RCS 2

2nd Rebooting Computing Summit

Summary Report

The Chaminade
Santa Cruz, CA
May 14-16, 2014

Prepared By:
Alan M. Kadin
And the IEEE Rebooting Computing Committee

http://rebootingcomputing.ieee.org/

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

June 2014

Contents

Foreword

The Future Directions Committee (FDC) is a committee of the IEEE Technical Activities Board (TAB). Through volunteers from IEEE 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 eight IEEE Societies/Councils and two professional IEEE staff directors. The RC committee organized a 1st Rebooting Computing Summit (RCS 1) in December 2013 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.

In order to implement this vision, the RC Committee identified four initial technologies for further discussion, a mainstream approach of **Augmenting CMOS**, together with alternative approaches of **Neuromorphic**, **Approximate**, and **Adiabatic/Reversible Computing**. These provided the basis for the 2nd Rebooting Computing Summit (RCS 2) held in Santa Cruz, CA, May 14-16, 2014. RCS 2 followed a similar format to RCS 1, with about 50 invited experts in a variety of fields, breaking up into smaller groups to discuss each of these technologies. This Summary Report is intended not as a definitive technical report on RCS 2, 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 is a third Summit, RCS 3, now being planned for San Jose, CA on 24-25 October, 2014.

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, including slides and videos of RCS 2 presentations and developing plans for RCS 3.

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. Can we continue to project further improvements in computing performance in coming decades? Do we need to review the entire basis for computer technology, starting over again with a new set of foundations (hence “Rebooting Computing”), or are the current efforts in the computer industry sufficient to maintain progress?

Participating Societies and Councils

Circuits and Systems Society (CAS), Computer Society (CS), Council on Electronic Design Automation (CEDA), Council on Superconductivity (CSC), Electron Devices Society (EDS), Magnetics Society (MAG), Reliability Society (RS) and Solid-State Circuits Society (SSC); also, coordination with the 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:

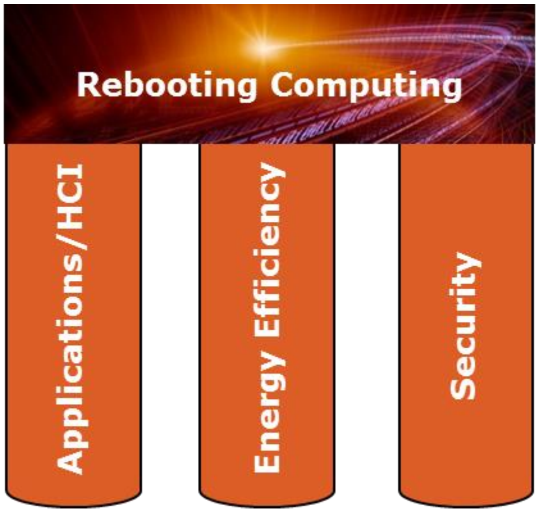
- Dan Allwood (MAG), Univ. of Sheffield, UK
- David Atienza (CEDA), Ecole Polytechnique Federale, Lausanne, Switz.
- Jonathan Candelaria (EDS), Semiconductor Research Corp.
- Erik DeBenedictis (CS), Sandia
- Paolo Gargini (ITRS), Intel
- Glen Gulak (SSC), Univ. of Toronto, Canada
- Bichlien Hoang, RC Program Director, IEEE Future Directions
- Subramanian (Subu) Iyer (EDS, CPMT, SSCS), IBM
- Yung-Hsiang Lu (CS), Purdue University
- Scott Holmes (EDS), Booz Allen Hamilton
- Alan M. Kadin (CSC), Consultant
- David Mountain (EDS, CS), NSA
- Oleg Mukhanov (CSC), Hypres, Inc.
- Vojin G. Oklobdzijja (CAS), U. Cal. Davis
- Angelos Stavrou (RS), George Mason Univ.
- Bill Tonti (RS), 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 limiting the use of 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 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 2-day workshop spread over 3 days, from Wednesday afternoon May 14 to Friday morning May 16, at the Chaminade in Santa Cruz, CA. The agenda is included here as Appendix A, and the list of attendees as Appendix B. The main theme of RCS 2 was on mainstream and alternative computing technologies for future computing, with 4 approaches identified before the Summit by the RC Committee. The format was similar to that for RCS 1, with a set of 4 plenary talks, followed by 4 separate breakout groups culminating in outbrief presentations and concluding in a final plenary discussion. The groups and interactions were coordinated by Facilitator Scott Holmes (Booz Allen Hamilton), and Alan Kadin (consultant) assisted as "scribe" for the Summit.

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 can this 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.

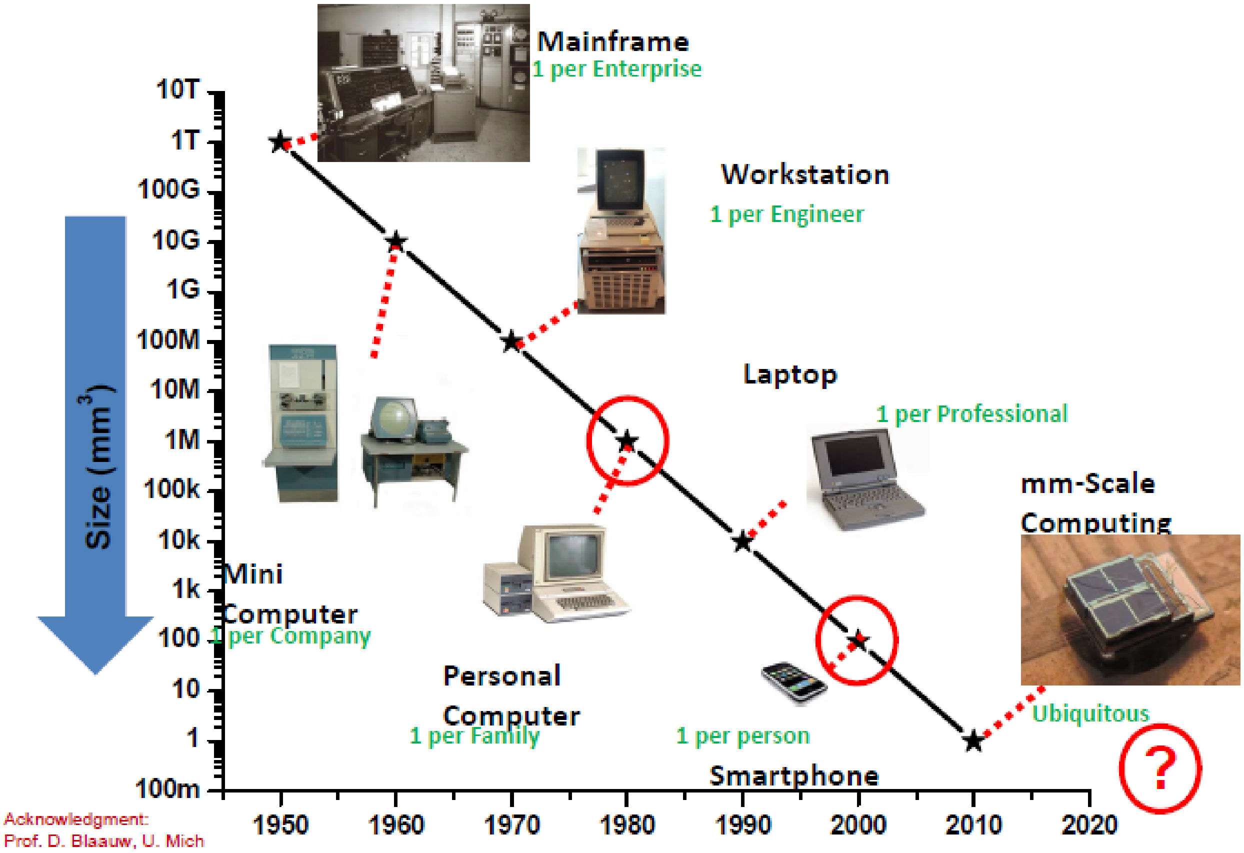
Plenary Talks

Four plenary talks were given, addressing each of the identified approaches. The videos and slides from these talks are available on the Rebooting Computing Website, http://rebootingcomputing.ieee.org.

