

**Report to the National Science Foundation on
The Workshop for Energy Efficient Computing
Held April 14-15, 2015 at the Holiday Inn Arlington Ballston,
Arlington, VA**

Workshop Organizers

Sankar Basu, NSF CISE
Dimitris Pavlidis, NSF ENG
Thomas Theis, SRC-NRI (Chair)
Keren Bergman, Columbia University (Co-Chair)
Jon Candelaria, SRC-GRC

Workshop Participants

Neal Anderson, University of Massachusetts at Amherst
Christian Binek, University of Nebraska-Lincoln
Clive Bittlestone, Texas Instruments
Jeffrey Bokor, UC-Berkeley
Leonid Butov, UC-San Diego
Suman Datta, Penn State University
Alexander Demkov, University of Texas-Austin
Y. Shaya Fainman, UC-San Diego
David Frank, IBM
Andrew Kent, New York University
Alexander Khitun, UC-Riverside
Steve Kramer, Micron
Roger Lake, UC-Riverside
Steven Levitan, University of Pittsburgh
Mahesh Mehendale, Texas Instruments
Azad Naeemi, Georgia Tech
Dmitri Nikonov, Intel
Marios Papaefthymiou, University of Michigan
Wolfgang Porod, Notre Dame University
Shriram Ramanathan, Harvard University
Jaijeet Roychowdhury, UC-Berkeley
John Shalf, Lawrence Berkeley National Laboratory
Paul Solomon, IBM
Kang Wang, UCLA
Ian Young, Intel

Participating Remotely

Debdeep Jena, Cornell University
Supriyo Datta, Purdue University
John Bowers, UC-Santa Barbara
Joerg, Appenzeller, Purdue University

Contents

Introduction	3
Commission to Workshop Participants	3
Workshop Agenda and Methodology.....	4
Questions Posed to Workshop Participants and Key Conclusions from the Breakout Discussions.....	6
<i>Topic 1: What are the most promising research directions in the exploration of new devices for computing, and how should a research program be structured to accelerate progress?.....</i>	6
<i>At the device and circuit level, what are the key factors limiting progress in computing – particularly the energy efficiency of computing?.....</i>	6
<i>Are their promising new device and circuit concepts that should be explored? What are the missing pieces of knowledge or key experimental demonstrations that would allow the true potential to be understood?</i>	9
<i>To explore and understand the promise of a truly new device concept, what disciplines should be brought together? What is the “right size” for a university-based team with this mission? What resources would be needed by the team?</i>	14
<i>Topic 2: What are the most promising research directions in the exploration of new circuit and system architectures based on exploratory or emerging devices, and how should a research program be structured to enable the exploration of such novel architectures?.....</i>	14
<i>From the point of view of circuit and higher level system architecture, what are the key factors limiting progress in computing -- particularly the energy efficiency of computing?.....</i>	15
<i>Are current research directions and funded research efforts sufficient to address these factors? If not, what new research directions would address these "gaps"?</i>	15
<i>Are their promising new architectural concepts based on emerging or exploratory devices that should be explored? What are the missing pieces of knowledge or key experimental demonstrations that would allow the true potential to be understood?</i>	17
<i>What is the best way to explore and understand the promise of novel architecture based on an exploratory device concept? What disciplines should be brought together? What is the “right size” for a university-based team with this mission? What resources would be needed by the team?</i>	20
Summary of Workshop Conclusions	21
References.....	23
Appendix.....	24

Introduction

To quote a widely acknowledged report sponsored by the National Research Council of the United States¹, “The implications of a dramatic slowdown in how quickly computer performance is increasing—for our economy, our military, our research institutions, and our way of life — are substantial. The essential engine that made that exponential growth possible is now in considerable danger. Thermal-power challenges and increasingly expensive energy demands pose threats to the historical rate of increase in processor performance.”

Performance improvements in computing are slowing because both the underlying device technology and the long-established von Neumann computer architecture are rapidly maturing. Research to address this challenge must therefore be cross-disciplinary, encompassing the devices that process, store, and communicate information, as well as the architectures in which those devices are organized. Having recognized this challenge, The National Science Foundation (NSF) and the Semiconductor Research Corporation (SRC) convened thought leaders from the relevant disciplines at the NSF – SRC Workshop on Energy Efficient Computing, April 14-15, 2015 at the Holiday Inn Arlington Ballston, Arlington, VA. Dr. Thomas Theis (IBM Research, on assignment to Semiconductor Research Corporation as Executive Director, SRC Nanoelectronics Research Initiative) led the effort. Prof. Keren Bergman (EE Department Chair, Columbia University) served as co-chair with Dr. Theis. Workshop participants explored device concepts for logic, memory, and communication, and associated circuits and higher level architectures. Looking well beyond currently funded research, they strove to identify the most promising scientific and technical research goals with the potential to greatly extend the practical engineering limits of energy efficient computing. They also discussed critical characteristics of a multidisciplinary research effort that would effectively address those goals.

This report outlines the research needs and opportunities for energy efficient computing as developed by workshop participants. It is provided to NSF to assist in planning future programs for exploratory research aimed at critical long-term needs of the semiconductor industry in particular, and the information technology sector in general.

Commission to Workshop Participants

Workshop participants were challenged to identify and recommend promising future research directions for energy efficient computing. They were asked to consider device concepts for logic, memory, and communication, and associated circuits and higher level architectures. They were asked to look well beyond currently funded research, and propose research directions and advances that could result in orders of magnitude improvements in the energy efficiency of computation. These might be entirely new research directions or outstanding problems to be

solved within established approaches. Finally, all workshop participants were tasked with developing crisp recommendations that could be incorporated in this Workshop Report.

Workshop Agenda and Methodology

The workshop began with introductory remarks by NSF leaders and the workshop Chairs, and self-introductions by workshop participants. Dr. Theis then presented a view of current research aimed at new devices and architectures for computing. He outlined the increasing perception within the microelectronics industry that such research is needed, and the increasing research investments over the last decade by industry working with NSF and other federal government funding agencies. He emphasized the resulting progress and learning. While acknowledging that the research has not yet produced a clear “winner”, he pointed to the current rapid emergence of new device concepts, suggesting that more invention is to come. He argued that a vast and promising landscape of research possibilities remains to be explored, gave some examples of that unexplored territory and challenged workshop participants to propose better examples.

