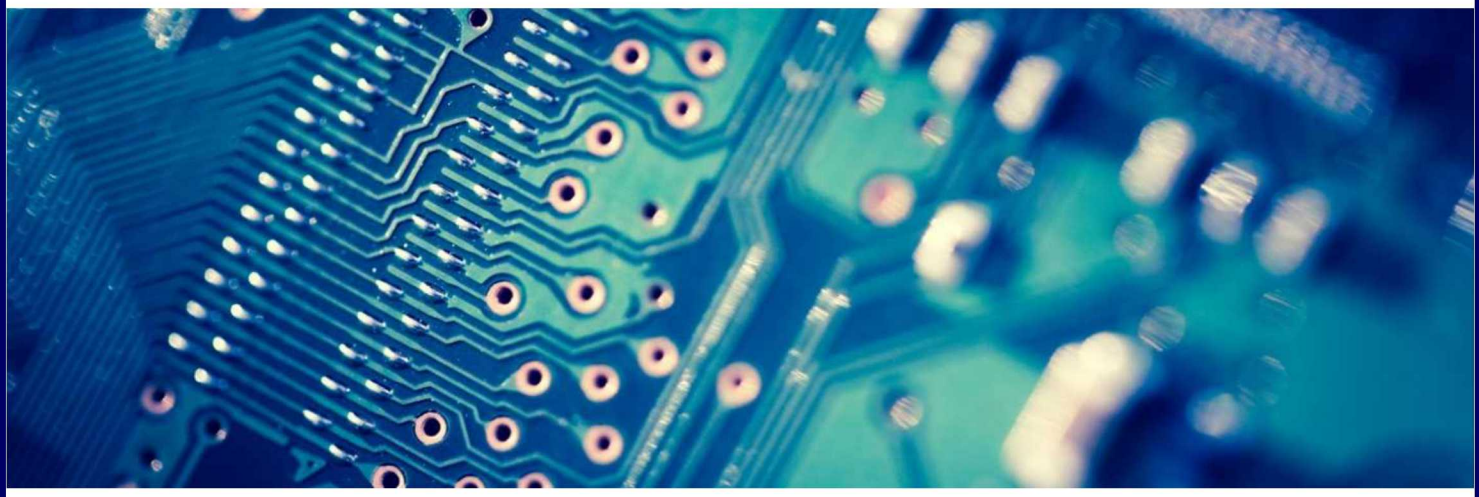
SAND2018-1579R

2/9/2018



Solving the Information Technology Challenge Beyond Moore's Law

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Table of Contents

Abstract

Moore's law summarizes an observation made in 1965 by Intel co-founder Gordon Moore that the number of transistors per square inch on integrated circuits had doubled every year since they had been introduced. He predicted that the trend would continue well into the foreseeable future. In 1975, he updated his prediction to double every two years. The doubling phenomenon held true for quite some time, enabling the global Information Technology industry to reach upwards of $4 trillion; however, existing technologies are reaching their limit. Within the next decade, it will become possible to create lithographically produced devices with characteristic dimensions in the 3nm–5nm range. This range corresponds to a dozen or fewer Si atoms across critical device features and will therefore be a practical limit for controlling charge in a classical sense. At this point, Moore's law will no longer be applicable.

Yet, society has come to rely on the benefits provided by Moore's Law. In the past decade, microelectronic devices and associated software have progressed from supporting individual consumer products to providing vital elements of the infrastructure that is now critical for the nation's quality of life and government function. The tapering of Moore's Law in the coming decade will result in greater energy need in order to maintain an equivalent increase in computing power. Information Technology, already the fastest growing consumer of energy worldwide, will consequently require even more energy if computational ability is to continue advancing; otherwise U.S. computing growth will be significantly restricted, severely threatening the nation's ability to solve pressing scientific and national security problems. This transition will require effort on a decadal time scale, so it is critical to start laying the strategic foundation for change now.

We are proposing to apply unique DOE capabilities to Public-Private programs for basic/applied research to accelerate the development of energy efficient IT beyond the end of current roadmaps, and to maintain an advanced manufacturing base in the economically critical semiconductor space. This partnership will allow DOE to leverage significant industry investments with the goal of enabling low-power computing and suitably low cost smart grid and building electronics.

This report provides an overview of a workshop held on July 27-28, 2016 at Sandia National Laboratories in Albuquerque to itemize the DOE laboratory capabilties and provide a high level organization of those capabilties into a full evaluation framework for new computing paradigms that spans from fundamental breakthroughs in materials and devices to full system architectures and software environments (see Appendix A: Workshop Agenda and Breakout Group Topics for agenda).

Executive Summary

Moore's law summarizes an observation made in 1965 by Intel co-founder Gordon Moore that the number of transistors per square inch on integrated circuits had doubled every year since they had been introduced. He predicted that the trend would continue well into the foreseeable future. In 1975, he updated his prediction to double every two years. The doubling phenomenon held true for quite some time, enabling the global Information Technology industry to reach upwards of $4 trillion; however, existing technologies are reaching their limit. Within the next decade, it will become possible to create lithographically produced devices with characteristic dimensions in the 3nm–5nm range. This range corresponds to a dozen or fewer Si atoms across critical device features and will therefore be a practical limit for controlling charge in a classical sense. At this point, Moore's law will no longer be applicable; and the growth rate will flatten by 2025.

Yet, society has come to rely on the benefits provided by Moore's Law. In the past decade, microelectronic devices and associated software have progressed from supporting individual consumer products to providing vital elements of an infrastructure that is now critical for the nation's quality of life and government function. The tapering of Moore's Law in the coming decade will result in greater energy need in order to maintain an equivalent increase in computing power. Information Technology, already the fastest growing consumer of energy worldwide, will consequently require even more energy if computational ability is to continue advancing; otherwise U.S. computing growth will be significantly restricted, severely threatening the nation's ability to solve pressing scientific and national security problems. This transition will require effort on a decadal time scale, so it is critical to start laying the strategic foundation for change now.

As mentioned above, information technology (IT) represents the fastest growing consumer of energy. Uncontrolled, this demand will have significant implications on the U.S. energy landscape. In one example of projected IT energy growth, Cisco reports (Cisco Global Cloud Index, 2013–2018) that data center traffic (a useful metric for energy demand) is projected to grow at a compound annual rate (CAGR) of 23% from 2013-18. With no improvement in computing efficiency (i.e., the end of Moore's Law), this growth is expected to be directly reflected in increased energy demand, from 91 billion kilowatt-hours in 2013 to 252 billion kilowatt-hours in 2018. Simply *meeting* this increased demand would require 60 new 500-megawatt power plants. The energy requirement is likely to be exacerbated in the next decade with the end of conventional Moore's Law technology scaling.

