RCS 3

3rd Rebooting Computing Summit

"Rethinking Structures of Computing"

Summary Report

Hilton Santa Cruz/Scotts Valley, California
October 23-24, 2014

Prepared By:

Alan M. Kadin

And the IEEE Rebooting Computing Committee

http://rebootingcomputing.ieee.org/

http://rebootingcomputing-ieee.blogspot.com/

November 20, 2014

Contents

Foreword

The Future Directions Committee (FDC) is a committee of the IEEE Technical Activities Board (TAB). Through volunteers from IEEE's Societies and Councils, FDC seeks to identify multidisciplinary topics in which IEEE can play a unique role for catalyzing and crystallizing goals and activities which increase the efficiency of developing the needed technologies of the future. Rebooting Computing (RC) is an ongoing initiative of the FDC, initiated in 2012, which proposes to rethink the computer through a holistic look that addresses all aspects of computing, both software and hardware, and make recommendations for future development. The RC Committee consists of volunteers from nine IEEE Societies/Councils and two professional IEEE staff directors. The RC committee organized a 1st Rebooting Computing Summit in December 2013 (RCS 1), bringing together a selection of thought leaders and decision makers from government, industry, and academia, to brainstorm ideas and lay initial foundations for Rebooting Computing. This generated a vision of future computing based on three pillars of Energy Efficiency, Security, and Human-Computer Interface. The 2nd Rebooting Computing Summit in May 2014 (RCS 2) focused on four initial technologies for further discussion, a mainstream approach of Augmenting CMOS, together with alternative approaches of Neuromorphic, Approximate, and Adiabatic Computing.

Together, RCS 1 and RCS 2 provided the basis for the 3rd Rebooting Computing Summit (RCS 3) held in Santa Cruz, CA, October 23-24, 2014. The theme was **"Rethinking Structures of Computation"**, and focused on the topics of **Parallelism**, **Security**, **Approximation**, and **Human-Computer Interface**. RCS 3 followed a similar format to RCS 1 and RCS 2, in which about 45 invited experts in a variety of fields heard plenary talks on each of these topics, and then broke up into smaller groups to discuss each topic in parallel. There was also a poster session, and a pre-announcement of a new government initiative in future computing research.

This Summary Report is intended not as a definitive technical report on RCS 3, but rather it reflects the presentation and discussions that took place at the Summit. The intention of the RC Committee is to engage the technical and scientific communities in a conversation about the best collaborative plans forward, and through IEEE activities of meetings, publications, and related events, to provide the key ingredients to accelerate the realization of the future of computing. The next step may be a conference on Rebooting Computing to be held in the next year.

The RC Committee also created a Web Portal (http://rebootingcomputing.ieee.org) and Blog (http://rebootingcomputing-ieee.blogspot.com), and we encourage interested parties to view these for additional information, videos of RCS 3 presentations and developing plans for a Rebooting Computing conference next year.

Elie Track and *Tom Conte*

Co-Chairs, IEEE Rebooting Computing

What Is "Rebooting Computing"?

Early computers required an initialization process to load the operating system into memory, which became known as "booting up," based on the old saying about "pulling yourself up by your own bootstraps." Even now, if a computer freezes up or overloads, a power cycle or "reboot" may be necessary to reinitialize the system. Can we apply this concept metaphorically to the entire computer industry?

"IEEE Rebooting Computing" is an inter-society initiative of the IEEE Future Directions Committee to identify future trends in the technology of computing, a goal which is intentionally distinct from refinement of present-day trends. The initiative is timely due to the emerging consensus that the primary technology driver for almost 5 decades, Moore's Law for scaling of integrated circuits, is finally ending. How can we continue to project further improvements in computing performance in coming decades? We need to review the entire basis for computer technology, starting over again with a new set of foundations (hence "Rebooting Computing").

Participating Societies and Councils

IEEE Computer Society (CS), Circuits and Systems Society (CAS), Council on Electronic Design Automation (CEDA), Council on Superconductivity (CSC), Electron Devices Society (EDS), Magnetics Society (MAG), Reliability Society (RS), Nanotechnology Council (NTC), and Solid-State Circuits Society (SSC). Also, coordination with International Technology Roadmap for Semiconductors (ITRS).

Co-Chairs of RC Committee:

- Elie K. Track, President CSC, nVizix LLC
- Tom Conte, President-Elect CS, Georgia Tech

Other Committee Members:

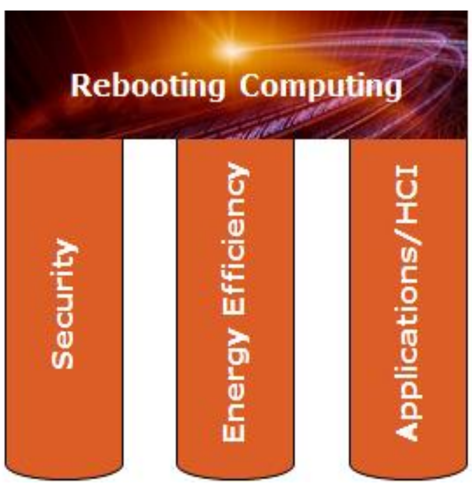
- Dan Allwood (MAG), University of Sheffield, UK
- Neal Anderson (NTC), University of Massachusetts, Amherst.
- David Atienza (CEDA), Ecole Polytechnique Federale, Lausanne, Switzerland
- Jonathan Candelaria (EDS), Semiconductor Research Corporation
- Erik DeBenedictis (CS), Georgia Institute of Technology
- Paolo Gargini (ITRS), Intel
- Glen Gulak (SSC), University of Toronto, Canada
- Bichlien Hoang, RC Program Director, IEEE Future Directions Staff
- Subramanian (Subu) Iyer (EDS, CPMT, SSCS), IBM
- Yung-Hsiang Lu (CS), Purdue University
- Scott Holmes (EDS), IARPA
- Alan M. Kadin (CSC), Consultant
- Arvind Kumar (EDS), IBM
- David Mountain (EDS, CS), NSA
- Oleg Mukhanov (CSC), Hypres, Inc.
- Vojin G. Oklobdzija (CAS), University of California at Davis
- Angelos Stavrou (RS), George Mason University
- Bill Tonti (RS), FDC Director, IEEE Future Directions
- Ian Young (SSCS), Intel

RCS 1: Future Vision and Pillars of Computing

The first Rebooting Computing Summit was held at the Omni Shoreham Hotel in Washington, DC, Dec. 11-12, 2013. This was an informal gathering of 37 invited leaders in various fields in computers and electronics, from industry, academia, and government, with several plenary talks and subsequent smaller breakout groups on several topics. The summary is available online at http://rebootingcomputing.ieee.org/RCS1.pdf. The consensus was that there is indeed a need to "reboot computing" in order to continue to improve system performance into the future. A future vision and three primary pillars of future computing were identified. While RCS 2 has moved on to address key technology issues, the vision and pillars remain central to the Rebooting Computing efforts.