To further challenge the thinking of workshop participants, six keynote speakers briefly shared their views of ongoing research and suggested promising directions for future research:

- Steep Slope Devices: Suman Datta (Penn State)
- Spintronics: Kang Wang (U.C.L.A.)
- Nanophotonics: Y. Shaya Fainman (U.C. San Diego)
- Nano Oscillators for non-Boolean Computation: Steven Levitan (U. Pittsburgh)
- Spin-Wave Computation: Alexander Khitun (U.C. Riverside)
- Digital Phase Logic: Jaijeet Roychowdhury (U.C. Berkeley)

These personal views helped to spark discussion and sharing of alternative views by members of the audience.

The key methodology employed for the remainder of the workshop is evident from the Workshop Agenda. (See the Appendix.) Panelists were charged with answering two broad topical questions:

Topic 1: What are the most promising research directions in the exploration of new devices for computing, and how should a research program be structured to accelerate progress?

Topic 2: What are the most promising research directions in the exploration of new circuit and system architectures based on exploratory or emerging devices, and how should a research program be structured to enable the exploration of such novel architectures?

Each major topic was elaborated and clarified by a list of sub-topical questions which were presented to all workshop participants. In order to facilitate intimate technical discussions, participants were divided into three breakout groups of equal

size for discussion of Topic 1, with each group focusing on a different broad class of exploratory devices. Group 1 focused on “Steep Slope” Devices. Group 2 focused on “Hysteretic” or Memory-like Devices. Group 3 focused on “Oscillatory” or Energy-conserving Devices. This classification scheme is not scientifically rigorous. For example, steep slope devices can display hysteresis. However, it did promote the building of discussion groups of the right size (~ 10 – 12 participants) for productive discussions and composed of individuals with some shared research interests. After each group had extensively discussed and developed consensus answers to the appropriate sub-topical questions under its assigned topic, everyone took a break. Workshop participants then reconvened to address Topic 2 which focused the discussion on new architectures. To promote interactions among participants with diverse interests, the groups were reorganized so that a two-thirds majority of individuals in each group had not been together for the discussion of topic 1. The next day, all workshop participants convened and representatives of each of the six breakout groups presented a concise summary of their group’s conclusions regarding each Topic. Questions and discussion with the audience helped to clarify the conclusions. Conclusions under both Topic 1 and Topic 2 were found to generally agree or to represent complementary points of view. No major disagreements were noted between the conclusions of the two independent breakout groups.

The questions discussed by the breakout groups and the resulting consensus answers and related recommendations to NSF are found in the following section of this report.

Questions Posed to Workshop Participants and Key Conclusions from the Breakout Discussions

Topic 1: What are the most promising research directions in the exploration of new devices for computing, and how should a research program be structured to accelerate progress?

At the device and circuit level, what are the key factors limiting progress in computing – particularly the energy efficiency of computing?

Workshop participants identified the limited ability (with current devices and circuit architectures) to further reduce operating voltage as a very important constraint on continued progress in the energy efficiency of computing. The field effect transistor (FET) has been the dominant device for digital computation for over three decades. To maximize switching speed, ease of design, and reliability of operation, FET circuits are invariably designed to operate in the irreversible or “abrupt” switching limit. In this limit, the switching of a transistor driven by another transistor may be modeled as a time-varying resistance $R(t)$ in series with a capacitance C (the load) connected to a constant voltage V (the power supply or operating voltage). With a constant voltage power supply, it is easy to show that the active power is proportional to CV^2f , where f is the switching frequency. For decades, miniaturization (scaling transistors and wires to ever smaller dimensions) allowed reduction of both C and V while f was increased. However, around 2003 – 2005, the semiconductor industry confronted a fundamental limit to further reduction of transistor operating voltage. Ever since, microprocessors designers have had to severely limit clock frequencies in order to keep total power, area, and power density within economically acceptable and physically achievable bounds.² Workshop participants noted that constraints on allowable power dissipation also contribute to the problem of “dark silicon” – the increasing inability to simultaneously exploit all of the computational resources on a chip or in a system. With stagnating clock frequencies, the evolution toward multicore microprocessors has placed daunting challenges on programmers aiming to optimally utilize and thus obtain meaningful performance gains from these systems.

Workshop participants identified additional key factors which limit progress based on today’s technology, and which must be addressed in the introduction of any new devices and the associated circuits and architectures.

- The increasing resistance, R , of the ever smaller interconnections (local wiring and contacts) between devices is an increasingly important limiter for device and circuit performance. While this parasitic resistance does not directly impact active power, it limits performance by introducing additional RC delay. Reduction of parasitic resistance must continue to be a goal of materials and device research.

- Device-to-device variation, reliability, endurance, and susceptibility of circuits to electromagnetic interference also limit current engineering approaches, and will be important issues to address in the introduction of any new device.
- Today's devices operate at 170 C or higher, balancing the conflicting demands of long term reliability (lower operating temperature) and efficient heat removal (higher operating temperature). Any new device must also operate at some elevated temperature which balances these conflicting demands, and any new device integrated with current silicon CMOS technology (for example, a memory device) must operate at 170 C or higher.

Are current research directions and funded research efforts sufficient to address these factors? If not, what new research directions would address these "gaps"?

Workshop participants were well aware that new device concepts, switching by physical mechanisms different from that of the field effect transistor, may avoid some fundamental limits of the field effect transistor. Further substantial reductions in operating voltage may be possible for *steep slope devices* which switch with a smaller voltage swing. Devices which do not represent digital state by a voltage or a charge packet may not require power proportional to CV^2f . More generally, devices and circuits designed to switch in a reversible (energy-conserving) way would not be not subject to any fundamental lower limit on power consumption.³

Workshop participants generally agreed that current research on exploratory devices and associated circuits and architectures is insufficient to properly address the limiting factors of current technology and explore the broad landscape of possible future technologies. Some observed that government research funding (for basic research relevant to the future of the semiconductor industry) is much greater in Europe and Asia than in the US. Many promising new research directions were suggested.