DOE has a **unique opportunity** to apply unique DOE capabilities to Public-Private programs for basic/applied research to accelerate the development of energy efficient IT beyond the end of current roadmaps as well as to maintain an advanced manufacturing base in the economically critical semiconductor space. These programs will allow DOE to leverage significant industry investments with the goal of enabling low-power computing and suitably low cost smart grid and building electronics.

Solving the daunting technological, economic competitiveness, and energy challenges described above will require both *manufacturing technology* advances, allowing the continuation of Moore's Law from the device patterning perspective, as well as groundbreaking advances in *device technology, going beyond*

CMOS, system architecture, and programming models, to allow the energy benefits of scaling to be realized. Only by considering these design aspects (see Conclusion and Appendix B) in combination with the manufacturing challenges can we expect to make cohesive progress in all areas to bring about the necessary advances to reduce energy requirements.

In addition to containing the growth of IT related energy demand (Figure 1), the output of this work will provide a path to sustaining exponential growth in computing capabilities to enable new scientific discoveries, maintain U.S. competitiveness in all segments of the computing market (from IoT, to datacenters, to supercomputing), and thus guarantee U.S. economic competitiveness and national security.

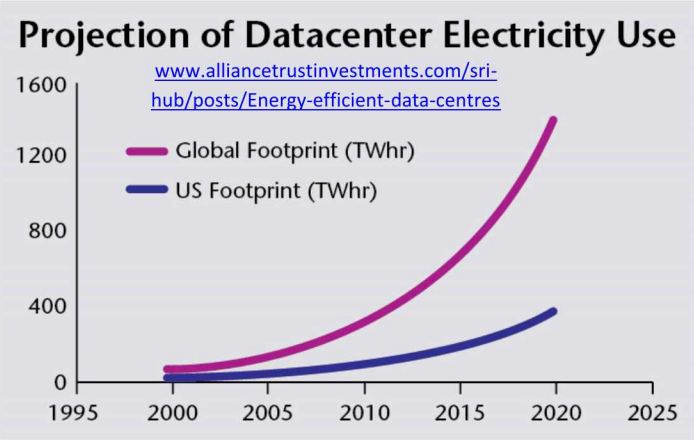


Figure 1: Projected datacenter electric use is projected to become a dominant consumer of worldwide electricity production.

To meet the goal of broad societal impact, we must not only ensure the transition of basic research to high volume manufacturing but more fundamentally shape basic research from the start, with an eye to manufacturability. This will be achieved through the development of a multi-lab ecosystem serving as a unified facility that can evaluate and demonstrate the manufacturing and energy savings feasibility of next generation technology options. Technologies will be rigorously evaluated for potential benefits on energy, implications on architecture, and programming paradigms. The most promising technologies will be evaluated for issues around high-volume manufacturing followed by ramp-up demonstration and their ability to deliver on the energy requirements. This phase will depend heavily on identifying specific manufacturing/device materials, leveraging the capabilities of the Materials Project and HPC to accelerate development using modeling and "virtual cycles of learning." Manufacturing feasibility will also include demonstration of patterning techniques needed to support the various technologies and scaling of those technologies. Delivering on this broad vision will require not only significant advances in the following four major thrust areas individually, but also the smooth and complete the integration of the successful outcomes from each of the four thrusts to achieve the final goal.

1. The ***Devices and CMOS Technology*** thrust will explore, identify, model, and demonstrate the new materials and devices for ultra-efficient computing. Examples include low voltage

transistor concepts such as the TFET, photonic devices, spintronics, and novel memory devices. *The goal is to use DOE's state-of-the-art materials synthesis, characterization and modeling capabilities in conjunction with materials development frameworks, such as the Materials Project and HPC high-throughput search for new materials, to increase throughput for discovering new electronic materials and devices by a factor of 1000x* over current methods.

2. The ***Advanced Manufacturing and Integration*** thrust will leverage DOE's expertise in EUV lithography and materials to develop novel nanomanufacturing methods, including EUV lithography, heterogeneous integration of advanced photonics and wide bandgap devices, and 3D stacking, all of which will contribute to higher density and will enable memory layers on top of logic layers, or even multiple memory and logic layers interleaved. This radical change challenges assumptions embedded in current architectures, and will provide a new dimension to extend Moore's Law scaling.
3. The ***Architecture*** thrust will apply DOE expertise in advanced computing to exploit the new device and materials systems and packaging technologies developed in the first two thrusts. Components will include accelerators, on-chip wide-bandgap devices, photonic blocks, and emerging memory devices. The objective of the new architectures will be to remove overhead in current designs and also offer hardware, which will provide more efficient support for important functionality like security and resiliency.
4. The ***Algorithms and Software Environments*** thrust seeks to create new paradigms integrated into the new systems; these new systems will then define how application designers interact with the machine. Existing programming models are designed with old architectures in mind. New programming models and runtimes are necessary that both expose the fundamental changes in relative costs of each operation, as well as break abstraction barriers such that the heterogeneity of future machines can be both exposed and exploited.

Society has relied heavily on Moore's law providing efficient, affordable technology and thus far, advances have continued to scale without the need for conceptual redesign of computing principles or materials. Evolving technology in the post-Moore's law era will require an immediate investment in basic sciences, including materials science, to study candidate replacement materials and alternative device physics to foster further technology scaling. Based on the history of the silicon FinFET, an advance in basic device physics takes about 10 years to reach mainstream use. Any new technology will require a long lead time and sustained R&D of one to two decades to come to fruition. Options abound, the race outcome is undecided, and the prize is invaluable. The winner will not only influence chip technology, but will also define a new direction for the entire computing industry.

Introduction

Hardware Scenarios

Because hardware plays such a fundamental role in the operation and performance of applications, we have assessed three unique scenarios of possible future hardware development. These (non-exclusive) scenarios are characterized by the expected change in the underlying technology as shown in Figure 2.