Future Vision of Intelligent Mobile Assistant

One future vision for 2023 suggested in RCS 1 consisted of ubiquitous computing that is fully integrated into the lives of people at all levels of society. One can think of future generations of smartphones and networked sensors having broadband wireless links with the Internet and with large computing engines in "the Cloud". More specifically, one may envision a wireless "intelligent automated assistant" that would understand spoken commands, access information on the Internet, and enable multimedia exchange in a flexible, adaptive manner, all the while maintaining data security and consuming little electric power. And of course, such a wireless assistant should also be small and inexpensive. Such a combination of attributes would be enormously powerful in society, but these are not yet quite achievable for the current stage of computer technology.



Three Pillars of Future Computing

RCS 1 further identified 3 "pillars" of future computing that are necessary to achieve this vision: Energy Efficiency, Security, and Human-Computer Interface.

Human/Computer Interface and Applications

A better Human/Computer Interface (HCI) is needed that makes more efficient use of natural human input/output systems, particularly for small mobile units. Improved natural language processing is just one key example. While there is a long history of text I/O, this is not really optimal. Wearable computers analogous to Google Glass may provide a glimpse into future capabilities.

Energy Efficiency

The small wireless units operate on battery power, and it is essential that they consume as little power as possible, so that the recharging is relatively infrequent. Some computing can be shifted to the Cloud, but enhanced performance requires substantial improvements in energy efficiency. In contrast, the data centers and servers in the cloud are wired, but their power consumption can be quite large, so that electricity bills are a major cost of operation. Improved energy efficiency is critical here, as well.

Security

With data moving freely between wireless units and computers in the cloud, encryption and computer security are critical if users can expect to operate without fear of data diversion and eavesdropping.

RCS 2: Future Computer Technology – The End of Moore's Law?

RCS 2 consisted of a 3-day workshop May 14 - 16, at the Chaminade in Santa Cruz, CA. The summary is available online at http://sites.ieee.org/rcsummit/rcs2/ . The main theme of RCS 2 was on mainstream and alternative computing technologies for future computing, with four possible approaches identified. The format was similar to that for RCS 1, with a set of four plenary talks, followed by four parallel breakout groups culminating in outbrief presentations and concluding in a final plenary discussion. The primary conclusions were that focusing on energy efficiency and parallelism will be necessary to achieve the goals of future computing, with complementary roles for both mainstream and alternative technologies.

Augmenting CMOS

Silicon CMOS circuits have been the central technology of the digital revolution for 40 years, and the performance of CMOS devices and systems have been following Moore's law (doubling in performance every year or two) for the past several decades, together with device scaling to smaller dimensions and integration to larger scales. CMOS appears to be reaching physical limits, including size and power density, but there is presently no technology available that can take its place. How should CMOS be augmented with integration of new materials, devices, logic, and system design, in order to extend enhancement of computer performance for the next decade or more? This approach strongly overlaps with the semiconductor industry roadmap (ITRS), so RCS 2 coordinated this topic with ITRS.

Neuromorphic Computing

A brain is constructed from slow, non-uniform, unreliable devices on the 10 μm scale, yet its computational performance exceeds that of the best supercomputers in many respects, with much lower power dissipation. What does this have to teach us about the next generation of electronic computers? Neuromorphic computing studies the organization of the brain (neurons, connecting synapses, hierarchies and levels of abstraction, etc.) to identify those features (massive device parallelism, adaptive circuitry, content addressable distributed memory) that may be emulated in electronic circuits. The goal is to construct a new class of computers that combines the best features of both electronics and brains.

Approximate Computing

Historically computing hardware and software were designed for numerical calculations requiring a high degree of precision, such as 32 bits. But many present applications (such as image processing and data mining) do not require an exact answer; they just need a sufficiently good answer quickly. Furthermore, conventional logic circuits are highly sensitive to bit errors, which are to be avoided at all cost. But as devices get smaller and their counts get larger, the likelihood of random errors increases. Approximate computing represents a variety of software and hardware approaches that seek to trade off accuracy for speed, efficiency, and error-tolerance.

Adiabatic/Reversible Computing

One of the primary sources of power dissipation in digital circuits is associated with switching of transistors and other elements. The basic binary switching energy is typically far larger than the fundamental limit ~kT, and much of the energy is effectively wasted. Adiabatic and reversible computing describe a class of approaches to reducing power dissipation on the circuit level by minimizing and reusing switching energy, and applying supply voltages only when necessary.

RCS3 Introduction and Setting Goals

The Summit opened with a welcome and review of earlier Summits (*Rebooting Computing: Changing Computing*) by Co-Chairs Elie Track and Tom Conte. This was followed by a presentation and group discussion on the Summit goals (*Rebooting Computing: Goal Setting*), led by Scott Holmes. Both of these are available in video format on IEEE.tv.

RCS3 Plenary Talks

Four plenary talks were given, addressing each of the identified approaches of Security, Parallelism, Approximation, and Human-Computer Interface. The videos from these talks are available on the Rebooting Computing Web Portal http://rebootingcomputing.ieee.org and IEEE.tv.

Trust and Security in Future Computing Systems, Neal Ziring, NSA

Neal Ziring is the Technical Director of the Information Assurance Directorate at the National Security Agency, in charge of defending the nation's public and private computer systems and networks from attack and intrusion. His talk provided an overview of the requirements to maintain secure and trusted communication and control throughout a worldwide network of computers and devices.

Computer Security is a broad field that has increasing significance in an age when everything is networked and computation is delocalized and distributed, with an exponentially increasing number of users and devices. There are a wide variety of threats to security, from computer hackers to commercial and government surveillance. One solution is pervasive encryption of data and communication, but more is needed on both the hardware and software levels. Mobile computing on smartphones and embedded devices presents special challenges, given that security provisions must be automatic and largely invisible to the end users.

The current state of computer security is not satisfactory. The key challenge is how to assure trust in computer and data systems. We need to establish standards and practices for authenticating identity and privilege in platforms and computation, including networks and the Cloud. These need to be easy to implement universally and automatically, as well as viable both economically and politically. The desire for privacy and anonymity must be balanced against the need for accountability. Finally, no security system is perfect, and flaws will inevitably be exploited as they are discovered. Standard protocols must be devised for automatic updating of security software on both recent and legacy systems, and for removing devices from the network if they become compromised. Trust and security can be assured in future computer networks only if these are incorporated as standard building blocks in all devices and networks.

HCI: What does the Future Hold for Human Experience?, Gregory Abowd, Georgia Tech

Prof. Abowd is with the School of Interactive Computing at the College of Computing at Georgia Tech. His talk surveyed how people have interacted with computers in prior generations, and how that is changing now and in the future.