New *steep slope* devices (beyond tunneling field effect transistors) could allow continued reductions in operating voltage and thus greatly relax constraints on performance due to energy dissipation. To have maximum impact, such devices must have the following attributes:

- small operating voltage (<100mV)
- large conductance ratio (>10⁵ desired)
- operating temperature well above 300K
- input output matching
- hysteresis < 10% of supply voltage
- a viable path to miniaturization which is comparable to or beyond the foreseeable limits of miniaturization of silicon CMOS technology.

Steep slope device concepts explored to date do not simultaneously exhibit all of these attributes, but new concepts continue to emerge and the most promising should be vigorously pursued.

Devices exhibiting *non-volatility* (memory devices or memory-like logic devices) could allow more energy-efficient circuits and architectures, particularly when the system, or parts of the system, must be frequently power cycled. Exploration of such architectures, particularly architectures based on new or emerging device concepts, is just beginning. Ongoing exploratory device research suggests that large improvements in switching energy and switching speed over existing non-volatile devices may be possible. Such advances would greatly increase the interest in architectures which optimally exploit non-volatility. Research aimed at such advances should be strongly supported.

Oscillatory devices and circuits allow new approaches to both Boolean and non-Boolean computation. Many of the proposed devices and circuits are inherently energy conserving, which could lead to energy conserving architectures. However, the value of this approach for energy efficiency is not yet clear. Practical systems must operate reliably despite device variability and environmental fluctuations, and engineering solutions, such as error correction, must, on fundamental grounds, increase energy dissipation. The research is at a very early stage, and many potential approaches may be viable, including micro- and nanoelectromechanical systems (MEMS/NEMS), nanophotonic systems, nanomagnetic systems, and more. There is a strong and immediate need to understand and quantify metrics for performance and energy efficiency and relevant device parameters for each approach. Beyond these initial pursuits lie many engineering research challenges.

Development of a technologically viable approach to *non-Boolean* computation could accelerate progress in performance and energy efficiency for many emerging applications of computing. Researchers should focus on device technologies that are highly compact and readily integrated with established technologies. These new devices may not be competitive with conventional field-effect transistor devices and circuits for computing Boolean functions, but they may offer ultra-low power for non-Boolean operations. While the non-Boolean operations might not form a functionally complete set, ultra-low power non-Boolean devices in conjunction with Boolean logic based on the transistor could result in systems that would have significant overall power-performance advantages over a traditional system solution.

In addition to these three broad areas of emerging device research, workshop participants pointed to emerging circuit design concepts aimed at improving the energy efficiency of computation. Examples include Approximate Computing (for applications such as image processing, where perfect accuracy is not required) and Computing with Unreliable (nondeterministic) Devices. Such concepts may be of particular interest in emerging neuromorphic and cognitive computing architectures. More generally, the suitability of various emerging device concepts (such as those discussed above) to these circuits and architectures should be explored. The co-design of devices and circuits was deemed essential to the optimization of energy efficiency.

Finally, it was noted that the development of multi-scale models for exploratory devices (bridging ab-initio, mean field, NEGF transport, TCAD, and SPICE) would accelerate research and development

Are there promising new device and circuit concepts that should be explored? What are the missing pieces of knowledge or key experimental demonstrations that would allow the true potential to be understood?

Workshop participants outlined many promising device and circuit concepts that are largely or entirely unexplored.

Steep Slope Device Concepts

One promising class of steep slope devices is based on **charge carrier cooling** in device structures which filter the energy or momentum distribution of carriers entering a conduction channel. Devices based on energy filtering by gated tunnel junctions (Tunneling Field Effect Transistors or TFETs) have been extensively studied in recent years, but related device concepts are still largely unexplored. For example, such devices might be based on gated tunneling in two-dimensional superlattices, one dimensional nanowires, and zero-dimensional quantum dot arrays.

Devices based on **gating of a phase transition** potentially offer the abrupt (highly non-linear) switching characteristic that is desired. But what physical system, what approach, will be best? Many phase transitions are possible. Examples include dipole moments, charge density waves, dipole excitons, room temperature excitonic condensates, structural degrees of freedom such as Peierl's transition, Mott-Hubbard transition, edge or interface disorder-induced Anderson localization, superconducting phase transitions, and ionic Coulomb blockade. Furthermore, the phase transition might be gated by voltage, electronic or ionic current, strain, or other variables.

Progress toward realization of such steep slope device concepts will require advances and insights in materials.

- Complex oxides including non-epitaxial oxides grown by conformal chemical methods such as ALD (ferroelectric, antiferroelectric, and multiferroic dielectrics)
- New gap engineered two and one dimensional materials
- Collective phenomena at interfaces ("even thinner" than 2D materials)
- Materials exhibiting electronic or other phase transitions which can be gated at temperatures well above room temperature.

It was noted that computational modeling may be a valuable tool for developing an understanding of growth mechanisms and the influence of defects and disorder on

electronic properties. A worthy goal for the Materials Genome initiative would be modeling-based discovery of new materials exhibiting desirable phase transitions.

Hysteretic Device Concepts

Spintronic (or nanomagnetic) devices are promising for memory and logic. Most devices studied to date are switched by magnetic precession, so that switching times cannot be less than the inverse magnetic precession frequency. In ferromagnetic materials, these switching times are very long (~ 1 nsec) compared to the picosecond time scale for switching of silicon FETs. **Device concepts based on antiferromagnetic or ferrimagnetic materials** could exhibit orders of magnitude faster switching. A viable fast-switching device concept would not involve exchange coupling to a ferromagnetic material. Presumably the “write” operation would be based on a magnetoelectric effect or perhaps by spin transfer torque induced by the spin-Hall or Rashba effects. The “read” operation might be based on anisotropic magnetoresistance, optical detection, resistance change related to a metal-insulator transition, or other possible mechanisms. Finally, any mechanism to increase the speed of magnetic switching (in addition to the above-mentioned approach through antiferromagnetic or ferrimagnetic materials) would also be of interest. For example, hot electron injection could disorder the spins in a magnet, possibly enabling switching on a picosecond time scale.