ITRS 2.0: System Drivers and More than Moore (MtM), Andrew Kahng, UCSD and ITRS

The International Technology Roadmap for Semiconductors (ITRS) has traditionally predicted technology trends in the semiconductor manufacturing industry, following Moore's Law scaling. ITRS is now rebooting itself as ITRS 2.0, going beyond Moore's Law ("More than Moore" or MtM) by focusing more on application pull on the system level, rather than merely on technology push on the device level. The next wave in applications in the coming decade may be ubiquitous computing in mobile systems. These changes will show up in the ITRS Roadmap issued in late 2015.

In addition to Moore's Law for IC scaling, Prof. Kahng described Bell's Law for volume scaling of complete computer systems from large rooms down to the mm-scale, combined with decreased cost and increased numbers of units sold. He also pointed out that in considering future system requirements vs. technology scale, the greatest opportunity for improvements lies in the intermediate regime where optimal packaging and system design can enhance system efficiency, particularly for mobile devices where minimizing power consumption may be critical. Furthermore, it will be necessary to include in the technology roadmap not only digital processors, but also broadband RF communication and emerging analog sensors. Cost is also critical, and will be included in future ITRS roadmaps.

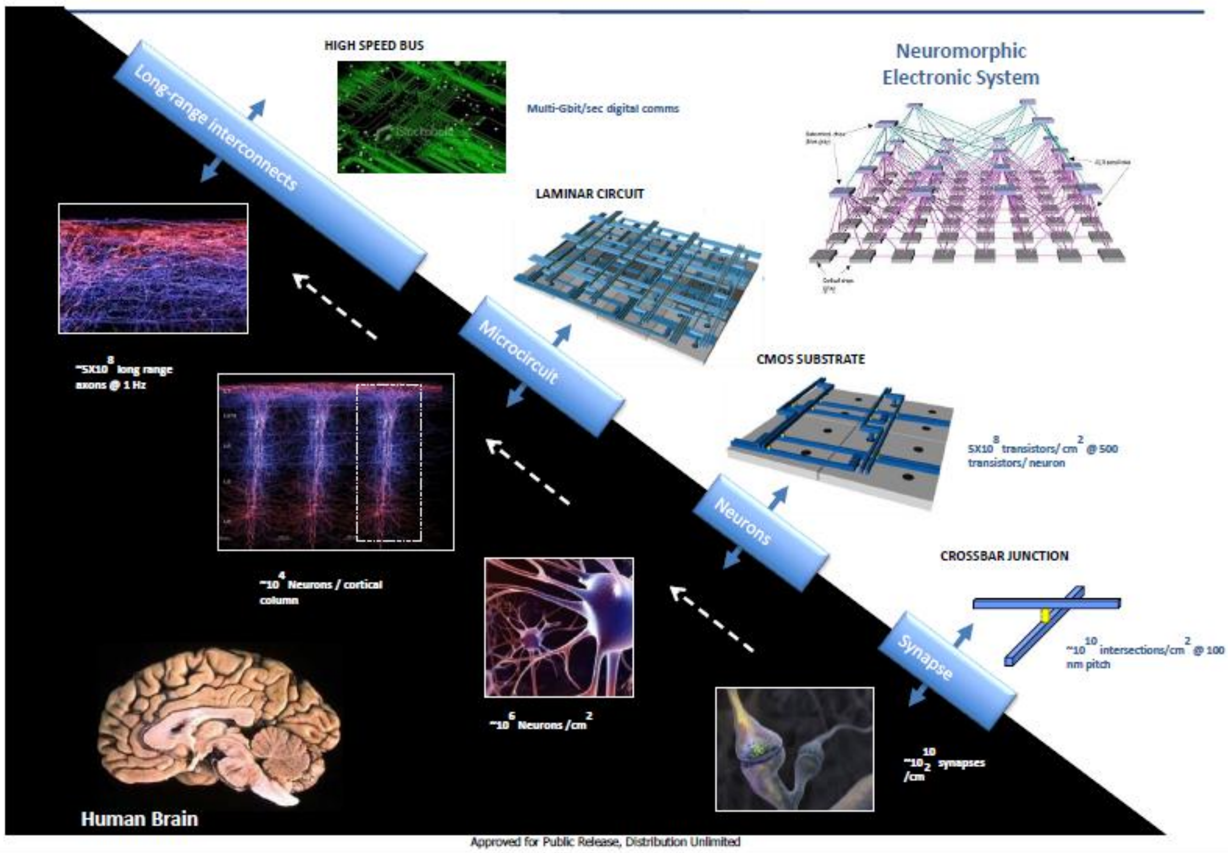


Introduction to Neuromorphic Computing: Insights & Challenges, Todd Hylton, Brain Corp.

Dr. Hylton was formerly the DARPA program director who initiated the SyNAPSE project on Neuromorphic Computing. He focused in his talk on the lessons learned during this project, including that building a computer from components that act as neurons is *not* the same thing as building a brain. We know how to do the former, but we don't really know how brains work, and we don't know how to develop the required software. The performance of brain structures on the largest scale may be only weakly dependent on the properties of neurons. Future R&D is necessary that combines algorithm development with network topology.

Dr. Hylton described the SyNAPSE neuromorphic architecture, comprising CMOS circuits and adaptive crossbar junctions that emulate the dynamic behavior of biological neurons and synapses, with high-speed buses corresponding to long-range axon bundles. These can be scaled up to millions or even billions of electronic neurons, with billions of synapses. These systems are expected to be very efficient for certain machine learning and pattern matching applications, provided that appropriate algorithms can be developed. However, designing a computer with the functionality of a brain, even a small-animal brain, will be much more difficult. Dr. Hylton emphasized that brains are evolved biological systems with a hierarchy of structures on different scales. On each scale, feedback among components tends to make the behavior insensitive to detailed structures at lower levels. This limits the utility of the reductionist approach to understanding brains. It is likely that true "intelligence" requires higher-level organization that is not yet understood. We should focus near-term efforts on designing machines and algorithms that can perform improved machine learning, for example for robotic systems.

