrypt attributes with different keys, enveloping the sensitive data and passing it on without revealing the key from the TA while revealing different attributes to different entities in the chain.

### CASE STUDY: ENCORE PROJECT

The processes and components designed for privacy management within the EnCoRe project demonstrate the feasibility of sticky policies. EnCoRe is a collaborative research effort undertaken by UK academic and industrial partners that uses consent and revocation management to give individuals more control over their personal information. In this context, revocation essentially means change of consent, potentially in a fine-grained way.


The project provides mechanisms for users to define consent policies and to change them. EnCoRe uses sticky policies to represent and enforce the consent and revocation preferences of end users. In general, EnCoRe supports the following:

- *Explicit management of consent and revocation.* Negotiating, setting, changing, and enforcing sticky policies are integrated with the management of security and

privacy policies. Compliance checking and auditing are integrated capabilities.

- *Bridging the disconnect between high-level and lower-level policies.* This includes mapping legal, business, social, and security requirements to high-level policies. We define an intermediate conceptual framework to model policies and reason on top of them. We then map these concepts into monitorable and enforceable policies driven by users' preferences.

Our solution is applicable in a variety of business contexts, and it is especially valuable where sensitive information is involved—for example, in healthcare scenarios such as biobanks and assisted-living facilities, providing third-party access to employee data, government scenarios, and cloud computing.



**The EnCoRe project provides mechanisms for users to define consent policies and to change them, as well as for enforcement of these policies.**

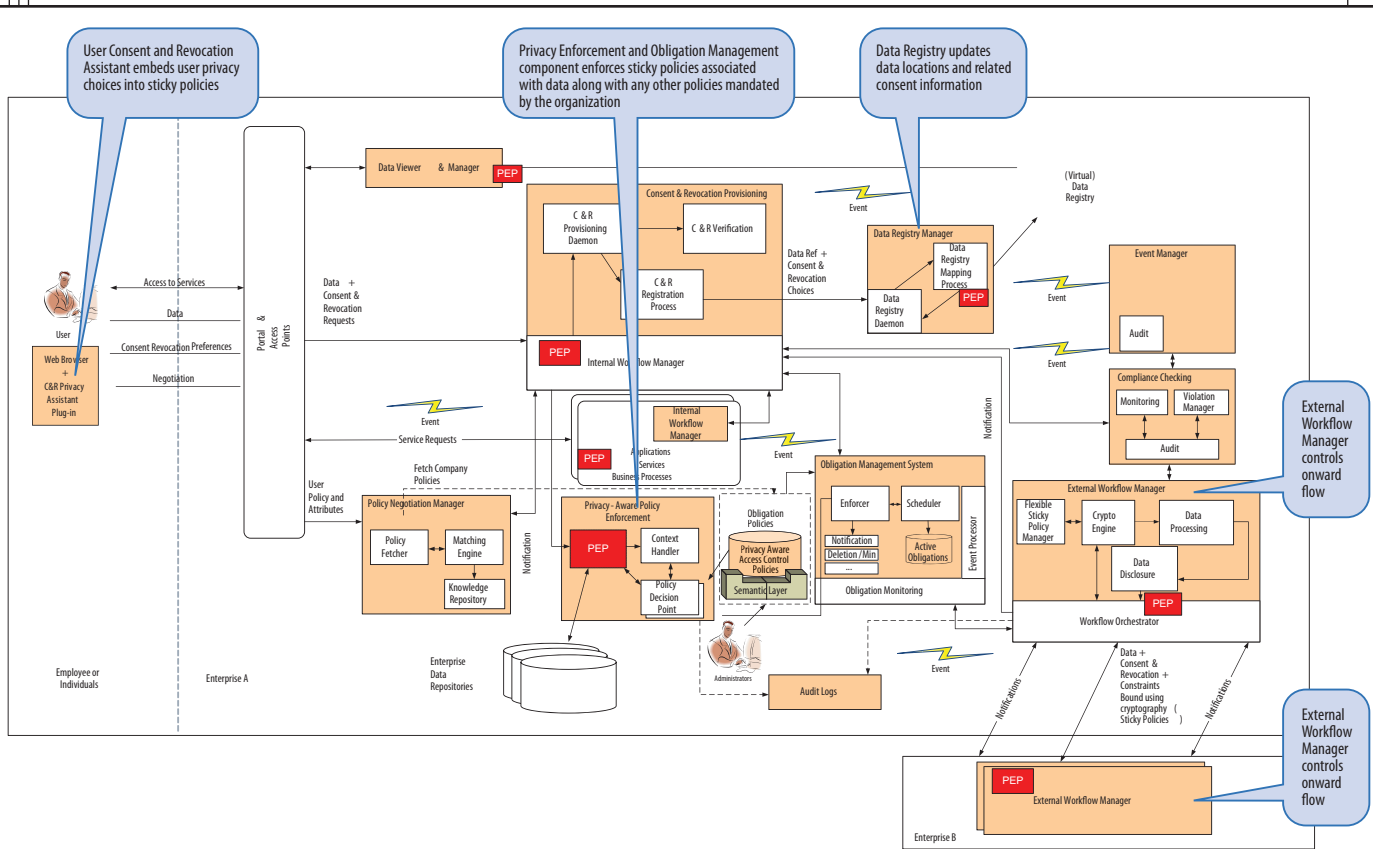
In the EnCoRe project, we have developed a flexible toolbox solution that can be customized and deployed consistently within the business processes of each involved SP. EnCoRe-compliant capabilities provide assurance about a given SP's privacy management practices and related management of consent and revocation.

Figure 3 illustrates the overall set of functionalities and capabilities that EnCoRe provides. The system can provide these components as a set of services in the cloud or it can deploy them as an overall stand-alone infrastructural solution. The components include the following:

- *Personal consent and revocation assistant.* This component provides user-side capabilities to help people express their consent by making privacy choices such as opt-in/opt-out, identifying preferences, and so on, and submitting revocation requests, along with the explanation of privacy practices that organizations provide. A Web browser plug-in can trigger this function during data-disclosure processes. The system can embed these privacy choices into sticky policies to ensure that third parties receiving the data will fulfill them.
- *Virtual data registry.* This repository—or an aggregation of synchronized repositories—keeps track of where each known individual's data has been stored within and outside the organization and identifies which type of data has been disclosed and to whom, along with any relevant associated sticky policies.



COVER FEATURE



**Figure 3. EnCoRe architecture.** EnCoRe components process personal data and enforce preferences. Users disclose their personal data with privacy preferences; the EnCoRe privacy-aware access control and obligation components enforce the preferences when third parties access the data; the data registry tracks the data's location; the external workflow manager creates and attaches sticky policies to data before the system discloses it to third parties. The system applies this approach recursively across chains of organizations.

- *Consent and revocation provisioning.* This component automatically updates the data registry every time there is a new expression of consent and revocation. It uses internal workflows to update an individual's preferences and identify constraints that affect the enforcement of access control and obligation policies.
- *Privacy-aware policy enforcement and obligation management.* Driven by consent, this component deals with access control over data and obligations. It enforces sticky policies associated with the data along with any other policies the organization mandates.
- *External workflow manager.* This component intercepts and tracks the flow of personal data, both within and between organizations, and propagates the associated consent information. Sticky policies ensure degrees of compliance with agreed policies and data subject's expressed preferences. Applications and services might need to be instrumented with agents that communicate with this component.
- *Auditing.* This component logs and tracks what happens to data, consent, and revocation during operational and administrative activities, includ-

- ing the flow of personal data within and beyond the organization.
- *Compliance checking and risk assurance.* The enterprise's privacy administrators use this key offline component to assess current risks and provide indications of compliance.

The sticky policies that the EnCoRe system sends to other organizations specify the purposes of using the data and any obligations and prohibitions, including notification and deletion after a certain time, that the user has specified in the consent and revocation preferences associated with that data. The TA is distributed in the sense that the EnCoRe external workflow manager component controls sharing of the information associated with the sticky policies, and the data registry records how it has been distributed. Optionally, an external TA can also perform some additional checks if the external workflow manager cannot perform them directly.

If the receiving party is EnCoRe-enabled, the system translates the high-level requirements expressed in the sticky policies into fine-grained access and obligation poli-

cies to be enforced along with the original privacy choices. To achieve this, mapping capabilities systematically translate high-level constraints (defined in the policy manifesto) into enforceable ones. If the receiving parties do not have EnCoRe-compliant systems, the external workflow manager assesses the extent to which the data can be released for a given purpose, sanitizing it before release if needed. EnCoRe administrators predefine the criteria for sanitizing data—for example, omitting some details or providing statistical information. The criteria for releasing data include evaluating the purpose for which the data was required and the outcome of risk assessment carried out on the receiving parties—for example, their ability to deliver the required privacy controls on specific data items.

To revoke consent, users edit their consent preferences through Web-based UIs. EnCoRe batches and automatically propagates these preferences throughout the system as well as beyond it to the other organizations involved, leveraging the information stored in the data registry. Organizations can apply this approach recursively to disclose information to one another.

## FUTURE DIRECTIONS

We have developed the core mechanisms for managing sticky policies within the EnCoRe project along with a PKI-based implementation of the required mechanisms. Next steps are to deploy them in a case study with a customer and provide advanced implementations of the protocols, including multiple verification and control capabilities.

In the longer term, we envision using the EnCoRe system to add information to the sticky policy regarding technical and process control mechanisms or boundaries that the receiving entity should have in place for it to be considered trustworthy or that it is EnCoRe-compliant. We are also researching better ways of propagating consent and revocation changes along the chain within which data is shared, including the external workflow managers of the other entities that periodically check, update, and trigger enforcement of relevant user preference options stored elsewhere.

Open issues that we are currently researching include stronger enforcement and trying to prevent SPs from cheating by breaking promises to TAs. A logical binding can easily be unbound, but even with a cryptographic binding, after the personal data has been decrypted, the binding is broken in the sense that the users' data is then fully available to the authorized party and subsequent actions could be taken that contravene the policy. The solution needs to protect data after it has been decrypted,<sup>11</sup> but current options result in stronger protection at the cost of poor scalability or unrealistic expectations regarding the hardware or operating system environment the SPs use.

Trusted computing also might be used to ensure that receivers act according to associated policies and con-

straints. However, the digital signature only proves the authenticity of a binding the data subject established in the past. If encryption is applied only to text files that adhere to a predefined structure, it can be relatively easy to corrupt policies; thus, a skilled hacker could tamper with the file and make the policy illegible. Watermarking schemes<sup>12</sup> and obfuscation techniques<sup>13</sup> also can provide content protection, but they do not ensure policy enforcement or offer protection for the data after access.

**S**ticky policies offer a promising approach for privacy management within and across organizational boundaries that can be leveraged in various contexts, including in the cloud. The user defines sticky policies when disclosing data to an organization. These policies dictate the preference conditions and ensure that appropriate constraints will be audited and degrees of assurance provided.

Using sticky policies allows tracing and auditing via TAs and enforcement of user preferences by SPs. In addition to advancing the state of the art by providing an end-to-end data management solution, the approach is scalable, provides different options to drive the interaction process between the SPs and TAs, and allows optional involvement of storage service providers.

Privacy advisors or client applications will mediate user interactions to mitigate the complexity of creating sticky policies and binding them to data. This solution could be used in several business areas, but would be particularly appropriate where sector-specific legislation or user concerns are strongest—for example, in domains relating to healthcare, finance, or defense.

We are working to extend and broaden this approach to achieve accountability by using contractual assurances along the service provision chain from SPs to organizations, enhanced on the technical side by enforcement of corresponding machine-readable policies propagated with data, integrated risk assessment, assurance, and auditing.<sup>14</sup> Thus, organizations can ensure that all who process data observe their obligations to protect it, regardless of where that processing occurs. **□**

## Acknowledgments

Several parties provided helpful input and feedback about this research, most notably Liqun Chen, Gina Kounga, and Archie Reed.

## References

1. S. Pearson, T. Sander, and R. Sharma, "Privacy Management for Global Organisations," *Data Privacy Management and Autonomous Spontaneous Security*, LNCS 5939, Springer, 2009, pp. 9-17.

## COVER FEATURE

2. Organization for Economic Cooperation and Development, "OECD Guidelines on the Protection of Privacy and Trans-border Flows of Personal Data," 1980; [www.oecd.org/document/18/0,3343,en\\_2649\\_34255\\_1815186\\_1\\_1\\_1\\_100.html](http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_100.html).
3. S. Pearson and D. Allison, "Privacy Compliance Checking Using a Model-Based Approach," *E-Business Applications for Product Development and Competitive Growth: Emerging Technologies*, IGI Global, 2011, pp. 199-220.
4. G. Karjoth, M. Schunter, and M. Waidner, "Platform for Enterprise Privacy Practices: Privacy-Enabled Management of Customer Data," *Proc. 2nd Workshop Privacy Enhancing Technologies (PET 02)*, LNCS 2482, Springer, 2002, pp. 69-84.
5. M. Casassa Mont, S. Pearson, and P. Bramhall, "Towards Accountable Management of Identity and Privacy: Sticky Policies and Enforceable Tracing Services", 2003; [www.hpl.hp.com/techreports/2003/HPL-2003-49.pdf](http://www.hpl.hp.com/techreports/2003/HPL-2003-49.pdf).
6. S. Pearson, M. Casassa Mont, and G. Kouna, "Enhancing Accountability in the Cloud via Sticky Policies," *Secure and Trust Computing, Data Management and Applications*, vol. 187, Springer, 2011, pp. 146-155.
7. D. Boneh and M.K. Franklin, "Identity-Based Encryption from the Weil Pairing," *SIAM J. Computing*, vol. 32, no. 3, 2003, pp. 586-615.
8. M. Casassa Mont, S. Pearson, and P. Bramhall, "Towards User Control and Accountable Management of Privacy and Identity Information," *Proc. 8th European Symp. Research in Computer Security (ESORICS 03)*, LNCS 2808, Springer, 2003, pp. 146-161.
9. H.C. Pöhls, "Verifiable and Revocable Expression of Consent to Processing of Aggregated Personal Data," *Proc. 10th Int'l Conf. Information and Communications Security (ICICS 08)*, LNCS 5308, Springer, 2008, pp. 279-293.
10. A. Shamir, "How to Share a Secret," *Comm. ACM*, vol. 22, no. 11, 1979, pp. 612-613.
11. Y. Zuo and T. Keefe, "Post-Release Information Privacy Protection: A Framework and Next-Generation Privacy-Enhanced Operating System," *Information Systems Frontiers*, vol. 9, no 5, pp. 451-467.
12. L. Perez-Freire et al., "Watermarking Security: A Survey," *Trans. Data Hiding and Multimedia Security*, LNCS 4300, Springer, 2006, pp. 41-72.
13. R. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymisation," *Proc. Int'l Conf. Data Engineering (ICDE 05)*, IEEE CS Press, 2005, pp. 217-228.
14. S. Pearson, "Toward Accountability in the Cloud," *IEEE Internet Computing*, July/Aug. 2011, pp. 64-69.

**Siani Pearson** is a senior researcher in the Cloud and Security Research Lab at HP Labs Bristol. Her research focuses on privacy-enhancing technologies, accountability, and the cloud. She received a PhD in artificial intelligence from the University of Edinburgh, UK. Pearson is a fellow of the British Computer Society, a senior member of IEEE, and a Certified Information Privacy/Information Technology Professional. Contact her at [siani.pearson@hp.com](mailto:siani.pearson@hp.com).

**Marco Casassa Mont** is a senior research scientist in the Cloud and Security Research Lab at HP Labs Bristol. His research interests include strategic aspects of risk management, security and privacy, and technologies applied to business contexts and emerging scenarios, including the cloud. Casassa Mont received an MSc in computer science from the University of Turin, Italy. He is a senior member of IEEE and a member of the UK Institute of Information Security Professionals. Contact him at [marco.casassa-mont@hp.com](mailto:marco.casassa-mont@hp.com).



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

# computing | now

ACCESS | DISCOVER | ENGAGE

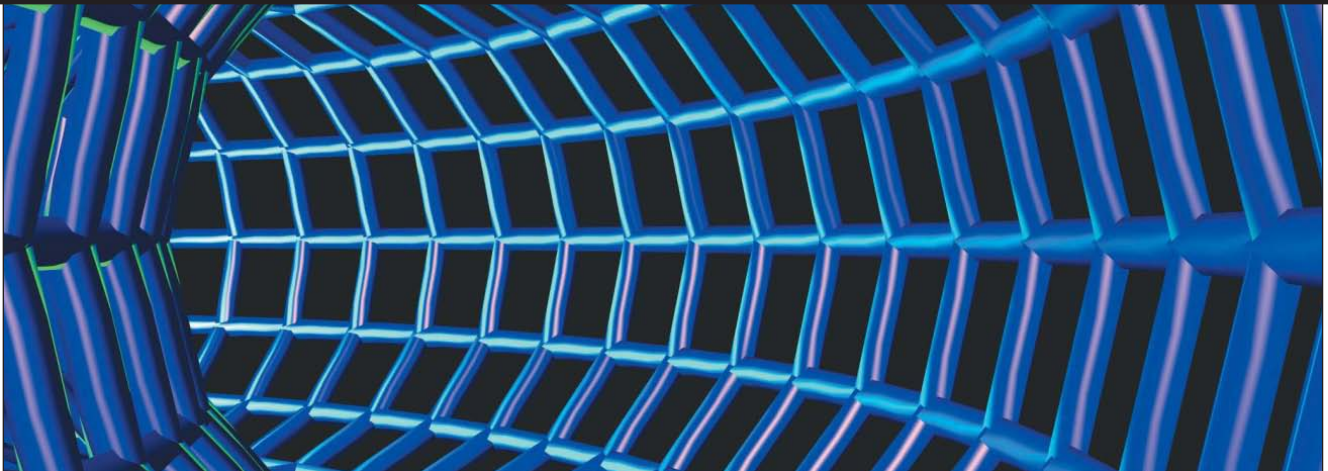
Let us bring technology news to you.



<http://computingnow.computer.org>  
Subscribe to our daily newsfeed



## PERSPECTIVES



# Trends in Server Energy Proportionality

Frederick Ryckbosch, Stijn Polfliet, and Lieven Eeckhout, *Ghent University, Belgium*

**Server energy proportionality, as quantified by the proposed EP metric, has improved significantly, from 30-40 percent in 2007 to 50-80 percent today, but much more can be done to move systems closer to ideal.**

**E**nergy efficiency has emerged as a major design driver in servers, as it impacts capital costs as well as powering and cooling expenses.

Luiz Barroso and Urs Hölzle<sup>1</sup> demonstrated that server capital costs dominate the total cost of ownership in a classic datacenter, with 69 percent of the monthly TCO related to server purchase and maintenance costs. In contrast, the cost of all infrastructure and power to host a contemporary datacenter with commodity-based lower-cost servers or higher power prices is more than twice the purchase and maintenance costs. The researchers concluded that, with electricity and construction expenses rising, datacenter facility costs—which are proportional to power consumption—will become an increasingly larger part of the TCO. In other words, a datacenter's TCO primarily will be a function of its power consumption, and the purchase and maintenance costs will matter less. Besides cost considerations, improving energy efficiency reduces carbon dioxide emissions, leading to greener IT.

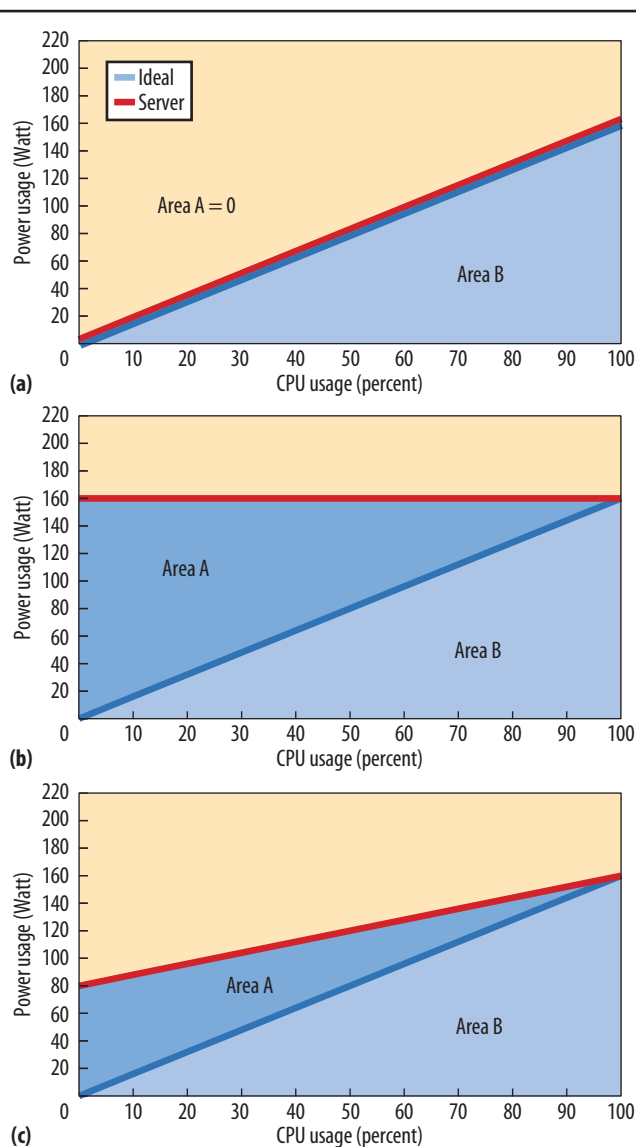
Improving a server's energy efficiency is nontrivial and presents many challenges. Peak power consumption affects capital costs for power distribution, the supply units, and the cooling infrastructure. High power consumption leads to increased power density and temperature, which affects

cooling costs and possibly hardware reliability and availability. Total energy consumption impacts the operating expense for powering the servers. Further, servers in large datacenters rarely are completely idle and seldom at or near maximum utilization—they typically operate in the 10-50 percent utilization range.<sup>2</sup> This insight motivated Barroso and Hölzle<sup>3</sup> to make the case for servers that consume energy proportional to their utilization level or load. The key is that servers should be optimized both to reduce peak power consumption and to maximize performance at lower utilization levels.