Hysteretic materials and device structures exhibiting a large change in resistance upon switching would open dramatic new possibilities in energy efficient devices and circuits. Today’s magnetic random access memory (MRAM) devices may exhibit magnetoresistance (MR) ratios ~ 1.5 or 2 , while an MR ratio $>10^4$ is desired. Therefore, research on materials (such as Heusler alloys) with improved spin-filtering characteristics should be encouraged. Perhaps such research could be connected to computational studies under the Materials Genome initiative. Other systems which could enable devices with large on/off resistance ratios include metal insulator transitions in various materials and ferroelectric tunnel junctions (FTJs). Fatigue (wear-out upon repeated switching) is concern for FTJs.

Another broad class of potential devices would be based on manipulation of **spin orbital effects or topological objects**. Skyrmions are topological defects in the magnetization of thin films. They can be created and moved by electrical currents, and offer the potential advantage for domain-wall devices of not being easily pinned by edges or edge defects. Workshop participants pointed out that electric field control of skyrmions would be highly desirable for energy efficiency of skyrmion-based devices. Topological insulators are of current research interest because they offer the possibility of large spin Hall torques for more energy efficient magnetic switching. For such an approach to be effective, thin films of topological insulators must be demonstrated with good bulk insulating properties at room temperature and above. Most materials being investigated use heavy atoms from the lower rows

of the periodic table to get the high spin-orbit interaction necessary for the band inversion associated with the topological edge or surface states. However, such materials tend to have small bulk bandgaps. It was pointed out that materials with large built-in electric fields offer a mechanism for the creation of topological edge or surface states in materials composed of lighter atoms and thus wider bulk bandgaps. Such materials should be explored for their room-temperature bulk insulating properties. Finally, it was noted that magnetic states can also be used to control current on surfaces of topological insulators – a possible approach to magnetic devices with large on/off resistance ratios.

Propagation of spin-waves (magnons) through ferromagnetic metals or insulators is a potential mechanism for **communication of information over short distances**. Short wavelength (exchange energy dominated) spin-waves are of interest because they travel faster. Moreover, the combination of nanomagnets and spin waves enables the concept of holographic memory⁷ with exciting prospects for digital or non-Boolean pattern recognition and content-addressable memory functions.

In addition to the above high-priority research needs, workshop participants cited other potentially productive research paths. Citing a need for information storage approaches that scale to greater densities, they suggested the exploration of approaches that can **exploit three dimensions**. This might be a holographic approach, perhaps based on plasmonics and therefore less limited in density by the wavelength of light. Magnetic storage might also exploit the vertical dimension. The development of a superior **selector** – a device with a steep IV characteristic, like a diode, that can be integrated into a cross-connect memory array – would reduce the cost and increase the achievable density of three-dimensional arrays of memory devices. **Molecular memory**, based on charge trapping in molecular systems remains attractive, but it has been difficult to find systems in which energy barriers and tunneling distances provide the desired stability of memory state. Exploration of self-assembled molecular films exhibiting ferroelectric or anti-ferroelectric properties was mentioned as one potential path to achieving the necessary stability. Can such a system, or any molecular memory, meet stability requirements while being switched by a voltage on the order of ~100 mV?

Oscillatory Device Concepts and Circuits

A variety of oscillatory systems have been proposed for computation. Relaxation and spiking oscillators have been explored in the context of neural network applications, but much of the workshop discussion focused on arrays of coupled near-harmonic oscillators. Although they do not “spike”, such arrays are believed to exhibit key attributes of biological neural networks for tasks such as pattern recognition. Energy conserving oscillatory systems have also been explored to some extent for digital (Boolean) computation.⁴

Four different approaches to oscillatory computation were considered at the workshop. These include:

- 1) The use of weakly coupled non-linear oscillators to perform pattern matching and filtering operations by means of their tendency to synchronize in phase and frequency when starting from nearby states
- 2) The use of neuromorphic oscillatory networks to implement computing structures like convolutional neural networks and Boltzmann machines. Here, networks of spiking oscillators reinforce or inhibit each other's activities leading to convergence to a functional value.
- 3) The use of networks of oscillators to perform Boolean digital phase logic. Here, logic levels are encoded as phase (0, 180) relative to a reference and simple circuits can be used to perform both logic and memory operations.
- 4) The use of photonic modal oscillatory networks. For example, recent advances in multi-wavelength and mode locked laser arrays may offer platforms for phase inter-locking. Optical ring resonators use the path length of the ring to selectively process information. Single optical waveguides can carry multiple frequencies thus enabling the "broadcasting" patterns of information to be processed in parallel by arrays of resonators.

The first three of these are "technology agnostic" the concepts translate to many devices and even modalities (spintronics, MEMS, nanophotonics, etc.) The last one takes advantages of the unique properties of information encoding in optical signals.

For both Boolean and non-Boolean computations, many potential technological approaches may be viable, including micro- and nanoelectromechanical systems (MEMS/NEMS), nanophotonic systems, nanomagnetic systems, spin-wave systems, and more. With current capabilities in nanofabrication, many of these systems can be made very compact. However, the scaling properties and ultimate scaling limits of various oscillator technologies are not well understood. Moreover, conference participants emphasized the need to understand and quantify other metrics for performance and energy efficiency for various possible technologies. For non-Boolean applications such as pattern recognition, these metrics should encompass both sensing (interfacing with the outside world) and control (how oscillators interact with one another or are set to a stored pattern). It was noted that the potential for the same oscillators to both sense and compute may be particularly valuable when the sensing modality matches the signal from the environment – for example, when the oscillators sense light for the purpose of image recognition. Another promising path can utilize optical oscillators with linear and nonlinear feedback coupled with fibers to provide inputs as sensing devices and outputs for actuation. Devices based on propagation of waves (for example, spin waves or Nano photonics) can utilize interference, holographic processing, and non-linear dynamics for non-Boolean computation.