SyNAPSE Architectural Concept



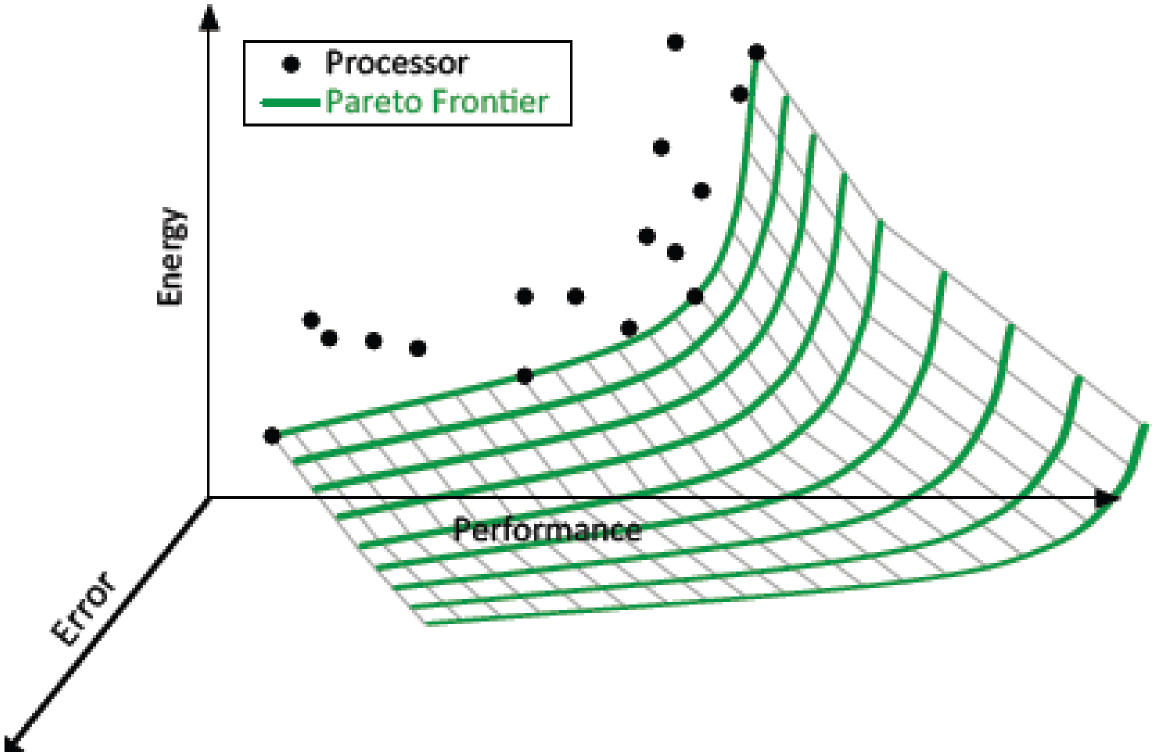
HIGH SPEED BUS
Multi-Gbit/sec digital comms
Neuromorphic Electronic System
LAMINAR CIRCUIT
CMOS SUBSTRATE
5X10^{8} transistors/ cm^{2} @ 500 transistors/ neuron
CROSSBAR JUNCTION
~10^{10} intersections/cm^{2} @ 100 nm pitch
Long-range interconnects
Microcircuit
Neurons
Synapse
~5X10^{8} long range axons @ 1 Hz
~10^{4} Neurons / cortical column
~10^{6} Neurons /cm^{2}
~10$^{10}_{2}$ synapses /cm
Human Brain

Approved for Public Release, Distribution Unlimited

Approximation: Beyond the Tyranny of Digital Computing, Hadi Esmaeilzadeh, Georgia Tech

Prof. Esmaeilzadeh addressed a set of approaches to reducing power based on obtaining acceptable answers from components that may themselves be inaccurate or unreliable. For many applications such as searching, image processing, and analog sensor processing, an approximate result may be quite sufficient, and can be obtained with reduced power on mobile devices. This requires innovation on both the software and the hardware levels, and involves a change from the traditional paradigm where maximum accuracy was assumed to be essential.

It is customary to represent the tradeoff between computer performance and power. Dr. Esmaeilzadeh proposed thinking of Error as another dimension in this tradeoff, where tolerance of error may in some cases permit enhanced performance at decreased power. Error in this context comprises two key aspects. First, there are classes of applications where exact calculations are not always necessary, such as those with analog inputs or outputs, searching, optimization, or learning algorithms. A computation that is only as accurate as needed will be lower in energy (and time) than one which always achieves maximum accuracy. Second, as the transistor scale continues to shrink, variability in basic device performance and reliability are becoming increasing problems. An alternative computer architecture that is designed to be tolerant of device error may be able to use these hardware resources more efficiently. For example, neural systems in the brain are composed of basic elements that are highly variable, yet the system performance is nevertheless highly robust. In applying approximate computing approaches to electronic systems, co-design of hardware and software (algorithms and compilers) is necessary for optimum performance.

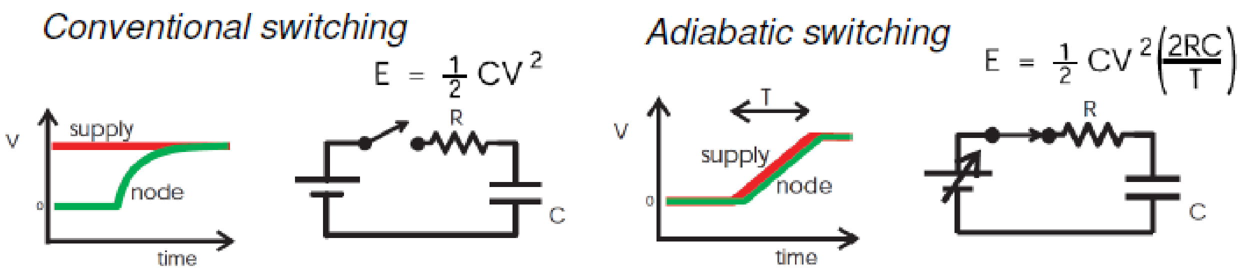


Adiabatic/Reversible Classical Computation: An Overview, David Frank, IBM

Dr. Frank reviewed approaches to classical computing (i.e., non-quantum computing) known as Adiabatic and Reversible Computing. Both of these are general approaches to reducing the switching power of computer circuits, trading off speed and complexity. Adiabatic computing reduces power dissipated in resistors, instead temporarily storing power in inductors and capacitors. While adiabatic computing can be applied to conventional CMOS technology, power reductions may be somewhat limited. Novel alternative technologies such as Quantum Dot Cellular Automata (QCA) and superconducting circuits (such as nSQUID) may be more amenable to power reduction.

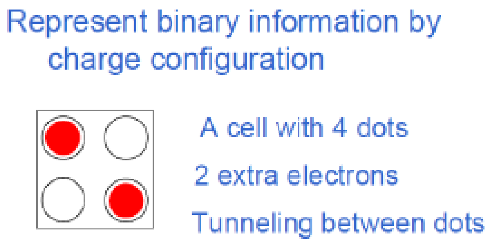
Many years ago, it was derived theoretically that the minimum switching energy per gate should be kT *ln*(2). However, CMOS typically has switching energy >100,000 kT, so there is a lot of room for improvement. One way of understanding adiabatic computing is via the static charging energy of the capacitance, $E = ½ CV^2$, all of which is dissipated in the resistor R in conventional switching. By slowly ramping the power supply, the energy dissipated may be reduced by a factor of $t/(RC)$, where t is the switching time. This requires high-Q inductors and an ac power supply, but it can offer significant advantages. (Reversible computing can in principle go further to reduce power dissipation to arbitrarily low levels, at the expense of additional circuit and timing complexity.) It may seem ironic that one would deliberately slow a computation, but it turns out that this tradeoff may be favorable for improved performance, if parallelism is effective and devices are cheap and small.