In December 2007, the Standard Performance Evaluation Corporation released SPECpower\_ssj2008, an industry-standard benchmark for server power and performance.<sup>4</sup> SPECpower measures performance and power at different utilization levels, from which it computes an overall metric. The metric thus includes some notion of energy proportionality but does not explicitly quantify it.

We propose a metric to quantify a server's energy proportionality. Using the EP metric on published power and performance data, we evaluated how energy proportionality has evolved over time, examined how the EP metric relates to SPECpower, and quantified how further improving servers' EP can reduce total energy consumption.

## PERSPECTIVES



**Figure 1.** Energy proportionality (EP) is computed as 1 minus area A divided by area B; area A is defined as the area between the server and ideal curves, and area B is defined as the area below the ideal curve. (a) Perfect energy-proportional system (EP = 1). (b) Non-energy-proportional system (EP = 0). (c) 50 percent energy-proportional system (EP = 0.5).

## EP METRIC

An analogy can easily explain the EP concept. A car's engine consumes fuel when the car is being driven, and even more fuel when the car accelerates. However, the engine also consumes fuel when it is running but idle—for example, when the car is stopped at a traffic light or stop sign. In these situations, the energy consumption is not proportional: the engine consumes energy (fuel) even though the car does not make physical progress. Similarly, a computer system consumes energy when it is idle—the system is powered on but does not do any useful work.

Ideally, an energy-proportional system would consume zero power when completely idle, and it would consume power proportional to its utilization level when doing useful work—for example, when operating at 30 percent of its peak performance, the system should consume 30 percent of its peak power. In practice, however, power consumption is higher than what the ideal scenario would suggest.

Informed by the seminal work on energy proportionality by Barroso and Hölzle,<sup>3</sup> we define a server's EP as 1 minus the integral of the relative delta in power consumption for the ideal energy-proportional server, across a range of utilization levels. As Figure 1 shows, EP is computed as 1 minus the area between the server's power consumption and the ideal power consumption curve (area A) divided by the area under the ideal curve (area B). An EP of 1 means that the server consumes power proportional to its load (Figure 1a). An EP of 0 indicates that the server consumes a constant amount of power irrespective of its load (Figure 1b). Figure 1c represents a server with an EP of 0.5; this server consumes 50 percent of its peak power at zero load. Intuitively, EP can serve as a quantitative metric for how closely a server's energy proportionality approaches perfect scaling across different server utilization levels.

Figure 2 quantifies system energy proportionality of 213 systems under test from 20 vendors between the fourth quarter of 2007 and early January 2011 ([www.spec.org/power](http://www.spec.org/power)). It is encouraging to observe that server manufacturers are designing increasingly energy-proportional systems. Whereas older systems (circa 2007) had an EP in the 30-40 percent range, current systems have an EP in the 50-80 percent range; some servers even have an EP close to 90 percent.

## EP VERSUS SPECPOWER

The SPECpower\_ssj2008 benchmark defines a server's power efficiency as the sum of the performance measured at each utilization level divided by the sum of the average power at each utilization level, including active idle. Performance is measured as the number of transactions completed per second over a fixed period of time. The benchmark starts its execution with a calibration phase to determine the system's maximum throughput; it then measures performance (throughput) and power at each utilization level starting at maximum load and decreasing in 10 percent increments.

Because SPECpower includes power and performance numbers at different utilization levels, it arguably already includes some notion of energy proportionality. Why, then, is there a need for the EP metric? Figure 3 plots EP versus SPECpower score for the same 213 systems under test. Although SPECpower correlates well with EP, the correlation is not perfect: a system with a high EP does not necessarily have a high SPECpower score, and vice versa. The key difference is that EP focuses on energy propor-

tionality only and quantifies how a server compares to the ideal energy-proportional system, whereas SPECpower quantifies average power and performance across different server load levels.

### ROOM FOR IMPROVEMENT

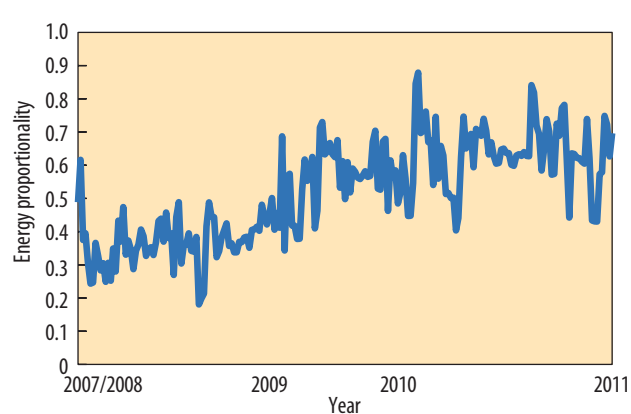
Although energy proportionality has improved dramatically in recent years, a gap remains between contemporary servers and the ideal energy-proportional system. How much more energy (and cost) can be saved by making servers even more energy proportional?

To address this question, we compared power consumption at different utilization levels for the server with the highest EP and the server with the highest SPECpower score, as Figure 4 shows. Assuming that servers operate in the 10-50 percent utilization range most of the time<sup>3</sup>—in fact, we assumed that server operation is uniformly distributed between 10 and 50 percent—we derived how much total energy can be saved by making the server more energy proportional. For the server with the highest EP score, ideal energy proportionality can potentially reduce total energy consumption by 34 percent; for the server with the highest SPECpower score, total energy consumption can be reduced by around 50 percent. Improving energy proportionality further should significantly reduce energy and cost.

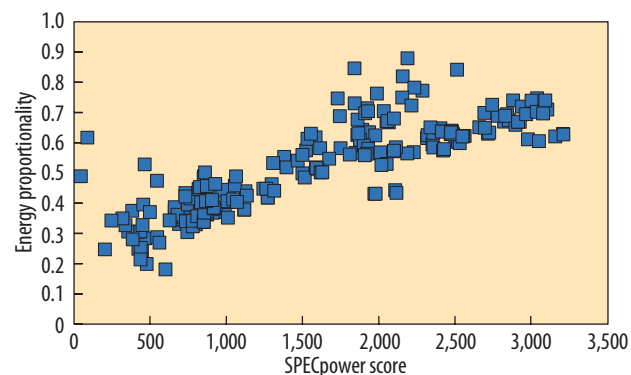
Although we likely will never achieve 100 percent energy proportionality—there will always be some overhead for keeping a server running—improving a system's energy proportionality remains an important goal. Barroso and Hölzle<sup>3</sup> reported that the CPU accounts for 50 percent of total system power at peak performance for a Google server. At lower utilization levels, however, it accounts for less than 30 percent, with the remaining 70 consumed by DRAM, hard drives, power supplies, and so on.

Looking at a large-scale datacenter, the building's mechanical and electrical infrastructure consumes considerable power, including chillers, computer room air conditioning, uninterruptible power systems, power distribution units, humidifiers, and so on. This suggests that achieving energy proportionality at the system level will require improvements across the entire system. Researchers are accordingly pursuing ways to increase energy proportionality in the CPU,<sup>5</sup> the network,<sup>6</sup> and storage devices;<sup>7</sup> others advocate rapidly transitioning an entire server between a high-performance active state and a near-zero-power idle state in response to instantaneous changes in load.<sup>8</sup>

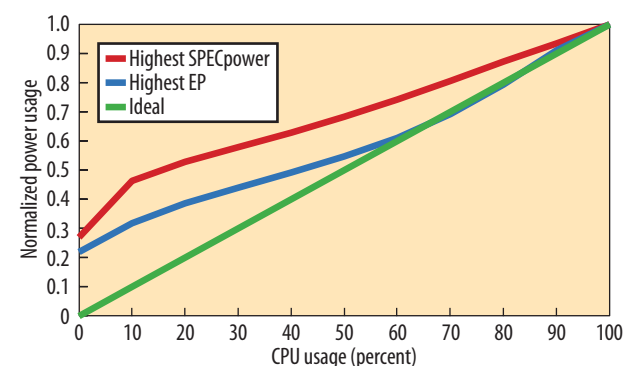
**E**nergy proportionality should be a key target in future server design. Although energy proportionality, as quantified by the proposed EP metric, has improved significantly in recent years, much more can be done to move systems closer to ideal energy proportionality. Toward this end, we are pursuing a workload



**Figure 2.** EP over time for 213 systems under test from 20 vendors. Servers are becoming increasingly energy-proportional, with some having an EP close to 90 percent.



**Figure 3.** EP versus SPECpower score for the systems under test. EP focuses on energy proportionality only and quantifies how a server compares to the ideal energy-proportional system, whereas SPECpower quantifies average power and performance across different server load levels.



**Figure 4.** Power consumption, as a function of CPU load, for the system under test with the highest EP and the system with the highest SPECpower score. For the server with the highest EP score, ideal energy proportionality can potentially reduce total energy consumption by 34 percent; for the server with the highest SPECpower score, total energy consumption can be reduced by around 50 percent.



## PERSPECTIVES

characterization and large-scale system simulation methodology for advising datacenter system integrators to move toward cost- and power-efficient infrastructures. **□**

### Acknowledgments

The authors thank the anonymous reviewers for their thoughtful comments and suggestions. Frederick Ryckbosch is supported through a doctoral fellowship by the Research Foundation—Flanders (FWO). Stijn Polfliet is supported through a doctoral fellowship by the Agency for Innovation by Science and Technology (IWT). Additional support is provided by the FWO projects G.0232.06, G.0255.08, and G.0179.10, the UGent-BOF projects 01J14407 and 01Z04109, and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) ERC Grant agreement no. 259295.

### References

1. L. Barroso and U. Hölzle, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, Morgan & Claypool, 2009.
2. X. Fan, W.-D. Weber, and L.A. Barroso, "Power Provisioning for a Warehouse-Sized Computer," *Proc. 34th Ann. Int'l Symp. Computer Architecture* (ISCA 07), ACM Press, 2007, pp. 13-23.
3. L.A. Barroso and U. Hölzle, "The Case for Energy-Proportional Systems," *Computer*, Dec. 2007, pp. 33-37.
4. K.-D. Lange, "Identifying Shades of Green: The SPECpower Benchmarks," *Computer*, Mar. 2009, pp. 95-97.
5. Y. Watanabe, J.D. Davis, and D.A. Wood, "WiDGET: Wisconsin Decoupled Grid Execution Tiles," *Proc. 37th Ann. Int'l Symp. Computer Architecture* (ISCA 10), ACM Press, 2010, pp. 2-13.
6. D. Abts et al., "Energy Proportional Datacenter Networks," *Proc. 37th Ann. Int'l Symp. Computer Architecture* (ISCA 10), ACM Press, 2010, pp. 338-347.
7. J. Guerra et al., "Energy Proportionality for Storage: Impact and Feasibility," *ACM SIGOPS Operating Systems Rev.*, Jan. 2010, pp. 35-39.
8. D. Meisner, B.T. Gold, and T.F. Wenisch, "PowerNap: Eliminating Server Idle Power," *Proc. 14th Int'l Conf. Architectural Support for Programming Languages and Operating Systems* (ASPLOS 09), ACM Press, 2009, pp. 205-216.

**Frederick Ryckbosch** is a PhD student in the Electronics and Information Systems Department at Ghent University, Belgium. His research interests include computer architecture in general, and simulation, modeling, and optimization of high-end computer systems in particular. Ryckbosch received an MS in computer science and engineering from Ghent University. Contact him at [frederick.ryckbosch@elis.ugent.be](mailto:frederick.ryckbosch@elis.ugent.be).

**Stijn Polfliet** is a PhD student in the Electronics and Information Systems Department at Ghent University. His research interests include computer architecture in general, and simulation, modeling, and optimization of high-end computer systems in particular. Polfliet received an MS in computer science and engineering from Ghent University. Contact him at [stijn.polfliet@elis.ugent.be](mailto:stijn.polfliet@elis.ugent.be).

**Lieven Eeckhout** is an associate professor in the Electronics and Information Systems Department at Ghent University. His research interests include computer architecture and hardware/software in general, with a focus on performance analysis, evaluation and modeling, and workload characterization. Eeckhout received a PhD in computer science and engineering from Ghent University. He is a member of IEEE and the ACM. Contact him at [leeckhou@elis.ugent.be](mailto:leeckhou@elis.ugent.be).

**IT Professional**  
TECHNOLOGY SOLUTIONS FOR THE ENTERPRISE

## CALL FOR ARTICLES

IT Professional seeks original submissions on technology solutions for the enterprise. Topics include

- emerging technologies,
- cloud computing,
- Web 2.0 and services,
- cybersecurity,
- mobile computing,
- green IT,
- RFID,
- social software,
- data management and mining,
- systems integration,
- communication networks,
- data center operations,
- IT asset management, and
- health information technology.

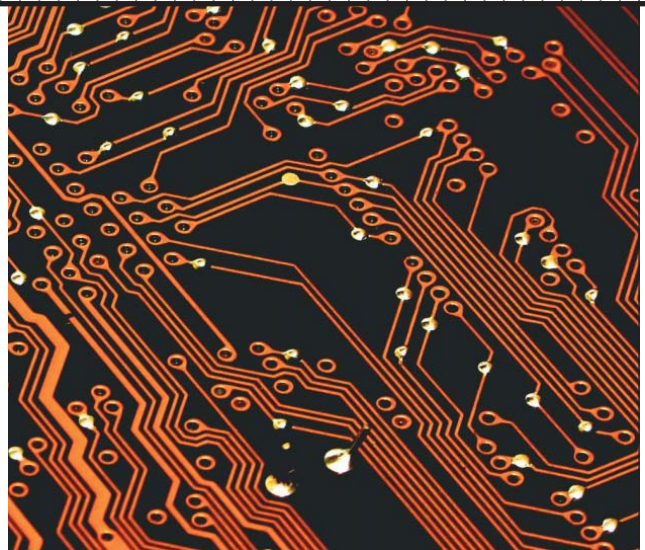
We welcome articles accompanied by Web-based demos. For more information, see our author guidelines at [www.computer.org/itpro/author.htm](http://www.computer.org/itpro/author.htm).

**[WWW.COMPUTER.ORG/ITPRO](http://WWW.COMPUTER.ORG/ITPRO)**



## RESEARCH FEATURE

# The Three Rs of Cyberphysical Spaces



Vivek Menon, *Amrita Vishwa Vidyapeetham, India*

Bharat Jayaraman and Venu Govindaraju, *State University of New York at Buffalo*

**The ability to identify people and answer questions about their whereabouts in a cyberphysical space is critical to many applications. Integrating recognition with spatiotemporal reasoning enhances the overall performance of information retrieval.**

**A** *cyberphysical space* is embedded with intelligence, providing a natural interface with humans using vision, speech, gestures, and touch, rather than keyboard and mouse. Key to realizing this paradigm is identifying and tracking people in the space. Indeed, the ability to identify and track people and answer questions about their whereabouts is critical to many applications.<sup>1</sup>

The scenarios range from environments in which most of the individuals are known or preregistered, as in health-care monitoring, to those in which most of the individuals are unknown, as in homeland security. Consider two real-life scenarios:

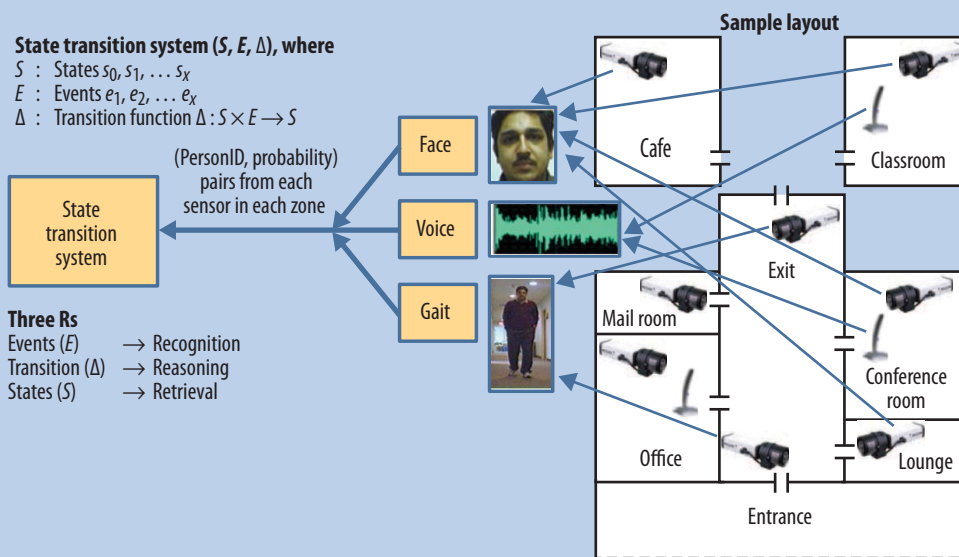
- An elderly resident in an assisted living facility wears a radio-frequency identification (RFID) badge for continuous monitoring of his presence. On one occasion, he enters the elevator alone and is trapped due to a power failure. The RFID signals that his badge transmits aren't in the range of any receiver. Only much later, when the elevator resumes its service, do the attendants discover him.
- An intruder has gained illegal entry into a secure facility that surveillance cameras monitor. After an alert

sounds, security personnel set out to find the intruder. The search team relies on input from the control room personnel monitoring the facility through multiple video feeds. The intruder no longer appears on any of the video feeds. The search team has no other recourse but to search each room.

These scenarios illustrate the need for automated approaches to transforming multimedia data into a form suitable for information retrieval, a challenge that spans video and audio processing, computer vision, spatiotemporal reasoning, and data models. The scenarios also highlight the need for unobtrusive data gathering; people should be able to go about their normal activities without being subject to a pause-and-declare routine or the burden of RFID tags or badges.<sup>2</sup> Identifying people from their face, gait, or voice is more natural and less obtrusive and hence more suited for a cyberphysical space.

We postulate three fundamental operations for a cyberphysical space: *recognition*, *reasoning*, and *retrieval*—the three Rs. Biometric recognition based on modalities such as face, gait, and voice is inherently inexact and error prone; thus, a probability distribution typically represents the output of such a recognizer.<sup>3</sup> Information retrieval in a

## RESEARCH FEATURE



**Figure 1.** Using a state transition system to identify and track people in cyberphysical spaces. The system comprises states, events, and a transition function. The events abstract approaches to recognition; the transition function abstracts approaches to reasoning; and the states provide a basis for defining data models to support information retrieval.

cyberphysical space is concerned with the location of people at various points in time. Consequently, the queries will be probabilistic and spatiotemporal—for example, where was X last seen? or what's the probability that Y and Z met in the high-security zone between 6 p.m. and 7 p.m.? To alleviate the shortcomings of a purely recognition-based approach, we show that integrating recognition with spatiotemporal reasoning enhances the overall performance of retrieval in a cyberphysical space.

The extensive literature on ubiquitous or pervasive computing<sup>4</sup> abounds in applications<sup>5</sup> in which a physical device's actuation is at the paradigm's core. Actuation-based cyberphysical spaces tend to use simpler sensors, such as for temperature, pressure, and motion, and are more concerned with networking sensors and computing devices. In contrast, in our novel paradigm, information retrieval based on spatiotemporal queries is the main driver for identification and tracking.

### A SCALABLE, UNIFIED APPROACH

Unlike earlier approaches to the problem of tracking people,<sup>6</sup> our paradigm doesn't require massive numbers of sensing devices to cover the monitored space completely; rather, in a more realistic scenario, biometric devices embedded in only the main zones—such as hallways, entrances, and exits—capture data from a distance.<sup>7</sup> This is a significant departure from current methods that tag people based on cues such as clothing or height and establish correspondences between adjacent fields of view interspersed by blind spots. Such approaches don't scale. Because we focus on indoor environments such as homes and offices,

which don't suffer the power or battery-life problems of outdoor environments, we can deploy and maintain sensors and other infrastructure with greater ease.

As Figure 1 shows, our approach provides a unified treatment of the three Rs using a novel state transition system that comprises states, events, and a transition function. This system's strength is that it accommodates different approaches to recognition, reasoning, and retrieval: the events abstract approaches to recognition; the transition function abstracts approaches to reasoning; and the states provide a basis for defining data models to support information retrieval.

This article extends our previous work, which focused on a purely recognition-based approach.<sup>8</sup> Our state transition system is fundamentally probabilistic because the biometric recognition that underlies events is inexact. To evaluate the performance of identification and tracking in a cyberphysical space, we also formulate quantitative metrics based on two information-theoretical concepts: *precision* and *recall*. These concepts are standard performance measures in the information-retrieval literature, but we adapt their definitions to suit our context.

### ABSTRACT MODEL FOR CYBERPHYSICAL SPACES

We abstract a cyberphysical space's behavior as a state transition system ( $S, E, \Delta$ ), where  $S$  is the set of states labeled  $s_0, s_1, \dots, s_x$ ;  $E$  is the set of events labeled  $e_1, e_2, \dots, e_x$ ; and  $\Delta : S \times E \rightarrow S$  is a function that models the state transition when an event occurs.<sup>8</sup> We can depict the state transitions as follows:



$$s_0 \stackrel{e_1}{\rightarrow} s_1 \stackrel{e_2}{\rightarrow} s_2 \dots \stackrel{e_x}{\rightarrow} s_x.$$

A *state* records for each zone and each occupant  $o_i$ ,  $i = 1 \dots n$ , the probability of presence in that zone,  $p_s(o_i)$ . For each occupant, the sum of probabilities across all zones equals 1. The states abstract the information necessary to support information retrieval.

An *event* abstracts a biometric recognition step and is represented as a set of person-probability pairs,  $\langle o_i, p(o_i) \rangle$ , where  $p(o_i)$  is the probability that the system recognized occupant  $o_i$  at this event. We also have

$$\sum_{i=1}^n p(o_i) = 1.$$

The *transition function* abstracts the reasoning necessary to affect state transitions. In the zone of occurrence, we define  $p_s(o_i) = p(o_i) + x_i * p'_s(o_i)$ , where  $x_i = 1 - p(o_i)$  and  $p'_s(o_i)$  is the probability of the occupant in the previous state. For all other zones, we define  $p_s(o_i) = x_i * p'_s(o_i)$ . This ensures that the sum of probabilities for an occupant across all zones in the resultant state equals 1.<sup>7</sup>

Events are assumed to be independent, but the transition function captures the dependency on the previous state, as in a Markov process. The “Hidden Markov Models” sidebar explains the difference between HMMs and our state transition system.