Specific device metrics for various oscillator technologies would include:

- Frequency of oscillation
- Quality factor, Q , (or some equivalent for systems such as ring oscillators)

- Response time (time to lock to a new state)
- Usable resolution in carrier frequency
- Method(s) for reading information into and out of each device.
- Mechanism(s) for communication between devices
- Range of interaction between devices (potentially limiting the kinds of computations that can be realized).
- Effects of linear and nonlinear interaction between devices
- Mechanisms and utilization of amplitude and phase dynamics

The oscillators in these systems are not “functionally complete” gates. While they accept multiple inputs and exhibit non-linearity and thresholding behavior, they exhibit symmetry between input and output and *cannot* be directly cascaded without auxiliary circuits. Thus devices cannot be treated separately from circuits and applications. Indeed, workshop participants identified several important research issues that can only be fully addressed by considering devices, circuits, architecture, and applications as a whole.

Thus the following questions for exploratory research were raised. What are the promising computational models, and what are the promising representations of information in oscillatory networks? How will variability from device to device and in coupling between devices affect the predictability and repeatability of results? Engineering decisions must be made regarding the coupling between oscillators – the topology, impedance matching, whether or not the coupling is static or dynamic, and whether or not a reference oscillator is used. For non-Boolean applications, the quality or reliability of pattern match detection will depend on the way output signals are processed. The energy efficiency of the system will depend in part on the choice of support circuitry for storing and retrieving pattern templates. Are there advantages to using dynamical principles of winnerless competition as a mechanism for sequential information processing and robust representation of the transient cognitive mode dynamics? Is it possible to achieve robustness and reproducibility over a wide range of control parameters and incoming signals and noise immunity? Could networks be arranged in hierarchical levels to perform complex tasks? Are there physical mechanisms other than electronics for realizing such systems?

For both Boolean and non-Boolean computations, photonics technology may play a unique and important role because it not only enables high speed communications on a silicon chip, but also provides a fundamental and unique interface between electrons and photons (e.g., lasers, nonlinear optical materials, etc.). These applications may benefit from development of new nonlinear optical metamaterials and light emitters based on novel nanoscale engineered materials (e.g., electro-optic transition metal oxides and their heterostructures integrated with Si in combination with metal-dielectric- semiconductor (gain) metamaterials). The integration and co-optimization of the newly discovered nonlinear optical metamaterials with nanoscale device structures is poised to enable new physical layer capabilities in photonic systems.

To explore and understand the promise of a truly new device concept, what disciplines should be brought together? What is the “right size” for a university-based team with this mission? What resources would be needed by the team?

Recommendation: Research teams should bring together top experts from the various disciplines needed to explore and demonstrate a particular device concept. For example, a team might include materials scientists, condensed matter physicists, device engineers, experts in metrology, circuit designers and system architects. Of course, choice of team members and relevant disciplines should be done in the context of the particular concept to be explored. For example, in the exploration of oscillator-based computing, expertise in architecture and applications was seen as absolutely essential. Capability to pursue first-principles design of new materials, interfaces, and nanostructures, and processes for synthesis would be valuable in many cases. For this reason among others, research teams should have connections to the national labs. Connections to industry would also bring highly desirable perspectives.

Oscillator based computing represents, perhaps, the clearest example of the need for a vertically integrated research team that can relate the advantages and capabilities (and limitations and challenges) of new devices and circuits to the architectural and algorithmic level. In the past, the “abstraction hierarchy” (device, circuit, logic-element, module, component, system) has been the way in which circuit designers, logic designers, and system designers could compartmentalize their work while sharing well understood interfaces between levels. However, for oscillatory systems, new abstractions are needed so that designers can perform competitive trade-off analyses and develop efficient CAD tools for simulation, synthesis and optimization. For example, there is no “RTL level” representation for a network of coupled oscillators. Therefore, in the development and proof of concept phases of this research, tightly integrated groups of researchers from all levels will be essential.

The optimum size of the research team is a question of balance. In some cases, one or a few individual investigators might pursue fundamental issues or rapidly explore truly new ideas. Most of the discussion, however, focused on the optimum size of larger multidisciplinary teams. A small team of, say, five individuals will be more agile, and can more directly interact across disciplinary boundaries. A larger team (perhaps a multidisciplinary, multi-university Center) can explore a broader class of research options when it is unclear which will win.

Topic 2: What are the most promising research directions in the exploration of new circuit and system architectures based on exploratory or emerging devices, and how should a research program be structured to enable the exploration of such novel architectures?

From the point of view of circuit and higher level system architecture, what are the key factors limiting progress in computing -- particularly the energy efficiency of computing?

Workshop participants observed that architectures suited for new and emerging applications of computing, such as the Internet of Things (IoT), may differ from architectures suited for established “scalable” applications (from cell phones on up to server farms and numerically intensive computing centers). A primary challenge for high-performance computing is the energy and performance cost of data movement and data storage, a reflection of the ongoing drive to increase memory capacity without adversely impacting performance. Another important challenge is the inefficient use of resources (e.g. dark silicon) resulting from bursts of intense computation followed by long idle times with power consumption dominated by transistor leakage current. On the other hand, devices in the IoT space must be even more energy efficient, driving greater emphasis on standby power, and the ability to frequently and instantly switch the system on and off. Volatile memory and logic devices are not well suited for this type of computing. Nevertheless, many important circuit and higher-level architectural issues are of concern across all application categories; the slow and energy-inefficient memory hierarchy, the lack of a dense and energy efficient non-volatile memory, and leakage control in low-voltage circuit design. Every application must balance the conflicting demands of dynamic power versus static power and energy efficiency versus the need for margins to allow for noise immunity and device-to-device variability. The drive for energy efficiency also demands consideration of application specific circuits and architectures (accelerators) versus the programmability of less energy-efficient instruction set processors.

Summing up, two fundamental drivers, energy efficiency and cost, are relevant to both computing in the small (IoT) and computing in the large (server farms and numerically intensive computing). Solutions in one application space could very well impact others.

Are current research directions and funded research efforts sufficient to address these factors? If not, what new research directions would address these "gaps"?