Specifically, the earliest computers (starting in the late 40's) were large mainframes with batch processing used for scientific calculations, with many users for each computer. The personal computer revolution started in the late 60's, with spreadsheets as the "killer app" and a one-to-one paradigm. Twenty years later, as microprocessors started to proliferate in a variety of devices on a variety of scales (inch/foot/yard), ubiquitous computing was born, with human-to-human communication as the killer app. We have already started to move to the 4th generation of computing, known as "complementary computing". This takes ubiquitous computing one step further, and integrates computing devices into the human environment in a seamless way, and even blurs the distinction between humans and computers. A clever way to identify the key aspects of this new generation is in terms of the rhyming trio of *"The Cloud, The Crowd, and The Shroud"*, where cloud computing and crowd computing are already well known. "The shroud" refers collectively to wearable sensors and devices, of which Google Glass and the Apple Watch may represent present-day examples.

Looking to the future, the presence of this 4G computer network can enable an increased level of individual self-sufficiency, where any individual can have complete information on himself, his environment, and the world. This is the "self-sufficient genius" identified as a possible killer app in the table below. Alternatively, the same complementary computing paradigm may also enable large organizations to anticipate the needs of a given individual, leading to the "symbiotic corporation" identified as another possible killer app. More generally, complementary computing may have applications in navigation, health, education, and sustainability. There are major challenges to achieve this, such as security and privacy, but this offers the potential for technology to augment human capabilities and well-being.

Generation	Start Timeframe	People to Device ratio	Canonical device	Killer app
1G: Batch	Late '40's	Many – 1	Mainframe	Scientific calculation
2G: Personal	Late '60's	1 – 1	PC	Spreadsheet
3G: Ubiquitous	Late '80's	1 – Many	Inch/foot/yard	Human-human communication
4G: Complementary	Late '00's	Many – Many	Cloud/Crowd/Shroud	Self-sufficient genius OR Symbiotic Corporation

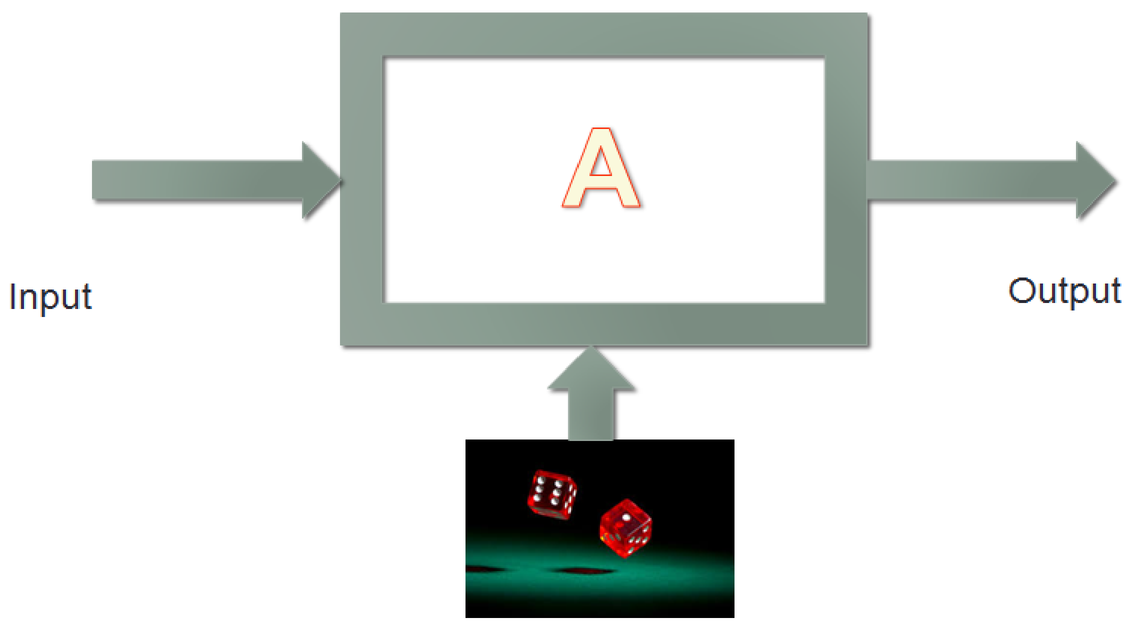
Randomness and Approximation, Dick Lipton, Georgia Tech

Prof. Lipton holds the Storey Chair of Computer Science at Georgia Tech, and spoke about the importance of exploiting randomness and approximation in achieving more efficient algorithms. This is in contrast to classic computing algorithms, which focused on exact solutions, at least to the precision specified for input and output data. Many problems in modern computation do not require exact solutions, but rather require good enough results quickly.

In particular, in dealing with large data sets that are an increasing part of modern computing, sampling provides a means for analyzing only a small part of the data, yet obtaining almost as much useful information as if the entire data set were analyzed. The process for selecting representative data is generally random, but proving randomness can be problematic. Similarly, testing of complex algorithms may involve a very large number of inputs, but randomness may also be employed to provide a drastic reduction in the number of representative tests. Applications for sampling include problems in economics, weather simulation, molecular dynamics, population genetics, and load balancing in parallel computing systems.

Approximation is another important approach to decreasing computation time. This requires a specification of how much precision is necessary for a given application. Approximation is already used in certain applications such as video compression, but there has not been a systematic method to incorporate approximation in software.

In order to progress further, we need to create a computer culture that uses randomness/approximation more widely, and permits it to be easily implemented. That can provide a big enhancement in algorithm efficiency, without a significant increase in software development effort.



Parallelism and Future Supercomputing, Pete Beckman, Argonne

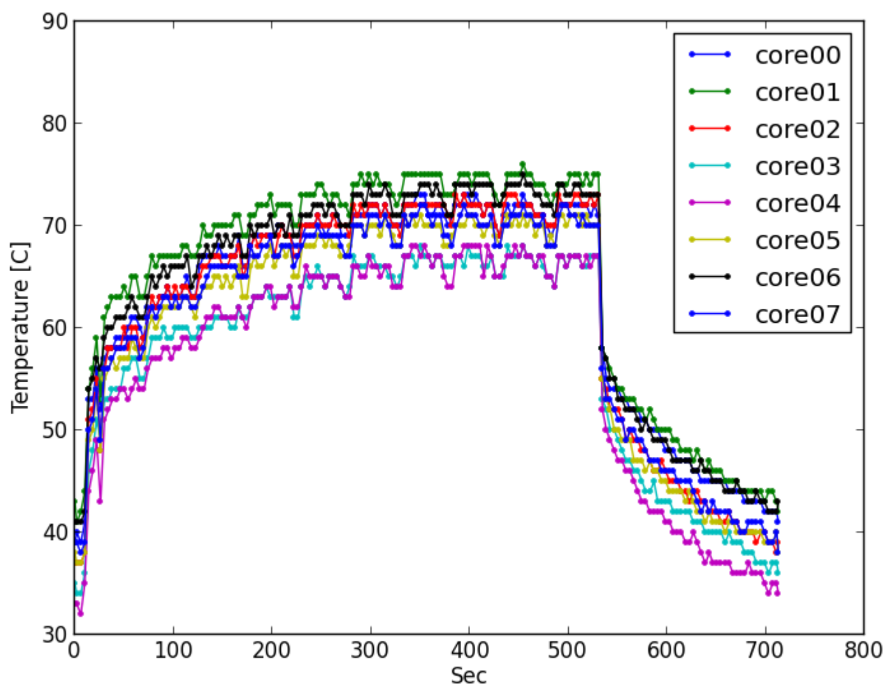
Dr. Beckman is the Director of the Exascale Technology and Computing Institute at Argonne National Laboratory. He presented an overview of the current state of parallel supercomputers, and the needs looking forward to future exascale performance and beyond.