Table 1 illustrates this dependency, showing a sample state transition in a hypothetical four-zone cyberphysical space with five occupants. Event  $e_{11}$  occurs at zone 2 ( $z_2$ ) and corresponds to the movement of occupant  $o_5$  from  $z_1$  to  $z_2$ . The states  $s_{10}$  and  $s_{11}$  reflect the probability of the five occupants’ presence in each of the four zones before and after event  $e_{11}$ . The occupants listed in the last row correspond to the ground truth ( $G$ ).

Because we don’t continuously monitor the environment, we record a discrete set of biometric recognition events corresponding to an occupant’s movements from one zone of the environment to another. The choice of biometric sensors for a zone can vary and depends on various factors. For example, face recognition might not be suitable in some zones for privacy

## HIDDEN MARKOV MODELS

**H**idden Markov models (HMMs) and their variants, such as factorial HMMs and coupled HMMs, are examples of dynamic Bayesian networks.<sup>1</sup> In these models, transition probabilities derive from empirical data gathered about people’s movements through the space over a period of time. Because we can’t assume a predictable pattern of movement of people through various cyberphysical space zones, we don’t adopt this approach.

In our state transition system, biometric capture devices provide direct information on event occurrences in specific zones. Given any event in a zone, the next state is unambiguously determined, although the state information is probabilistic. Furthermore, a state with  $n$  occupants and  $m$  zones would require only  $m \times n$  storage, because for each of the  $m$  zones we record the probabilities of each of the  $n$  occupants being present in that zone.

### Reference

1. Z. Ghahramani, “Learning Dynamic Bayesian Networks,” *Adaptive Processing of Sequences and Data Structures*, LNCS 1387, C.L. Giles and M. Gori, eds., Springer, 1998, pp. 168-197.

reasons, and voice recognition might not work well in a noisy zone.

Biometric recognition based on a single modality can be error prone, so fusing multiple modalities can improve the recognition process’s overall accuracy. When recognition is based on more than one biometric modality, the system fuses together the outputs from the individual recognizers to derive a single set of person-probability pairs. To detect an outsider’s presence, when a person bearing little or no resemblance to any of the registered occupants arrives at the entry zone, the biometric recognition step would produce low probabilities for all occupants.

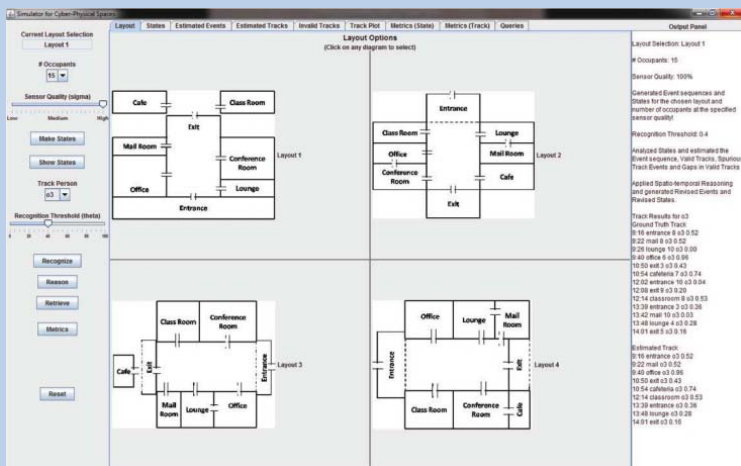
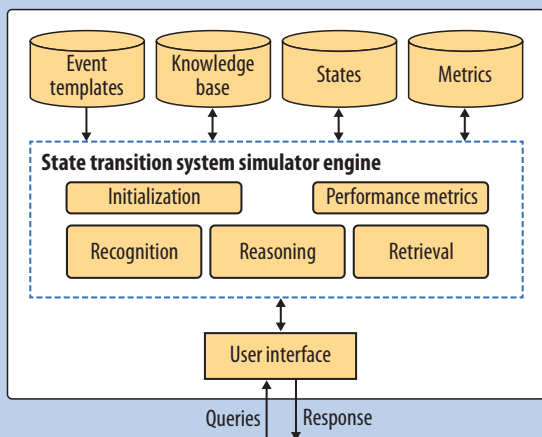
## EXPERIMENTAL TESTBED AND PERFORMANCE METRICS

We use an experimental testbed to validate our abstract model. Figure 2a shows the overall architecture, and Figure 2b shows the main interface of our simulator for a cyberphysical space. In a simulation run, a script randomly generates a set of trajectories for a given set of occupants. Each trajectory corresponds to one occupant’s movements

**Table 1. Sample state transition.**

Occupants	State $s_{10}$				Event $e_{11}$	State $s_{11}$			
	Zone 1 ( $z_1$ )	Zone 2 ( $z_2$ )	Zone 3 ( $z_3$ )	Zone 4 ( $z_4$ )		Zone 1 ( $z_1$ )	Zone 2 ( $z_2$ )	Zone 3 ( $z_3$ )	Zone 4 ( $z_4$ )
$o_1$	0.06	0.19	0.62	0.13	0.06	0.06	0.24	0.58	0.12
$o_2$	0.05	0.10	0.34	0.51	0.01	0.05	0.11	0.34	0.50
$o_3$	0.10	0.50	0.25	0.15	0.18	0.08	0.59	0.21	0.12
$o_4$	0.04	0.09	0.22	0.65	0.20	0.04	0.27	0.17	0.52
$o_5$	0.52	0.09	0.19	0.20	0.55	0.23	0.60	0.08	0.09
$G$	$o_5$	$o_3$	$o_1$	$o_2, o_4$	$o_5$	—	$o_3, o_5$	$o_1$	$o_2, o_4$

RESEARCH FEATURE



**Figure 2. Simulator for cyberphysical spaces. (a) Overall architecture, including the state transition system simulator engine and the event templates, knowledge base, states, and metrics. (b) User interface.**

and comprises a totally ordered sequence of events (where an event consists of a zone an occupant visited at a particular time) and the probabilities the biometric recognizer generated for this event. The user can choose a layout from a set of available options, specify the number of occupants in the space, and adjust parameters such as sensor quality ( $\sigma$ ) and recognition threshold ( $\theta$ ) before a simulation run. The simulator also generates performance metrics associated with recognition before and after spatiotemporal reasoning.

We derived the simulation data using the face images of 45 individuals and three video cameras of varying image quality. We customized an OpenCV ([www.intel.com/technology/computing/opencv/index.htm](http://www.intel.com/technology/computing/opencv/index.htm)) implementation of the eigenface algorithm for this purpose. The video camera quality, as well as variations in pose, illumination, and expression, can cause fluctuations in the recognition process's overall accuracy, especially in unconstrained settings. Our formulation of sensor quality  $\sigma$  abstracts intrinsic and extrinsic factors that can affect recognition output. We used 10 different event templates (face images) for every individual. When the sensor quality is reduced (using the slider bar), the system chooses a lower-quality image such that the event probability for the person recognized is correspondingly lower. A varying number of false positives across these event templates accounts for the variability, typical of unconstrained biometric recognition.

To evaluate a cyberphysical space's performance, we define the concepts of precision ( $\pi$ ) and recall ( $\rho$ ) for a cyberphysical space in terms of the ground truth  $G$ :

$$\pi = t_p / (t_p + f_p), \text{ where } t_p \text{ is the set of true positives and } f_p \text{ is the set of false positives. The set } t_p = \{o_i : p_s(o_i) \geq \theta \wedge o_i \in occ(G)\}, \text{ while the set } (t_p + f_p) = \{o_i : p_s(o_i) \geq \theta\}.$$

$$\rho = t_p / (t_p + f_n), \text{ where } t_p \text{ is defined as above, and } f_n \text{ is the set of false negatives. The set } (t_p + f_n) = \{o_i : o_i \in occ(G)\}.$$

For a given input event sequence, ground truth  $G$  is a sequence of states wherein the presence or absence of any occupant in any zone is known with certainty (0 or 1). Precision captures the extent of false positives, and recall captures the extent of false negatives. These definitions are stated in terms of a recognition threshold  $\theta$ ; only those people with a probability  $\geq \theta$  are assumed to be present. When a person's probability in two or more zones is  $\geq \theta$ , the zone with the highest probability is taken as the zone of that person's presence. We refer to the set of people occurring in a ground truth  $G$  as  $occ(G)$ .

We plotted performance metrics across multiple runs for 15 occupants. Figure 3a plots average precision and average recall for varying values of recognition threshold  $\theta$  at sensor quality  $\sigma = 1.0$ . Note that the average precision increases up to  $\theta = 0.6$  and then declines. The average precision is low at low values of  $\theta$ , because a high proportion of false positives is present in the set of recognized occupants. As  $\theta$  increases, the proportion of false positives diminishes until reaching a point of inflexion. From this point, the average precision begins to decline because the true positives also fail to get recognized. Average recall decreases with increasing  $\theta$ , because the proportion of false negatives steadily increases with  $\theta$ .

Figures 3b and 3c show the average precision and average recall curves for varying sensor quality  $\sigma$ . The dependence of precision and recall on the recognition threshold  $\theta$  causes the precision curves in Figure 3b to assume a bell shape. Depending on the application, the

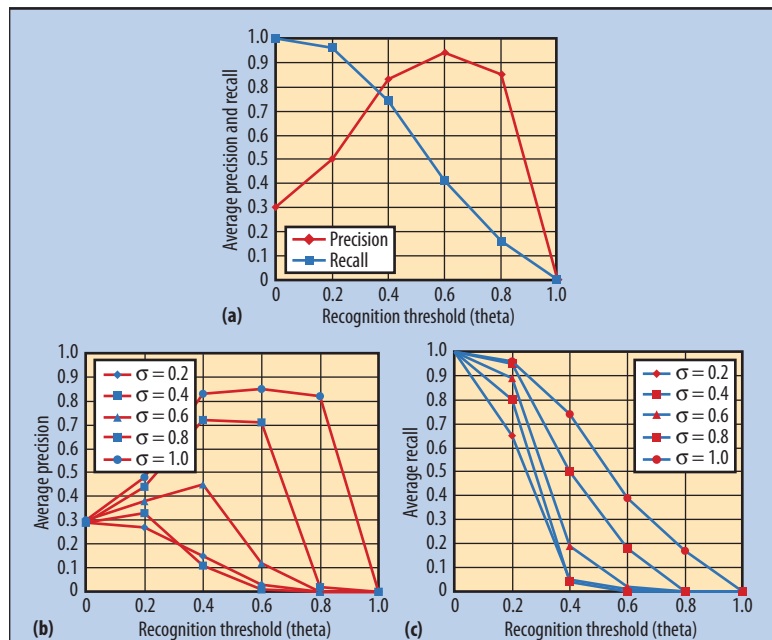
user can choose the recognition threshold for a cyberphysical space. For example, in security-related applications, the user might minimize the number of false positives, whereas in assisted living scenarios, the user might want to minimize the false negatives. In the absence of any additional information, a reasonable operating point is one in which false positives equal false negatives. At any given  $\sigma$ , we can obtain such a recognition threshold  $\theta$  from the intersection of the average precision and average recall curves, as Figure 3a shows.

### SPATIOTEMPORAL REASONING

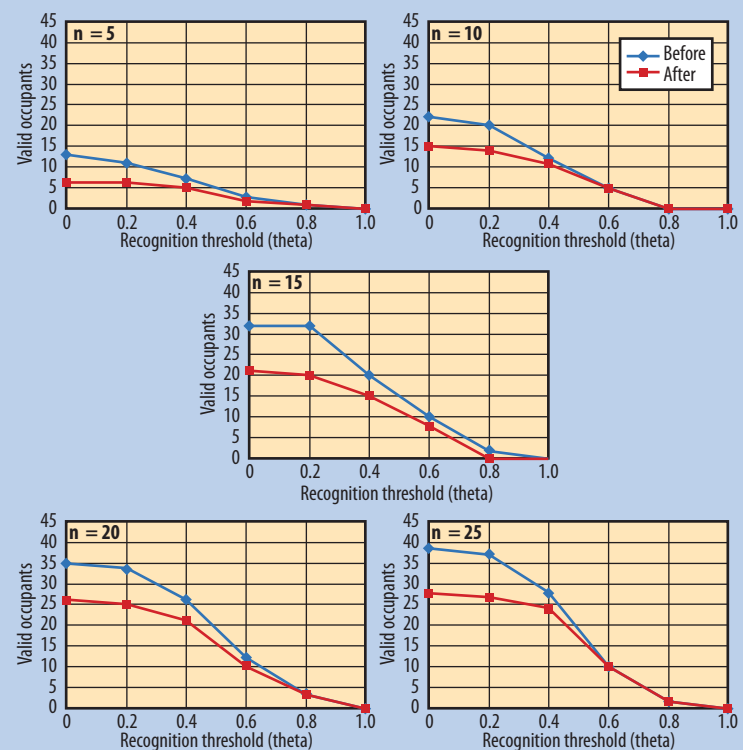
Our experimental results for a purely recognition-based approach (the blue curve in Figure 4) show that as the number of people in the cyberphysical space increases, so does the number of spuriously identified people. To minimize recognition errors' impact, the key is to determine a person's identity on the basis of information from a track of events and corresponding states rather than a single event. The reason is that consecutive track elements of a valid track will mostly obey the zone adjacencies in the physical environment, whereas spurious tracks will mostly violate the zone adjacencies. A transition function of the form  $\Delta : P(S) \times E \rightarrow S$  can capture this track-based reasoning. The transition function takes as input a set of previous states, computes the occupant tracks from these states, and determines the next state when an event occurs.

To determine the tracks from a set of states, the transition function determines for each event occurrence the individual who moved between two zones of the cyberphysical space. From this information, the system immediately obtains the set of all tracks (both valid and spurious). The criterion for determining which occupant moved is defined in terms of the maximum difference in occupant probabilities between two consecutive states in the event's zone of occurrence. It might appear that the person with the highest probability in an event is the one who moved. However, we don't adopt this approach for two reasons: event information could be erroneous; and comparing consecutive states gives due importance to both event and historical information, which the system captures and retains only in the states.

The initial states of a cyberphysical space are likely to have more errors because track-based



**Figure 3.** Precision and recall metrics. (a) Average precision and average recall versus theta shows that average precision increases up to  $\theta = 0.6$  and then declines. (b) In average precision curves for varying sigma, the dependence of precision on the recognition threshold  $\theta$  results in the bell-shaped precision curves. (c) Average recall curves for varying sigma show that average recall decreases with increasing  $\theta$ .



**Figure 4.** Estimated number of occupants before and after reasoning. The benefits of reasoning are more pronounced at a lower value of  $\theta$ , in which the number of false positives is higher.



## RESEARCH FEATURE

## RELATED WORK ON SPATIOTEMPORAL REASONING

**S**patiotemporal reasoning over occupant tracks is similar to a higher-order Markov process, because the next state depends on multiple previous states. When the transition function also updates the information in previous states, the resulting inference is closer to that of a Markov Random Field (MRF) analysis.

In the MRF approach, we can model the operation of a cyber-physical space with an undirected graph whose nodes correspond to space-time (or zone-event) points and edges capture space-time adjacency. Spatiotemporal reasoning with MRF is based on a neighborhood analysis around an event's zone of occurrence. Although it's more general in principle, it's also computationally more complex than track-based reasoning, which is more specialized and hence can more efficiently incorporate a global view of the system. Researchers have also investigated spatiotemporal reasoning from a logic and constraint perspective, with applications in geographical information systems, computer vision, planning, and so on.<sup>1</sup>

## Reference

1. A. Gerevini and B. Nebel, "Qualitative Spatiotemporal Reasoning with RCC-8 and Allen's Interval Calculus: Computational Complexity," *Proc. European Conf. AI (ECAI 02)*, IOS Press, 2002, pp. 312-316.

## RELATED WORK ON SPATIOTEMPORAL DATABASES

**T**he data in cyberphysical spaces is fundamentally probabilistic and spatiotemporal because people are moving between zones over a period of time, and we're interested in their trajectories. Hence, the data models and query languages of interest to us in a cyberphysical space are also probabilistic and spatiotemporal.

Considerable research has focused on spatiotemporal databases during the past two decades. For example, location-based systems have been a major driver for the interest in moving object databases and their associated data models, query languages, indexing, and uncertainty.<sup>1,2</sup>

In addition to the challenges involved in spatiotemporal databases, research into probabilistic databases has gained momentum over the years. This growth is owing to the emergence of a broad range of applications that must manage large and imprecise data sets in domains such as sensor networks, information extraction, and business intelligence.<sup>3</sup> Our research in cyberphysical spaces makes crucial use of probabilistic and temporal concepts, while we treat the spatial issues more qualitatively (symbolic) than quantitatively (geometric).

## References

1. R.H. Guting and M. Schneider, *Moving Object Databases*, Morgan Kaufmann, 2005.
2. N. Pelekis et al., "HERMES: Aggregative LBS via a Trajectory DB Engine," *Proc. 28th Int'l Conf. Management of Data (SIGMOD 08)*, ACM Press, 2008, pp. 1255-1258.
3. N. Dalvi, C. Re, and D. Suciu, "Probabilistic Databases: Diamonds in the Dirt," *Comm. ACM*, July 2009, pp. 86-94.

reasoning on shorter tracks is less effective in mitigating the errors owing to recognition. Over a period of time, as longer tracks form, the reasoning process not only determines subsequent states with less error, it can also correct the errors in the initial states. Such a transition function would have the form  $\Delta : P(S) \times E \rightarrow P(S)$ . That is, it takes a set of states as input, computes the tracks from these states, and determines as output the next state along with a revised set of previous states.

Figure 4 shows the benefits of integrating recognition and track-based reasoning (indicated by the red curve) to reduce the extent of spuriously identified occupants. The benefits of reasoning are more pronounced at a lower value of  $\theta$ , in which the number of false positives is higher. The "Related Work on Spatiotemporal Reasoning" sidebar provides more details.

## INFORMATION RETRIEVAL

The state transition system provides a natural basis for retrieving answers to queries about occupants' whereabouts in the cyberphysical space. We show how to formulate spatiotemporal queries using the SQL database query language, focusing on computing probabilities, an aspect novel to our model. For more information, see the "Related Work on Spatiotemporal Databases" sidebar.

We also have developed more complex queries in a constraint-based extension of logic programs, called constraint logic programming (Real), or CLP(R). This extension permits general recursive queries and reasoning over real-valued variables and arithmetic operations.<sup>9</sup>


We define an occupancy relation, *occupancy(start, end, person, zone, probability)*, where *start* and *end* define a time interval comprising a discrete totally ordered set of points (because events are also discrete). The attribute *probability*  $\in \mathfrak{R}$ , the set of real numbers, is functionally dependent on the other four attributes. This relation captures the state information after the system has performed recognition and reasoning and determined a set of valid occupants. SQL queries' basic syntax is as follows: `SELECT attributes FROM relations WHERE condition`. The *condition* is typically a conjunction of simpler tests that serve as a basis for tuple selection. This basic syntax has numerous extensions for performing aggregate operations, grouping, ordering, and so on.

Consider the following query: What is the probability that occupant 7 was in the lounge at any time between 10 a.m. and 11 a.m.? Because multiple subintervals can exist from 10 a.m. to 11 a.m., while  $o_7$  was in the lounge (with different probabilities), the answer to the query is 1 minus the product of the probabilities that  $o_7$  wasn't in the lounge during every such subinterval:

```
(1 - PROD(
  SELECT (1-probability) as prob2
  FROM occupancy
  WHERE person = o7 and
         zone = lounge and
         10:00 <= start and end <= 11:00)
)
```

The probability that an occupant wasn't in the lounge at a given time is 1 minus the probability that he was in the lounge at this time, because the sum of the probabilities across all zones equals 1 at any given time.

Although query-independent performance characterization is holistic at a system level, characterizing a cyberphysical space's performance with respect to queries might be more meaningful and better cater to a user's interest. The performance metrics of any given query are defined in terms of the ground truth, which is a set of true answers associated with the query. The nature of the response set might vary depending on the type of query the user poses. It might also comprise entities such as occupants and zones, or attributes such as probabilities of presence and time of occurrence. The response set can likewise include derived attributes such as duration of presence, tracks, and so on, on the basis of relations defined as part of the data model. From an information-retrieval perspective, precision is the fraction of retrieved answers relevant to the query, and recall is the fraction of the answers relevant to the query that the system successfully retrieved.

**O**ur state transition model serves as an effective basis for developing practical deployments; in particular, the precision recall curves help determine a suitable operating point for fine-tuning the cyberphysical space to suit the application at hand. Our current focus has been on relatively small indoor spaces, such as offices and nursing homes, where it is possible to know registered occupants in advance. We plan to explore larger indoor spaces, such as airport terminals, where our interest is in intruder identification and tracking. Such spaces also require us to address the system issues that arise out of networking a large number of cameras. 

## References

1. J. Hightower and G. Borriello, "Location Systems for Ubiquitous Computing," *Computer*, Aug. 2001, pp. 57-66.
2. A. Pentland and T. Choudhury, "Face Recognition for Smart Environments," *Computer*, Feb. 2000, pp. 50-55.
3. D. Bouchaffra, V. Govindaraju, and S. Srihari, "A Methodology for Mapping Scores to Probabilities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Sept. 1999, pp. 923-927.
4. M. Satyanarayanan, "Pervasive Computing: Vision and Challenges," *IEEE Personal Comm.*, Aug. 2001, pp. 10-17.
5. DJ. Cook and S.K. Das, "How Smart Are Our Environments? An Updated Look at the State of the Art," *Pervasive and Mobile Computing*, Mar. 2007, pp. 53-73.
6. J. Krumm et al., "Multi-Camera Multi-Person Tracking for EasyLiving," *Proc. 3rd IEEE Intl. Workshop Visual Surveillance (VS 00)*, IEEE CS Press, 2000, pp. 3-10.
7. M. Tistarelli, S.Z. Li, and R. Chellappa, eds., *Handbook of Remote Biometrics for Surveillance and Security*, Springer, 2009.
8. V. Menon, B. Jayaraman, and V. Govindaraju, "Multimodal Identification and Tracking in Smart Environments," *Personal and Ubiquitous Computing*, Dec. 2010, pp. 685-694.
9. V. Menon, "Integrating Recognition and Reasoning for Tracking and Querying in Smart Environments," doctoral dissertation, Amrita University, Mar. 2010.