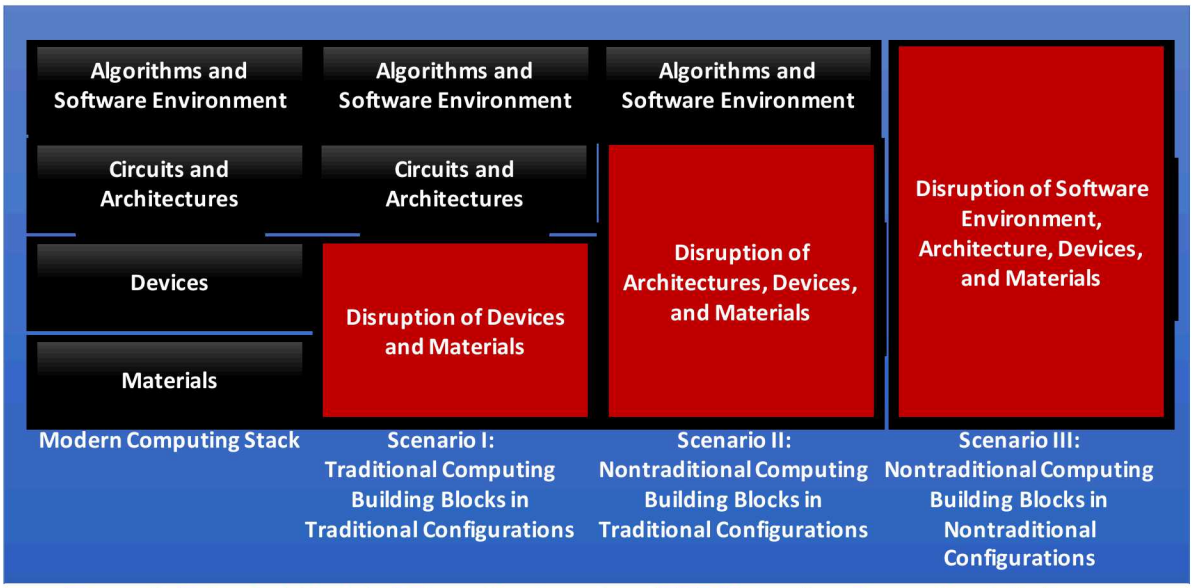


Figure 2: Extended Model of Disruption based on IEEE Rebooting Computing Initiative

Scenario I: Traditional Computing Building Blocks in Traditional Configurations

This scenario corresponds to changes at the device level that do not change the fundamental computing model. In this scenario, we may expect new materials and device designs resulting in parallel architectures based on the classical computing model. The devices will improve density, performance, and power consumption. An example is carbon nanotube-based FETs.

Scenario II: Nontraditional Computing Building Blocks in Traditional Configurations

This scenario corresponds to innovative methods of using new materials and devices to improve performance and/or energy efficiency that still follow the basic von Neumann computing model but that cannot be exploited without additional information flowing from the program or generated code to the hardware. Examples include approximate computing, reversible computing and specialized non-CMOS "conventional" accelerators.

Scenario III: Nontraditional Computing Building Blocks in Nontraditional Configurations

In this scenario, we may expect new device designs that operate using non-von Neumann computation models and additionally enable new capability not expressible in non-Boolean logic. Examples of this include ternary logic, analog computation, quantum computing, neuromorphic computing, and novel accelerators.

Based on the model shown in Figure 2, the Beyond Moore Computing challenge was initially broken into four Working Groups to most effectively approach the issues associated with each component of the modern computing stack. It was later realized that Advanced Manufacturing capabilities underlie the success of each of these four working groups, and as such, it was included as a fifth working group that crosscuts the entire spectrum of research.

1. Algorithms and Software Environments
2. Hardware and Circuit Architecture
3. Communications and Logic Devices
4. Materials
5. Advanced Manufacturing

The following sections describe the project plans for each working group outlined above.

Working Group 1: Algorithms and Software Environments

Motivation

Beyond Moore's Law devices could cause revolutionary perturbations to algorithms and system software environments. These effects have already been seen in quantum and neuromorphic computing, in which new mathematical models drive algorithms, related software tools and environments in fundamentally different ways from traditional parallel computing. Beyond Moore Computing (BMC) is giving rise to additional new inexact computing models such as approximate, stochastic, probabilistic, and analog computing. Even seemingly evolutionary advances such as large capacity persistent memories, near memory logic, and extreme device-level heterogeneity result in shifting ratios of familiar parameters that ultimately require new ways to approach data structures, threads of control, synchronization, and communication.

A key concern is the extent to which existing software paradigms will be required to change in order to accommodate these hardware changes. The run-time and operating system environments that manage resources with conventional computing devices have undergone decades of innovation and refinement. Adopting fundamentally different models of computation will certainly offer new opportunities for algorithms and applications, but it may also require significant changes in how we structure computations.

Software Perspectives: Rethink, Retune, Recompile

The Hardware Scenarios described above present varying challenges to the development of new algorithms and applications. These can be summarized as "rethink, retune, or recompile", which references the varying degrees of risk that each scenario presents.

Rethink Perspective

For the Rethink Perspective, the software developer must adopt a significantly new approach to computation due to drastic changes in hardware behavior. An example of this includes changes in the computational model or logical devices. This approach requires significant effort to identify the best algorithm within the new computational model for both existing and new problem sets. It will require new theories of computation (academia), new methods to translate computational theory to algorithms (labs), new methods of expressing algorithms (labs), and new performance and instrumentations tools (industry and labs).

Value proposition: maximally improved levels of performance and efficiency; new markets; high risk

Retune Perspective

In this perspective, the software developer must retune an existing algorithm to accommodate distinct changes in hardware performance. This approach requires effort to incorporate hardware characteristics into the application execution model. We expect the shift in key parameters such as memory capacity, persistence, and a multiplicity of heterogeneous compute elements will require new algorithms, potentially new programming models as well as algorithm and data structure partitioning tools, runtime libraries, and OS services. It will require new tools for performance modeling, measurement, and experimentation (labs).

Value proposition: leverage existing knowledge and code bases, medium to high improvement in performance; moderate risk

Recompile Perspective

In this perspective, the software developer must only recompile existing applications to accommodate evolutionary changes in hardware devices. An example of this is the development of devices using conventional CMOS technology with a smaller fabrication process. This perspective represents the simplest approach to adopting new hardware as it requires change to the translation stage of application development but not the execution stage. It requires innovation and investment in compiler technology and novel runtime systems to select appropriate optimizations and re-factoring strategies. To exploit customizable hardware structures, there will be a need for new compiler/runtime/operating system mechanisms to generate customized application configurations and to dynamically control customized processing blocks. The recompile perspective will need measurement and feedback tools, new R&D for compiling, automatic partitioning, dynamic compilation, and auto-tuning (industry and labs). There will be a need to improve scaling as the problem becomes more difficult with increased hardware complexity (labs).

Value proposition: leverage existing code bases, less validation of code required; low to medium improvement in performance; high risk for compiler technology; medium risk for debug/performance tools; low risk for application developers.

Unique Capabilities

The Labs have broad and deep expertise in application specific frameworks and libraries, parallel programming models, languages (including domain specific languages), compilers and translation tools, performance measurement and tools, and visualization tools. In the proposed program, these skills and tools will be brought to bear on two cross-cutting thrusts: Memory-integrated Computing; and Approximate Computing. Partnership with academia, industry, and Standards bodies will complement the Labs' efforts in some areas (e.g. theory of computing for novel architectures). For more details on existing capabilities that the national laboratories can leverage for addressing the problem statement see Appendix C: Existing Lab Capabilities for Software Technology.