Present supercomputer hardware is based on massive parallelism using ~ 25 MW power, with electricity cost ~ $25M. It may be possible to extend this to exascale in the near term, but this approach is too expensive and inefficient to be maintained in the longer term.

Programming of parallel computers is also inefficient, in that for example it does not include unpredictable effects of inhomogeneous heating that produce variable speeds and hence uncontrolled latency. For example, measured data (shown in the figure below) were presented on the dynamic temperatures of nominally identical processors in a multi-core system. Temperatures may vary by as much as 20 ^{o}C, corresponding to a substantial variance in clock speed.

We need to develop software that enables users to dynamically control access to processors and memory. Furthermore, we will need new latency-tolerant algorithms and methods, and new tools that will measure and predict distributions in latency and processor/memory execution. The response should be dynamic as well as adaptive, particularly if the software is designed to be portable between different supercomputer systems.

Looking ahead to the longer term, even more massive parallelism will require a new lower-power technology. In order to obtain improved performance from further parallel scaling, we need to change programming to be parallel everywhere, with improved integration with memory and inter-processor bandwidth.



The dynamic operating temperatures of nominally identical parallel cores in a multi-core system can differ, which can influence timing and latency in unpredictable ways.

Poster Presentations

5 posters were presented before dinner on Thursday, Oct. 23, on various topics including distributed memories, nanoprocessors, and cryogenic computing.

Self-Authenticating Chip Architecture Using Embedded DRAM, S. Rosenblatt et al., IBM

Memcomputing: Computing with and in Memory, M. di Ventra, UCSD and Y. Pershin, U.S.C.

Optimal Adiabatic Scaling and Processor-In-Memory-and-Storage, E. DeBenedictis, Sandia

Nanoprocessor Thermodynamics: Probing Nature's Boundary Conditions for Rebooting Computing, N. Anderson, U. Mass

Energy Efficient Cryogenic Computing for Exascale Systems, O. Mukhanov & A. Kadin, Hypres Inc.

Summaries of Group Outbriefs

Each of the four groups met separately and presented their conclusions to the entire Summit. These "Outbriefs" are included at the end of this report as Appendices C to F, with brief summaries below.

Outbrief on HCI – Integrate Human and Environment with Computer

Projecting current trends to the future, distinct recognizable computers will still be present in data centers, serving the Cloud. But most computing on the human scale will be integrated with communication and entertainment devices (like smartphones) and analog sensors (for navigation and health), all of which will be strongly interactive and networked. The vision here is similar to the personal digital assistant imagined in RCS 1, designed to serve the best interests of each individual, with control via voice and gestures. This is also compatible with the vision of Complementary Computing described by Prof. Abowd, characterized by "the Cloud, the Crowd, and the Shroud".

Outbrief on Security – Trust and Privacy

Computer network security is indeed a problem, and it needs an organized effort to rebuild security from the ground up. This will require a combination of new technology (hardware), economics, culture change, and new regulations. IEEE Standards and government procurement practices could help to establish universal hardware security primitives and open-source security software. With such a security system, each device would maintain identity, levels of trust, accessibility, and accountability.

Outbrief on Approximate Computing – Full Stack Effort

Approximate computing relaxes the abstraction of near-perfect precision in general-purpose computing, communication, and storage, providing many opportunities across the system stack for designing more efficient and higher performance systems. The novelty in this approach is embracing error holistically across the system stack and making unreliability explicitly exploitable. Both software and hardware need to be addressed. These same considerations also apply to alternative computing paradigms such as neuromorphic computing.

Outbrief on Parallel Computing – Ubiquitous Parallelism

Innovations are needed in areas such as integration of processors and memory, heterogeneous parallelism, virtual distributed parallelism, and closing the (hardware) custom-reconfigurable gap. In terms of improved software, languages and compilers should be better able to specify and take advantage of parallelism and constraints in problems, and to optimize code on a variety of platforms. Furthermore, the CS curriculum should expose students early on to ubiquitous parallelism in programming. Finally, parallel computing can benefit by advances in other computing trends such as neuromorphic (by taking inspiration from the brain) and random/approximate (by incorporating a golden library of hardware/software primitives).

New Federal Initiative in Future Computing – Randal Bryant, OSTP

Dr. Randal Bryant of the White House Office of Science and Technology Policy (on leave from Carnegie Mellon University) gave a brief presentation at the end of RCS3 about a major new government effort focused on the future of computing, known as the *National Strategic Computing Initiative*. Among the objectives is to foster research that will get around anticipated roadblocks in computer performance (such as the ending of Moore's Law scaling), enabling the US computer industry to remain world leaders in the coming decades. Future computing may use exascale parallelism and Big Data Analytics for applications that include scientific discovery, national security, and economic competitiveness.

Conclusions and Looking Ahead

RCS 3 addressed the theme of "Rethinking Structures of Computation", focusing on software aspects including HCI, Random/Approximate Computing, Parallelism, and Security. The speakers and participants agreed that these are all important aspects that must be part of the future Rebooting Computing. Several key conclusions are discussed below, and together with RCS 1 and RCS 2 may help to motivate future conferences and research in the field.

4th Generation Computing

Computing is entering a new generation, characterized by world-wide networks coupling the Cloud with a variety of personal devices and sensors in a seamless web of information and communication. This is more than just the Internet or the Internet of Things; it also encompasses Big Data and financial networks. This presents new challenges, and will require new sets of tools on every level, with contributions needed from industry, academia, and government.

Dynamic Security for Distributed Systems

One key challenge is in the area of computer security. Current security systems represent a patchwork of solutions for different kinds of systems. What is needed is a universal, forward-looking set of protocols and standards that can apply to all parts of the distributed network, with a combination of simple hardware and software building blocks. These must also be dynamic and capable of being updated to reflect newly recognized system features and threats.

Ubiquitous Heterogeneous Parallelism

Parallelism is a central feature of future computing, even if an alternative technology should take hold. This will be massive parallelism for high-performance computing, but even personal devices will be parallel in nature. In many cases, these parallel processors and memories will be heterogeneous and distributed. This represents a strikingly different paradigm than the conventional von Neumann machine, and may require rethinking many of the foundations of computer science.