**Vivek Menon** is an assistant professor of information systems in the School of Business at Amrita Vishwa Vidyapeetham (Amrita University), India. His research interests include intelligent systems and smart environments. Menon received a PhD in computer science from Amrita University. He is a member of IEEE, the ACM, and the Association for Information Systems. Contact him at [vivek\\_menon@cb.amrita.edu](mailto:vivek_menon@cb.amrita.edu).

**Bharat Jayaraman** is a professor in the Computer Science and Engineering Department at the State University of New York at Buffalo. His research interests include software systems and languages. Jayaraman received a PhD in computer science from the University of Utah. He is a senior member of IEEE and a member of the ACM. Contact him at [bharat@buffalo.edu](mailto:bharat@buffalo.edu).

**Venu Govindaraju** is a Distinguished Professor and directs the Center for Unified Biometrics and Sensors at SUNY Buffalo. His research interests include pattern recognition applied to biometrics and document analysis. Govindaraju received a PhD in computer science from SUNY Buffalo. He's a Fellow of IEEE, the ACM, the American Association for the Advancement of Science, and the International Association for Pattern Recognition. Contact him at [govind@buffalo.edu](mailto:govind@buffalo.edu).

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

**IEEE Intelligent Systems**

**THE #1 ARTIFICIAL INTELLIGENCE MAGAZINE!**

IEEE Intelligent Systems delivers the latest peer-reviewed research on all aspects of artificial intelligence, focusing on practical, fielded applications. Contributors include leading experts in

- Intelligent Agents • The Semantic Web
- Natural Language Processing
- Robotics • Machine Learning

Visit us on the Web at [www.computer.org/intelligent](http://www.computer.org/intelligent)

## CAREER OPPORTUNITIES

**SR. CONSULTANT**, Austin, TX, Ascendant Technology. WebSphere Commerce. Req. MA (or foreign equiv.) in Comp Sci., or related OR BA (or foreign equiv.) in Comp. Sci.+ 5 yrs. IT exp. Resume only to C. Jones, HR Mgr, ref. 111367, 16817 167th NE, Woodinville, WA 98072.

**PROGRAMMER ANALYST** - Evaluate, design, code, optimize, & modify d/bases & d/base applics, using Oracle 10g/11g, PL/SQL, SQL, SQL\*Plus, TOAD, Power

Designer, Informatica Developer, Linux/UNIX, WIN 95/98/2000/XP, Autosys, Subversion, Datawarehouse concepts & Java. Freq travel reqd. Reqs MS Comp Sci, Eng or rel. Mail resumes to Saiana Technologies Inc., 1 Newburgh Rd., Edison, NJ 08820.

**SENIOR ORACLE DEVELOPER** in Plano, Texas. Leads analysis of organizational needs & goals in the support and implementation of Oracle e-business

### TENURE-STREAM FACULTY POSITION DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

The Department of Computer Science and Engineering (CSE) at Michigan State University invites applications for a tenure-stream faculty position in the area of computer vision, image processing, and its applications to biometric recognition. Candidates at all ranks will be considered. The appointment starts in August 2012.

The CSE Department conducts leading-edge research in many areas, with particular strength in software engineering and formal methods, computer networks and security, computer graphics and visualization, bioinformatics and digital revolution, data mining, machine learning and pattern recognition, and natural language processing. The Department's external research awards have nearly doubled in the last couple of years. Multidisciplinary research across a broad range of disciplines is strongly encouraged and is being actively pursued by the faculty. Partnering with several other departments and universities, the CSE Department is a major contributor and plays an important role in the NSF Science and Technology Center for the study of Evolution in Action (BEACON) on our campus.

Candidates should have a Ph.D. in Computer Science or a closely related field with evidence of research accomplishments, teaching skills, and an ability to work effectively with other researchers. The successful candidate will be expected to develop an externally funded research program of national prominence that includes fundamental research, publications in high quality conferences and journals, and training graduate students. Leadership is expected in development of educational programs to provide state-of-the-art knowledge to both undergraduate and graduate students.

MSU enjoys a large, park-like campus with outlying research facilities and natural areas. The greater Lansing area has approximately 450,000 residents. The local communities have excellent school systems and place a high value on education. The University is proactive in exploring opportunities for the employment of spouses, both inside and outside the University.

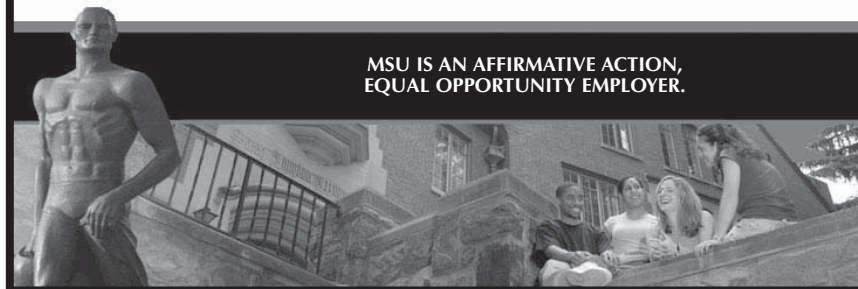
Candidates should submit an application for this position through: <https://jobs.msu.edu/>. Refer to posting #4905. Closing date is December 1, 2011. Applications will be reviewed on a continuing basis until the position is filled. For full consideration, applications should be received by the closing date.

#### Faculty Search Committee

Department of Computer Science and Engineering  
3115 Engineering Building  
Michigan State University  
East Lansing, Michigan 48824-1226  
<https://jobs.msu.edu/>

**MICHIGAN STATE  
UNIVERSITY**

MSU is committed to achieving excellence through cultural diversity. The University actively encourages applications and/or nominations of women, persons of color, veterans and persons with disabilities.



MSU IS AN AFFIRMATIVE ACTION,  
EQUAL OPPORTUNITY EMPLOYER.



Applied Materials, Inc. is accepting resumes for the following positions in **Santa Clara and Sunnyvale, CA**:

#### Technical Support Supervisor

Ref# SVBSE

Support all hardware technical issues for line of company's products.

#### Total Product Support Engineer

Ref# SCCKU

Provide customer site support to engineering, specifically related to installation, operation, calibration and services.

#### Project Support Office Leader

Ref# SCPVA

Manage complex project, products and programs, which may be enterprise-wide.

Please mail resumes with reference number to Applied Materials, Inc., 3225 Oakmead Village Drive, M/S 1217, Santa Clara, CA 95054. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

[www.appliedmaterials.com](http://www.appliedmaterials.com)



is seeking an

## Engineer, Staff – Electronic Design

Austin, TX

Reqs MS in Electrical Eng. Reqs Circuit topography for block level design; Block level specification; Physical implementation; Drafting and presenting technical design documents; Cadence Schematic/Layout editors and Matlab; Analog Design; and CMOS technology.

Mail resumes to:

HR Operations Coordinator  
5300 California Ave.  
Bldg. 2, #22108  
Irvine, CA 92617  
Must reference  
job code ENG7-TXSG.



applications. Will consult with users to determine reqs for corporate Oracle e-business related applications. Req'd: Bachelor's degree in Engineering & 3 yrs exp. in job or as Programmer Analyst. APPLY BY MAIL ONLY at Goodman Networks, Inc., 6400 International Pkwy, Ste. 1000, Plano, TX 75093. No Calls. No recruiters. Job applicants only.

**SR. SOFTWARE DEVELOPER II**, Colorado Springs, CO, Insurance Technologies. Req. BA (or foreign equiv.) in Econ., Math, Comp. Sci. or related + 5 yrs. software devel. exp. OR MA (or foreign equiv.) in Comp. Sci. Resume only to T. Shuminsky, COO, ref#061015, 2 S. Cascade Ave., #200, Colorado Springs, CO 80903.

**FLOWSERVE U.S. INC.** has an opening for Business Analyst, Sr. in Moosic, PA to design, develop, test, implement & maintain web applications. Requires Bachelor's degree in IT or Eng. & 5 yrs experience. Send resumes via email to [ttippen@flowserve.com](mailto:ttippen@flowserve.com). Must reference Job Code #12781BR in email subject line. EOE.

**SYSTEM ADMINISTRATOR:** Install, configure, deploy, test and maintain hardware and software IBM TSM server/client application on multiple OS, having experience with HSM, GPFS and AIX6.1, OLARIS 8/9/10, SUN Storagetek and IBM TS3500 Tape libraries, SAN switches, EMC Data Domain Disk libraries, Unix shell scripting and PERL. Certified TSM Administrator preferred. Frequent travel reqd. Reqs MS comp sci, eng or rel. Mail resumes to Sapphire Solutions Inc, 523 Green Street, 2nd Floor, Iselin, NJ 08830.

**SR. FUNCTIONAL ANALYST**, full-time, w/Fadel Partners, Inc. (New York, NY). Bus. process analysis; functional design implementations; configure/modify modules; root cause analysis. Need Bachelor's + exper. w/Oracle eBusiness Suite. May involve travel and/or temporary relocation to unanticipated/unknown sites in U.S. Resumes to [hr@fadel-partners.com](mailto:hr@fadel-partners.com) or by mail/courier to 38 E. 29th St., 9th fl., New York, NY 10016. Say for Job XSF.

**BUSINESS INTELLIGENCE CONSULTANT** (Altius Consulting, Inc - Houston, TX). Reqmts: BSc in Comp Sci (or foreign equiv) & 2 yrs exp in: 1) Software Platforms: FlexMI Platform & Flex MDM Studio; 2) Internet Technology: .NET based applics, C#; 3) Programming Languages: C, C++, Java; 4) Database & Related Tools: SQL based applics; 5) Oil & Gas industry namely, incl: Oil & Gas Accounting; Market Intelligence; Procurement & Supply Chain; & 6) Project mgmt. Full-time position. Send cvr ltr, resumé & employment



## SAMSUNG ELECTRONIC US R&D CENTER COMPUTER SCIENCE LABORATORY

### RESEARCHER/SENIOR RESEARCHER

Samsung Electronic's US R&D Center, San Jose, is looking for passionate individuals to join our growing Computer Science research team. We currently have a number of openings in systems-related research specifically in the areas of operating systems, distributed storage architectures, heterogeneous manycore programming and transparent distributed computing. Candidates should have a PhD in a related field together with a strong publication record. This is a great opportunity to be part of an advanced "free-thinking" research team creating state-of-the-art technologies for the future business of Samsung Electronics.

**By E-mail:** [b.gilmore@sisa.samsung.com](mailto:b.gilmore@sisa.samsung.com)



Juniper Networks is recruiting for our Sunnyvale, CA office:

**ASIC Engineer #20731:** Responsible for all aspects of physical design for blocks and/or full-chip design.

**Software Engineer #22666:** Develop features for Junos Pulse on Android including writing functional and design specifications.

**Senior Director, Product Management #1327:** Collect product requirements from customers and sales. Work with engineering on customer requirements to realize them as capabilities of company products. Lead new product introduction teams and drive the product through various stages of development.

**ASIC Engineer #10527:** Responsible for all aspects of physical design for blocks and/or full-chip design.

**Software Engineer #16312:** Design and develop software features for new and existing company switching platforms in a team environment. Support all EX platforms on respective platform specific features.

**Technical Support Engineer #10986:** Work with the customer to resolve technical and nontechnical problems related to router, protocols and network design. Troubleshoot complicated hardware and software issues, replicate customer environments and network problems in the lab.

**Test Engineer #18086:** Design and develop software engineering tools software on Web and Unix platforms.

**Software Engineer Staff #22667:** Design, develop and maintain networking applications and protocols on highly scalable networking platforms.

Mail single-sided resume with  
job code # to  
Attn: MS A1.2.1.435  
Juniper Networks  
1194 N. Mathilda Avenue  
Sunnyvale, CA 94089



i n v e n t

HP Enterprise Services, LLC is accepting resumes for the following position:

## Services Information Developer

West Lafayette, IN  
Reference: ESWLSID11

Conceptualize, design, develop, unit-test, configure, or implement portions of new or enhanced (upgrades or conversions) business and technical software solutions through application of appropriate standard software development life cycle methodologies and processes.

Mail resume to HP Enterprise Services, LLC, 5400 Legacy Drive, MS H1-6F-61, Plano, TX 75024. Resume must include Ref. #, full name, email address & mailing address. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

## Engineers Wanted

Irving, TX

IT Company in Irving, TX has openings for Soft. Engineer & Sr. Soft. Engineer to work as SAP Consultants focused on delivering projects around SAP products including design, configuration, testing and support. Travel to client sites may be needed. Master's/Bachelor's degree + 2-5 yrs of exp. & exp. with SAP Sales Distribution (SD) and Materials Mgmt (MM). Experience with Warehouse Management (WM) or Quality Management (QM) required. SAP ABAP BAPI and ECC 6.0 req'd.

Send resumes: Optimal Solutions Integration, Inc. 1231 Greenway Dr. Ste 900, Irving, TX 75038 / Ref: SE2011

refs to: [Job2011008X@hotmail.com](mailto:Job2011008X@hotmail.com). (Absolutely no phone calls.)

**NOKIA SIEMENS NETWORKS US LLC (NSN)** has a position in Irving, TX: Customer Service Engineer: Troubleshoot & implement of network telecom switches; commission, integrate, troubleshoot GSM NSS & GPRS network elements; manage GSM Circuit Core trouble tickets & other duties/skills required. [Job ID: NSN-TX11-CSRV]. Send resume to: NSN Recruiter, MS 4C-1-1580, 6000 Connection Dr, Irving, TX 75039 & note Job ID#.

**SIEMENS PLM SOFTWARE INC.** has an opening in Livonia, MI & various unanticipated worksites throughout the U.S. for a Software Product Consultant to install, configure, customize & deploy Teamcenter product suite at various customer sites. Requires BS degree & 6 yrs exp in PLM &/or Product Data Management. Requires 70% domestic travel. Send resumes to [PLMCareers@ugs.com](mailto:PLMCareers@ugs.com). Job code UG593 must be referenced in email subject line. EOE.

**EMPHASIS SOFTWARE DEVELOPMENT INC.** has an opening in Jersey City, NJ for Project Manager/Release/Change Management to perform project management within the hedge fund industry. Requires Bachelor's degree & 8 years experience. Send resumes to [employment@citco.com](mailto:employment@citco.com). Must reference Job code ESD46 in email subject line.



# ALLIANCE

DATA SYSTEMS

ADS Alliance Data Systems, Inc have positions in the following:

### Irving, TX

**Technical Director:** Duties include manage technical staff along with project timeless for builds & manage system environments; develop data warehouse on Unix/Oracle platforms; knowledge of Oracle database; perform Unix scripting/programming & other skills/duties required. [Job ID: AD-TXII-TECH]

**Senior Interactive Developer:** Duties include design, architect, prototype, development & implementation of software applications for online business to business customers; work with web development & related technologies using AJAX, SML, JQuery, Javascript, CSS styling, etc.; & other duties/skills required. [Job ID: AD-TXII-SRIND]

### Wakefield, MA

**Interactive Developer:** Duties include SQL to build reporting data marts, ad hoc user requests, & unit testing of reports; exp. with multiple databases such as Oracle, SQL Server, DB2, & Netezza; exp. with Cognos 8 or above; & other duties/skills required. [Job ID# AD-MAII-ID]

### Arlington, VA

**Senior Software Engineer:** Perform software testing & development responsibilities involving creating & implementing test strategies, test planning, test automation, & test mgmt.; functional testing of web-based applications & client based applications; & other duties/skills required. [Job ID: AD-VAII-SSWE]

Mail resume to Attn: S. Resler-HR Coordinator, Alliance Data, 601 Edgewater Dr., Wakefield, MA 01880 & note the Job ID #.

**Apple** is looking for qualified individuals for following 40/hr/wk positions. To apply, mail your resume to 1 Infinite Loop 84-REL, Attn: AEC Staffing-SA, Cupertino, CA 95014 with Req# and copy of ad. Job site & interview, Cupertino, CA. Principals only. EOE.

### **Senior Mac OS X Development Engineer [Req# 8896140]**

Responsible for performing software engineering specific to translation and localization of Apple's software user interfaces into foreign languages. Req's Associates, or foreign equivalent, in Engineering or related field plus eight (8) years of professional experience in job offered. Must have professional experience and/or academic background in: managing budget of localization projects, developing a budget management system for worldwide localization teams, FileMaker Server, FileMaker database system development, UNIX Shell scripting, Apache, PHP, XML/XSLT, Objective-C, Perl, and MacOS X programming and web application development.

### **Electronics Design Manager [Req# 8557840]**

Manage a team of DC/DC power systems design engineers to drive innovative power circuit designs and system level partitioning from concept investigation and design thru successful product level implementation. Requires 15 years experience in job offered or in a related occupation, including power supply design; DC-DC power conversion topologies including analog and digital controls; Spice based simulation tools; MathCad; Excel mathematical analysis; DC-DC voltage regulator chip architectures; magnetic design; PCB layout and state-of the art techniques for integration and efficiency optimizations; thermals, mechanical and packaging challenges; high volume manufacturing technologies and production variance. May have direct reports.

### **Camera Design Engineer [Req# 8558145]**

Evaluate, qualify, implement and tune camera subsystems in iPhone and iPod Nano products. Requires Master's degree, or foreign equivalent, in Engineering, or related field. 3 years professional experience in job offered or in a related occupation. Professional experience must be post-baccalaureate and progressive in nature. Must have professional experience with: interfacing and tuning CMOS image sensors; prototyping solutions and camera bench testing methodology; optics design and manufacturing processes, camera module design and manufacturing processes; imaging algorithms (AE, AWB, etc.).

### **Lead Engineer [Req# 8896305]**

Test iOS device integration and compatibility with Apple and 3rd party accessories and peripherals, specifically automotive head units, using analog and digital communications. Must have 5 years of professional experience in job offered. Must have experience and/or academic background in testing: multimedia, digital audio, video, automotive head-unit integration, interoperability with iPod devices, 3rd party peripherals, mobile audio devices, Mac OS, iTunes, USB signal analysis, and Bluetooth. May include managing direct reports.





**NVIDIA Corporation**, market leader in graphics & digital media processors, has professional engineering opportunities at various levels in **Santa Clara, CA**:

**ASIC Design Engr** to design and implement the industry's leading Graphics, Video / Media & Communications Processors; **Systems SW Engr** to support NVIDIA's new high performance chipset business; **HW Engr** to contribute to the development of very high speed clock generation/distributions; **System Design Engr** to participate and conduct various graphics cards qualification and release; **Portals/Java Architect** to design and develop Portals/Java software architecture; **ASIC Design Engr** to deliver comprehensive models for custom memory designs; **Lead Engr for ASIC/FPGA Design** to lead a team of four (4) distributed FPGA engineers in the development and maintenance of prototyping platform and other FPGA based projects; **Sr. Systems SW Engr** to work with NVIDIA's Tegra software graphics team in designing and developing the most advanced mobile computing technology; **DFT Engr** to implement and verify key DFT logic modules, including test mode controllers, IO Bist, Memory Bist, and Jtag; **Systems SW Engr** to design, implement and optimize all of the multimedia drivers for NVIDIA's processors; **ASIC Design Engr** as a member of the notebook chip team, bringup and test NVIDIA's graphic chips; **Sr. HW Engr** to engage in defining, documenting, designing, verifying and emulation of technical ASIC developments; **DFT Engr** to design and implement test methodologies for large, complex, high-volume Digital IC's; **System Level Test Development Engr** to develop test automation of system and software for System Level Test (SLT); **Sr. Systems SW Engr** to design, develop and implement software for state-of-the-art 3D graphics processors for next-generation computers, graphics platforms and other hardware configurations; **Technical Customer Program Mngr** to develop program schedules, milestones and deliverables; **SW Engr** to design and develop camera control and image signal processing software for the latest cutting-edge NVIDIA Tegra hardware; **Physical Design Engr** to be responsible for all aspects of physical design and implementation of Graphics Processors, Integrated Chipsets and other ASICs; **Architect Sr.** to develop algorithms and design hardware extending the state of the art in hardware support for computer graphics; **Sr. SW QA Engr** to maintain and execute driver test plan on a daily basis; **CUDA Developer Tools Engr** to work on our CUDA (NVIDIA's parallel computing architecture) developer tools team; **SW Engr** to design, implement and optimize all of the multimedia drivers for NVIDIA's processors; **Architect** to improve the current systems and develop new systems; **DFX HW Engr** to design and implement test methodologies for large, complex and high volume Digital IC's.

We also have an opening in **Fort Collins, CO** for **Sr. SW Engr** to design, implement, and optimize all of the multimedia drivers for NVIDIA's processors.

If interested, send resume to:

**NVIDIA Corporation**

Attn: MS04 (J. Goodwin)

2701 San Tomas Expressway

Santa Clara, CA 95050

Please no phone calls, emails or faxes.

# Seasonal Specials Are Back!

Act now and get a free membership  
 CSDP Bundle: Normally \$595, now \$495  
[www.computer.org/certification](http://www.computer.org/certification)



## Distinguish Yourself From the Crowd Earn Your CSDP

Earning the Certified Software Development Professional (CSDP) credential is the best way to prove your abilities, skills, and knowledge.