Industrial Partnership Opportunities

As primary end users, the national laboratories have always had strong partnerships with HPC vendors. Partnership activities include the development and tuning of software to match performance requirements. The national laboratories do not have a strong partnership with other information and communications technology (ICT) stakeholders who are more focused on consumer-driven computing, e.g., mobile devices and data centers.

Research Agenda

Recommendation for Software to Support Memory Integrated Computing

A proposed project for one BMC thrust is "Memory-Integrated Computing." Memory-integrated computing can be realized using several forms of NV-RAMs. Non-CMOS memory devices are being fabricated, and integrating logic with ReRAM has been proposed as a new architecture for computing

devices. This thrust will cross cut materials, device architectures, manufacturing techniques, hardware architectures, software tools and algorithms. There is also significant potential for industry (i.e., Intel, IBM, ARM, Micron, Cray *etc.*) participation thus facilitating public-private partnerships.

The software group will contribute: computing models; execution models; programming models; languages; translation tools; debug, visualization and performance tools; algorithms; and applications. Compute models include: Non-deterministic Finite Automata (NFA) such as the Micron DRAM embedded Automata processor in which the "regular expression" is a primitive operation; Ternary Content Addressable Memories (TCAM); systolic arrays; and pipelined custom processing arrays. Software tools needed for memory-integrated computing include languages to express algorithms in alternative computing models, tools to map logical representations to physical resources, and tools to evaluate correctness and performance.

The software group will work closely with the hardware and circuit architecture group, who will investigate logic and memory in the same device. Simulation and modeling infrastructure (software simulation and FPGA emulation) developed by the hardware group to evaluate candidate circuit and software algorithms and tools will target hardware architectures. Physical prototypes using current and emerging technologies built as part of this program using in-house fabrication and packaging will be targeted by the software team's artifacts.

Recommendation for Software to Support Approximate, Inexact, Lossy operators

Data analytics (recognition, mining, and synthesis) has become a pervasive component of computing in society and often requires even less accuracy than the simulation itself (visualization for example). The need to reduce energy consumption for computing operations and data movements pushes toward using inexact (approximate, probabilistic, stochastic, lossy) hardware operators to compute or transfer scientific data. Such operators use much fewer gates, latches, flip-flops and data paths from standard generic functional units. The consequence is that inexact operators provide different accuracy/performance/energy trade-offs.

One of the first studies on the applicability of inexact/approximate/lossy computing on HPC applications was published very recently [DJL14]. The study focused on climate/weather simulation, specifically on the Lorenz '96 toy model. The study compares the solution of a reference execution on classic hardware with the solution computed on an inexact hardware resulting from probabilistic pruning - pruning the paths in an operator (adder, multiplier) that have the lowest probability to be active. This approach managed to reduce the surface and power of the adder and multiplier to half and a third respectively of their reference counterparts. Climate scientists involved in this experiment believe the accuracy of the resulting inexact operators is acceptable with respect to the other errors affecting climate simulation (measurement, discretization, truncation, etc.).

Hence, for the same area and power, inexact/approximate/lossy hardware could include significantly more operators (adders, multipliers) than exact hardware. Data transfer energy can also be greatly reduced by using lossy compression. Climate research also deals with a large volume of data during the simulation and the post analysis. Estimates of the raw data requirements for the CMIP6 project exceed

10 PB [BHM16]. Recently a group of authors from seven institutions involved in climate simulation evaluated lossy data compression on climate simulation data within a large ensemble. That research followed another publication, in 2014, proposing a methodology for evaluating the impact of data compression on climate simulation data [BXM14]. State of the art lossy compression algorithms compress datasets by an order of magnitude. Such algorithms, implemented in hardware or software, will reduce significantly the energy required to move large datasets in the system network and through the storage hierarchy. However, our understanding on the implications of using inexact/approximate/lossy operators in scientific computing and data analytics is very limited. To leverage the gains in performance and energy provided by such operators, we need a thorough analysis (including V&V and UQ) of their applicability for a large variety of scientific applications and data analytics. Fundamental changes might be needed in all layers of the software stack from numerical methods, algorithms, compilers and runtimes.

Outcomes and Metrics for Success

Our goal is to make more efficient use of the hardware than is currently possible using today's hardware and programming models. Co-designing architectural features within the system software (including programming models and compilers in that category), will allow a greater fraction of the total performance to be delivered than if the petascale trajectories for hardware and software were continued. Metrics include: GOPS/W – that is, giga operations rather than only floating point operations; concurrency measures such as number of operations in flight per unit of time; and density measures – number of operation or storage units in a 3D volume. The latter metric is closely related to the hardware architecture.

Milestones

Years 1-2

1. Refine the metrics and tools for understanding and characterizing memory system requirements for relevant applications.
2. In cooperation with Circuits and Architectures, establish target workloads and exemplar applications to represent that workload. These form the basis for developing more detailed metrics to measure improvement, and to foster new architectural concepts to solve those problems.
3. As specific tasks, pursue two levels of memory/compute integration: **fine granularity** at the level of low level logic elements such as comparators, atomics, and accumulation operators co-located with storage elements; **coarse granularity** at the level of very simple CPUs co-located with "banks" of storage elements. Along those two tracks, compute models, algorithms, programming languages, runtime libraries, and co-existence with traditional compute systems will be studied.
4. Conduct a thorough analysis with application developers, computational scientists, numerical library developers, compilers and runtime experts to assess the implication (feasibility and performance/energy gains) of using inexact/approximate/lossy operators for scientific simulations and data analytics.
5. Initiate industry and academia partnerships to identify joint research opportunities.

The outcome of Year 2 would be:

1. Definition and initial prototypes of programming tools targeted to the memory integrated computing architectures developed by the Circuits and Architectures team.
2. Thorough analysis (including V&V and UQ) of the applicability of approximate computing for a large variety of scientific applications and data analytics.

Years 3-5

1. Spiral development and refinement of multiple programming models and associated language constructs, compilers, mapping tools for memory-integrated computing architectures produced by the Circuits and Architectures team.
2. Evaluate software tools according to metrics formalized in Years 1-2, with recommendations for hardware architecture co-design.
3. Establish interfaces to industry to evaluate usability of programming and mapping tools for more realistic industry-driven scenarios.
4. Work with the Circuits and Architectures team to develop programming language support for the taxonomy of approximate computing approaches to effectively map applications to architectures featuring approximate computing functions.
5. Work on techniques to re-structure popular algorithms to better fit emerging technologies and architectures.

Years 5-6

1. For Years 5-6, the algorithms and software tools would be transitioned to physical prototypes.
2. The algorithms and software group will also work with experts on the impacted software to develop demonstrators and prototypes targeting advanced devices and hardware architectures.