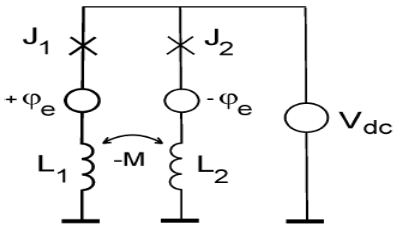
Several adiabatic and reversible logic families have been investigated. It seems that the benefit for conventional CMOS may be limited to about a factor of 10, but alternative technologies may be better matched to these approaches. QCA circuits are based on coupled cells, each cell containing 4 quantum dots, with 2 extra electrons that are located on two of the dots. A variety of QCA designs have been proposed and simulated, but rather few have been demonstrated. The nSQUID is a superconducting device with two Josephson junctions and two mutually coupled inductors. Small circuits with nSQUIDs have been operated at a cryogenic temperature T=4 K, with demonstrated switching energy ~3kT, close to the theoretical limit. Practical adiabatic and reversible circuits and systems remain to be fully demonstrated.



QCA

nSQUID





Poster Presentations

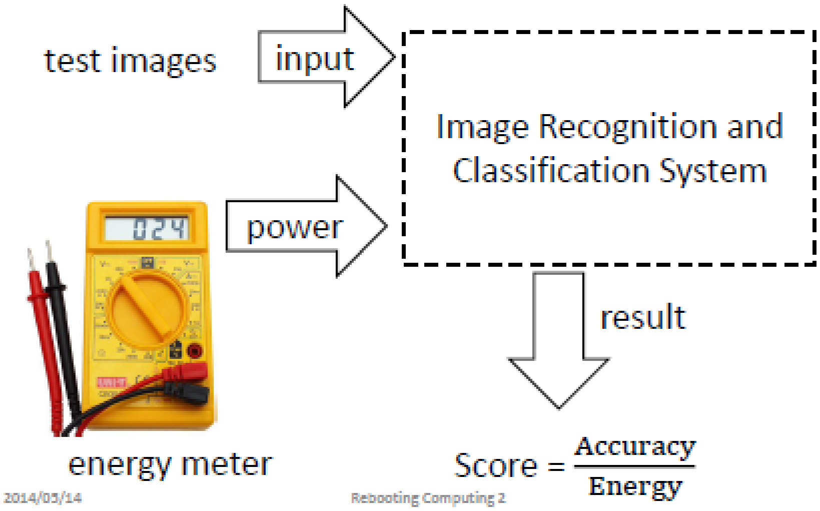
6 posters were presented after dinner on Wed., May 14, on topics including neuromorphic and approximate computing, and on computing with distributed memory.

- *Neurons and Synapses in a Superconducting Digital Architecture*, Ken Segall, et al., Colgate University
- *Intelligent Computing with Neuromemristive Systems*, Dhireesha Kudithipudi, Rochester Institute of Technology
- *Approximate Computing: Software and Applications*, Adrian Sampson, et al., University of Washington
- *Multi-level Control of RRAM for Neuromorphic Synapses*, Liang Zhao & Yoshio Nishi, Stanford University
- *Memristors for Neuromorphic Computing*, Stan Williams, HP
- *Memory Integrated Computing*, Maya Gokhale, et al., Livermore

Prizes for Rebooting Computing?

IEEE Competition for Low-Power Image Recognition, Yung-Hsiang Lu, Purdue

Prof. Lu proposed an IEEE prize competition, focusing on Low-Power Image Recognition using a mobile device, possibly for 2015. This would involve presentation of a set of test images to the device, and a limited time to accurately identify the images.


test images
input
Image Recognition and Classification System
power
result
energy meter
$\text{Score} = \frac{\text{Accuracy}}{\text{Energy}}$
2014/05/14
Rebooting Computing 2

XPRIZE for Rebooting Computing?, Mark Stalzer, Moore Foundation

Dr. Stalzer described the XPRIZE Foundation, and how a prize related to Rebooting Computing could be set up and funded. A sponsor who could provide ~$10M would need to be found. This could be a major challenge that would require several years to achieve, and could engage leading academic and industrial teams. An example was presented of the Qualcomm Tricorder XPRIZE competition, inspired by the hand-held diagnostic medical device from the Star Trek television series.

QUALCOMM TRICORDER XPRIZE®

Healthcare in the Palm of Your Hand.

Summaries of Group Outbriefs

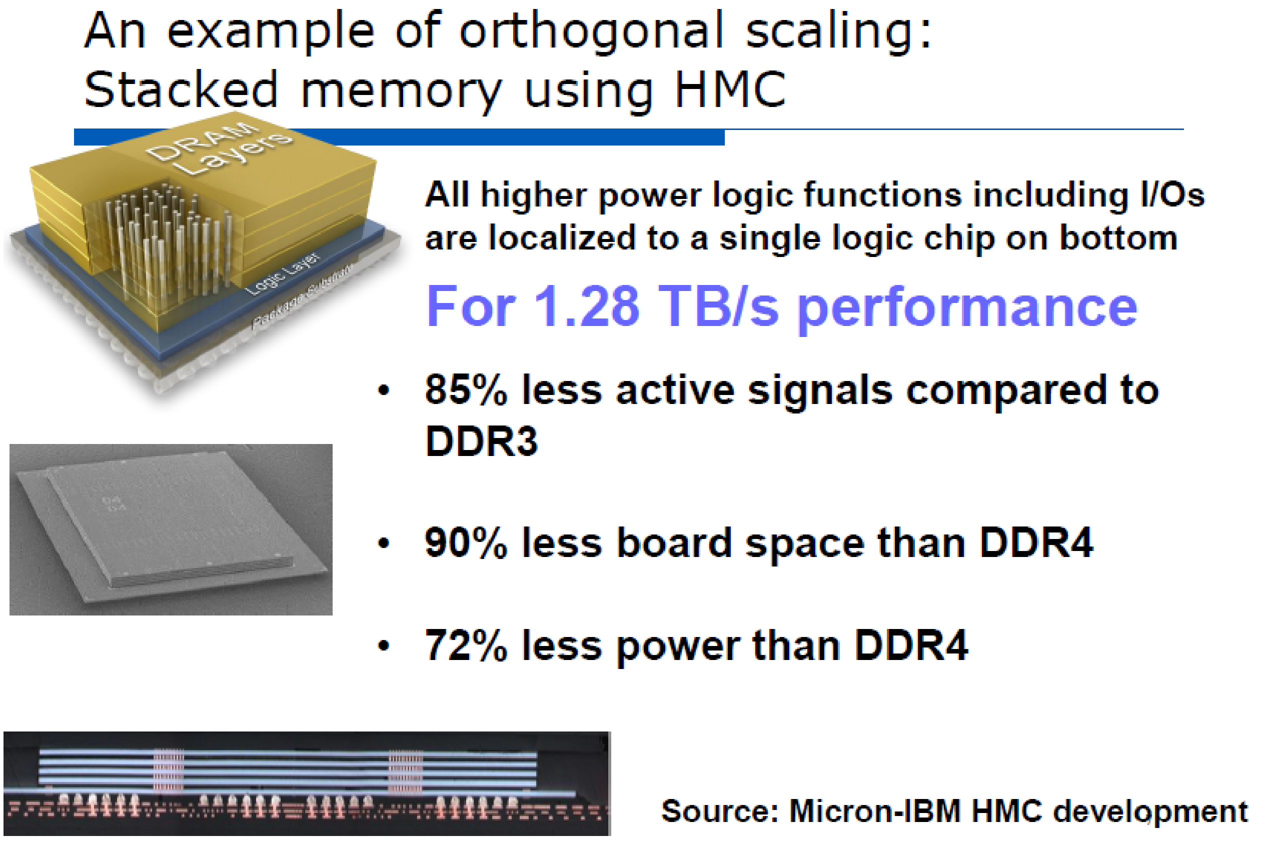
Each of the four groups met separately and presented their conclusions to the entire Summit. These “Outbriefs” are given at the end of this report as Appendices C to F, with brief summaries given below.