Adaptive Programming

High-level programming needs to operate efficiently on a wide variety of platforms. This may require providing high-level information (e.g., on parallelism, approximation, memory allocation, etc.) that can be properly optimized by the compiler or system software. Furthermore, the system should learn to become more efficient based on the results of repeated operations and appropriate user feedback, i.e., it should exhibit long-term adaptive learning.

Vision of Future Human-Centric Computing

Prof. Abowd identified the new generation of Complementary Computing, where the boundary between computer and human is blurred. Others have asserted that a personal computing device should be programmed to act in the best interests of each individual. Finally, for an optimum human-centric computing system, the computing devices should be adapted to the needs and preferences of the individual human user, rather than the human adapting to the needs of the computer or the programmer. We have already seen the start of this revolution, but the ending is still being imagined.

Appendices

Appendix A: Agenda for Rebooting Computing Summit 3 (RCS3)

23-24 October, 2014 – Hilton Hotel, Santa Cruz, CA

Thursday Oct. 23

8:30 – 9:00 AM — Welcome and Review of RCS 2 – Elie Track and Tom Conte

9:00 – 9:30 AM — Goals and Outcomes of RCS 3 – Scott Holmes, Facilitator

9:30 – 10:15 AM — Plenary Talk on Security – Neal Ziring

10:30 – 11:15 AM — Plenary Talk on HCI – Gregory Abowd

11:15 AM – 12:00 PM – Plenary Talk on Random & Approximate Computing – Dick Lipton

1:00 – 1:45 PM – Plenary Talk on Parallelism – Pete Beckman

1:45 – 2:00 PM -- Division into groups; Introduction of Discussion Leaders

Parallelism – Sudip Dosanjh

Security — David Mountain and LeAnn Miller

Approximate Computing —Hadi Esmaeilzadeh

HCI/Applications —Erik DeBenedictis

2:00 – 4:30 PM Breakout sessions of 4 Groups in Parallel (See above)

4:30 – 5:00 PM —Plenary gathering to review progress– Facilitator: Scott Holmes

5:00 – 6:30 PM — Poster Session

Friday Oct. 24

8:30 – 9:00 AM — Recap of Day 1 – Track/Conte/Holmes/Kadin

9:00 – 11:30 AM – Continuation of Breakout Sessions

11:30 AM – 12:00 PM – Outbrief on Parallelism – Dosanjh

1:00 – 1:30 PM – Outbrief on HCI -- DeBenedictis

1:30 – 2:00 PM — Outbrief on Approximate Computing – Esmaeilzadeh

2:00 – 2:30 PM — Outbrief on Security – Mountain and Miller

2:45 – 3:00 PM — Pre-Announcement of New Federal Computing Initiative – Randal Bryant, OSTP

3:00 – 3:30 PM – Summit Conclusions – Future Plans. Facilitator: Scott Holmes

Appendix B: RCS 3 Participants

Gregory Abowd	Georgia Tech
Neal Anderson	U. Mass Amherst
Pete Beckman	Argonne National Lab
Randal Bryant	Office of Science & Technology Policy
Tom Conte	Georgia Tech
Erik DeBenedictis	Sandia National Lab
Gary Delp	Mayo Clinic
Massimiliano Di Ventra	University of California at San Diego
Stephen Diamond	IEEE Cloud Computing Initiative
Sudip Dosanjh	Lawrence Berkeley Lab
Hadi Esmaeilzadeh	Georgia Tech
Paolo Gargini	ITRS
Carrie Gates	Dell
Kathy Grise	IEEE
Jennifer Hasler	Georgia Tech
Bichlien Hoang	IEEE Future Directions
Scott Holmes	Booz Allen Hamilton
Jiang Hong	National Science Foundation
Alan Kadin	Consultant
Andrew Kahng	UC San Diego, ITRS
David Kirk	NVIDIA
Arvind Kumar	IBM
Rakesh Kumar	University of Illinois
Yung-Hsiang Lu	Purdue Univ.
Martti Mantyla	Aalto University, Finland
Kathryn McKinley	Microsoft Research
LeAnn Miller	Sandia
David Mountain	NSA
Ravi Nair	IBM
Bryan Payne	Nebula

RCS3 Participants (continued)

Yuriy Pershin	University of South Carolina
Peter Petre	HRL
Wolfgang Porod	University of Notre Dame
Kathy Pretz	IEEE
Abbas Rahimi	University of California San Diego
Shishpal Rawat	CEDA
Sami Rosenblatt	IBM
Mark Stalzer	Moore Foundation
Bill Tonti	IEEE Future Directions
Elie Track	IEEE Council on Superconductivity
Neal Ziring	NSA

Appendix C: Group Outbrief on HCI – Erik DeBenedictis

with Gregory Abowd, Paolo Gargini, Scott Holmes, Martti Mantyla, Bill Tonti, Elie Track, Shishpal Rawat.

The HCI group dialog evolved to include the constituency served by computer applications and the connection to technology.

Over time, computers have evolved from stand-alone devices to increasingly connected systems. Initially, computers became networked into their environment through what is now known as the Internet. Later, the data on computers became a defining characteristic, such as the data in search engines. Most recently, computers have evolved into systems containing humans, both in the context of social networking applications and situations where humans are part of the computing platform itself (crowd-sourcing). This led to the "equation" in our group for the capacity of a computer of tomorrow:

$$\text{Comp}_{\text{tomorrow}} = N \times \text{Comp}_{\text{today}} + \text{World},$$

where N is the number of computing engines in the network, and the World consists of other humans as well as sensors and devices. Aside from the traditional capabilities such as fast arithmetic and large memory, $\text{Comp}_{\text{tomorrow}}$ has additional resources like:

Analog	Adaptive	Social	Creative
Its own laws	Self-powered	Notional	Learns
Dynamic	Resilient	Intuitive	

The $\text{Comp}_{\text{tomorrow}}$ will empower communications between the following classes of people:

IEEE constituencies:	Problem-oriented people
	Systems folks
	Algorithms
Non-computing people	

An application-oriented diagram of the computer system would be as follows:

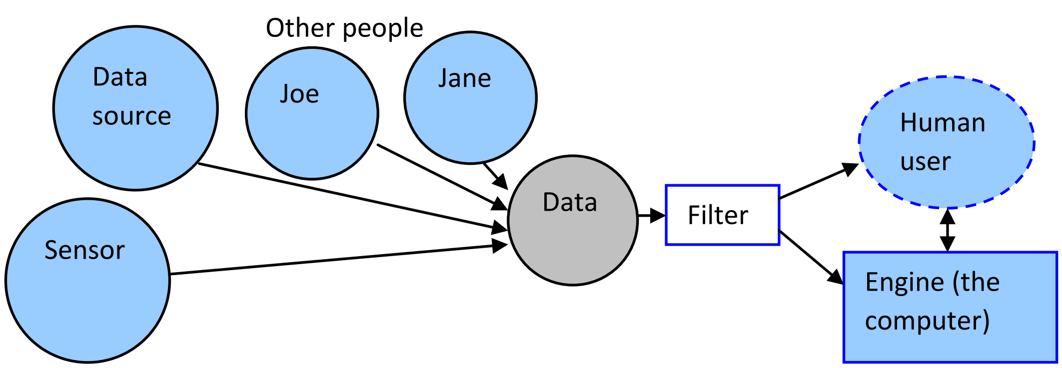
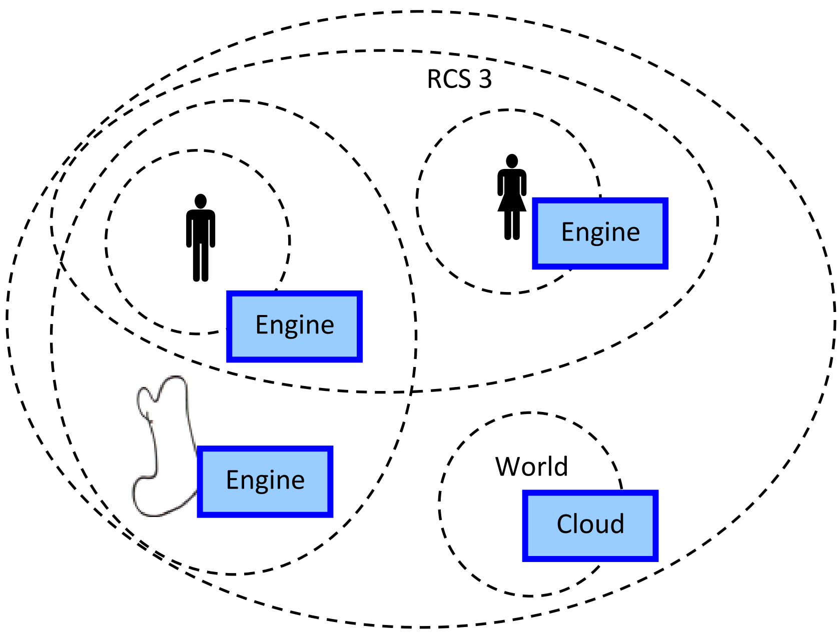


Diagram of components and inputs to a future computer system, with a smart filter that selects and prioritizes data by relevance to the human user.

This led to a view of how applications interact with their constituencies. The idea is that applications serve a single person at the lowest level, but that person may be part of a hierarchy of other groups of people, machines, and data.

In the diagram below, the man and woman are each served by their own computer, or engine. This computer represents the individual's interests. However, the man and woman would be groups such as RCS 3 (this workshop). The man (self-contributed by a specific member of our group from Finland) may be in the "Finland" group (referring to the shape Finland on the diagram), which crosses the boundary of RCS 3. The hierarchy continues until we are all seen as members of the world, with centralized cloud services such as search engines.



Sample relationships among networks of future computers.

This view has some connection to the Future Personal Digital Assistant discussed in RCS 1, yet different from current practice. In current practice, certain outspoken constituencies (software vendors, copyright owners, advertisers, search engines) place themselves at the top of a single hierarchy. The group discussion was specifically incompatible with the idea of single hierarchy.

Group discussion included technological change for the way data is organized. Currently, the major search engines index all data as a single group. For example, the search terms *blue sky* will not understand the difference between the sky's color on a non-cloudy day versus a person feeling blue while sitting under the sky. In conjunction with the multiple hierarchy of applications above, the specific idea discussed is that data would be organized in reference to the object in the real world that it applies to, thus allowing a search on colors associated with the sky in the example. This brief description generalizes to the concept of deeper semantic meaning in data storage and organization – which is known to require more computer capability and storage.

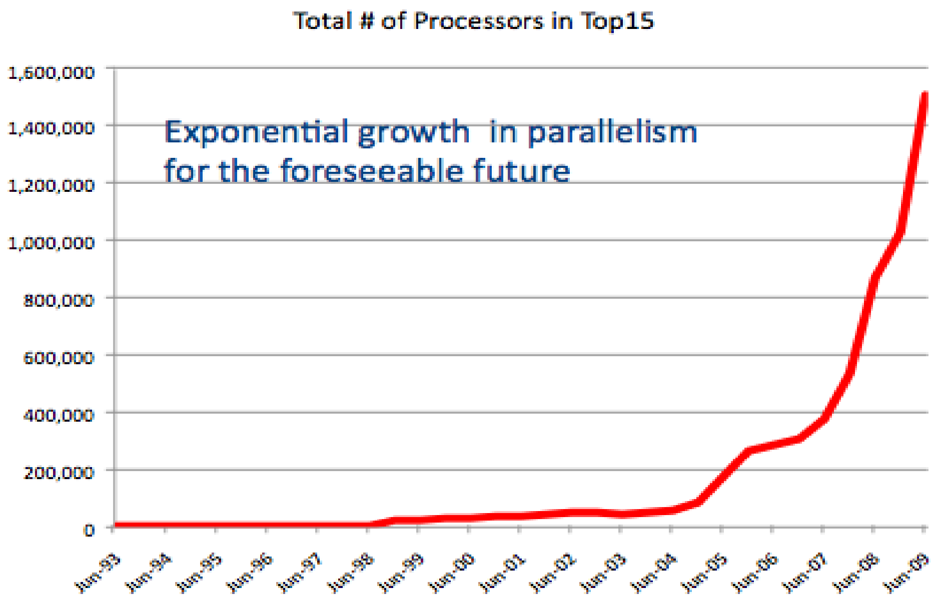
The idea of deeper meaning extended in group discussion to knowledge and context of the individual, as captured in part by the diagram above. Some of the computer's data would be the history and context of the man or woman user, RCS 3, Finland, the world, or any other level in the hierarchy. The future application would be able to answer questions based on the context. Asking the computer, "How many political parties are there?" might return the number of political parties in Finland, if the question was asked from within Finland.

Issues

- Scalability
- Interaction
- Business model
- Trust model
    - Defend user's interest as opposed to a web company

Appendix D: Group Outbrief on Parallel Computing – Sudip Dosanjh

With Pete Beckman, Randy Bryant, Massimiliano Di Ventra, Gary Delp, Hong Jiang, Alan Kadin, Andrew Kahng, David Kirk, Arvind Kumar, Yuriy Pershin, Wolfgang Porod



Parallelism has become ubiquitous and dominant in high performance computing, with no end in sight. But making efficient use of these large numbers of parallel processors is becoming increasingly difficult.