Working Group 2: Hardware and Circuit Architecture

Motivation

As new device technologies emerge to continue performance scaling beyond the end of MOSFET scaling and Moore's Law, each candidate solution will present tradeoffs (Figure 3) between metrics of increased parallelism, reliability, and performance in both time and energy.

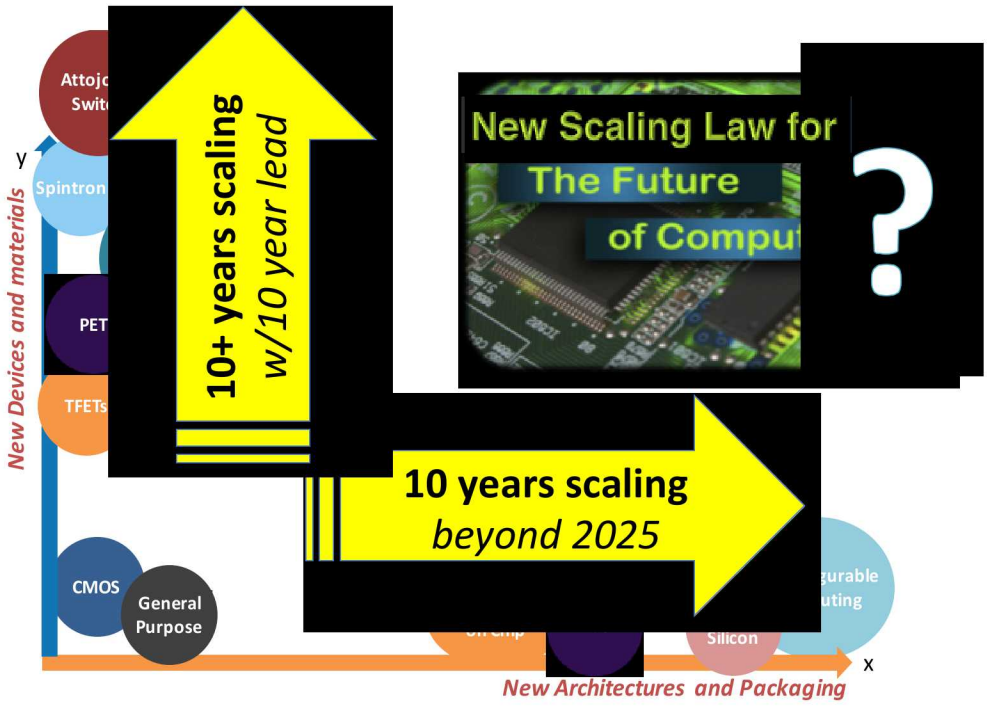


Figure 3: The architecture and circuits area must evaluate new device technologies in the context of circuits and system architectures (the vertical axis), as well as evaluate and develop new architectures and packaging technologies to provide guidance for the development of new devices. Both dimensions of research are essential to discovering a new scaling law for the future of computing.

New devices that arise from the device technologies or new materials areas must be evaluated in the context of circuits and full system architectures in order to determine how to most effectively those new devices and to determine if efficiency improvements at device scale can translate to delivered improvements to applications at chip and system scale. An integral dimension of this area will be addressing the packaging and integration challenges arising from new materials or technology improvements and taking the information and metrology from those studies to guide the device and materials research direction.

Likewise, the architecture and circuits area depends on insights from the applications and algorithms area to identify and develop improvements and innovations in architecture that advance delivered performance and usability of advanced hardware concepts.

In both cases, the DOE offers unique capabilities to create multi-scale models of advanced device architectures and concepts that have been utilized by our partners in the computing industry. These tools must also be extended to model the characteristics of emerging device technologies, advances in packaging, and to include new computational methods such as inexact computing (surveyed in [MIT16], [AAH13], [ACK06], [JHO13] and [KLP13], with some underlying theory explored in [WIG06]).

Hardware Perspectives: Bottom-up vs. Top-down Architecture Drivers

In order to facilitate evaluation of advanced computing concepts, the DOE laboratories have established a multiscale framework for creation, modeling, and evaluation of hardware technologies. These capabilities are open source and, in many cases, the technology development is shared and developed cooperatively with the computing industry. The high-level framework used to evaluate technology alternatives and their impact on architectures and system-scale performance is outlined in Figure 4.

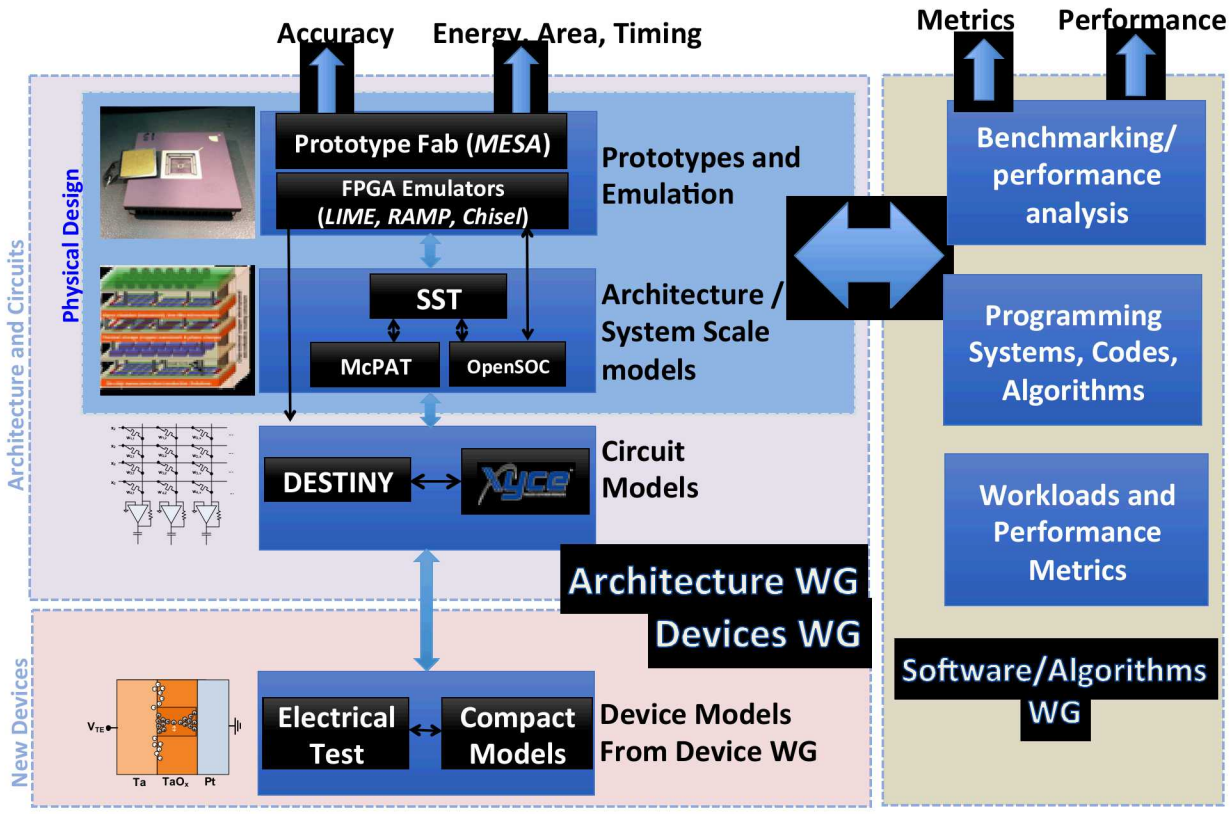


Figure 4: This is a schematic of the evaluation framework that would be utilized to both study new device technologies that come from the "bottom-up" and workloads and algorithms requirements that come from the "top down."