In guiding the group discussions, the participants were instructed to consider the set of Heilmeier questions, attributed to the late George Heilmeier (http://en.wikipedia.org/wiki/George_H._Heilmeier) , formerly head of DARPA as well as a top manager at Bellcore and Texas Instruments. Some of these questions are:

1) *What is the problem, and why is it hard?*
2) *How is it done today?*
3) *What is the new technical idea? Why can we succeed now?*
4) *What is the impact if successful?*

Augmenting CMOS: Performance Enhancement without Moore’s Law Chip Scaling.

Moore’s Law scaling seems to be ending for ICs, but the practical limit may be one of cost rather than physical limitations. If smaller devices are not cheaper (including the cost of a new generation of nanofabrication machines), they will not be adopted. However, we can continue to pursue Moore’s Law-type improvements in system performance by the use of “orthogonal scaling”. For example, when you can’t scale the chip anymore, focus more effort on scaling the package and the board. An example was presented of the Hybrid Memory Cube (HMC), which stacks multiple memory chips.



An example of orthogonal scaling:
Stacked memory using HMC

DRAM Layers
Logic Layer
Package Substrate

All higher power logic functions including I/Os are localized to a single logic chip on bottom

For 1.28 TB/s performance

- 85% less active signals compared to DDR3
- 90% less board space than DDR4
- 72% less power than DDR4

Source: Micron-IBM HMC development

Neuromorphic Computing: Dynamic Machine Learning

A major goal for neuromorphic computing is dynamic (unsupervised) machine learning, preferably in a low-power system with a small form factor such as a smartphone. A specific "killer app" might be real-time visual analysis (image recognition), virtual reality, or anomaly detection. A more complete understanding of the brain is not essential to short term progress in this field, but is critical for long-term advancement. So interdisciplinary collaboration with neuroscientists such as those working on the Brain Initiative would be helpful.

Approximate Computing: Harnessing Error Tolerance to Enhance Performance

Two complementary trends in software (applications) and hardware (devices) are tending to favor developing approximate computing. On the one hand, modern applications such as optimization, machine learning, and pattern recognition are intrinsically error-tolerant. On the other hand, both ultra-small CMOS and many alternative nanoscale technologies are prone to problems of variability and reliability. But traditionally, high-level software development has not paid attention to issues of variability of low-level devices, except perhaps to discard entire blocks of memory. What is needed is a systematic approach to approximate computing from top to bottom. This may require a "killer app" that can motivate initial collaborative efforts in this direction, such as an earplug real-time translator or a pocket IBM Watson machine.

Adiabatic Computing: Ultra-Low-Power Circuit Design

Dramatic reductions in power are the primary drivers for adiabatic computing. These reductions in power come from slowing down the clock speed of the circuits, which is opposite the traditional direction. However, one can compensate for this reduction in speed by increasing the integration scale, and applying massive parallelism. An example of a long-term vision was suggested whereby a computer could be scaled from a 2D array of 10^8 gates in 1 cm^2, to a 3D array of 10^{15} gates in 1 cm^3. Applying the ideal adiabatic scaling could lead to a power-efficiency improvement of $\sqrt{10^7} \sim 3000$, with the same total power, despite reducing the clock rate by the same factor ~ 3000. While this is likely not possible with conventional CMOS, one might envision a 3D molecular self-assembly with integrated logic and memory. Such a structure might apply to neuromorphic-type architectures, for example. Clearly, more R&D is needed.

Timeframe	**Today**	**Changes**	**Tomorrow**
Integration scale	10^8 logic transistors	$\times 10^7$	10^{15} logic transistors
Clock speed	3 GHz	3000× slower	1 MHz
Performance	Chip is 2D comprised of 100 nm^2 gates. 	3000× reduction in joules/op OR 3000× increase in energy efficiency	Chip is 3D comprised of 100 nm^3 gates.

Conclusions and Looking Ahead

CMOS Will Continue to Advance

The general consensus of the participants at RCS 2 was that although the computer industry will continue to advance in the near term, the end of Moore’s Law on the chip level will require at least a partial Rebooting of Computing. The primary focus of the semiconductor industry has been on scaling digital integrated circuits, but that is changing to a more system-level view that incorporates packaging, analog I/O, and consumer applications throughout the planning and design process. This change is recognized in the development of a new industry roadmap, ITRS 2.0.

Alternative Technologies Offer Complementary Approaches

The Summit considered one mainstream approach and three alternative technologies. Alternative computing technologies and approaches are not yet mature enough to supplant conventional CMOS, but they are sufficiently promising to recommend further R&D to identify how they may provide opportunities for Rebooting Computing in the future. Perhaps surprisingly, most of the participants agreed that these approaches are likely to be complementary rather than competitive, at least at the present stage. For example, the Approximate and Adiabatic approaches might be relevant to Neuromorphic architectures, and each of the alternative approaches may lead toward the development of special purpose processors that could help to extend the reign of CMOS, while leading to novel technologies in the longer run.

Energy Efficiency is Dominant Theme

All of these technological approaches are addressing to some degree the key pillar of energy efficiency from RCS 1, which is recognized as a central issue both in battery-powered mobile devices and grid-based servers. The ultimate goal for mobile devices is to reduce power to a level where energy harvesting may be generally sufficient, while the continued expansion of servers should not be impeded by the cost of electricity.

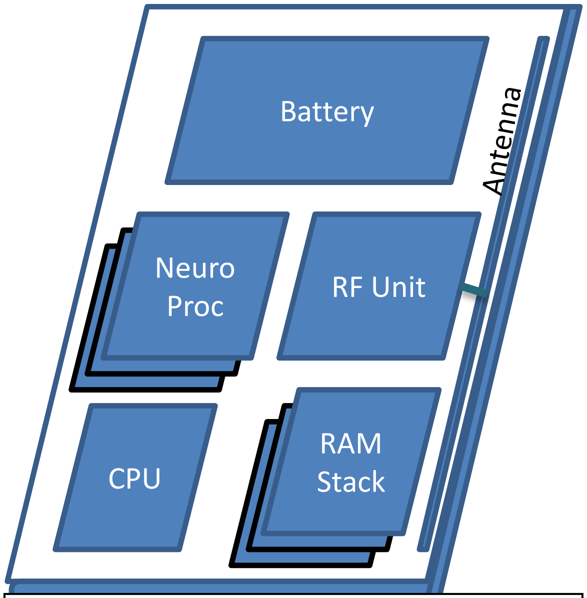
Exploiting Parallelism is Another Common Theme

All of the groups identified another key enabler for further scaling of computer performance: capturing, describing, and managing parallelism efficiently. Modern applications are increasingly parallel in nature, and both hardware and software performance scaling is contingent on exploiting this parallelism. Massive parallelism can achieve high performance even for relatively slow devices, as biological brains have shown. This parallelism is a possible focus topic during RCS 3.