Future Challenges:

- End of weak scaling
- Rate of change is accelerating
- Impacts everyone
- More levels of parallelism
- Indeterminate effects no longer ignorable (locality, reliability, impacts of temperature, aging)
- Deep memory hierarchies and greater sensitivities
- Need to rethink the storage hierarchy
- Need abstractions for parallelism

Discussion Questions:

- Innovation: What are the innovations needed for the exascale, zettascale, and beyond regimes? These could include (but are not limited to) energy efficiency, memory bandwidth, device scaling, and packaging.
- Programming: How can parallel programming be made simpler? Since Parallel computing is becoming ubiquitous, should parallel programming be taught at an earlier stage?
- Other Computing: How can other computing trends such as neuromorphic, approximate, and adiabatic computing affect the future direction of parallelism?

1. Innovation Grand Challenges for Exascale and Beyond

- Integrated Processor and Memory
    - More effective parallelism
    - Not moving so much data
    - Storage Class Memories
- Virtual Exascale
    - Getting away from paradigm of single large exascale system
    - More efficient use of Cloud and Edge Systems (and even internet of things)
    - Automated Software to partition given problem among resources, minimizing bandwidth required.
    - Security and economic issues need to be resolved.

2. Programming: How can we make parallel programming easier and improve education?

- Mathematical construct becomes running code on a variety of platforms within a few hours
    - Better mechanisms for expressing what the programmer knows (persistence, locality, affinity, ownership, privacy)
    - Alternative wording: To be able to convert problems to abstractions that express the parallelism and constraints clearly so that they are usable by a computer
- The computer optimizes the code better than a human (even a heroic programmer)
    - Build-up, remember and optimize for optimal behavior (continuous improvement)
- Invert the CS classroom so that algorithms are taught from the perspective of highly parallel resources
    - Avoid teaching last year's technology (e.g., data movement dominates and not FLOPS)

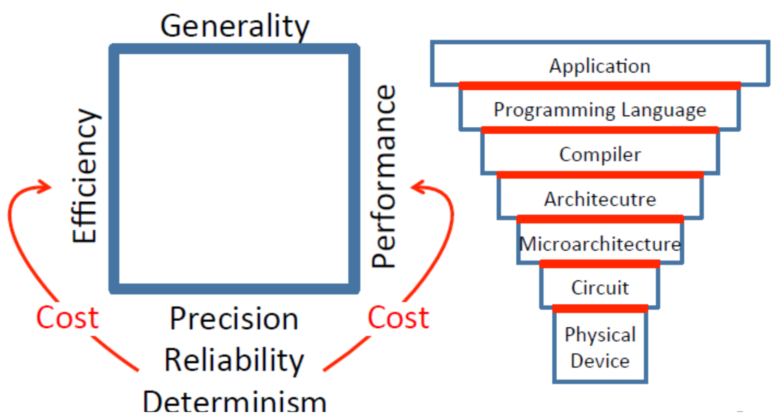
3. Other Computing Grand Challenges

- Intelligent memories. (See Q1.)
    - Combining compute and memory (including emerging memory devices). [inspiration from neuromorphic; opening up the Von Neumann CPU-memory bottleneck.]
- Build useful computers that are brain-inspired. Features including:
    - Continuous learning [distributed processing, 'intelligence']
    - Integrate ~1e11 processing elements (neurons) with ~1e4 fanout (connectivity)
    - Within 10x of brain's energy/op (= low-energy goal)
- Golden library (of hardware and software primitives) for random and approximate computing in parallel computing.
- Close the (hardware) custom-reconfigurable gap. Power, area, and performance of reconfigurable fabrics within 10x of full-custom in leading-edge technology.

Appendix E: Group Outbrief on Approximate Computing – Hadi Esmaeilzadeh

With Neal Anderson, Tom Conte, Jennifer Hasler, Rakesh Kumar, Dick Lipton, Yung-Hsiang Lu, Kathryn Mckinley, Ravi Nair, Peter Petre, Abbas Rahimi, Martin Rinard

Full-Stack Effort: Avoid overkill design throughout the entire software/hardware stack



Why Approximation?

- Performance growth hits the energy wall
- Conventional technologies anticipated to fall short of historical trends and projected demand.
- Radical departures from conventional approaches are necessary.
- New technologies are emerging that are variable.
- Emerging applications can tolerate inaccuracy.
- Lower layers work hard to expose a general, reliable, "precise" and (mostly) deterministic interface, <u>but at a big cost in efficiency</u>!

Opportunities

- Contribute to prolonging CMOS scaling
  - Embrace variability
  - Improve Yield
  - Expose parallelism
- Enable new technologies that are intrinsically variable
  - Small feature sizes (5 nm …)
  - Memristors, PCM
  - Magnetics
  - Chemical
  - Photonic
  - Analog
- Bridge non-von Neumann models with von Neumann modes
  - Allow interoperability, neuromorphic

- Increasingly, emerging applications are error resilient/inherently approximate
  - Optimization (e.g. approx. SAT, analytics)
  - Randomized algorithms
  - Machine learning
  - Pattern recognition
  - Decision making
  - Sensory applications
  - Financial application
  - Web search
  - Robotics
  - Augmented reality
  - Cyber-physical systems

Challenges

- Abstractions for algorithm design and programming
  - Exposing low level error in abstract way to the higher level
  - Exposing the knobs of the low level
  - Algorithm to hardware translation automation and design tools
- Converting component-based approximation to end-to-end solution
- Quality of results: understanding how to measure quality for each application and what level is acceptable
- Modeling low-level component behavior and composing into overall behavior (compounding)
- Introspective feedback loop for self-adapting systems: need both "knobs" for measurements and tuning.
  - Fallback strategies
- Full-stack effort: no layer wants to be the first to change, but somebody has to be first
  - Start with "killer app" that motivates effort at all layers simultaneously
  - Start with sufficiently universal program like SAT
  - Evolutionary path with augmenting current practices
- Incorporating imprecision into hardware design and synthesis process
  - e.g., Design tools
- Adoption by large body of programmers and designers
  - Common problem with other energy efficient techniques
- Users should be oblivious to approximation

Next Steps

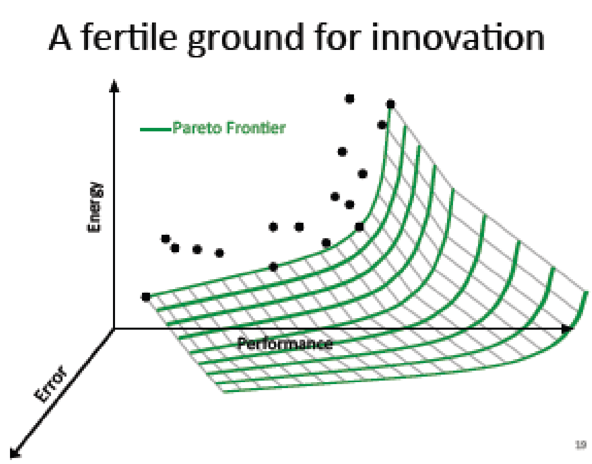
- Exploring software and algorithmic approximation
- Start collaborations between academic and semiconductor industry to being prototyping approximate hardware
  - Modeling variability and error
- Prize competitions that come with university funding, like Grand Challenges, including energy, network and privacy considerations
  - Scene reconstruction and object recognition on the Shroud
  - Face/voice recognition on the Shroud

Information Gaps

- Marketing input on key applications to start
- User requirements: what quality level is "good enough" for these applications?
  - Solutions for reusability across different operation conditions
  - How much testing is enough validation that you've met those requirements?
- How do errors from different components compound?
  - Composability and interoperability

What is the new technical idea; why can we succeed now?