Much more research will be needed to develop and understand circuits and architectures that would leverage the capabilities of emerging devices. Workshop participants pointed to the following broad research questions:

- What, for a proposed technology, will be required to beat “end of line” CMOS, and what are the fundamental limits and ultimate prospects of various emerging device technologies? For example, what architecture(s) can best leverage nonvolatile devices, even if slower switching? As another example, what architecture(s) can leverage adiabatic/reversible devices, even if slower switching? Clocking should be included in such studies. In these contexts, what is

the role synchronous logic and clocking in limiting energy efficiency? Are there viable asynchronous solutions?

- What architectures and emerging technologies will minimize communication costs? Optical data movement enabled by ultra-low power nanophotonic integration may offer energy-efficient high-bandwidth system-wide connectivity. How can system architectures be designed to take advantage of these unique capabilities?
- Architectures based entirely on local connections, such as convolutional neural networks and systolic arrays, are suited to specialized functions (image recognition and signal processing, respectively). Are there architectures with reduced long range connections that are suited for more general purpose computing? Steep slope devices might enable truly low-voltage signaling and communication, perhaps in conjunction with improved error coding or pre-compensation. Could such devices enable circuit architectures in which islands of higher voltage devices communicate over truly low-voltage (~ 10 mV?) links? Finally, could new device concepts make practical the coding of information in multiple dimensions, such as amplitude and phase, or multiple frequencies?
- How can the memory hierarchy be collapsed and/or made more efficient? More generally, how can logic and memory best be integrated? Can multivalued memory or logic enhance energy efficiency, and do any of the emerging device technologies lend themselves to such an approach? New architectures based on non-volatile memory and possibly, non-volatile logic, potentially address some pressing issues related to energy efficiency, but the benefits have not been quantified. Can spintronic or other hysteretic devices enable superior reconfigurable and adaptive circuits – a new architecture for reprogrammable logic?
- What architectures will maximize the value of 3D integration? Chip stacking based on through-silicon vias (TSVs) provides limited interconnection density or bandwidth between layers. Can new devices and integration processes enable stacking of active device layers with higher connectivity between layers?
- What truly new architectures are enabled by emerging devices? Associative memories based on coupled oscillators are an example, but more examples must be discovered and explored to understand the potential. What device properties would provide the most benefit to various neuromorphic circuit architectures?
- What is the best combination of device and (specialized) architecture for various important and hard computational tasks? (Examples might include optimization problems, multiply and accumulate as an operator in physical simulation, multi-agent control, signal processing, pattern recognition, and classification.)

Workshop participants also noted broad and important research issues that will have to be addressed in the introduction of any new device and associated architecture. Design tools, based on compact device models that are fast and scalable, could enable rapid exploration of broad classes of novel circuits and systems – facilitating the exploration of new architectures. It was noted that such tools would be particularly useful right now for simulation and modeling of arrays of coupled oscillators – facilitating the development of the new programming

models that will be needed for such systems. Since data movement at all scales plays such a pivotal role in the energy efficiency and performance of computation, design platforms should incorporate tools that can fully assess associated communication and interconnectivity metrics. Holistic exploration of novel materials and devices, design methodologies, and architectures is needed to realize new computational paradigms that address the data movement challenges. For example, scalable nanophotonic systems and 3D stacking have the potential to enable large improvements in the energy efficiency of data movement. Understanding the impact of these emerging technologies at the system level is critical to developing new architectures and assessing the associated gains in energy efficiency.

Another broad set of research issues that will have to be addressed is the development of circuits and architectures which comprehend and address issues such as device-to-device variation, sensitivity to noise, power supply variations, and other environmental factors, and the effects of aging and wear out on new devices and materials systems. It was noted that device concepts amenable to implementation as four-terminal devices may allow circuit architectures which are relatively immune to power supply variation and noise. In this respect, optically coupled devices may also have very desirable characteristics.

Finally, consideration of fundamental limits should be included in the co-exploration of emerging device and associated architectures. While sources of energy dissipation and limits to performance are well understood in established technologies, this is not the case for emerging technologies. Thus, the pursuit of new device/architecture combinations should increasingly seek evidence that the hoped-for performance and efficiency improvements are not precluded at the outset by physical law and other fundamental considerations

Are their promising new architectural concepts based on emerging or exploratory devices that should be explored? What are the missing pieces of knowledge or key experimental demonstrations that would allow the true potential to be understood?

Workshop participants identified a few areas of architectural inquiry that will be critical to the development and application of emerging devices.

Architectures for Heterogeneous Systems

It seems unlikely that a single device will optimally support all useful architectures. Research must identify strategies for integrating multiple application-specific sub-units, possibly involving very different devices. It may be desirable to integrate non-Boolean coprocessors or accelerators with general purpose digital processors. Interfaces between subsystems must not dominate system resources.

Architectures Exploiting Non-volatility

Microarchitectures that leverage the non-volatility of some emerging devices are needed. At least three levels of architectural innovation should be explored.

- Small changes to established processor architectures, in which non-volatile devices are utilized in latch circuits and register files to facilitate functions such as power gating and check-pointing
- Incorporation of non-volatile devices in reconfigurable architectures such as field programmable gate arrays (FPGAs) to promote attributes such as run-time reconfigurability and more efficient implementation of diverse functions
- Implementation in novel architectures such as processor-in-memory (smart pixels or thousands of small cores) or neuromorphic systems (storing connection weights)

It was noted that hysteresis in a properly designed device may confer an advantage in noise immunity. Furthermore, although hysteresis adds some energy dissipation, this dissipation will normally be small compared to the dissipation involved in charging and discharging interconnection capacitances. The exception would be energy-conserving circuits for which hysteretic devices are less likely to be suited.

Architectures optimizing data movement

Such architectures might, for example, be specifically designed for optical data movement, providing new computational capabilities by merging data processing and communication functions. Such architectures might utilize non-volatile devices which combine the functions of logic and memory, and might utilize interconnections compatible with emerging devices and new state variables. Such architectures might be dynamically adaptive. They might be cellular in organization, suited for functions such as image processing which require few global interconnections.