Relationship to Neighboring Areas (Devices and Software)

The evaluation framework enables a flow of information and metrics from the top-down, where the target workloads, algorithms, and performance metrics guide architectural decisions that, in turn, make demands that can be propagated down to device technologies development. This top-down flow requires

creation of more agile, flexible tools for circuit design, measurement, as well as accelerated design of new specialized hardware or accelerators,

The framework also facilitates the bottom-up evaluation flow, where new device technologies may inspire new architectural approaches or require evaluation in the context of new hardware architectures to determine if and how they can be made to meet the performance needs of the target applications. This design flow requires changes to our circuit models and simulators to incorporate compact models of new/novel devices. In Figure 4, the top-down approach largely explores the horizontal axis and the bottom-up explores the vertical axis. It is the sum-total of bringing these two approaches together that can result in defining a new Moore's law for scaling device performance for the coming decades (the upper-right quadrant of the graph).

Unique Capabilities

The DOE labs bring a wide variety of tools as well as architecture and circuits expertise to bear on prototyping, simulating and evaluating new architectures and devices technologies.

Circuit models enable low-level evaluation

The DOE laboratories have numerous tools in this area including ORNL's DESTINY (a 3D dEsign-Space exploraTIon tool for SRAM, eDRAM and Non-volatile memorY), and NVRAMsim, DRAMsim, and Sandia's CrossSim analog resistive crossbar computation simulator. Using Xyce, Sandia Labs scalable parallel electronic circuit simulator, it is possible to implement custom compact circuit models of Beyond Moore computation and logic devices.

System Level Prototyping/Simulation/Modeling tools

System-scale simulation requires tools that are able to model systems of all sizes, including even the largest scaled-up systems. For example, Sandia Structural Simulation Toolkit (SST) provides software based simulation tools that simulate systems at a functional level. Simian (LANL's parallel discrete event simulation engine) is useful at the system level. ExaSAT and ASPEN tools accelerate the development of analytic models for first-order accurate modeling of system concepts and algorithmic performance. FPGA-based emulators, such as LLNL's Logic in Memory Emulator (LIME) enable evaluation of novel heterogeneous IP blocks or RTL generated by other DOE lab tools (such as Chisel) in a full application level emulation. Chisel, out of Berkeley, is a high-level logic design DSL that can bridge the gap between logic-level design and circuit level design by outputting synthesizable RTL that can be run through the synthesis tool flow to get accurate power and area estimates for calibrating higher-level modeling tools. Chisel can also generate C++ simulator components that can be plugged directly in to SST to facilitate system-level modeling consistent with the synthesizable RTL that is used to create the power estimates. McPAT provides power and energy modeling and abstracts the circuit-level models to full integrated circuit architecture. Lastly, circuit-level prototyping tools such as Xyce enable calculation of power and area estimates that can then be fed back into the system-scale simulators.

Rapid Digital Design Prototyping Tools

Rapid circuit design is necessary to study the limits of accelerator technologies and specialization, but also to investigate power and area estimates for such novel designs. Effective research in this area benefits

from rapid prototyping of architectures at a logic/RTL for functional and performance studies. Beyond mature capabilities (i.e. Verilog and VHDL), Chisel promises a productive domain-specific language to accelerate the pace of digital design, and derived components such as RISC-V processor cores and the OpenSOC NoC fabric bridge the gap between architectural simulators and circuit-level designs. All of these tools can target simulation, emulation for large-scale study using FPGA emulators or Fab.

Prototype Production

All the labs have access to and expertise in FPGA emulation platforms for performing circuit level simulation and prototyping of novel architectures. FPGAs enable validation and evaluation platforms for computing architectures without going through the expense and risk of producing a full ASIC design at a foundry. For full ASIC designs, the MESA Fab at Sandia can be used to route, create masks, and tape out chip designs using leading-edge process technology – including fabrication at outside foundries such as TAPO or TSMC.

Validation

The veracity and quality of the tool chain requires a continuous process of validation under a controlled environment to ascertain its applicability and error bounds. An important aspect of this validation process is Baselining in which early circuit and architecture prototypes are exercised and compared to results produced by a predictive tool chain. NIST's Advanced Measurement Laboratory (AML) acts as a validation and proving ground for PNNL's Center for Advanced Technology Evaluation (CENATE). The AML can provide the required validation process and prime novel modeling approaches.

Research Agenda

Bottom-up (device technology driven) Research Directions

New transistor technologies

As the materials science and device technology area produces numerous alternative transistor technologies, it will be crucial to be able to evaluate the performance impact of these devices at a circuit and systems level. While a number of these candidate technologies fall under the same digital computing model as MOSFETs, many of the transistor technologies will have a dramatic impact on architectural and programming model choices. Many emerging transistor technologies are able to operate at lower voltages and thus operate with less energy, leading to even further expansion on explicit parallelism than current devices. Others have faster switching rates, increasing clock frequency, ultimately resulting in a completely different design point.

To better evaluate and take advantage of upcoming technologies, we will create performance and cost models for promising devices and technologies that incorporate compact models of these emerging devices into existing architectural simulation and modeling tools. Simulators with new device models will enable the software technology research area to evaluate new architectures based on these devices. Such tools will be critical to rapidly exploring the design space and thus estimating the tradeoffs between different devices and manufacturing technologies, as well as how these differences affect architecture and programming models.

New memory technologies

Along with new transistors, new memory technologies are making their debut in manufactured chips. Two prime examples are resistive RAM (RRAM) and magnetic RAM (MRAM). These memories not only offer lower energy for some accesses, but are also nonvolatile. This capability shows great promise for enabling architectural solutions to problems like dark silicon, where powering down sections of the chip may be necessary. In addition, RRAM and MRAM, combined with other mature memory technologies such as STT-RAM, challenge the traditional view of memory hierarchy where DRAM is the last level of main memory. These memories have diverse energy and latency costs for reading and writing, as well as manufacturing cost per bit, making it increasingly important to select the right type of memory for the proper level of the memory hierarchy and application workload.