Meeting the Vision of Future Computing

RCS 2 yielded a series of compatible ideas that could contribute to a vision of future computing, including the results of RCS 1 as new applications drivers. The Augmenting CMOS group contribution (which is similar in direction to ITRS 2.0), offered the Smartphone as an effective current product-class on which to base further innovations. The emerging vision suggested in the concept figure would be driven by continuing increase in device count like Moore’s Law, yet utilizing the third dimension. The adiabatic and reversible group offered a technology direction that could continue to reduce power per device, a key enabler of this vision.

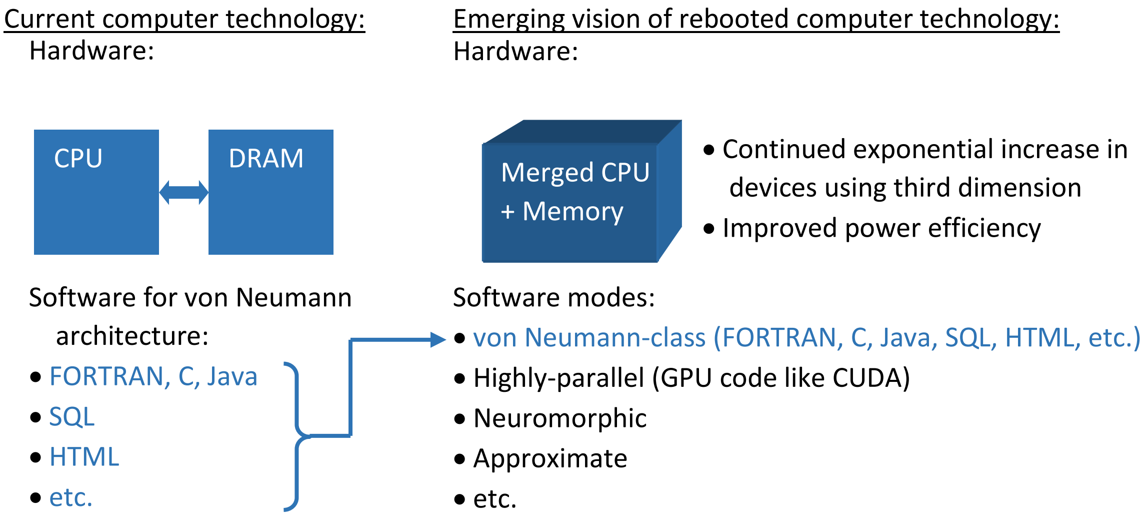
Software for the “rebooted computer” could be envisioned as putting all existing software into a “von Neumann mode” of the future computer and then adding additional modes. The neuromorphic and approximate computing groups offered emerging software paradigms that could advance programming efficiency and energy efficiency. Specifically, a neuromorphic computer learns and thus displaces programmer labor. Approximate computing essentially enables hardware to go closer to the edge of the envelope in energy efficiency while meeting reliability requirements.



Vision of Future Mobile Intelligent Assistant

For example, one vision of future computing that was suggested in RCS 1 was a mobile computer (with the form factor of a smartphone, perhaps) that doubles as an intelligent automated assistant. There was general agreement at RCS 2 that the 4 technological approaches can all work together to support such a vision. As suggested in this concept figure, this mobile computer could consist of processors optimized for different applications. This might include a general purpose CPU with a 3D stack of RAM chips providing “orthogonal scaling”. But it could also include a Neuromorphic Processor (with 3D stacking to increase the number of active neurons and connections) that might handle the intelligent user interface with dynamic learning. This suggests separate processors, but integration on the same larger chip is another possible option. Furthermore, one or more of these processors might include approximate and adiabatic computing architectures to further decrease power dissipation, enabling longer battery operation between recharging. The figure also shows an RF unit and antenna for a broadband wireless link to the Internet and the Cloud, which would also dissipate significant power.

Although the power and spatial constraints for data centers and high-performance computing are less restrictive, the same hardware methods and software paradigms could also be applied to these systems. For example, neuromorphic computing approaches offer great promise for efficiently analyzing large amounts of unstructured information under dynamic conditions. A generic representation of future computer technology on large and small scales is presented below.



Current computer technology:

Hardware:

Software for von Neumann architecture:

- FORTRAN, C, Java
- SQL
- HTML
- etc.

Emerging vision of rebooted computer technology:

Hardware:

- Continued exponential increase in devices using third dimension
- Improved power efficiency

Software modes:

- von Neumann-class (FORTRAN, C, Java, SQL, HTML, etc.)
- Highly-parallel (GPU code like CUDA)
- Neuromorphic
- Approximate
- etc.

The Next Summit: RCS 3

The IEEE Rebooting Computing Committee is now planning for the follow-up 3rd Summit, RCS 3, to be held October 24-25 near San Jose, CA, after the IEEE Technology Time Machine symposium (TTM). The agenda of RCS 3 is still being developed, and may include additional topics not addressed in RCS 2, as well as connecting back to the other key pillars from RCS 1 of Security, Energy Efficiency, and Human/Computer Interface. Further information on RCS 3 will be available soon on the Rebooting Computing Website http://rebootingcomputing.ieee.org.

RCS Publications and Future Conferences

The goal of the RC Committee and the participants is to move toward publication of a White Paper or article summarizing the conclusions of the RCS series of Summits. The venue of such a report might be in a journal such as *IEEE Computer*, or alternatively in a new journal such as the *IEEE Journal of Exploratory Solid-State Computational Devices and Circuits*. In addition, these summits could lead to the establishment of an annual international conference on Rebooting Computing, which will bring together engineers and computer scientists from a wide variety of disciplines, to help promote a new vision of Future Computing.

Appendices

Appendix A: Agenda for Rebooting Computing Summit 2 (RCS2)

14-16 May, 2014 – The Chaminade, Santa Cruz, CA

Wednesday May 14

2:00 – 2:30PM — Review of RCS 1, and introduction to RCS 2 agenda – Elie Track and Tom Conte

2:30 – 3:30 PM — Group discussion on goals and outcomes of RCS2. Facilitator: Scott Holmes.