By adding the CSDP credential to your resume, you will demonstrate you are:

- **Current** with best software practices
- **Connected** with industry's brightest minds
- **Career-minded** and ready for that next promotion
- **Committed** to advancing the software engineering profession



To read how the CSDP credential has helped employers and employees, go to:

[www.computer.org/getcertified](http://www.computer.org/getcertified)

## Engineers Wanted

Irving, TX

IT Company in Irving, TX has openings for Soft. Engineer & Sr. Soft Engineer to work as SAP Consultants focused on delivering projects around SAP products including design, configuration, testing and support. Travel to client sites may be needed. Master's/Bachelor's degree + 2-5 yrs of exp. & exp. with SAP Netweaver Basis, Enterprise Portal, Security Administration, Solutions Manager and SAP ECC 6.0 required.

Send resumes: Optimal Solutions Integration, Inc. 1231 Greenway Dr. Ste 900, Irving, TX 75038 / Ref: SE2011

## Nokia Siemens Networks US, LLC (NSN)

has the following open position in  
**Arlington Heights, IL**

## Customer Support Engineer

**Job ID# NSN-AH11-CUST**

Work with data networks, including analysis of customer syst. requirements & participate in proposal preparation, specification review, & point-by-point response utilizing industry standards for data networks & data security protocols; LAN & TCP/IP; network security & other skills/duties required.

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

## CLASSIFIED LINE AD SUBMISSION DETAILS

Rates are \$400.00 per column inch (\$500 minimum). Eight lines per column inch and average five typeset words per line. Free online listing on [careers.computer.org](http://careers.computer.org) with print ad. Send copy at least one month prior to publication date to: Marian Anderson, Classified Advertising, Computer Magazine, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720-1314; (714) 816-2139; fax (714) 821-4010.

**Email: [manderson@computer.org](mailto:manderson@computer.org)**

## ADVERTISER INFORMATION • SEPTEMBER 2011

### ADVERTISER

Alliance Data Systems  
APC/Schneider Electric  
Apple  
Applied Materials, Inc.  
Broadcom  
HP Enterprise Services, LLC  
ICPP 2011  
Juniper Networks  
Michigan State University  
Nokia Siemens Networks  
NVIDIA Corporation  
Samsung Electronics  
Seapine Software, Inc.  
UMUC

### PAGE

82  
17  
83  
80  
80  
82  
Cover 2  
81  
80  
86  
84  
81  
Cover 4  
13

### Advertising Sales Representatives (display)

Western US/Pacific/Far East:  
Eric Kincaid  
Email: [e.kincaid@computer.org](mailto:e.kincaid@computer.org)  
Phone: +1 214 673 3742  
Fax: +1 888 886 8599

Eastern US/Europe/Middle East:  
Ann & David Schissler  
Email: [a.schissler@computer.org](mailto:a.schissler@computer.org), [d.schissler@computer.org](mailto:d.schissler@computer.org)  
Phone: +1 508 394 4026  
Fax: +1 508 394 4926

### Advertising Sales Representatives (Classified Line)

Greg Barbash  
Email: [g.barbash@computer.org](mailto:g.barbash@computer.org)  
Phone: +1 914 944 0940 | Fax: +1 508 394 4926

### Advertising Sales Representatives (Jobs Board)

Greg Barbash  
Email: [g.barbash@computer.org](mailto:g.barbash@computer.org)  
Phone: +1 914 944 0940 | Fax: +1 508 394 4926

### Advertising Personnel

Marian Anderson: Sr. Advertising Coordinator  
Email: [manderson@computer.org](mailto:manderson@computer.org)  
Phone: +1 714 816 2139 | Fax: +1 714 821 4010

Sandy Brown: Sr. Business Development Mgr.  
Email: [sbrown@computer.org](mailto:sbrown@computer.org)  
Phone: +1 714 816 2144 | Fax: +1 714 821 4010



## BOOKSHELF

**C**omputer Architecture: A Quantitative Approach, 5th ed., John L. Hennessy and David A. Patterson. The computing world today is in the midst of a revolution in which mobile and cloud computing have emerged as the dominant paradigms driving programming and hardware innovation. In an updated edition that covers the mobile computing revolution, the authors explore the ways in which cell phones, tablets, laptops and other mobile computing devices access software and technology in the cloud. Each chapter includes two real-world examples, one mobile and one data-center, to illustrate this revolutionary change.

Morgan Kaufmann; 978-0-12383-872-8; 708 pp.

**B**uilding Software for Simulation: Theory and Algorithms, with Applications in C++, James J. Nutaro. Written for those new to the field of modeling and simulation as well as experienced practitioners, this book explains how to design and implement simulation software used to engineer large systems while presenting relevant mathematical concepts and algorithms for code development. The author covers those elements of Zeigler's theory of modeling and simulation that are most important for building simulation tools and provides comprehensive examples of their use in robotics, control and communications, and electric power systems. Readers will explore the design of object-oriented simulation programs, simulation using multi-core processors, and the integration of simulators into larger software systems.

John Wiley & Sons; 978-0-470-41469-9; 347 pp.

**N**umber-Crunching: Taming Unruly Computational Problems from Mathematical Physics to

Science Fiction, Paul J. Nahin. The author demonstrates how the power of modern computing can be applied to unusual scientific problems. He describes how the art of number-crunching has changed since the advent of computers and explains how high-speed technology helps to solve conundrums such as the three-body, Monte Carlo, leapfrog, and gambler's ruin problems. The book provides a historical background for the problems presented, offers numerous examples and challenges, supplies Matlab codes for the theories discussed, and includes detailed solutions.

Princeton University Press; 978-0-691-14425-2; 408 pp.

**S**emantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL, 2nd ed., Dean Allemang and James Hendler. Semantic Web models and technologies provide information in machine-readable languages that enable computers to access the Web more intelligently to perform tasks automatically without the direction of users. Focused on developing useful and reusable models, this book explains how to build semantic content and applications that access that content. It surveys the latest Semantic Web tools for organizing, querying, and processing information and includes detailed information about the current ontologies used in key Web applications including e-commerce, social networking, and data mining.

Morgan Kaufmann; 978-0-123-85965-5; 384 pp.

**C**yber Warfare: Techniques, Tactics and Tools for Security Practitioners, Jason Andress and Steve Winterfeld. This book explores the battlefields, participants, and tools and techniques used in today's digital conflicts. The authors provide concrete examples of cyber attacks and



offer real-world guidance on how to identify threats and defend networks against malicious attacks, offering an insider's point of view that details the ethics, laws, and consequences of cyber warfare and how computer criminal law might evolve. The concepts discussed in this book will give those involved in information security a better idea of how cyber conflicts are carried out now, what they might look like in the future, and how to detect and defend against them.

Elsevier; 978-1-59749-637-7; 289 pp.

**D**igital Forensics with Open Source Tools, Cory Altheide and Harlan Carvey. As a definitive resource on the use of open source tools to investigate computer systems and media, this book details core concepts and techniques for forensic file system analysis on both Linux and Windows systems. The authors demonstrate both well-known and novel forensic methods using command-line and graphical open source tools to examine a wide range of target systems and artifacts.

Syngress; 978-1-59749-586-8; 264 pp.

Send book announcements to [newbooks@computer.org](mailto:newbooks@computer.org).

## COMPUTER SOCIETY CONNECTION

# Barrett Wins Pre-College Education Award



**T**he UK's Tom Barrett was recently honored with the IEEE Computer Society's Distinguished Contributions to Public Service in a Pre-College Environment Award. Barrett is a teacher, speaker, and curator of ideas who employs cutting-edge educational technology to inspire and engage children in their learning. His award citation reads, "For inspired leadership and dedication in promoting the use of modern technology in education both locally and internationally."

Barrett has a passion for helping teachers connect and learn together through the use of online tools. He has been instrumental in supporting and encouraging thousands of teachers worldwide to connect and to build networks that help support their professional development.

During the past 10 years, Barrett has taught at three different Nottinghamshire primary schools as a specialist



**Tom Barrett is considered a leading voice on the application of new technologies in the classroom.**


ICT teacher, floating staff member, and full-time classroom teacher in grades ranging from Nursery to Year 6. In addition, he has played a major role in successful Office for Standards in Education, Children's Services and Skills inspections, one of which was judged to be "Outstanding."

Barrett was the first educator in the world to work with and develop learning applications for a multi-touch surface device designed by

Philips. He currently participates in the steering group for Durham University's SynergyNet, a project that is researching the future path of multi-touch pedagogy. Barrett worked with Google to bring its teacher academy event to London in 2010, the first time Google held the event outside the US. He has acted as curator for numerous training resources including the Interesting Ways series and Maths Maps. Barrett has used crowdsourcing methods to redefine the way educational resources are created, raising the bar for high-quality, practical resources developed by teachers for teachers. A principal consultant to Notosh Limited, Barrett continues to share his expertise on his [edte.ch](http://edte.ch) blog and via 140 characters on Twitter (@tombarrett).

## PRE-COLLEGE ENVIRONMENT AWARD

The IEEE Computer Society's Distinguished Contributions to Public Service in a Pre-College Environment Award honors outstanding individuals who further the professional and technical goals of the IEEE Computer Society in K-12 primary and secondary schools.

Nominations are due 15 October. To learn more about Computer Society awards, including the Pre-College Environment Award, visit [www.computer.org/awards](http://www.computer.org/awards). 

Engineering and Applying the Internet

## Internet Computing

IEEE Internet Computing reports emerging tools, technologies, and applications implemented through the Internet to support a worldwide computing environment.

For submission information and author guidelines, please visit [www.computer.org/internet/author.htm](http://www.computer.org/internet/author.htm)

## COMPUTER SOCIETY PUBLS MAKE JCR TOP-10 LISTS

The IEEE Computer Society publishes seven magazines and journals that are among the top 10 in their respective categories, according to the latest *Journal Citation Reports* from Thomson Reuters.

*IEEE Transactions on Pattern Analysis and Machine Intelligence* was the top-ranked journal in the computer science artificial intelligence category in the 2010 *JCR*, continuing its long history of being among the most highly ranked technical journals. The journal also ranked fourth in the electrical engineering category in the 2010 reports.

In the computer science software engineering category, *IEEE Micro* ranked sixth, *IEEE Internet Computing* seventh, and *IEEE Transactions on Software Engineering* ninth. Two IEEE Computer Society publications, *IEEE Transactions on Mobile Computing* and *IEEE Pervasive Computing*, were among the top 10 in the telecommunications category. *TMC* and *PvC* ranked fourth and seventh, respectively, in the 2010 *JCR* telecommunications category.

Two Computer Society publications were also among the top 10 in the computer science hardware and architecture category. *IEEE Micro* ranked fourth, and *Computer*, the Society's flagship magazine, ranked tenth. *Computer* focuses on all areas of computing.

*JCR* metrics define journal performance across disciplines and institutions worldwide. The reports present quantitative data that provides an objective way to evaluate the world's leading journals and their impact and influence in the global research community. *JCR* covers more than 9,100 of the most highly cited, peer-reviewed journals from 78 nations. By compiling articles' cited references, the reports help to measure research influence and impact at the journal and category levels and show the relationship between citing and cited journals.

Learn more at [http://thomsonreuters.com/products\\_services/science/science\\_products/a-z/journal\\_citation\\_reports](http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports).

## Software Process Award Named for Watts Humphrey

The IEEE Technical Activities Board has approved changing the name of the Software Process Achievement (SPA) Award to recognize the outstanding achievements of the late Watts S. Humphrey, a software engineering process pioneer at Carnegie Mellon University's Software Engineering Institute who was an early supporter of the award.

The award's new official name will be the IEEE Computer Society/Software Engineering Institute Watts S. Humphrey Software Process Achievement Award. The name change was also endorsed by the SPA Award Selection Committee, the IEEE Computer Society Awards Committee, and the IEEE Computer Society Board of Governors.

Humphrey's legacy includes development of the Software Capability Maturity Model, the Software Process Assessment and Software Capability Evaluation methods, and the Personal Software Process and Team Software Process methodologies.




Software engineering pioneer Watts S. Humphrey was awarded the National Medal of Technology in 2005.

"Watts was a strong proponent for the establishment of this prestigious award and continued to be an advocate in the selection process for many years," said SEI Director and CEO Paul D. Nielsen in requesting that the award's name be changed to honor Humphrey's accomplishments.

Humphrey founded the Software Engineering Institute's Software Process Program in the 1980s and served as its director from 1986 until 1996. He was the author of 11 books.

Cosponsored by the IEEE Computer Society and the Software Engineering Institute, the Watts S. Humphrey Software Process Achievement Award recognizes outstanding achievements in improving an organization's ability to create and evolve software-dependent systems. The award may be presented to an individual, group, or team. Nominees are most often employees of an organization that produces, supports, enhances, or provides software-dependent systems.

Anyone can nominate candidates for the honor, and organizations can nominate themselves. All nominations must be seconded by a senior executive of the organization in which the nominated individual or team works, and supported by a 10-page nomination package detailing how the nominee's software engineering or process improvement work is, to an exceptional degree, significant, sustained, measured, and shared. For more information, visit [www.computer.org/awards](http://www.computer.org/awards). 



## COMPUTER SOCIETY CONNECTION

# IEEE Design & Test of Computers Moves to CEDA

**T**he IEEE Council on Electronic Design Automation is taking over ownership of *IEEE Design & Test of Computers* from the IEEE Computer Society.

An agreement transferring ownership of the magazine was signed in June 2011 by IEEE Computer Society President Sorel Reisman and Andreas Kuehlmann, president of the IEEE Council on Electronic Design Automation. The ownership change is effective 1 January 2012.

“Over the past decade, the IEEE members who form the *D&T*


community have congregated around CEDA. Hence, it makes sense to get the periodical into the hands of the organization that is most interested in its future,” said David Alan Grier, IEEE Computer Society vice president of publications. “We think this move is right for both CEDA and *D&T*.”

Volunteers and staff from both organizations have been working together since last fall to formulate an agreement that would allow CEDA to own a well-established magazine targeted to the design and test communities.

“We are excited to take over this high-quality magazine. We are planning to join forces with two of our member societies, the Circuits and Systems Society and the Solid-State Circuits Society, as well as the Test Technology Technical Council to further strengthen this publication,” said Kuehlmann.

Published since 1984 by the IEEE Computer Society (and copublished with CASS since 2002), *D&T* has a loyal subscriber base. Transferring the magazine to CEDA will provide additional resources to promote it to a wider audience. *D&T* covers the tools, techniques, and concepts used to design and test electronic product hardware and supportive software. The magazine is a leader in analysis of current and near-future practices.

Said Rajesh Gupta, vice president of publications for CEDA, “*D&T* complements our existing portfolio of publications, which consists of *IEEE Transactions on Computer-Aided Design* and *IEEE Embedded Systems Letters*, to provide a comprehensive benefit to our community, bringing together not only the latest advances in theory and practice, but also the personalities and views that shape our industry and profession.”

As part of the agreement, *D&T* issues produced by CEDA will be available to Computer Society Digital Library subscribers through 2014 at [www.computer.org/portal/web/csdl/magazines/dt](http://www.computer.org/portal/web/csdl/magazines/dt). 

**B. Ward, Editor; [bnward@computer.org](mailto:bnward@computer.org)**

## SC11

International Conference for High Performance Computing, Networking, Storage and Analysis

**12-18 November 2011**  
Seattle, Washington, USA

The SC11 conference continues a long and successful tradition of engaging the international community in high performance computing, networking, storage and analysis.

*Register today!*

<http://sc11.supercomputing.org/>



## CALL AND CALENDAR

### CALLS FOR ARTICLES FOR COMPUTER

*Computer* seeks submissions for an April 2012 special issue on interaction beyond the keyboard.

Interaction with computers has become an integral part of daily life for most people. As computing technologies proliferate, simple user interfaces and ease of use become key success factors for a wide range of products.

Although the keyboard and mouse are still the dominant user interfaces in home and office environments, with the massive increase in mobile device usage and the many new interaction technologies available, the way we interact with computers is becoming richer and more diverse. Touch-enabled surfaces, natural gestures, implicit interaction, and tangible user interfaces mark some of these trends.

Authors are encouraged to submit original research that describes groundbreaking new devices, methods, and approaches to human-computer interaction in a world of ubiquitous computer use. Suitable topics include interactive surfaces and tabletop computing, tangible interaction and graspable user interfaces, and user interfaces based on physiological sensors and actuators.

Direct inquiries to guest editor Albrecht Schmidt of the University of Stuttgart at [albrecht@computer.org](mailto:albrecht@computer.org).

Paper submissions are due **1 November**. For author guidelines and information on the electronic submission process, visit [www.computer.org/portal/web/peerreviewmagazines/computer](http://www.computer.org/portal/web/peerreviewmagazines/computer).

*Computer* seeks submissions for a September 2012 special issue on



modeling and simulation of smart and green computing systems.

Sustainable and efficient utilization of available energy resources is perhaps the fundamental challenge of the current century. Academic and industrial communities have invested significant resources in developing new solutions to address energy-efficiency challenges in several areas including IT and telecommunications, green buildings and cities, and the smart grid.

Modeling and simulation methodologies are necessary for the comprehensive performance evaluation that precedes costly prototyping activities for such complex, large-scale systems. This special issue aims to disseminate the latest advances in modeling and simulation of smart and green computing systems, which are critical from the perspective of sustainable economic growth and environmental conservation.

Topics of interest include modeling and simulations of energy-efficient computing systems, green communications systems, and smart grid applications. For author guidelines

and information on how to submit a manuscript electronically, visit [www.computer.org/portal/web/peerreviewmagazines/computer](http://www.computer.org/portal/web/peerreviewmagazines/computer).

Articles are due by **1 March 2012**. Visit [www.computer.org/portal/web/computingnow/cocfp9](http://www.computer.org/portal/web/computingnow/cocfp9) to view the complete call for papers.

### CALLS FOR ARTICLES FOR IEEE CS PUBLICATIONS

*Computing in Science & Engineering* plans a May/June 2012 special issue on scientific computing with graphics processing units.

GPUs aren't just for graphics anymore. These high-performance many-core processors are used to accelerate a wide range of science and engineering applications, in many cases offering dramatically increased performance compared to CPUs. Computer architects also use them to build the world's largest supercomputers. However, the use of GPUs in scientific computing comes with added risks. The effort needed to port applications can be substantial, and not every application benefits equally well from GPU acceleration.

The guest editors seek contributions covering all aspects of using GPUs to solve challenging computational science problems. Of special interest are articles presenting the results of porting efforts of large-scale scientific applications on large-scale, GPU-based, high-performance computers.

## SUBMISSION INSTRUCTIONS

The Call and Calendar section lists conferences, symposia, and workshops that the IEEE Computer Society sponsors or cooperates in presenting.

Visit [www.computer.org/conferences](http://www.computer.org/conferences) for instructions on how to submit conference or call listings as well as a more complete listing of upcoming computer-related conferences.

**CALL AND CALENDAR**

**EVENTS IN 2011**

**October**

- 4-7 ..... LCN 2011
- 4-7 ..... SRDS 2011
- 10-14 ..... PACT 2011
- 22-25 ..... FOCS 2011

**November**

- 6-10 ..... ASE 2011
- 6-13 ..... ICCV 2011
- 7-9 ..... ICTAI 2011
- 12-18 ..... SC 2011
- 21-23 ..... NCCA 2011

**December**

- 5-8 ..... E-Science 2011
- 7-9 ..... ICPADS 2011
- 11-14 ..... ICDM 2011
- 18-21 ..... HiPC 2011

**HIPC 2011**

**T**he 18th IEEE International Conference on High-Performance Computing serves as a forum where top experts from around the world can present cutting-edge research results. Conference events highlight high-performance computing activities in Asia. HiPC 2011 focuses on all aspects of high-performance computing systems and their scientific, engineering, and commercial applications.

Conference organizers have solicited contributions on topics that include cluster, cloud, and grid cloud computing; peer-to-peer algorithms and networks; scalable servers and systems; and parallel languages and programming environments

HiPC 2011 takes place 18-21 December in Bangalore, India. Visit [www.hipc.org](http://www.hipc.org) for complete conference details.

tems, Madrid, Spain; <http://lsd.ls.fi.upm.es/srds2011>

10-14 Oct: PACT 2011, 20th Int'l Conf. on Parallel Architectures and Compilation Techniques, Galveston, Texas; <http://pactconf.org>

22-25 Oct: FOCS 2011, 52nd IEEE Symp. on Foundations of Computer Science, Palm Springs, California; [www.cs.ucr.edu/~marek/FOCS11](http://www.cs.ucr.edu/~marek/FOCS11)

**NOVEMBER 2011**

6-10 Nov: ASE 2011, 26th IEEE/ACM Int'l Conf. on Automated Software Eng., Lawrence, Kansas; [www.continuinged.ku.edu/programs/ase](http://www.continuinged.ku.edu/programs/ase)

6-13 Nov: ICCV 2011, 13th Int'l Conf. on Computer Vision, Barcelona, Spain; [www.iccv2011.org](http://www.iccv2011.org)

7-9 Nov: ICTAI 2011, 23rd IEEE Int'l Conf. on Tools with Artificial Intelligence, Boca Raton, Florida; [www.cse.fau.edu/ictai2011](http://www.cse.fau.edu/ictai2011)

12-18 Nov: SC 2011, ACM/IEEE Int'l Conf. for High Performance Computing, Networking, Storage, and Analysis, Seattle; <http://sc11.supercomputing.org>

21-23 Nov: NCCA 2011, First IEEE Symp. on Network Cloud Computing and Applications, Toulouse, France; <http://sites.google.com/site/ieeencca2011>

**DECEMBER 2011**

5-8 Dec: E-Science 2011, 7th Int'l Conf. on e-Science, Stockholm; [www.escience2011.org](http://www.escience2011.org)

7-9 Dec: ICPADS 2011, IEEE Int'l Conf. on Parallel and Distributed Systems, Tainan, Taiwan; <http://conf.ncku.edu.tw/icpads>

11-14 Dec: ICDM 2011, IEEE Int'l Conf. on Data Mining, Vancouver, Canada; <http://webdocs.cs.ualberta.ca/~icdm2011/index.php>

18-21 Dec: HiPC 2011, IEEE Int'l Conf. on High-Performance Computing, Bangalore, India; [www.hipc.org](http://www.hipc.org)

Articles are due by 14 September. Visit [www.computer.org/portal/web/computingnow/cscfp3](http://www.computer.org/portal/web/computingnow/cscfp3) to view the complete call for papers.