- Embracing approximation, holistically, in general-purpose computing
- Relaxation abstraction of near-perfect accuracy
- Increasingly, emerging applications are error resilient
- Conventional CMOS does not seem to outpace other approaches
- Difference in cost between precise and approximating output is increasing



Appendix F: Group Outbrief on Security – David Mountain and LeAnn Miller

With Sami Rosenblatt, Neal Ziring, Bryan Payne, Steve Diamond, Carrie Gates, Tom Conte, Dick Lipton

"It needs to change, but there is hope!"

During his plenary talk at RCS3, Neal Ziring pithily stated "Security stinks!" In the past, this primarily affected the security of information. In the future, as exemplified by the Internet of Things (IoT), this will affect both physical and information security, Moreover, security failures in these systems pose potentially severe consequences for humans – driverless cars would be one example. The overall risks to society associated with inadequate security can be expected to dramatically increase. The time to prepare is now (perhaps yesterday is more accurate). When a disruptive cybersecurity event occurs, the existence of carefully considered solutions and ideas will be essential – they will also prevent or delay these events, and minimize their consequences.

The RCS1 report envisioned a 2023 security future, described as follows.

> *"In the year 2023 and beyond, the concepts of security and trust are closely intertwined. Computer systems will be capable of doing what a human individual would do if that person had all the information and understanding needed to make an informed decision. Systems follow the executive assistant mode. It knows when to follow rules, when to break the rules and when to ask a human operator for more information. Trust is implicit in this new model of security because users trust that the device has his or her best interests at heart and trust the device to do the right thing."*

The foundational elements of this goal are:

- Identity – ensuring both the human and the device "know" each other using mechanisms that do not compromise fundamental human attributes (hard to replace your DNA).
- Levels of trust – a minimum of security is automatically enforced, most likely in hardware, with flexibility to adapt residing in software. Fail properly is a prominent design theme.
- Accessibility – "best of breed standards" in security practices and implementations are widely used and enable scalable integration utilizing a building block approach.
- Accountability – Poor security practices have consequences that will be harmful; developers and providers need to be responsible for doing things correctly.

The status quo of poor security has evolved over time due to the diversity of actors involved (individuals, organizations, adversaries, etc.) Changing the status quo is possible, and the solutions will involve a synergistic combination of:

- Technology – research challenges need to be addressed
- Economics – a combination of carrot and stick will be involved; government buying practices could make a difference
- Culture Change – security information is available and can be used to make decisions; minimum levels of trust will be expected
- Legislative/regulatory action – clear, enforceable rules are needed

At RCS3, our group identified ways to disrupt the status quo in order to achieve the 2023 goal defined in RCS1. Breaking the overall problem down into distinguishable components, we were able to articulate research challenges, next steps, key concepts, and roles for the IEEE and government.

Hardware can and should provide a minimum level of security/trust. The key challenges in this area:

- To define the security primitives needed for this minimum level
- To develop methods to ensure hardware integrity (such as physically unclonable functions)
- To identify mechanisms that enforce the "fail properly" design goal

The IEEE can be a catalyst for making this happen. Developing standards, methods for evaluation, and communication/advocacy are inherent roles for the IEEE. As a non-partisan, international organization, the IEEE is perhaps uniquely suited to provide this service.

The software environment can naturally provide flexibility in creating multiple the levels of trust offered by the hardware. There are a number of good ideas that can be drawn from academic work and industry to support the notion of interoperable building blocks that have inherent security built in. As one example, Django (an open-source web application framework) has a system that can prevent most XSS attacks, see https://docs.djangoproject.com/en/dev/topics/security/.

While the cloud is a "centralized target" for cyber attacks, it also enables widespread adoption of solutions due to its relative homogeneity of hardware and software, and the ability to standardize practices with the buy-in of a few key adopters. By choosing wisely, improving the inherent security of a small set of widely used software components (Google App Engine, MongoDB, Python, etc.) could increase security significantly.

Key research challenges include:

- Scalability of the security solutions (key management and trusted platform modules are difficult to scale)
- Automated red teaming to specify the level of security demonstrated

The government can play an important role in this arena by:

- Developing and sharing open source security building blocks (SELinux is a good example).
- Defining minimum security requirements for procurement to support the use of standards and best practices.

A possible role for the IEEE is in the development of a set of threat models to drive the research, implementation, and evaluation of these building blocks.

The Internet of Things (IoT), which includes existing products such as smartphones as well as emerging devices such as smart toasters, will provide new avenues of attack, but also creates an opportunity "to get things right the first time". Key research challenges are:

- Low power security
- Fail properly techniques

- Learn from existing fault tolerant ideas
- Identifying proper use of resets that are not cyber accessible would be beneficial
- Devising and promoting basic evaluation techniques, e.g., formal methods appear practical in simple devices

- Separation of logical networks sharing the same physical networks

This component of security seems naturally suited to utilizing advances in neuromorphic computing, due to the diversity of products, multiple levels of trust required, requirements of personal adaptability, and need for energy efficient solutions.

Inevitably, humans will be part of the security ecosystem, as designers, developer, and users (or abusers). Security must be made easy, and privacy policies should be flexible and retroactively enforced. We must acknowledge the fact that instant gratification/delayed risk scenarios create security problems. International regulatory or legal enforcement of a minimum level of security is necessary. An overarching issue crossing both the cloud and the IoT is reduced privacy due to aggregation of information. A fundamental concept is that control of information resides with the owner (the individual). Technical solutions should provide:

- Verification of identity with minimal use of passwords but protection of personal attributes
- Ubiquitous crypto
- Customization of privacy/security rules
- Pervasive use of data tagging to enable permanent delete

The discussion of this dimension of security overlaps significantly with the Rebooting Computing pillar of the human-computer interface. In particular, the grand challenge of an executive assistant, and the possibilities for using neuromorphic computing, are similar.

An interesting question was posed – what is the real cost of cryptography? Looking at it in terms of power needed, if ubiquitous crypto reduced email spam by 20%, would the power cost of implementing this solution be offset by the reduced need for sending and storing the spam? What level of reduction is sufficient?

While improving privacy and security in future computing systems is challenging, it is possible. RCS1 defined a vision for future security, and RCS2 articulated models of computing that are likely to emerge. The RCS3 group has identified key concepts, research challenges, and next steps that provide a path forward.