We will study how each of these technologies affects architecture and programming models, and determine for which level of the memory hierarchy each technology is best suited. This will include a study of how to best exploit the unique capabilities of new memories (such as non-volatility) in order to seamlessly power down portions of future large-scale chips, or avoid frequent external storage accesses to facilitate big-data applications and produce more efficient mobile devices.

3D monolithic integration

3D monolithic integration not only reduces the average distance of data movement, but also increases the density of interconnection between layers. Together, these characteristics can both reduce energy and increase bandwidth. In fact, recent advances enable multiple layers where each one can be logic or multiple types of memory. Therefore, it is possible to manufacture a layer of general-purpose logic, fast and small memory, a layer of accelerators, a layer of larger non-volatile memory, etc. All this combined with dense inter-layer connections translates to dense logic with multiple types of memory, all with low-latency and high-bandwidth connections. To fully exploit monolithic integration, we need to optimally design architectures that both take advantage of new capabilities and also deal with the shortcomings resulting from this manufacturing technology. Current chip stacking techniques, commonly called 2.5D are beneficial but deliver a small fraction of these benefits because the limitations of the interconnect densities between layers.

Top-Down (Application-Driven) Research Directions

Specialized architectures

Current architectures consume substantial amounts of energy and area because they are general-purpose and programmable. For instance, a typical processor spends only about a quarter of its power budget for actual computation, 10% to access memory, and the rest for the pipelining and logic necessary to make the processor programmable. Specialized architectures attack this overhead by designing special-purpose circuits that reduce data movement, addressing, and instruction overhead to implement a function that many applications can use (such as FFT). These specialized architectures, typically in the form of accelerators, have been shown to provide 100x speedups in a wide variety of computations for constant energy consumption. Dark silicon and reconfigurable computing promise further energy efficiency gains using different approaches to incorporating specialized functions. However, aggressive use of specialized

architectures will increase system heterogeneity, and require efficient methods of integrating and programming these specialized architectures to spur broader adoption by the IT industry.

ASCR can use existing expertise across the DOE to identify the most important applications, methods, or kernels relevant to DOE that are in critical need of performance beyond exascale computing. DOE can take the lead in developing specialized architectures and exemplar programming frameworks to be used as accelerators in future computing systems. In conjunction with the software technology area, this study will extend to making the best use of accelerators by programmers, and more generally, ways to make new accelerators more accessible. GPU accelerators have already created disruptions in programming environments just to deal with two different kinds of processing elements. Mobile systems often incorporate a dozen accelerators. Systematically studying these systems will produce a strategic plan that can avoid the programmability challenges associated with the use of dozens of different kinds of processors (or even 100 different kinds).

Inexact computing

Improved energy efficiency and robustness to performance fluctuations (inherent in post-Moore devices and immature design toolchains) may be gained by carefully relaxing accuracy requirements. Inexact computing is an active research area that has not been widely adopted yet in HPC despite the large number of applications that can tolerate a small amount of error in their outputs. In addition, inexact computing is a cross-cutting optimization technique that can be applied to processor architectures, memory hierarchies, boards, or even the interconnects for scalable computing systems in datacenters. Inexact computing can lead to both increased performance and accelerate the adoption of novel technologies that offer better energy efficiency but with higher error rates.

In close collaboration with the software technologies area and industry, DOE architecture evaluation tools will be used to identify candidate applications and quantify the error they can tolerate. Even the error sensitive portions of an algorithm may be amenable to iterative improvements in results produced by inexact computing. If so, the relevant tradeoffs in various levels can then be studied. For instance, we can design inexact accelerators, memories, on-chip networks, all of which will play a role in increasing the energy efficiency and performance of future systems, despite statistical variability in characteristics of individual post-Moore devices. Architecting systems to deliver only the required computing capabilities may reduce complexity, cost, and energy use of computation. This study can extend to programming models in order to give the programmer the ability to explicitly specify tolerable error and communicate that to the hardware that can then make scheduling or optimal reconfiguration decisions.

Outcomes and Metrics for Success

Metrics for success need to be chosen to serve the desired overall project outcomes, with periodic coordination between thrust areas to maximize chances of success despite largely independent progress. It is well known that energy efficiency has become a primary limiter to improving the performance and integration density of CMOS logic devices. The highest-level efficiency target is to improve the energy efficiency of digital computation by a factor of 100 to 1000x within 10 years. Contemporary microprocessor-based computers consume approximately 400 picojoules per operation when you include

all of the overheads (instruction and operand fetch, decode across the entire memory hierarchy). The target of the DOE Exascale Computing Project is to improve this by a factor of 20x to 20 picojoules per operation. Our strategy is to use new architectures and new device technologies to increase the efficiency to 20 femtojoules per equivalent operation – a factor of 1000x from an exascale baseline. We say "per equivalent operation" because specialization and inexact computing may blur the definition of what is considered an "operation" on current conventional instruction processors.

It will also be crucial to improve integration density so that we can increase functionality per unit area (or unit volume) by 100 to 1000x within a ten-year span. Architectures that make aggressive integration of memory technologies within logic can increase integration densities by a factor of 100x or more even in the absence of reductions to logic area density while still delivering compelling capabilities for big data applications. In fact, energy-efficiency and integration density are synergistic: 1000x better energy efficiency reduces energy densities (per device) such integration densities for computing logic could also be improved by an equivalent factor, and improved integration often improves energy efficiency of data movement.

Recognizing that much power is lost through leakage in modern CMOS devices, another metric will be to create zero-leakage-power system on chip – an extremely valuable capability for low-power embedded applications. Architectures that exploit nonvolatile memory technologies could provide an approach to instant-on and instant-off (zero standby power) electronics.

Milestones

Years 1-2

1. Develop metrics and tools for understanding and characterizing memory system requirements for relevant applications.
2. In cooperation with Software & Algorithms, establish target workloads and exemplar applications to represent that workload. These will form the basis for developing more detailed metrics to measure improvement, and will foster new architectural concepts to solve those problems. For example, the EMBCC benchmarks for embedded and mobile applications, TPC for database, and BigDataBench for datacenters.
3. Initiate interactions academic communities and other research organizations such as SRC to generate list of technologies (bottom up) and architecture concepts (top-down) that offer opportunities for substantial impacts. In years 3-5 we will start using our evaluation framework to winnow down the large number of initial concepts to a smaller field of choices that have demonstrated greater impact capability.
4. Initiate industry partnerships to identify precompetitive joint research opportunities as well as better define the interface between entities in the public/private partnership. DOE needs to solidify its role as a neutral evaluator to independently review of the technology options (separate the wheat from the chaff).
5. In cooperation with the Device technology group, establish interfaces and data exchange for tools within Arch/Circuits effort and also determine the type of

information (and what form) needed from the Devices effort that can be used as inputs to circuit and architecture modeling tools.