Neuromorphic Architectures

Many neuromorphic architectures have been proposed and studied, but the optimum coupling of architecture and associated devices for a given function is an unsettled and active area of research. Much research has been aimed at Hopfield networks, where each node takes a binary value depending on whether or not the inputs to the node exceed its threshold. Devices which allow continuous tuning of a resistance, perhaps based on metal-insulator or other phase transitions, may allow more compact and energy-efficient implementation of the weighting of network connections. Spiking neural networks and the implementation of spike timing dependent plasticity (STDP) are more recent developments. In such networks, the relative timing (or phase) of inputs to nodes is important, since it makes possible long term potentiation and inhibition (learning) during operation, rather than in a separate training process. The use of systems of coupled near-harmonic oscillators to implement associative memories is a related approach. Optimization of device characteristics for both spiking network architectures and for networks of near-harmonic oscillators (see below) is a largely unexplored problem.

Architectures enabled by oscillatory devices

The energy-conserving nature of many oscillatory systems suggests the possibility of highly energy efficient computing for both Boolean and non-Boolean applications. However, practical engineering solutions for reliable system operation will involve energy dissipation that is not encompassed by the quality factor, Q , of the oscillators. First, any error correction or feedback control process will result in additional dissipation. Furthermore, it will be wasteful to have oscillators running while doing no computation, regardless of the system Q . Thus systems might be designed for piecewise continuous operation. In the non-Boolean domain, a set of inputs would be provided to an oscillator array, and the oscillators would be run only long enough to settle into a stable state so that the output can be read. A more challenging research problem would be the design of systems for continuous operation, in which the time evolution of the system is essential to the computation. The Hierarchical Temporal Memory (HTM) of George and Hawkins⁵ is an example of a computational model that might be implemented in such an architecture.

More generally, workshop participants emphasized the intertwining of device- and circuit-level considerations with architecture and applications, particularly for non-Boolean applications. Fundamental issues, such as the number of patterns that can be stored by an associative memory consisting of N coupled oscillators, are understood only in special cases. The optimum representation of information and the value of various computational models may depend on the characteristics of the oscillators and the topology of the interconnections. Furthermore, these systems appear to be well-suited for a set of valuable but specialized functions. Can the domain of applications be substantially broadened? For functions such as associative memory, how do these systems compare with other approaches? To quantify such comparisons, a set of challenge problems (benchmarks) representing both Boolean and non-Boolean applications should be developed.

Architectures enabled by photonic devices

The established trend of using photonic devices and interconnections to replace wires on shorter and shorter length scales will, no doubt, continue. One promising possibility is based on the emergence of devices for direct conversion of photons to excitons. These excitonic devices⁶ might enable detection, storage, processing and transmission of data packets without the need for power-hungry electronic circuits such as drivers or trans-impedance amplifiers.

Looking beyond the ongoing replacement of copper interconnections, emerging compact non-linear optical devices may open new architectural possibilities such as “computation in the network”. Co-design of devices and architecture will be necessary to explore the possibilities. Right now, most optical “switches” are electrically gated. The demonstration and development of a compact all-optical switch with gain would make possible an all-optical memory, greatly expanding the

architectural possibilities. The demonstration and development of optical amplifiers with broad gain bandwidth would make possible new switching fabrics for wavelength division multiplexed communication and signal processing.

Additional architectures

Workshop participants also pointed to ongoing research in circuits and architectures for Stochastic Computing, Approximate Computing and Cognitive Computing. Research to date has been largely based on established device technology, but emerging device concepts introduce new possibilities. For Neuromorphic or Cognitive Computing, devices which compactly implement the functions demanded by artificial neural network (ANN) algorithms, may be of great value. One specific and well-known example is the synaptic weighting function. This requires many transistors in a conventional digital design, but device exhibiting a resistance which varies *linearly* in response to cumulative current or voltage spikes should enable a more compact and energy efficient analog implementation of the weighting function.

What is the best way to explore and understand the promise of novel architecture based on an exploratory device concept? What disciplines should be brought together? What is the “right size” for a university-based team with this mission? What resources would be needed by the team?

Workshop participants affirmed their view, already discussed under Topic 1, that research teams must be large enough to include the various technical disciplines necessary to consider the entire “stack”. Teams might include experts on materials and devices, devices and circuits, and circuits and microarchitecture. While it was again noted that small agile teams can be ideal for quick explorations, it was generally agreed that the focus should be on “device aware architecture research” and “architecture aware device research”. Finally, it was noted that progress is likely to be slow, and will require research sponsors to take the “long view”.

Summary of Workshop Conclusions

Broad Conclusions

Research teams should address interdisciplinary research issues essential to the demonstration of new device concepts and associated architectures. Any new device is likely to have characteristics very different from established devices. The interplay between device characteristics and optimum circuit architectures therefore means that circuit and higher level architectures must be co-optimized with any new device. Devices combining digital and analog functions or the functions of logic and memory may lend themselves particularly well to unconventional information processing architectures. For maximum impact, research should focus on devices and architectures which can enable a broad range of useful functions, rather than being dedicated to one function or a few particular functions.

Prospects for New Devices

Many promising research paths remain relatively unexplored. For example, the gating of phase transitions is a potential route to “steep slope” devices that operate at very low voltage. Relevant phase transitions might include metal-insulator transitions, formation of excitonic or other electronic condensates, and various transitions involving structural degrees of freedom. Other promising mechanisms for low-power switching may involve transduction. Magnetoelectric devices, in which an external voltage state is transduced to an internal magnetic state, exemplify the concept. However, transduction need not be limited to magnetoelectric systems.

In addition to energy efficiency, switching speed is an important criterion in choice of materials and device concepts. For example, most nanomagnetic devices switch by magnetic precession, a process which is rather slow in the ferromagnetic systems explored to date. Magnetic precession switching in antiferromagnetic or ferrimagnetic materials could be one or more orders of magnitude faster. Other novel physical systems could be still faster. For example, electronic collective states could, in principle, be switched on sub-picosecond time scales.

More generally, devices based on computational state variables beyond magnetism and charge (or voltage) could open many new possibilities.