4:00 – 4:45 PM — Introduction to Augmentation of CMOS –Andrew Kahng

4:45 – 5:30 PM — Introduction to Neuromorphic Computing – Todd Hylton

5:30 – 6:00 PM— IEEE Competition for High Efficiency Micro-CPU – Yung-Hsiang Lu

8:00 – 9:30 PM — Poster Presentations by attendees

Thursday May 15

8:30 – 8:45 AM — Introduction and re-cap of Wednesday Track/Conte/Holmes/Kadin

8:45 – 9:30 AM — Introduction to Adiabatic and Reversible Computing – David Frank

9:30 – 10:15 AM — Introduction to Approximate Computing – Hadi Esmaeilzadeh

10:45 – 11:15 AM — Introduction and Considerations for XPRIZE – Mark Stalzer

11:15 AM – 12:00 PM Division into groups (facilitator)

Augmentation of CMOS – Subu Iyer

Neuromorphic Computing — David Mountain

Approximate Computing —Hadi Esmaeilzadeh

Adiabatic/Reversible Computing —Erik DeBenedictis

1:00 – 3:00 PM – Breakout sessions (See above)

3:30 – 4:00 PM — Plenary gathering to check progress– Facilitator: Scott Holmes

4:00 – 5:30 PM — Return to Breakout sessions (See above)

Friday May 16

8:30 -9:00 AM — Outbrief on Neuromorphic Computing – David Mountain

9:00 – 9:30 AM — Outbrief on Approximate Computing – Hadi Esmaeilzadeh

9:30 – 10:00 AM — Outbrief on Adiabatic/Reversible Computing – Erik DeBenedictis

10:30 – 11:00 AM — Outbrief on Augmentation of CMOS – Subu Iyer

11:00 AM – 12:00 PM –Plenary Discussion – Conclusions – Future Plans. Facilitator: Scott Holmes

Appendix B: RCS 2 Participants

John Aidun	Sandia National Labs
Neal Anderson	UMass Amherst
Colin Cantlie	Defence Research and Development Canada
Juan-Antonio Carballo	ITRS
An Chen	Global Foundries
Fred Chong	UC Santa Barbara
Tom Conte	Georgia Tech
Thomas Coughlin	Consultant
Erik DeBenedictis	Sandia National Labs
Andre DeHon	Univ. of Pennsylvania
Gary Delp	Mayo Clinic
Hadi Esmaeilzadeh	Georgia Tech
David Frank	IBM
Paul Franzon	North Carolina State Univ.
Mike Garner	ITRS
Maya Gokhale	Lawrence Livermore National Lab.
Kevin Gomez	Seagate
Bichlien Hoang	IEEE Future Directions
Scott Holmes	Booz Allen Hamilton
Todd Hylton	Brain Corp.
Subu Iyer	IBM
Alan Kadin	Consultant
Andrew Kahng	UC San Diego, ITRS
Tracy Kimbrel	National Science Foundation
David Kirk	NVIDIA
Dhireesha Kudithipudi	Rochester Inst. of Technology
Arvind Kumar	IBM
Fabrizio Lombardi	Northeastern Univ.
Yung-Hsiang Lu	Purdue Univ.

RCS2 Participants (continued)

Matthew Marinella	Sandia National Labs
David Mountain	NSA
Siddhardtha Nath	UC San Diego
Michael Neimier	Univ. of Notre Dame
Vojin Oklobdzija	U. Cal Davis, IEEE Circuits & Systems
Shishpal Rawat	Intel
Samar Saha	IEEE Electron Devices Society
Adrian Sampson	Univ. Washington
Ken Segall	Colgate Univ.
Horst Simon	Lawrence Berkeley Lab
Marc Snir	Argonne National Lab
Mark Stalzer	Moore Foundation
Bill Tonti	IEEE Future Directions
Elie Track	IEEE Council on Superconductivity
David Tuckerman	CMEA Ventures
Jeffrey Voas	NIST
Mary Ward-Callan	IEEE Technical Activities
Michael Wengler	Qualcomm
Stan Williams	HP
Ian Young	Intel

Appendix C: Group Outbrief on Augmenting CMOS

Summary by Subramanian S. Iyer

Background:

The relentless scaling of CMOS technology over the last five decades appears to be slowing down in cost per transistor, power-performance, and density.

Goals:

This RCS 2 subgroup was tasked with identifying alternative and additional hardware-based technology strategies that have the potential to meet historical expectations of Moore's "law". This is also referred to as "Orthogonal Scaling."

Keynote talk by Kahng:

Traditionally, the ITRS roadmap for semiconductors has provided direction for semiconductor technology development. Of late however, this classical roadmap has been less relevant. The ITRS is developing a new methodology dubbed ITRS 2.0 where the roadmap is driven by the dominant application: mobile.

Deliberations:

The workgroup deliberated at length and came to the following conclusions:

- Mobile and "Stabile" applications will both drive the roadmap. The latter will likely be more important in the long run as bandwidth to the "cloud" becomes unlimited and "free." Mobile technology will be light, low power and be dominated by diverse data acquisition. Most heavy lifting will be done in the cloud. Cloud processing centers will be memory centric and both power-performance and heterogeneous processing will be important.
- Interconnect scaling has stalled, and transistor parasitics have increased significantly. These are the key detractors of power performance improvements in advanced nodes. Focus needs to shift towards interconnects.
- Orthogonal scaling:
    - Heterogeneous integration is tremendous value – add for mobile applications.
    - Package and board scaling have not happened for the last several decades. The new economic realities of semiconductor scaling makes package scaling a viable option with potentially large returns.
    - 3-dimensional integration using interposers, die and wafer stacking offer potential to reduce board footprint, reduce communication power, decrease latency and increase bandwidth.
    - Both hardware and system based fault tolerance and redundancy need to be incorporated to much greater lengths than today.

The workgroup felt that addressing these concerns and the addition of orthogonal features would indeed allow us to meet the historical expectations of Moore's "law."

Appendix D: Group Outbrief on Neuromorphic Computing

Summary by David Mountain

The explosion of available data from sensors, multiple information sources, and large numbers of people, creates programming and analysis problems for traditional computing. Neuromorphic computing, where the processing elements are trained using this data, can naturally exploit this situation. Coupled with advances in technology and neuroscience, this computing approach has tremendous potential. Our group identified potential "killer apps" that could help drive this development:

- Real time visual analysis for immediate decision making
- Visual analysis of large amounts of data
- Individualized virtual reality
- Personalized computing environment
- "Replace" people with machines

However, a number of barriers also exist:

- Current work is driven by neuroscience, not applications
- Multidisciplinary teams are difficult to create
- Our level of understanding of the brain, particularly the critical abstractions it uses, is limited
- Standardized data sets and tools to support research are incomplete

The group identified three increasing levels of training/learning in a neural net. The first is static, or training once with no or infrequent updates to the weights in the neural net. This approach is used widely. The second is adaptive, where the weights are updated on a very frequent or continuous basis. There is a significant amount of research in this area, with some implementation. The third level is dynamic learning, where the structure of the neural net itself is changed in response to the data; it is unclear how much work and progress is being made at this level.

An important question, discussed but not answered by the group, was the extent to which our limited understanding of the brain will hinder development. While progress in the implementation of neuromorphic computing is possible with our current understanding, we will be unable to fully utilize the technology without identifying the key operating principles and abstractions used by the brain.

The key next steps for this approach are the following:

- Minimize the information gap about the current state of neuroscience/psychology by inviting speakers in these areas
- Identify natural partners, such as the Institute for Neuromorphic Engineering
- Create a list of available tools and data sets to identify gaps
- Invite someone (Jeff Hawkins was suggested) to speak about current abstractions being utilized, their limitations and their successes

We also brainstormed a bit about grand challenges and/or X-prize candidates. Ideas included generalized anomaly detection, model induction, and "find my daughter playing soccer on my iPhone without accessing the cloud."

Appendix E: Group Outbrief on Approximate Computing

Summary by Hadi Esmaeilzadeh

We have entered the era at which performance growth hits the energy wall, and conventional computing technologies are anticipated to fall significantly short of historical trends and the projected demand for computing. Radical departures from conventional approaches are necessary to provide performance and efficiency across a large class of applications. Approximate computing is such a radical departure that 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.