**CALENDAR**

**OCTOBER 2011**

4-7 Oct: LCN 2011, 36th IEEE Conf. on Local Computer Networks, Bonn, Germany; [www.ieeeln.org/index.html](http://www.ieeeln.org/index.html)

4-7 Oct: SRDS 2011, 30th IEEE Int'l Symp. on Reliable Distributed Sys-

**COMPUTING THEN**

Learn about computing history and the people who shaped it.

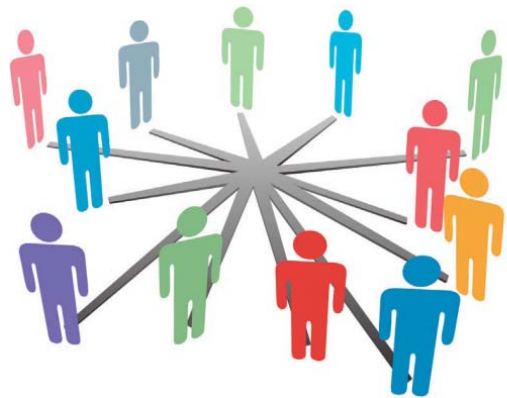
<http://computingnow.computer.org/ct>



## SOCIAL COMPUTING

# Let's Gang Up on Cyberbullying

Henry Lieberman, Karthik Dinakar,  
and Birago Jones, MIT Media Lab



The novel design of social network software can help prevent and manage the growing problem of cyberbullying.

Cyberbullying has emerged as a major problem in recent years, afflicting both children and young adults. Tragic stories in the news about suicides of bullied teens have drawn public attention to the issue, and statistics indicate that its prevalence is growing. A 2006 survey commissioned by the National Crime Prevention Council showed that more than 43 percent of US teens were subjected to cyberbullying at some point in the previous year ([www.ncpc.org/cyberbullying](http://www.ncpc.org/cyberbullying)), while a 2008 survey by UCLA researchers reported that nearly three-quarters of teens had been bullied online at least once in the past 12 months ([www.safeinyourspace.org/2008juvonengross.pdf](http://www.safeinyourspace.org/2008juvonengross.pdf)).

Social networks provide many benefits for youth, like helping to start and maintain friendships and providing a personally meaningful context for practicing reading and writing (D. Boyd, *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*, MIT Press, 2007). But if too many kids are bullied too often on social networks, kids will feel scared to join them and parents won't permit their children to participate.

In the Internet's early days, many people were surprised and discouraged by receiving spam in

their e-mail. If no measures had been taken to combat the rapidly growing volume of spam, the Internet as we know it today wouldn't exist. Fortunately, a technical solution—spam filters—managed to get the problem under control, even if it hasn't totally eliminated spam. Are technical solutions available to likewise manage cyberbullying?

The MIT Media Lab's Time Out project is investigating a range of innovations in social network software to help prevent cyberbullying and mitigate the consequences when it does occur. Our efforts fall into two broad categories: detecting possible cases of cyberbullying by using machine learning to better understand cyberbullying language; and intervention technologies for participants as well as network providers and moderators. *Reflective interfaces* encourage participants to carefully consider their behavior and choices, while visualization tools can help providers and moderators control the escalation of cyberbullying.

## DETECTION

Detecting cyberbullying, which is personalized and contextual, is much more difficult than detecting spam, which is sent identically to large numbers of people. However, our analysis indicates that most

cyberbullying occurs around a small number of topics: race and ethnicity, sexuality and sexual identity, physical appearance, intelligence, and social acceptance and rejection. If we can understand whether a message is about those topics, and whether its tone is positive or negative, we can identify many possible cyberbullying messages.

One class of bullying messages we have studied involves accusations of being gay or lesbian, with a negative intent. Often this takes the form of ascribing stereotypically female characteristics to a male or stereotypically male characteristics to a female. For example, a comment addressed to a male might be "You'd look great in lipstick and a dress."

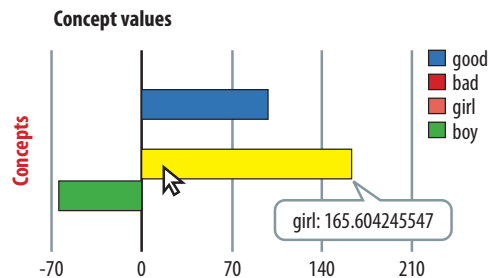
This doesn't suggest people ought to feel bad if such things are said about them—we certainly don't endorse any stereotypes. But in practice, such statements are often used in cyberbullying and so are among several clues as to whether it might be occurring.

Computers still can't fully understand English, but progress in natural-language processing means that sometimes we can partially understand some aspects of a text. Active areas of research include *topic detection*, a mainstay of search and database engines; and *affect analysis*,

## SOCIAL COMPUTING

You look like a fashion model!!!!

Profanity: None Concepts model fashion look like  
Intuitive reasoning



**Figure 1.** The term “fashion model” is commonly associated with females; thus, when directed at a straight male, the sentence “You look like a fashion model!” might be an example of cyberbullying.

determining whether a message conveys a positive or negative emotion.

Among the statistical tools we’re using are machine learning classifiers, which are effective for many topic detection problems. We train these classifiers on a set of cyberbullying messages identified by humans and analyze the messages for statistical regularities (K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the Detection of Textual Cyberbullying,” *Proc. 2011 Social Mobile Web Workshop, Assoc. for the Advancement of Artificial Intelligence*, 2011, pp. 11-17).

Our “secret ingredient” in this process is a commonsense knowledge base with associated reasoning techniques. We’ve collected about one million sentences describing everyday life that provide the kind of background knowledge AI programs need to go beyond simple word matching and word counting. We use these to simulate the kind of vague, informal reasoning that people do, rather than reasoning with mathematical precision.

The knowledge base contains the kind of statements that help a detection system decide whether a sentence might be referring to stereotypical male or female concepts—for example, “lipstick is a kind of makeup” or

“women wear dresses.” As Figure 1 shows, the term “fashion model” is commonly associated with females; thus, when directed at a straight male, the sentence “You look like a fashion model!” might be an example of cyberbullying.

Similarly, commonsense knowledge about what people typically eat might indicate that the comment “You must have eaten six hamburgers for dinner tonight” was intended to insult someone for being overweight.

We would like to avoid directly accusing an individual of being a bully in any situation. Our goal isn’t to achieve 100 percent certainty in detecting cyberbullying but to call out the possibility of its occurrence. If a pattern is repeated over time, seems to be escalating, or has a consistently negative tone, our confidence in estimation might increase.

## INTERVENTION

There are many participants in the cyberbullying process: the perpetrator, the victim, friends, family, teachers, and so on. We can design interventions specifically for each role. As Figure 2 shows, when a possible cyberbullying message is detected, we could unobtrusively provide a link to educational material appropriate to the user’s situation.

For potential cyberbullies, the material could encourage empathy for the victim and warn of possible damage to the bully’s social reputation. The intervention could exhort victims to seek emotional support, learn how others have dealt with similar situations, give suggestions for appropriate responses (such as humor), and discourage them from retaliating. The material could induce friends to defend the victim rather than join in with the bully.

The key is to offer advice that is personalized, specific, and actionable. The material can take many forms including written stories, video, or interactive narratives.

Other measures could subtly change the social network interface to encourage reflection, or to slow the spread of a potentially insulting message that has been sent. Instead of just a simple “Send” message, for example, the button could be changed to remind the user of the consequences: “Send to 350 people in your network.” Likewise, an “Are you sure?” confirmation could be added to potentially problematic messages. Or delivery could be delayed overnight to give the sender a chance to rescind the message in the morning before it’s actually delivered.

Social network providers and moderators also have a role to play: they’re obligated to provide a safe and welcoming environment for their participants, especially newcomers. To that end, we propose a kind of “air traffic control” interface called SpeedBump, shown in Figure 3, that helps a moderator visualize the community’s connections, history, and topics.

SpeedBump’s goal isn’t to catch every instance of cyberbullying but to prevent incidents from escalating into major outbreaks. Social network providers inform us that such incidents tend to occur in clusters: typically they spread rapidly throughout a particular group, like students at a school, or are triggered



Figure 2. When a possible cyberbullying message is detected, an automated intervention system could unobtrusively provide a link to educational material appropriate to the user's situation (offensive terms redacted by authors).

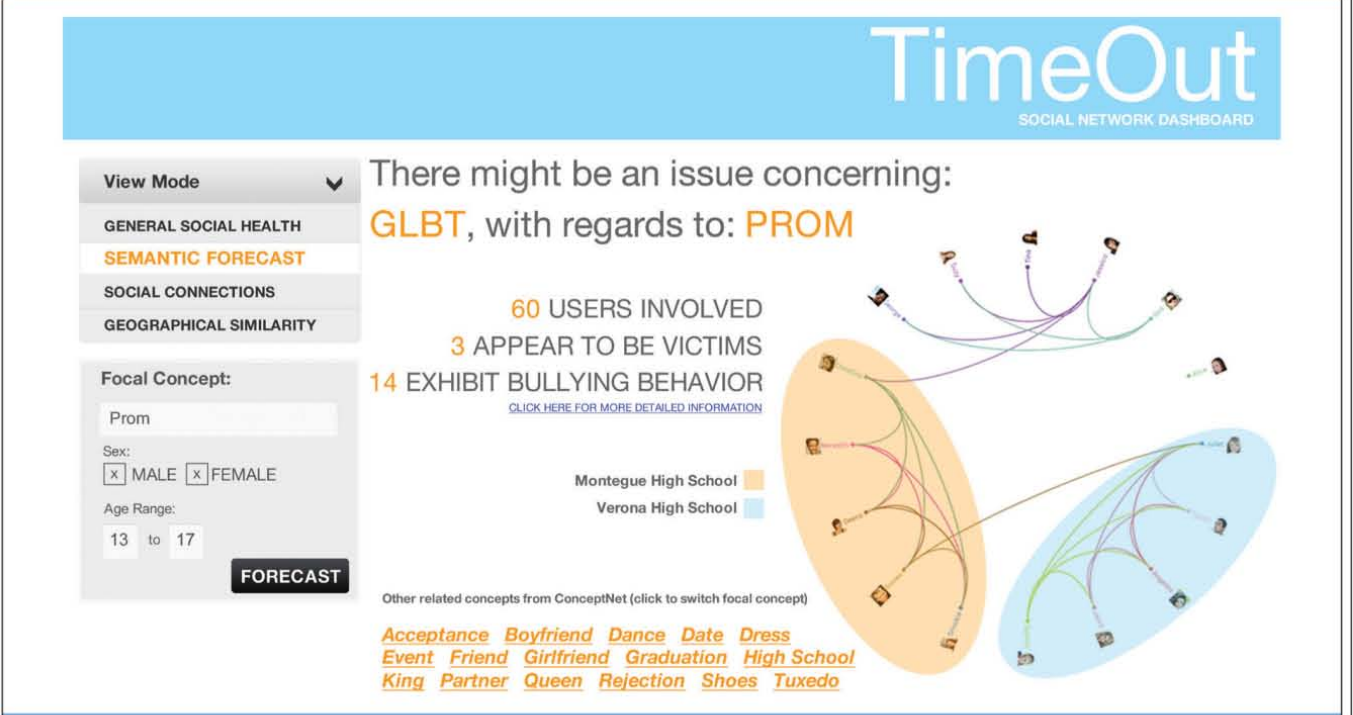


Figure 3. SpeedBump helps a social network moderator visualize the community's connections, history, and topics.



## SOCIAL COMPUTING

by a particular event, like a high school prom.

**A**t the March 2011 White House Conference on Bullying Prevention, President Obama said, “Today, bullying doesn’t ... end at the school bell—it can follow our children from the hallways to their cell phones to their computer screens. If there’s one goal of this conference, it’s to dispel the myth that bullying is just a harmless rite of passage or an inevitable part of growing up. It’s not.” ([www.whitehouse.gov/blog/2011/03/10/president-obama-first-lady-white-house-conference-bullying-prevention](http://www.whitehouse.gov/blog/2011/03/10/president-obama-first-lady-white-house-conference-bullying-prevention))

At its core, cyberbullying is really a people problem: no software can substitute for teaching kids how to have healthy personal relationships. But the novel design of social network software can help prevent and manage the problem.

*Henry Lieberman is a principal research scientist at the MIT Media Lab, where he heads the Software Agents Group. Contact him at [lieber@media.mit.edu](mailto:lieber@media.mit.edu).*

*Karthik Dinakar is a research assistant at the MIT Media Lab. Contact him at [kdinakar@media.mit.edu](mailto:kdinakar@media.mit.edu).*

*Birago Jones is a research assistant at the MIT Media Lab. Contact him at [birago@media.mit.edu](mailto:birago@media.mit.edu).*

**Editor: John Riedl, Department of Computer Science and Engineering, University of Minnesota; [riedl@cs.umn.edu](mailto:riedl@cs.umn.edu)**

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



IEEE Software offers pioneering ideas, expert analyses, and thoughtful insights for software professionals who need to keep up with rapid technology change. It’s the authority on translating software theory into practice.

[www.computer.org/software/SUBSCRIBE](http://www.computer.org/software/SUBSCRIBE)

# SUBSCRIBE TODAY

## GREEN IT

# Software Bloat and Wasted Joules: Is Modularity a Hurdle to Green Software?

Suparna Bhattacharya and K. Gopinath, *Indian Institute of Science*  
 Karthick Rajamani and Manish Gupta, *IBM Research*



Adopting an integrated analysis of software bloat and hardware platforms is necessary to realizing modular software that's also green.

System administrators have long observed how software bloat tends to negate increases in computing hardware capacity. While functionally richer and more flexible, newer software packages often incur larger resource overhead in typical execution scenarios. This trend is a consequence of the need to rapidly develop applications with complex business logic and integration requirements while addressing a wide range of operational considerations.

Coincident with the emphasis on application functionality and flexibility, there has been a declining focus on the efficient use of computing resources. During the past two decades, software design paradigms have evolved to prioritize programmer productivity over runtime efficiency, in part due to dramatic improvements in performance enabled through CMOS technology.

However, the gradual reduction in energy efficiency improvements over successive CMOS technology

generations, which has limited performance growth, combined with the rising cost of energy, has renewed interest in getting software to better utilize hardware resources.

## ENERGY WASTE DUE TO RUNTIME BLOAT

In contrast to programs tuned for a specific use, large software systems are standardized around deeply layered frameworks (or modules) that facilitate rapid development. Each layer is designed to ensure composability of its functions to support a high degree of flexibility for reuse and interoperability. In a typical execution scenario, the system uses only a small subset of functions but still pays the overhead for supporting full functionality. With more layers, the number of potential function combinations grows exponentially, compounding the hidden burden of largely unused combinations.

Software engineering researchers have noted many forms of such

runtime bloat—runtime resource consumption disproportionate to the actual function being delivered—including the execution of excess function calls, the generation of excess objects, and the creation of excessively large data structures. For example, Nick Mitchell, Edith Schonberg, and Gary Sevitsky cited hundreds of thousands of method calls and new objects to service a single request and the consumption of a gigabyte of memory per hundred users in an application that needs to scale to millions of users (“Four Trends Leading to Java Runtime Bloat,” *IEEE Software*, vol. 27, no. 1, 2010, pp. 56-63).

Our review of several case studies supports these findings: a document-exchange gateway that creates six copies or transformations for each input document processed, a telecom application generating a megabyte of temporary objects per transaction, and so on.

Overutilization of resources due to software runtime bloat can result

## GREEN IT

in higher power consumption and wasted energy.

### ZERO-BLOAT SOFTWARE DESIGN

To date, server energy-optimization efforts have largely focused on the design of energy-efficient hardware and energy-aware thermal and power management techniques. Energy-proportional design has gained interest as a principled approach to achieve significant energy savings. However, reducing energy waste due to software inefficiencies requires a new perspective.

There's a parallel between energy waste due to hardware overprovisioning and that arising from the functional overprovisioning of software. Enterprise applications must support extremely demanding levels of variability and interoperability, but they actually exploit only a small fraction of this versatility in a typical deployment situation. Minimizing the runtime energy expended on built-in provisional generality is critical to achieving zero-bloat (green) software.

Of course, some costs in provisioning for generality can't be completely eliminated. Further, we don't yet fully understand how to model all such costs to enable systematic approaches for measuring and eliminating associated overhead.

It might be possible to draw a lesson from architecture research, where similar problems have been studied in a more structured setting. For example, researchers with Stanford's ELM (Efficient Low-Power Microprocessor) project discovered that it's mostly data and instruction supply overhead in general-purpose programmable embedded processors, not inefficiencies in the core logic, that account for the 50× energy-efficiency gap between these processors and hard-wired media ASICs (W.J. Dally et al., "Efficient Embedded Computing," *Computer*, July 2008, pp. 27-32). In this project, optimizing

instruction and data-supply energy costs improved energy efficiency by 23×, closing the gap with ASICs to within 3×.

Is the overhead expended in getting to the desired data and logic also a key source of bloat in software? Can we address these inefficiencies to help bridge the gap between large framework-based applications and custom-built programs?

### MODELING AND MEASURING BLOAT

Software doesn't come with built-in labels that indicate which portions of computation are necessary for a given

**Current state-of-the-art techniques can't measure how much overall excess resource consumption and energy waste are attributable to bloat.**

application and which lead to bloat. Estimating the amount of resources a nonbloat implementation would've consumed for a specific execution is also difficult.

Understanding the nature and sources of different types of software bloat is the first step to addressing the issue. The second is to quantify the magnitude of excess resource consumption attributed to each type of bloat. While the latter enables an estimate of how much room for improvement exists, the former provides insights on how to fix the problem.

In their study of large framework-based applications, Mitchell, Sevitsky, and Harini Srinivasan observed that information-flow patterns for a data record involved a sequence of expensive transformations with multiple levels of nesting ("Modeling Runtime Behavior in Framework-Based Applications," *Proc. 20th European Conf. Object-Oriented Programming*

(ECOOP 06), LNCS 4067, Springer, 2006, pp. 429-451). For example, moving a single date field from a SOAP message to a Java object in a stock-brokerage benchmark involved 58 transformations and generated 70 objects. Many of these were facilitative transformations for reusing existing parsers, serializers, and formatters. The observations highlight the considerable overhead expended in supplying data to the application's core business logic.

Measures of bloat rely on different heuristics to distinguish incidental overhead due to a specific category of bloat from strictly necessary resource usage.

When analyzing Java heap snapshots, for example, it's easy to differentiate data-structure representation overhead such as bytes expended by the JVM object headers, pointers (object references), collection glue, and bookkeeping fields from the actual application data contained in the structures. Data-structure health signatures, a relative measure of memory consumed by actual data versus associated representational memory overhead, reveal that data-structure bloat can increase the memory footprint of long-lived heap data structures by anywhere between a factor of two and five (N. Mitchell and G. Sevitsky, "The Causes of Bloat, the Limits of Health," *Proc. 22nd Ann. ACM SIGPLAN Conf. Object-Oriented Programming Systems and Applications* (OOPSLA 07), ACM Press, 2007, pp. 245-250).

From a power-consumption standpoint, execution bloat and associated excess temporary-object generation are even more interesting. Researchers have made good progress in diagnosing some signs of potential bloat in the use of temporary objects—for example, excessive or expensive data copies and the creation of expensive data structures with low utility (G. Xu et al., "Finding Low-Utility Data Structures," *Proc. 2010 ACM SIGPLAN*



*Conf. Programming Language Design and Implementation (PLDI 10)*, ACM Press, 2010, pp. 174-186). Even so, current state-of-the-art techniques can't measure how much overall excess resource consumption and energy waste are attributable to bloat.

## MITIGATING AND AVOIDING BLOAT

Tackling the source of bloat usually involves manual source-code fixes and assumes some domain knowledge about the application. In a few cases—for example, when the bloat originates in inefficient data structures—advisory tools can minimize manual effort.

Automatic code-optimization techniques can mitigate the symptoms of bloat. For example, static object reuse transformations help reduce the generation of excessive temporary objects due to bloat by amortizing the overhead of repeated object creation (S. Bhattacharya et al., “Reuse, Recycle to De-bloat Software,” *Proc. 25th European Conf. Object-Oriented Programming (ECOOP 11)*, LNCS 6813, Springer, 2011, pp. 408-432). We increased energy efficiency by up to 59 percent with this transformation using the SPECpower\_ssj2008 benchmark on an IBM HS21 blade server.

The traditional maxim for creating lean software, as advocated by David Parnas and Niklaus Wirth, among others, is to engineer it right by adopting minimalist design principles that avoid bloat. Such software is built in a series of stepwise refinements carefully crafted to provision each potential use case without sacrificing extensibility or reuse. The Linux kernel illustrates the successful adoption of this principle to efficiently satisfy diverse environments and requirements.

In framework-based environments, however, this approach is impractical. Many redeployable components must be dynamically programmable by business analysts and integrate with

dozens of heterogeneous systems and information sources. It's thus not easy to anticipate usage of a component at design time, nor is it feasible to incrementally change intermediate interfaces later.

Consequently, we propose designing software, programming models, and runtime systems in a way that makes it easier to detect and mitigate bloat. Programmers are often unaware of the overhead that systems might incur during actual deployment, and a low-level runtime optimizer can't “know” their intentions. Improving cross-layer line of sight into high-level functional intent and interoperability overhead—for

**The impact of bloat strongly depends on physical resources' energy proportionality.**

example, data-supply inefficiencies such as transformations and copies to facilitate reuse—will help both programmers and runtime systems deliver better energy-optimization solutions.

## BLOAT AND ENERGY PROPORTIONALITY

Understanding the causes of software bloat and addressing the challenges to eliminate it are important. However, bloat is unlikely to be completely eradicated where the design focus isn't just on efficiency but also on flexibility and function. Consequently, it's equally imperative to understand the exact impact of bloat on system power and energy consumption. Our work using the SPECpower\_ssj2008 benchmark (*Power-Performance Implications of Software Runtime Bloat: A Case Study with the SPECpower\_ssj2008 Benchmark*, tech. report RC25150, IBM Research, 2010) revealed several interesting observations in this regard.