Another relatively unexplored path to improved energy efficiency is the implementation of adiabatically switched devices in energy-conserving circuits. In such circuits, the phase of an oscillation or propagating wave may represent digital state; devices and interconnections must together constitute circuits that are non-dissipative. Nanophotonic, plasmonic, spin wave or other lightly damped oscillatory systems might be well-suited for such an approach. Researchers should strive to address the necessary components of a practical engineering solution, including mechanisms for correction of unavoidable phase and amplitude errors.

Networks of coupled non-linear oscillators have been explored for non-Boolean computation in applications such as pattern recognition. Potential technological approaches include nanoelectromechanical, nanophotonic, and nanomagnetic oscillators. Researchers should strive for generality of function and should address the necessary components of a practical engineering solution, including devices, circuits, and architectures that allow reliable operation in the presence of device variability and environmental fluctuations.

Prospects for New Architectures

While appropriate circuits and higher level architectures should be explored and co-developed along with *any* new device concept, certain novel device concepts may demand greater emphasis on higher-level architecture. For example, hysteretic devices, combining the functions of non-volatile logic and memory, might enhance the performance of established architectures (power gating in microprocessors, reconfiguration of logic in field programmable gate arrays), but perhaps more important, they might play an enabling role in novel architectures (compute in memory, weighting of connections in neuromorphic systems, and more). As a second example, there has been great progress in recent years in the miniaturization and energy efficiency of linear and non-linear photonic devices and compact light emitters. It is possible that these advances will have their greatest impact, not in the ongoing replacement of metal wires by optical connections, but rather in enabling new architectures for computing. Computation “in the network” is one possible direction. In general, device characteristics and architecture appear to be highly entwined in oscillatory or energy-conserving systems. Key device characteristics may be inseparable from the coupling (connections) between devices. For non-Boolean computation, optimum architectures and the range of useful algorithms will depend on these characteristics.

In addition to the examples above, many other areas of architectural research might leverage emerging device concepts to obtain order of magnitude improvements in the energy efficiency of computing. Research topics might include architectures for heterogeneous systems, architectures that minimize data movement, neuromorphic architectures, and new approaches to Stochastic Computing, Approximate Computing, Cognitive Computing and more.

References

1. "The Future of Computing Performance: Game Over or Next Level?" S. H. Fuller and L. I. Millett, eds., National Academies Press, Washington, D.C. (2011).
2. T.N. Theis and P.M. Solomon, "In Search of the 'Next Switch': Prospects for Greatly Reduced Power Dissipation in a Successor to the Silicon Field-Effect Transistor," Proc. IEEE **98**(12), pp. 2005-2014 (2010).
3. R. Landauer, "Dissipation and Noise Immunity in Computation and Communication", Nature **335**, pp. 779-784 (1988).
4. T. Wang and J. Roychowdhury, "PHLOGON: PHase-based LOGic using Oscillatory Nano-systems," in Unconventional Computation and Natural Computation, Proc. 13th International Conference, UCNC 2014, London, ON, Canada, July 14-18, 2014, pp. 353-366 (Vol. 8553 of Lecture Notes in Computer Science).
5. D. George and J.C. Hawkins, Trainable Hierarchical Memory System and Method, US Patent No. 7,739,208, June 15, 2010.
6. P. Andreakou *et al.*, "Optically Controlled Excitonic Transistor," Applied Physics Letters **104**, 091101 (2014).
7. A. Kozhevnikov, *et al.*, "Pattern Recognition with Magnonic Holographic Memory Device," Applied Physics Letters **106**, 142409 (2015).

Appendix

WORKSHOP AGENDA

Tuesday, April 14, 2015

Introductory Session (8:00 – 9:00)

- 8:00 Welcome:
Dr. James Kurose, Assistant Director of the CISE Directorate and
Dr. Pramod Khargonekar, Assistant Director of the ENG Directorate
- 8:20 Welcome from the Chairs:
Thomas Theis (SRC-NRI); Keren Bergman (Columbia)
- 8:30 Introductions: all participants and attendees
- 8:40 Commission and Methodology: Theis and Bergman

Key Research Areas (9:00 – 11:20): Six speakers share their view of ongoing research and suggest promising directions for future research. It is hoped that these personal views will spark discussion and sharing of alternative views by members of the audience.)

- 9:00 Steep Slope Devices: Suman Datta (Penn State)
- 9:20 Spintronics: Kang Wang (U.C.L.A.)
- 9:40 Nanophotonics: Y. Shaya Fainman (U.C. San Diego)
- 10:00 Break
- 10:20 Nano Oscillators for non-Boolean Computation: Steven Levitan (U. Pittsburgh)
- 10:40 Spin-Wave Computation: Alexander Khitun (U.C. Riverside)
- 11:00 Digital Phase Logic: Jaijeet Roychowdhury (U.C. Berkeley)

Charge to the breakout groups (11:20 – 12:00)

Review and clarify proposed discussion topics and subtopics
What additional subtopics should be addressed?

Lunch (12:00 – 1:00)

Breakout Sessions (1:00 – 5:45)

(Theis and Bergman will each join one group.)

- 1:00 Session 1
Breakout Groups independently discuss Topic 1: What are the most promising research directions in the exploration of new devices for computing, and how should a research program be structured to accelerate progress?
- 3:00 Break
- 3:15 Session 2
Breakout Groups independently discuss Topic 2: What are the most promising research directions in the exploration of new circuit and system architectures based on exploratory or emerging devices, and how should a research program be structured to enable the exploration of such novel architectures?

6:30 Working Dinner

WORKSHOP AGENDA (continued)

Wednesday, April 15, 2015

Summarizing the Recommendations (8:00 – 12:00)

- 8:00 Reports from Breakout Groups on Topic 1
- 9:00 Discussion and Summary (facilitated by Bergman)
- 9:15 Reports from Breakout Groups on Topic 2
- 10:15 Discussion, and Summary (facilitated by Theis)

- 10:30 Break

- 10:45 General discussion of Recommendations (facilitated by Theis and Bergman)
 - Identify primary and cross-cutting themes from Breakout Groups.
 - Reconcile any competing recommendations from Breakout Groups
- 11:30 Summary Discussion

12:00 – 1:00 Working Lunch

1:00 Adjourn