Motivation.

Increasingly, emerging applications of interest are error resilient. Across these large class of applications, maintaining the current abstraction of near-perfect accuracy is unnecessary, overkill, and wasteful. At the same time, the traditional cadence of benefits from CMOS scaling is diminishing. The difference in the cost of providing an approximate versus a precise output has grown and is increasing. The rate at which data is being generated is growing overwhelmingly. Approximate computing provides an opportunity to deal with these problems by utilizing the inherent error-resilience of the emerging applications that are increasingly gaining prominence. These applications include but are not limited to sensory data processing, multimedia, optimization, big data analytics, randomized algorithms, machine learning, pattern recognition, cyber-physical systems, web search, and many more.

Opportunities.

Approximate computing can contribute to the continued CMOS scaling to very small technology nodes by providing a means to embrace the ever-increasing variability and reduce the cost of mitigating it. It may also enable new technologies that are intrinsically variable, including memristive, magnetic, chemical, photonic, and others. Approximate computing can also bridge non-von Neumann models with von Neumann models, including and specifically neuromorphic models of computing. Providing approximate algorithmic transformations that can establish this bridge is an essential requirement. These approximate algorithm transformations can enable hardware specialization and provide significant gains. Approximate computing can also provide mechanisms to utilize non-RAM machine models (e.g., functional programming, lambda calculus, or cellular automata). It can also allow interoperability between different models of computing. When conventional techniques are running out of steam, approximate computing can contribute to the IT-based economic growth. It even has the potential to enable new capabilities and markets. As an instance, for the Internet of Things, where energy is a scarce resource, efficiency gains from approximation can provide capabilities that can never be possible without approximation. The benefits from approximation can contribute to make computing green and sustainable.

Challenges.

Providing intuitive abstractions for design, programming, debugging, and validation of approximate systems is one of the main challenges. These abstractions need to expose low-level errors to the programmers and system designers in a high-level manner. There is also a need for exposing the knobs for controlling the tradeoffs between efficiency, performance, and quality. Providing abstractions and mechanisms to monitor and control approximation is another component that makes these systems self-adapting toward changes in data, context, and constraints. Understanding how to measure quality for each application and what level is acceptable is another important aspect of approximation. It seems essential to provide modeling tools that capture the low-level error behavior of system components and show how including approximate components affects final quality. At the hardware level, incorporating approximation effectively into hardware design, synthesis, and layout is necessary to provide large gains with small quality degradations. Solutions that also provide reusability across different operating conditions are essential for prevalent adoption.

Roadblocks.

Approximation is a full-stack effort, while even changes to one layer are challenging. One way to overcome this issue is to provide an evolutionary path by augmenting current practices. Another way is to demonstrate large gains with a "killer app" that motivates effort at all layers simultaneously. Adoption by a large body of programmers is another challenge that is a common problem with other energy efficient techniques. Other important issues include marketing and educating consumers about relying on approximate systems, and making such systems appealing to the consumers. However, with the ever-increasing reliance on web services that are less deterministic and reliable and already use some form of approximation in communication or computation, the market entry point is already in place.

Appendix F: Group Outbrief on Adiabatic and Reversible Computing

Summary by Erik DeBenedictis

Some points from David Frank's talk will be reemphasized here because they apply to the group consensus. The talk was videotaped and will be available through the Rebooting Computing Website.

Adiabatic and reversible computing have been "ahead of their time." Prior to around 2003, users preferred the high performance of single-core processors to multi-core processors with power efficiency improvements that were qualitatively similar to those offered by adiabatic and reversible computing. In retrospect, the fact that adiabatic and reversible methods for saving power have not been widely used may be due to the fact that users put power consumption low on the list of requirements. However, adiabatic and reversible techniques have a unique approach to reducing energy consumption that may lead to their becoming a technical prerequisite for maintaining the information revolution.

Adiabatic methods trade off greater power efficiency for more circuit complexity and lower clock rate. A shift of CMOS to adiabatic methods today would impose a circuit overhead of 27× (a rough estimate by David Frank) and require changes in design tools. While power costs are a concern and some of the increased circuit complexity could be placed on "dark silicon," the tradeoff is not economically viable today. However, the consensus opinion from the "augmented CMOS" group and ITRS as presented at RCS 2 is that manufacturing cost per gate will continue to decline exponentially (albeit not necessarily for standard CMOS). If this happens, the case for using adiabatic and reversible methods will get exponentially stronger over time and inevitably tip the economic balance in its favor.

The distinction between adiabatic and reversible computing is not widely appreciated, and will be emphasized here. Adiabatic and reversible computing aim to reduce power dissipation at the circuit level. In adiabatic computing, reductions are achieved by minimizing and reusing switching energy and applying supply voltages only when necessary. A reversible computer would use adiabatic methods, but would achieve greater power reduction using an additional method. The additional reductions in reversible computing are achieved by preserving input information during computation, thus avoiding dissipation of energy resulting from irreversible information loss.

The group listed a series of "indicators" of application or problem areas where adiabatic and reversible computing approaches would be especially applicable:

- The problem is amenable to a highly parallel solution, such as applications now using graphics processing units (GPUs); neural networks; or as a base for processor architectures that can generate high instruction-level parallelism.
- The cost of energy is high.
- The cost of hardware is low or will decline quickly with evolution in manufacturing.
- It was noted that CMOS could be used for adiabatic and reversible computing, but the group believes CMOS will be less efficient than other approaches.

The group identified four top technology areas that will require improvement to bring adiabatic methods into production:

1. High-Q resonators, which are used for "hot clock" resonant power supplies.

2. New device functionality over and above transistors. Example: The fourth terminal on the four-terminal variant of the proposed Piezo-Electronic Transistor (PET) makes this device awkward as a transistor replacement, but the slightly different function compared to a MOSFET makes it especially efficient in adiabatic and reversible circuits.
3. Devices that can function with lower static power.
4. Continued lowering of manufacturing costs.

The group identified research devices that could be applied to adiabatic and reversible computing. These include

- CMOS (although it is seen as the current approach upon which improvement is desired)
- Piezo-Electronic Transistors (PET)
- Superconducting devices (Josephson Junction; nSQUID, AQFP, …)
- Quantum Cellular Automata (QCA, quantum dot, atomic, molecular, …).
- Other devices, such as a Superconducting FET.

The group proposed two related demonstration milestones for adiabatic and reversible computing:

1. a 64-bit adder and/or
2. a 1 GFLOPS processor

These would use adiabatic circuits to reduce power to 1% of the equivalent circuit constructed using then-current CMOS.

The group decided on a before/after diagram illuminating a vision of how adiabatic and reversible technology could improve computer energy efficiency by large amounts. If industry succeeds in the continuation of Moore's Law for reducing device manufacturing cost (but not energy per CMOS logic gate), at some point the reduction in manufacturing cost will reach 10^7 (which is illustrated in the table in the outbrief section). The obvious structural scenario would be to extend current logic gates that are 100 nm on each of two sides to a 3D configuration where gates are 100 nm on each of three sides. This would correspond to a 1 cm^2 chip being layered to a depth of 10^7 and becoming a 1 cm^2 cube. According to adiabatic and reversible technology, the cube would have $\sqrt{}10^7$ = 3000× lower energy per gate operation, $\sqrt{}10^7$ = 3000× higher throughput, yet the chip and cube would dissipate the same power.