The impact of bloat is closely tied to its effect on the utilization of physical resources and whether each such resource is an execution bottleneck. It also strongly depends on the resources' energy proportionality—that is, their utilization-to-power characteristics.

Reducing bloat that affects a nonbottleneck resource usually decreases system power and energy consumption, the extent of which depends on the change in the resource's usage and its utilization-to-power characteristics. Reducing bloat that affects a bottleneck resource can increase system throughput, thereby improving energy efficiency and sometimes increasing system power. However, it's theoretically possible for energy efficiency at the increased throughput to be lower if there's a disproportionately high cost in power with increased usage of the previously underutilized resources.

In assessing the impact of bloat reduction on a bottleneck resource's power and energy consumption, it's important to compare the resource's efficiency at

- *peak achievable performance* with and without bloat, and
- *equiperformance*—that is, at the same performance point—with and without bloat.

The “Power-Performance Impact of Bloat” sidebar illustrates these metrics in more detail. An even more comprehensive analysis by the authors can be found in “The Interplay of Software Bloat, Hardware Energy Proportionality, and System Bottlenecks,” to appear in *Proc. 4th Workshop Power-Aware Computing and Systems (HotPower 11)*, 2011.

In the equiperformance case, bloat can cause a steep increase in power consumption if the underlying hardware has superlinear energy proportionality. For example, we found

## GREEN IT

## POWER-PERFORMANCE IMPACT OF BLOAT

**F**igure A illustrates the power-performance impact of software bloat in the presence of resource bottlenecks and different utilization-to-power characteristics of resources. Figure A1 shows the scenario without bloat, Figure A2 shows the impact of bloat when the bottleneck resource is at its peak utilization, and Figure A3 shows the impact of bloat when the system is at equiperformance.

The diagrams plot the execution time, power, and energy for software that uses three types of resources. R1 is the most power-hungry resource, with a significant utilization-to-power characteristic—for example, a CPU that uses dynamic voltage and frequency scaling. R2 is the bottleneck resource for this workload. R3 is a resource with a bimodal power characteristic—for example, aggressively power-managed memory—using on/off power modes at discrete intervals of coarse granularity. The x-axis represents execution time, with the length of the blue bars indicating the resources' service-time demands. The z-axis represents power, with the areas over the bars in light and dark brown shading, respectively, indicating energy consumed (power × time) in the scenarios without bloat and incrementally with bloat.

Bloat in bottleneck resource R2 reduces peak achievable perfor-

mance. Figure A2 shows how this slowdown causes the superlinearly energy-proportional resource R1 to be underutilized (increasing available slack), steeply decreasing its power consumption. The bloat in R3 causes it to be switched on for more time, rounded up to the granularity of its power-management interval. Peak power consumption decreases. The net change in energy consumption can be computed as the difference between the shaded areas for the two scenarios A2 and A1. Note that the energy can increase or even decrease depending on the steepness of power scale down and the granularity of power-switching decisions for R3.

In Figure A3, the bloated resource R2 has been scaled up in performance and power—for example, by changing its operating mode—to have the same execution time as the software without bloat. In this equiperformance scenario, there's no slowdown due to bloat and thus no change in power consumed by R1. Peak power increases and so does energy consumption; the change can be computed by summing the areas that are shaded dark brown. In the event the power for R2 scales significantly higher with its usage (as with R1), the energy cost of bloat—denoted by the increased shaded area—will be even higher than shown.

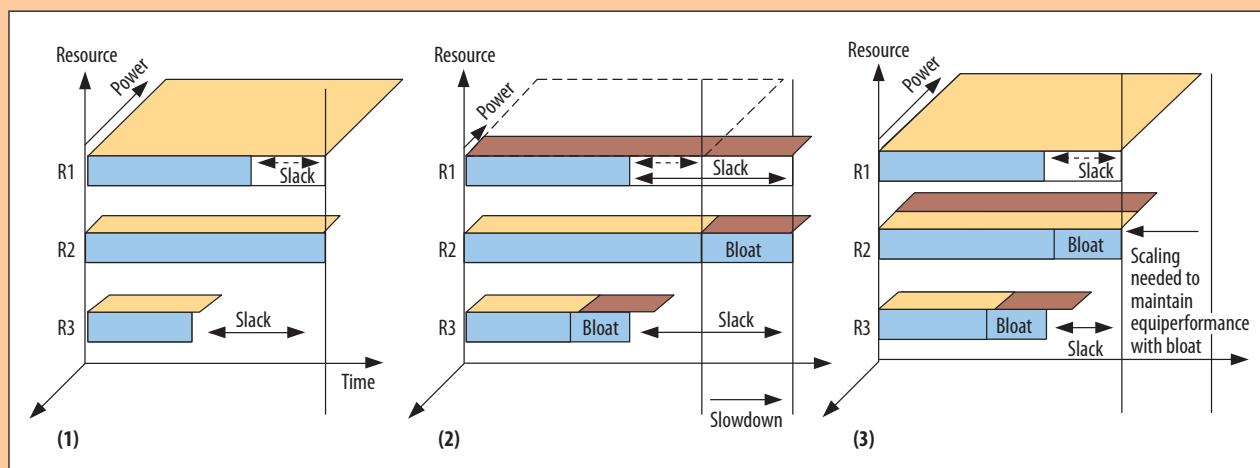


Figure A. Power-performance impact of bloat: (1) software without bloat; (2) bottleneck resource at peak utilization; (3) bottleneck resource with room to scale to maintain equiperformance.

that on an IBM Power 750 system with an aggressive load-based energy-saving solution (dynamic voltage and frequency scaling), a reduced-bloat implementation is 1.8× more energy-efficient than a heavily bloated one at equiperformance, though it's only 1.28× more energy-efficient at peak performance.

Both peak power and equiperformance power have important implications in a datacenter context. In general, peak-power comparisons reflect changes in power-provisioning costs, while equiperformance-power

comparisons are good indicators of operational energy cost impacts, particularly in systems with energy-aware management that would run the system at the lowest operating point needed to meet the service-level agreement (SLA). However, equiperformance power can also play a role in power-provisioning considerations. Consider a cluster that is sized for a certain aggregate throughput in which the power delivery limits the individual server's performance. In this situation, when there's a change due to bloat reduction, equiperformance

power would have a direct impact on power provisioning for the cluster.

**M**odularity is fundamental to the composability of software packages and to their rapid development and deployment. However, the prevalent approach to achieving it can lead to significant software bloat, which is detrimental to power, performance, and energy efficiency. The real issue isn't modularity itself but that, because of the difficulty in modularizing functions exactly as

needed, programmers inadvertently introduce superfluous processing and data overhead for reuse.

Understanding the sources and nature of bloat is an important start in addressing the problem. Techniques for measuring, managing, and mitigating bloat face significant challenges but there has also been progress in these areas during the past few years.

Our work demonstrates that the energy impact of bloat isn't trivial either. It also shows that relations between bloat, bottlenecks, and hardware power characteristics determine the exact impact of bloat on energy efficiency. Consequently, adopting an

integrated analysis of software bloat and hardware platforms is necessary to realizing modular software that's also green. **□**

*Suparna Bhattacharya is a PhD student in the Department of Computer Science and Automation at the Indian Institute of Science, Bangalore, as well as a senior technical staff member at the IBM India Systems and Technology Lab. Contact her at [suparna@csa.iisc.ernet.in](mailto:suparna@csa.iisc.ernet.in).*

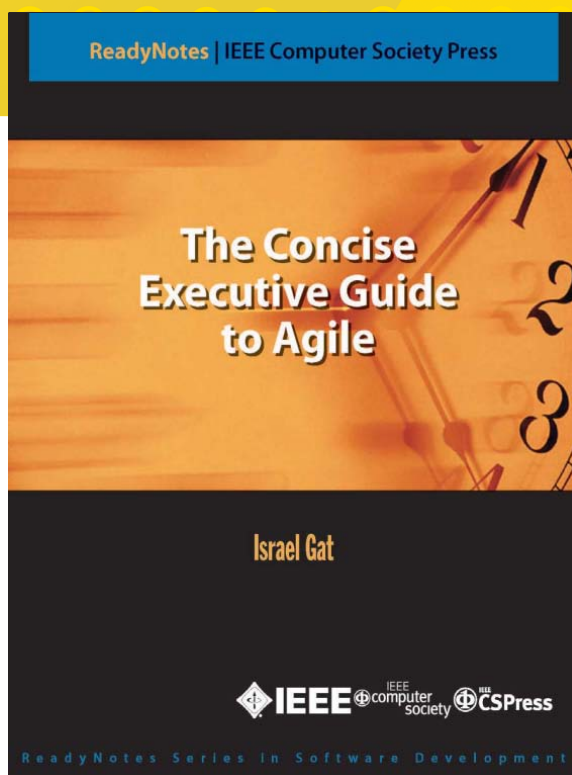
*K. Gopinath is a professor in the Department of Computer Science and Automation at the Indian Institute of Science. Contact him at [gopi@csa.iisc.ernet.in](mailto:gopi@csa.iisc.ernet.in).*

*Karthick Rajamani is a research staff member and manager of the Power-Aware Systems Department at IBM Research—Austin, Texas. Contact him at [karthick@us.ibm.com](mailto:karthick@us.ibm.com).*

*Manish Gupta is director of IBM Research—India and chief technologist of IBM India/South Asia. Contact him at [manishgupta@in.ibm.com](mailto:manishgupta@in.ibm.com).*

**Editor: Kirk W. Cameron, Dept. of Computer Science, Virginia Tech; [greenit@computer.org](mailto:greenit@computer.org)**

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



**NEW** from  **CSPress**

## THE CONCISE EXECUTIVE GUIDE TO AGILE

by Israel Gat

Get the tools and principles you need to lead an Agile transformation at your organization in this short and practical handbook, delivered digitally right when you need it.

PDF edition • \$15 list / \$12 members • 21 pp.

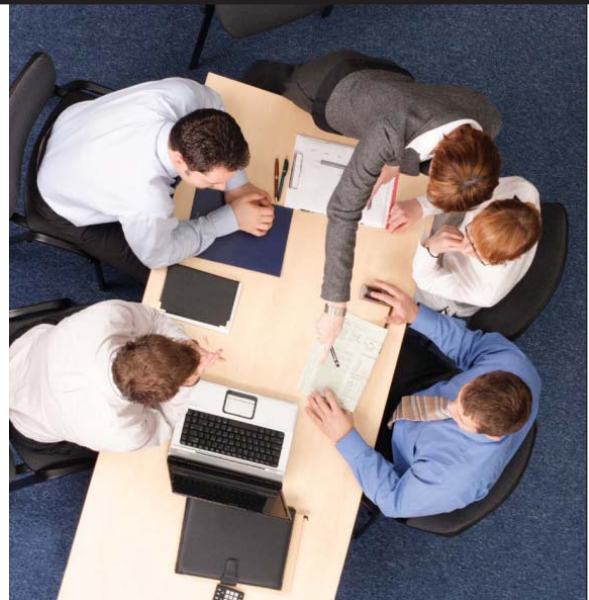
Order Online:  
[COMPUTER.ORG/STORE](http://COMPUTER.ORG/STORE)



## IN DEVELOPMENT

# Onshore Mobile App Development: Successes and Challenges

Christopher L. Huntley, *Fairfield University*



Developing mobile apps on multiple platforms requires innovative designers who can mold them into a native and immersive experience.

**D**uring the past two years, Applico, a mobile app development company based in New York City, has grown from a dorm room start-up to a successful business with 45 employees (75 by the end of 2011) and a client list that includes several household names, all without seeking outside venture funding.

An interview with Applico CTO Matt Powers reveals the factors contributing to this young company's remarkable growth and success.

**Huntley:** What does Applico do, and what distinguishes it from its competition?

**Powers:** Applico builds custom mobile applications on connected devices. We work with many large, multinational companies, including GM, AT&T, and Pearson Publishing. All of our resources are in the US. We don't offshore or outsource any of our work. We also do work for the US government, and our senior engineers have experience working on software for the Department

of Defense and the Office of Naval Research.

Applico has evolved into a mobile consulting firm as well. Our strategy and design assets are being utilized to a greater extent on every project. Our clients see a great deal of value in our knowledge of how to apply mobile technology to their respective business models.

## DEVELOPING AN ONSHORE BUSINESS

**Huntley:** Most of your developers are fresh out of school, but I assume that you've been at it a bit longer. How did you come to be CTO at a tech start-up?

**Powers:** My background is primarily in the defense industry, going back as far as working on missile defense systems for the US Army. I was introduced to the Android mobile platform while working on a situational awareness tool for the Office of Naval Research and the US Navy. As a result, I was tasked to grow the mobile aspect of the R&D group at Textron Systems from 2008 to 2010. During that time, I was exposed to a

lot of edge-case scenarios and had the opportunity to push the limits of these mobile platforms.

However, while working in this R&D environment was great, I didn't want to be locked into working in one problem space for an extended period of time. I began to explore other career paths so that I could leverage my mobile experience in the commercial arena. Applico presented the perfect opportunity for me to use my development skills as well as my formal software engineering training in a space in which I could not only be a technical lead, but also grow a development team and culture, set around Applico's core business values.

**Huntley:** With so many shops subcontracting work to offshore providers to manage costs, how do you justify the cost of doing the work in-house in New York City?

**Powers:** The offshore versus onshore debate has been going on for more than a decade. However, the discussion is completely different with regard to mobile development.

There are five primary reasons why our model is more beneficial to our clients:

- improved communication,
- potential for agile software development,
- decreased risk—clients can get a much better feel for the project's current status,
- timely resolution of questions since everyone is in a comparable time zone, and
- quality control.

In an environment in which time to market is so critical, these factors add up to make Applico's onshore model outperform the competition.

As an aside, some of these points come directly from our existing clients who have already tried going offshore and have stated that they would never do it again. We've heard many horror stories about our clients having offshored a project and finding that it comes back in a shambles.

Sometimes we can use the existing code base, but it's usually more cost-effective to start from scratch. We already have core groups in Boston, Massachusetts, and Austin, Texas, because of the large supply of programming talent available in those areas, and we don't see any need to go offshore to find more talent.

**Huntley:** You also do most of your graphic and application design in New York City. Will that likely change as the company grows? Also, were you able to leverage the 1990s "Silicon Alley" talent, or does that just not exist anymore?

**Powers:** We expect to keep our base operations in New York City because our core business is centered there and we have had great luck in finding unbelievable mobile talent in this area. However, we also have tentatively scheduled plans for expansion into other parts of the country in 2012.

In terms of leveraging old design talent, we haven't really gotten any traction on that front. Most of our in-house creative talent is young developers who have specific mobile experience. We need designers who have fundamental knowledge of all the platforms. There is a uniqueness to each platform, and our design team needs to be intimate with each one.

### CURRENT PROJECTS

**Huntley:** How many projects do you have in development right now? How do you manage it all?

**Each project is put through the same rigorous quality assurance tests regardless of complexity or size.**

**Powers:** Currently, we have 20 to 30 mobile apps actively under production. In addition to having multiple teams for each operating system, we also have some client-specific teams that comprise cross-platform engineers, designers, and product and project managers.

At first, it certainly was a challenge, but Applico got to a critical stage in its growth that required thinking about formal processes across the business as a whole. Having exposure to CMMI level 4/5 companies like Textron and Raytheon provided me with the experience and exposure I needed to bring a similar type of structure to Applico.

**Huntley:** What can you tell us about your latest client project?

**Powers:** We've recently been working with a global relocation and travel company to develop an internal app. It's one of the first instances in which a global, multibillion-dollar company is taking a step back when thinking about mobile. The company wants

to evaluate how mobile can change its entire business model instead of just a few processes. Thus far in the industry, we've seen many businesses move a few specific tasks to mobile, like using an app to cash a check or scan a boarding pass. It's our responsibility to formulate a strategy for moving our clients into mobile, looking not at just a handful of processes but at tens of thousands.

**Huntley:** As you indicated, your clients have included numerous big companies, such as Texas Instruments, Buick, AT&T, and *Men's Health Magazine*. How is working with large clients different from working with smaller ones?

**Powers:** Our marketing efforts are geared toward large, multinational companies as they're our primary clientele. Compared to them, we're a very small company. Thus, our marketing needs to show them how and why we're the best at what we do, as well as display our experience to back that up.

From a development standpoint, each client is treated exactly the same. For every project, unless dictated otherwise by our client, we create wireframes, mock-ups, and detailed requirements before even starting to code. Our internal project plans are all created with the same internal goals in mind. Additionally, each project is put through the same rigorous quality assurance tests regardless of complexity or size.

Adhering to this standard baseline ensures that quality won't suffer across projects for our smaller clients. For larger clients, it's a matter of putting the right team and resources together to build bigger systems properly.

**Huntley:** Among the projects you've completed so far, which was the most challenging?

**Powers:** In software, there are two things that scare me: the unknown

## IN DEVELOPMENT

and open source. I'm a proponent of leveraging as much open source as possible. But sometimes we don't know exactly how the software is constituted until we start using it.

One particular project was based on three open source XMPP frameworks in which we were trying to create a communication system spanning multiple platforms. The challenge came in the immaturity of the libraries with us filling in the gaps, all while trying to manage our clients' expectations and their tight deadlines. There were many start-up growing pains on this project, but I think it forced us to look at our internal policies and mature as a company.

**Huntley:** Given your experience with open source and the predilections of your larger corporate clients, what's your opinion of agile development?

**Powers:** We actually encourage our clients who are new to mobile to use agile development on their projects. These clients usually don't know much about mobile and really aren't sure what they want.

I generally ask our clients these questions: Would you rather save a little cost and get your product to market marginally faster, but not have something that you are totally happy with? Or would you rather be involved in the product from start to finish, have a more active role, and really get to participate in creating your vision?

Our clients usually choose the second option. Cartus is a perfect example of this. They're changing their entire business model, and

they need to be hands-on throughout the project. Agile provides us that flexibility.

### THE BIG PICTURE

**Huntley:** In your role as CTO, what do you see as the biggest challenges facing Applico today and going forward?

**Powers:** From a technical standpoint, as for any software company, our primary challenge is going to be finding the right talent. Our CEO, Alex Moazed, has been deeply involved in hiring and has been building our recruitment team very aggressively. Finding engineering talent stateside is always going to be an issue, and we must be smart in making hiring decisions.

Experienced engineers are important to us, but what we're really looking for are experienced engineers who live and breathe mobile. Mobile is still in its infancy, so finding that seasoned veteran with exposure to a mobile platform is difficult. That's why we've expanded our net, focusing not just on mobile experience but on finding talent where the problem space is the same—limited resources, poor Internet connectivity, and so on.

We've found that radio engineers, sensor network engineers, and, to some extent, back-end legacy Web developer skill sets translate well into the mobile space. We've had good success with training these people on developing mobile platforms and will probably continue doing so until mobile becomes even more mainstream.

Dealing with an incomplete and immature development environment presents unique challenges. Again, this emphasizes the importance of the diversity of the people we hire at Applico. We all have unique backgrounds—audio engineers, radar engineers, neural scientists—and chances are that one of us has seen some of the issues we encounter on Android/iOS/BB in a “past life” while working on an entirely different system. This perspective gives us an edge when working through the growing pains associated with developing a new platform.

Additionally, we're extremely active in developer groups across the major platforms, and we regularly host mobile meet-ups in which developers are encouraged to share some of their war stories in a social setting. This collaborative environment tends to shed light on common problems and reveal potential workarounds on any given platform.

Developing on multiple platforms is always going to be a challenge. In addition to the technical aspects such as device fragmentation, different hardware, different OSs, it's up to us to educate our clients about what makes a good app.

The user experience with the iPhone, Android, Blackberry, Windows Phone, and various tablet implementations is fundamentally different. Therefore, we not only need rock-star mobile developers, but also amazing designers who can mold the app into a native and immersive experience for each platform. **■**

*Christopher L. Huntley, In Development column editor, is a professor and chair of the Department of Information Systems and Operations Management at the Charles F. Dolan School of Business at Fairfield University. Contact him at [chuntley@mail.fairfield.edu](mailto:chuntley@mail.fairfield.edu).*

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

build your career  
**IN COMPUTING**

[www.computer.org/buildyourcareer](http://www.computer.org/buildyourcareer)





TIMELY, ENVIRONMENTALLY FRIENDLY DELIVERY

# DIGITAL EDITIONS

Keep up on the latest tech innovations with new digital editions from the IEEE Computer Society. At **more than 65% off regular print prices**, there has never been a better time to try one. Our industry experts will keep you informed with the latest technical developments in a format that's timely and environmentally friendly.

- Easy to Save. Easy to Search.
- Email notification. Receive an alert as soon as each digital edition is available.
- Two formats. Choose the enhanced PDF edition OR the web browser-based edition.
- Quick access. Download the full issue in a flash.
- Convenience. Read your digital edition anytime—at home, work, or on your mobile.
- Digital archives. Subscribers can access the digital issues archive dating back to January 2007.

From software architecture to security and networking, these magazines help you stay ahead of the competition:

- *Computer*—our flagship magazine with broad technical coverage
- *IT Professional*—critical IT topics
- *IEEE Software*—practitioner-oriented trends and applications
- *IEEE Security & Privacy*—computer security and data privacy
- *IEEE Internet Computing*—emerging and evolving Internet technologies

Interested? Go to [www.computer.org/digitaleditions](http://www.computer.org/digitaleditions) to subscribe and see sample articles.



## IDENTITY SCIENCES

# What Are Soft Biometrics and How Can They Be Used?

**Karl Ricanek Jr. and Benjamin Barbour**  
 University of North Carolina Wilmington



Face characteristics can be used for gender identification and age estimation in a wide range of biometric applications.

**T**oday it's difficult to find anyone who isn't aware of biometrics and its growing role in everyday life. Many countries have implemented or plan to implement biometric systems for national ID cards, border security, immigration control, and law enforcement, and biometric systems are now used in many retail stores, banks, government facilities, and even school cafeterias. The January 2011 Identity Sciences column ("Dissecting the Human Identity," pp. 96-97) discussed the most ambitious biometric endeavor to date: India's universal identification (UID) program, which will link a 12-digit number to a biometric trait for automatic identification of more than 1 billion citizens.

The emerging field of *soft biometrics* is exploring exciting new questions about biometric traits: Can certain traits be leveraged to determine age or ethnicity? What about health and mental state? Can they reveal intent or deception? Two especially promising application areas are gender identification and age estimation.

## SOFT BIOMETRICS

Soft biometrics are "characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals" (A.K. Jain, S.C. Dass, and K. Nandakumar, "Soft Biometric Traits for Personal Recognition Systems," *Proc. Int'l Conf. Biometric Authentication (ICBA 04)*, LNCS 3072, Springer, 2004, pp. 731-738).

The face, in particular, can provide rich information about a person: the size and geometry of the chin, lips, nose, eyebrows, and other face components can be used to distinguish gender, race, and ethnicity, while creases, lines, sagging, and wrinkles can reveal clues about age. Researchers are exploring other characteristics and features such as gait, the hand, the iris, and the periocular region (area around the eye) to identify soft biometrics, but most work to date has focused on the face.

Cognitive biases sometimes make it difficult for human observers to "read" a person's face. For example, people are generally better at recognizing and characterizing those of their own race (own-race

bias) and approximate age (own-age bias). Further, context can erode human performance on age, gender, and race/ethnicity determination tasks. For example, in experiments conducted by the Face Aging Group at the University of North Carolina Wilmington (UNCW), observers misclassified the gender of young children in photos due to the color of their clothing or the style and length of their hair (Y. Wang et al., "Gender Classification from Infants to Seniors," *Proc. 4th Int'l Conf. Biometrics: Theory, Applications, and Systems (BTAS 10)*, IEEE Press, 2010; doi: 10.1109/BTAS.2010.5634518).

People in certain occupations are often called upon to make such determinations. For example, store clerks must judge whether a customer who for some reason isn't carrying identification is old enough to purchase cigarettes or other age-controlled retail products. Likewise, political asylum seekers are typically treated differently based on their age, but immigration officials often can't obtain or verify credentials for this group.

For these reasons, the development of automated recognition systems

based on soft biometrics is a growing area of research.

### GENDER CLASSIFICATION

Automatic gender classification from face images has enjoyed renewed interest due to the availability of public face-image datasets, with ground truth and government face-image datasets disseminated as part of face-recognition challenges (P.J. Phillips, “Improving Face Recognition Technology,” *Computer*, Mar. 2011, pp. 84-86).

Current gender-classification systems can correctly identify the gender of 80 to 90 percent of young and middle-age adult faces, with some systems achieving 95 percent success on small data sets (S. Baluja and H.A. Rowley, “Boosting Sex Identification Performance,” *Int’l J. Computer Vision*, Jan. 2007, pp. 111-119; E. Makinen and R. Raisamo, “Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, Mar. 2008, pp. 541-547). However, these systems are less effective at identifying the gender of children and seniors. Children tend to have a generic face devoid of the sharp and soft facial features that distinguish gender, while wrinkles, creases, the loss of muscle tone, and fat-pad sagging soften and blur gender features in seniors.

To help identify someone’s gender, human observers often draw on contextual information such as hair length and style, clothing, jewelry, and cosmetics. However, classifying the gender of young children is difficult even with contextual clues, as Figure 1 shows.

An automated gender-classification system developed at UNCW impressively outperformed human observers (Y. Wang et al., “Gender Classification from Infants to Seniors”). In one experiment, 91.3 percent of 278 subjects correctly classified the gender of eight young children from face images only half



**Figure 1.** Example images of children from the FG-NET Aging Database used for human and computer gender classification. Correct gender from left to right: top row—male, female, female, male; bottom row—male, female, male, male.

**Table 1.** Mean absolute error (MAE) rate for age-estimation algorithm on FG-NET Aging Database.

Age range	On training database		On testing database	
	No. of photos	MAE	No. of photos	MAE
0-20	588	1.09	140	1.93
21-69	214	1.37	60	5.80
0-69	802	1.78	200	4.37

the time, while the system correctly classified 67.5 percent of the images. On the FG-NET Aging Database ([www.fgnet.rsunit.com](http://www.fgnet.rsunit.com)), a longitudinal dataset of scanned and native digital images of 82 subjects, the system successfully classified 78.1 percent of children 0-10 years old, 88.9 percent of seniors aged 56 and older, and 92.1 percent of adults 19-55 years old.

### AGE ESTIMATION

Researchers at the University of Illinois at Urbana-Champaign, West Virginia University, and UNCW are leading efforts in automated age estimation from face images. The performance of these systems has improved dramatically during the past decade: for queries evaluated on the FG-NET database, for example, the mean absolute error (MAE) between true age and estimated age has tumbled from nearly 10 years to under 4 years.

Age-group estimation involves determining the general age group to which a subject belongs—for

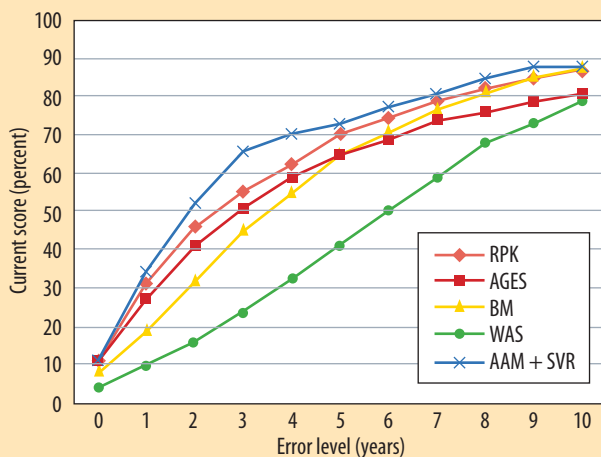
example, children, teens, adults in their twenties, seniors, and so on. Table 1 compares the MAE of one age-estimation algorithm—active appearance models with support vector regression (AAV+SVR)—for different age groups using the FG-NET database (K. Lu et al., “Age Estimation Using Active Appearance Models and Support Vector Machine Regression,” *Proc. 3rd Int’l Conf. Biometrics: Theory, Applications, and Systems (BTAS 09)*, IEEE Press, 2009, pp. 314-318).

Note that the MAE for adults is significantly higher than that for children. The MAE of 1.93 for children pulls the overall MAE down to 4.37 years from 5.8 years for adults. When evaluating age-estimation systems, it’s important to know the algorithm’s sensitivity for all age groups; hence, MAE per decade is a better measure than overall MAE.

Cumulative score is a good graphical measure of the performance of an age-estimation system in comparison to other systems against the same data set. The cumulative score



## IDENTITY SCIENCES



**Figure 2.** Cumulative score graph comparing various age-estimation algorithms on the FG-NET database. The curve of the best-performing system—AAM+SVR—is to the left of all others.



**Figure 3.** Comparison of true age of several subjects from the PAL database with their age as determined by an age-estimation algorithm.

is defined as  $CS(j) = N_{e \leq j} / N \times 100$  percent, where  $N_{e \leq j}$  is the number of images for which age estimation makes an MAE no larger than  $j$  years. The best-performing system's curve will be to the left of all others.

Figure 2 plots the cumulative score for AAM+SVR and four other age-estimation algorithms on the FG-NET database: regression from patch kernel (RPK), AGing pattErn Subspace (AGES), a bilinear model (BM), and Weighted Appearance Specific (WAS). In this case, AAM+SVR is clearly the most effective algorithm.

Perceived-age estimation attempts to mimic the human approach to classifying people by how old they look, as opposed to identifying their true chronological age. Figure 3 compares the true age of several subjects from the original Productive Aging Lab (PAL) database to their

age as determined by a UNCW age-estimation algorithm. As the figure shows, the algorithm is extremely accurate (W. Yang et al., "Ensemble of Global and Local Features for Face Age Estimation," *Proc. 8th Int'l Conf. Advances in Neural Networks* (ISNN 11), LNCS 6676, Springer, 2011, pp. 251-257).

Age-group and perceived-age estimation are especially useful to corporate researchers developing automatic demographic systems for deployment in marketing and retail sectors, and in most cases a robust age-estimation system is effective for these purposes.

As interest in age estimation has grown, new research databases have emerged. The Center for Vital Longevity Face Database at the University of Texas at Dallas (<https://pal.utdallas.edu/facedb/request/>

[index](#)) consists of studio-quality photo images of an ethnically diverse population of adults ranging from 18 to 93 years, with ground-truth age and gender. UNCW's Craniofacial Longitudinal Morphological Face Database ([www.faceaginggroup.com/projects.html#morph](http://www.faceaginggroup.com/projects.html#morph)) consists of ethnically diverse, longitudinal mug-shot images (scanned and native digital) of adults from 18 to 77 years of age and is the largest publicly available face-image dataset with ground-truth age and gender.

**S**oft biometrics for automated gender identification and age estimation have potential applications in a multitude of areas.

Gender identification can be used as a filter for digital photo albums and in digital signage to customize advertisements. Iris-based soft biometrics could also be used to minimize the size of comparisons for iris identification, which would benefit UID programs for large populations like the one being deployed in India.

Age estimation for access control is another promising application area. Manufacturers are interested in adapting such systems for age-restricted products in vending machines. Law enforcement and online service providers are also actively pursuing age-estimation technology to help identify and eradicate child pornography. **C**

**Karl Ricanek Jr.**, *Identity Sciences* column editor, is an associate professor in the Department of Computer Science and director of the Face Aging Group at the University of North Carolina Wilmington. Contact him at [ricanekk@uncw.edu](mailto:ricanekk@uncw.edu).

**Benjamin Barbour** is a graduate student in the Department of Computer Science at the University of North Carolina Wilmington. Contact him at [beb1512@uncw.edu](mailto:beb1512@uncw.edu).

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

# Take the CS Library wherever you go!



All 2011 issues of IEEE Computer Society magazines and Transactions are now available to subscribers in the portable ePub format.

Just download the articles from the Computer Society Digital Library, and you can read them on any device that supports ePub, including:

- Adobe Digital Editions (PC, MAC)
- Aldiko (Android)
- Bluefire Reader (iPad, iPhone, iPod Touch)
- Bookworm Online Reader (Online)
- Calibre (PC, MAC, Linux)  
(can convert ePub to .MOBI format for Kindle)
- ibis Reader (Online)
- ePUBReader (FireFox add-on)
- iBooks (iPad, iPhone, iPod Touch)
- Nook (Nook, PC, MAC, Android, iPad, iPhone, iPod, other devices)
- Sony Reader Library (Sony Reader devices, PC, Mac)
- Stanza (iPad, iPhone, iPod Touch)

[www.computer.org/csdl/epub\\_info.html](http://www.computer.org/csdl/epub_info.html)



IEEE  computer society

## THE PROFESSION

*Continued from page 112*

of interest and got the resources they needed to do so.

In companies such as AT&T, innovation was delegated to an internal research center, often the pride of the company. They were willing to invest heavily in the hope that at least some of the projects would be a success. But centralized innovation has some drawbacks, especially in getting the inventions to the marketplace. A good illustration is Xerox's failure to exploit the GUI research at Xerox PARC.

The open source software movement is an interesting alternative. While the big companies adopt a monolithic approach, open source

for innovation. Google invests in an internal entrepreneurial model that encourages employees to innovate. The company, with its flat organizational structure, supports innovative employees by offering services, data resources, and tools. It offers free time for entrepreneurship and provides additional resources for the most interesting projects. Most importantly, through Google Labs and similar facilities, it offers a venue where new products can be tried in the marketplace.

### VIRTUAL GARAGES

HP's birth took place in a garage in 1939. Many other electronics,

### THE SANDBOX

Traditionally, in engineering terms, a sandbox environment consists of a controlled set of resources for trying a new app without the risk of damaging critical parts of the system.

As an innovation model, we can take this further, including the phase of getting the product on the market. This is what Google offers its employees. We call this a "tightly coupled extended sandbox." "Tightly coupled" because the entire process is internal to the company, and "extended" because the market is a part of the sandbox and entrepreneurs are taking advantage of the infrastructure already in place for the company itself.

For Apple, the sandbox innovation environment is the App Store, together with the developer network and all the tools that go with it. The App Store mechanism controls the quality of published apps and also provides a platform for reaching out to the end user. Moreover, Apple provides easy integration with iAds, an advertising framework that includes a payment mechanism for iPhone and iPad apps. This is a "loosely coupled sandbox"—Apple owns the infrastructure, but the entrepreneurs can be you or me or anyone else. The marketplace for Android apps offers similar services.

For all these models, the short time from idea, via implementation and quality control, to the market is crucial. This could explain why some feel that "everything has been invented." Many of us might previously have had ideas for new apps, but they usually stay in the inventor's head or never come out of the lab. Even when a new invention reaches the market, only a fraction of the potential user community notices most of them.

Today, with all the available tools, development is simpler, and nearly any app can be put on the market immediately. Marketing is an important factor. Through centralized channels, such as iAds or Android

### The sandbox environment offers tools, APIs, and other resources that simplify development.

is based on a decentralized model. In this model, entrepreneurs with good ideas can set up a development project, then they invite the rest of the world to participate.

The motivation is to develop something in common, something innovative that is more flexible and better than what the private companies offer, and also something that's free for anybody to use. Amazingly, many excellent products have come out of this movement, including OSs (Linux), Web servers (Apache), programming languages (PHP), and browsers (Firefox). Even the basic parts of the Internet can be attributed to open source development.

While the open source movement has demonstrated its efficiency in large systems and software modules, it might not be as effective for developing end user apps. The problem could lie in marketing—how to get apps to the customer.

In "Entrepreneurial Innovation at Google" (*Computer*, Apr. 2011, pp. 56-61), Alberto Savoia and Patrick Copeland describe another model

computer, and software companies had similar humble beginnings in which the entrepreneurs themselves put in the time and money required to get a company off the ground.

Today, we've moved from the physical to the virtual garage. A virtual garage is provided with extensive toolkits in the form of OSs, development packages, and open source code. In this way, device manufacturers have opened their systems up for third-party developers, giving them access to all the underlying functionality of the PC, tablet, or smartphone.

Thus, implementing our proposed reminder app is no big deal. An entrepreneur with a Java development tool has access to everything that's needed to develop the app, including access to display, files, and location coordinates. In principle, there's nothing new here. Apple invited third-party developers to contribute to the Macintosh nearly 30 years ago. Other manufacturers have also opened their systems for independent developers. What's new is the support that's offered for getting new apps to the end user.



Market, it's easy to find what we're looking for. Thus, it's not only a matter of the time to market, but also the time before the world at large knows what we've invented.

Clearly, the sandbox environment has much to offer developers. The tools, APIs, and other resources simplify the development. The marketing support and predefined tools for downloading and installation help with the last and most crucial part of the development process—reaching out to the customer.

The models also may offer an opportunity for generating revenue. But, if we exclude the few success stories, making money on apps might not be so easy in the long run. There are already half a million apps in the App Store and 200,000 on the Android market. Newcomers to this market could be forced to offer apps for free to be recognized, thus reducing the potential for revenue.

## FROM THE USER'S PERSPECTIVE

This development environment also might not be sustainable in the long run from the user's perspective.

For a given function, the user must choose from that overwhelming number of apps. For example, there are about 400 wakeup apps and 2,500 calendar apps in the App Store, with around the same number of similar apps on the Android market. Although the most popular apps are presented first, popularity might not be a very good indicator of quality.

After choosing an app, the user must agree to let it have access to phone resources. Few users will be able to evaluate the security risks, and most hit the okay button without thinking. Then the user must download the app, perhaps pay for it, and install it on the device. When moving to a new device, especially if the native OS is changed, the user must locate the apps, make selections, pay for them, and download them again.

On a particular smartphone or

tablet, each of the various apps is represented as an icon in the apps window. It's now up to the user to remember which app did what, locate the icon, and remember how to use it. Updating apps is the user's task. Even if they're developed under the same guidelines, the apps could have different user interfaces.

Donald Norman and Jakob Nielsen complain about "the misguided insistence by companies (e.g., Apple and Google) to ignore established conventions and establish ill-conceived new ones" ([tinyurl.com/37yyrtv](http://tinyurl.com/37yyrtv)). When replacing an app, the user might therefore find that the new version offers a different user interface and different functionality.

Thus, while many enthusiastic users have welcomed the app idea, it might not be a good solution for the less technologically oriented. What might suit these users much better is a device in which all the necessary apps are embedded in the OS. We can expect this to happen in the long run.

## THE FUTURE

The app market is arguably only a testbed for the major companies, an implementation of a digital ecosystem in which the apps can compete, using a survival-of-the-fittest scheme to find the most interesting functions and the most popular approaches to them.

By letting third-party developers create the initial versions, and offering them to the marketplace, the major companies have a very simple and profitable way to determine the functionality to include in the next version of the basic software. They can then use their resources to ensure that each of these apps, or each function, is offered in a high-quality version. As an example, the new iOS 5, Apple's mobile OS, will clearly replace several of the more popular apps.

An even more competitive approach than including app functionality as a part of the native

OSs is offering this functionality through browser-based systems (T. Mikkonen and A. Taivalsaari, "Reports of the Web's Death Are Greatly Exaggerated," *Computer*, May 2011, pp. 30-36). The advantages of centralized dynamic applications, true platform-independence, ubiquitous access, and no local installation or update of software could lead to the demise of both apps and large native OSs.

In the software world, will the browser, the dinosaur that can do everything, be the survivor?

Looking at a clear, dark, night sky, we see myriad stars. There are as yet not as many apps as there are stars, but each one might shine as brilliantly, showing the extent of human invention and covering every "dark spot," every possible function. But in the long run, numbers and brilliance might not be enough. In our everyday lives, we seldom take the time to look at the stars. Instead we focus on functionality, practicality, and effectiveness. While the stars will always be there, we aren't so sure that apps will survive in a practical world. 

*Alessio Malizia is an associate professor in the Computer Science Department at the Universidad Carlos III de Madrid, Spain. Contact him at [alessio.malizia@uc3m.es](mailto:alessio.malizia@uc3m.es).*

*Kai A. Olsen is a professor at the University of Bergen and Molde University College in Norway and an adjunct professor at the School of Information Services, University of Pittsburgh. Contact him at [kai.olsen@himolde.no](mailto:kai.olsen@himolde.no).*

**Editor: Neville Holmes, School of Computing and Information Systems, University of Tasmania; [neville.holmes@utas.edu.au](mailto:neville.holmes@utas.edu.au)**

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

## THE PROFESSION

# Has Everything Been Invented? On Software Development and the Future of Apps

**Alessio Malizia**, *Universidad Carlos III de Madrid, Spain*

**Kai A. Olsen**, *University of Bergen and Molde University College, Norway*



Because we're continually offered an abundance of new apps, we may fall into the trap of thinking that everything in software has been invented.

“Everything that can be invented has been invented.” Although this infamous 1899 quote has been attributed to Charles H. Duell, then director of the US Patent Office, he actually never said such a silly thing. What Duell said to the US Congress was something quite different, that America's future success depends on invention (<http://tinyurl.com/3pjas6u>). This clearly remains the case today, not only for America but for most countries. Invention and innovation continue to prosper, not least in the software industry.

The past decade has given us new tools and new ways of disseminating applications. As a consequence, we may fall into the trap of thinking that everything in software has been invented.

## SOFTWARE INNOVATION

Some time ago, we had a creative idea for a new app, a location-dependent “reminder.”

The idea was to attach a message to a location, instead of to the date and time. With this app, the next time you go near a hardware store, your smartphone would beep and tell you to get a new hammer, a message you may have entered some time ago. Or, when you read about a famous restaurant in Madrid, you store its homepage URL in the reminder. A year later, when you exit a train in the center of Madrid, the information about the restaurant will pop up on the phone.

To develop this app, we searched the Web for relevant information on APIs and so on, and started programming. Then we found that a similar app was already available for Google's Android operating system. It just hadn't been out when we first got the idea for our app.

Well, you can't beat them all, so we initiated work on a new idea, another breakthrough app for smartphones. We developed an initial design, but then we found an iPhone app that did

something similar. Very annoying! A contributor to DanTech, a blog on tablet and smartphone innovations, relates a similar experience. “I have independently come up with several ideas last year, and all of them are either patented by large PC companies like Apple and Microsoft, or similar ideas have already appeared in new startups ...” (<http://tinyurl.com/3jdlbu4>).

What has happened? Has everything that can be invented been invented?

## INNOVATION MODELS

Traditionally, large companies have used a centralized approach to maintain control over innovation. Bell Labs is a good example. In “Bell Labs and Centralized Innovation,” Tim Wu describes how Bell Labs was a “scientific Valhalla” for researchers and engineers working there (*Comm. ACM*, May, 2011, pp. 31-33). They were free to pursue their own areas

*Continued on page 110*

IEEE  computer society  IEEE

# Authoritative Cutting-Edge Comprehensive

With over 414,000 articles covering the spectrum of computer science and engineering, CSDL is the definitive resource for academic, corporate, or government libraries. Whether your users are looking for the latest research on today's hot topic, foundational information, or quick answers to a problem, they will find what they need.

**Learn more!** [www.computer.org/library](http://www.computer.org/library)

Your institution may qualify for special subscription discounts. Your institution may also qualify for a **FREE 30-day trial of the CSDL.**

Email [csdl@computer.org](mailto:csdl@computer.org) for more details.



**IEEE CS  
DIGITAL  
LIBRARY**





# We Live Quality

Software quality is in our DNA. For over 15 years, we've lived and breathed it. The reason is simple: Your software affects our friends, our families, and ourselves.

Whether it's the latest video game or a secure banking web site or the software that analyzes medical test results, we want it to work right because we rely on it.

From our expert Consulting and Agile Services teams, to our award-winning application lifecycle management (ALM) solutions, to our world-class customer support, Seapine has helped thousands of companies worldwide build, test, and deploy quality software.

Go with Seapine, and get serious about software quality.

[www.seapine.com](http://www.seapine.com)



© 2011 Seapine Software, Inc. All rights reserved.