6. Explore our options for verification methodology and toolchain. Need to create a framework to identify the error contributions of the entire tool chain and determine where the effort to reduce errors needs to be placed.

Capstone for Year 2 is to demonstrate the hardware architectural evaluation framework tool-flow from end-to-end for a simple baseline scenario such as ncFET or inexact computing. The purpose is to determine how the evaluation tools will work together, establish interfaces between tools, identify gaps that require new development, and establish baseline measures for uncertainty and error in our models.

Years 3-5

1. Initiate spiral development and refinement of the evaluation framework that is initially demonstrated as capstone to first two years of effort. The proposed evaluation framework will act as a funnel (operate the framework to evaluate and prioritize opportunities based on their delivered performance). The framework would need to be developed in phases (first demo in year 2) with rough estimates of a minimum set of performance metrics. By the 3rd year, obtain higher fidelity, reduce errors, and increase scope of metrics that can be evaluated. Development of the next generation framework must occur in parallel with the utilization of the current version framework (spiral development pattern) with the expectation that experiments can be repeated with later versions of the framework to improve fidelity of understanding.
2. Make recommendations for which technologies are the most promising and efficient to continue digital computing performance scaling.
3. Establish and test validation strategy using a use case and first-generation of prototypes. For example, 3DI will be a big issue as most current tools cannot model 3DI. We must establish a path to creating prototype hardware to validate design, but need measurement of those prototypes to validate our models.
4. Establish verification methodology and toolchain. Integration of the power and area estimation models we develop into comprehensive frameworks we choose such as SST: Adding additional architectures, components, technology nodes and routing strategies.
5. Establish interfaces to industry to evaluate scalability of framework for more realistic industry-driven scenarios.
6. Develop a taxonomy of inexact computing approaches for the dominant technologies we identified with the dual goal of (i) mapping applications to architectures featuring inexact computing functions, and (ii) architecting functionally specialized systems to meet application accuracy requirements.
7. Inform the algorithms community on how to re-structure popular algorithms to better fit emerging technologies and architectures.

Years 5-10

There will be more of an emphasis on productization in 5-10 years Milestones.

1. Work with EDA tools for the winning technologies.
2. Working closer with industry on path to productization will be central to the 5-10 year roadmap. Milestones in 3-5 therefore focus on using our framework to narrow down a wide range of options to those that are most promising. Milestones for 5-10 focus more on manufacturability at volume and industry productization (industry interactions).

Working Group 3: Communications and Logic Devices

Motivation

As current CMOS devices approach the 5-nm scale, their reliability will decrease, leading to an overall reduced ability to compute. For instance, if ten Si atoms were aligned center to center in one dimension, the length of the line would be 5 nm long. At this point, thermal fluctuations and quantum mechanical effects, such as tunneling through junctions, are believed to be major drivers of noise in CMOS systems. Furthermore, power density would be higher at this scale (the power density reductions that came with Dennard scaling came to an end long ago), causing problems for heat generation and dissipation, placing limits on clock rates and gate voltages. CMOS devices as they currently are understood and manufactured can be made no smaller than this size scale due to these limitations, meaning that the power and size reductions yet still increasing computational power that every new process node has thus provided will effectively end in the very near future.

The goal of the Logic and Communications effort is to develop mitigation strategies for this upcoming challenge arising from the failing of Moore's law. The intentions set out at the Beyond Moore Computing summit, with eight national labs in attendance, are to foster the new technological era of ever-increasing computational power by harnessing the major capabilities at all laboratories in order to simultaneously squeeze the last possible advancements out of CMOS while studying new devices and exploiting new materials yet to be discovered.

Device Categories

Several potential logic devices or fabrication methods have been identified, including **nonvolatile memory, 3-D chip stacking, two-terminal logic and memory devices, 2D material device, spintronics, in-memory and reprogrammable processing -** such as in **magnetic tunnel junctions, memristors and single-electronics**, and **photonics and plasmonics** for interconnects.

Error! Reference source not found. Table 1 shows device taxonomy by computing principle (charge, spin, etc.) and lists the corresponding prospective materials to enable each device category. These categories are non-exhaustive, and each may contain a large number of new devices made from different materials. For instance, **magnetic tunnel junctions** (MTJ), which are gaining attention due to their promise of very low (~aJ/switch) power consumption, may be fabricated with transition metal (TM) alloys and thin insulating barriers (e.g., MgO). **Memristors** may be built from transition metal oxides or from nanoscale plasmonic nanodisc arrays. **Photonic interconnects** may be directly integrated onto chips using 3-D stacking to produce high density opto-electronic circuits. **Metallic nanophotonics**, could be exploited to perform both passive and active functions on a single level enabling electronic-photonic-plasmonic integration. **Single electron transistors** (SET) may bring the ultimate no-leakage, low-power operation, help to overcome thermal noise, and potentially achieve higher than CMOS logic densities, while being able to exploit fab technologies compatible with CMOS of the near future. In addition, single-electronics can serve as a bridge between quantum, neuromorphic and classical computing.

It is clear that new materials will be required to build many of these devices, with the devices being integrated into non-conventional architectures. In order to be successful at such an endeavor, device development work should be closely integrated with material discovery and architecture exploration. This

will require a synergistic effort aimed at a multiscale synthesis, modeling and characterization to understand and control device performance from the molecular level to actual logic and memory devices.

Table 1: Device taxonomy by computing principle

COMPUTING PRINCIPLE	CHARGE	WAVE-COMPUTING		SUPER-CONDUCTING	MOLECULAR, ORBITAL	PHASE, TOPOLOGIC. STRONGLY CORRELATED
		SPIN	*PHOTONIC*			
Devices	SET, TFET, memristors	MTJ, SPIN-WAVE	Photonic logic, wave-guides, Ising machine	JJ (SQUID, RSFQ, etc), sc-SETs	van-der-Waals/ 2D transistors	MottFET, Majorana switch
Switching speed **Power dissipation**	High Low	Very low	Low	Very high Very low	High Low	Possibly high Low
New functionalities	Memory integrated computing (SETs, memristors)	Memory integrated, enables analog	Enables analog, robustness	Enables analog	High densities	Ionic control
Risks	Fundament. physical limits of charge trans.	Relatively low gain	Low device density, challenging conversion	Cryogenic temperatures, low device densities	Low maturity	Low maturity
Materials	Si (atom. fab.), TMO, MOF, organic	Magnetic, TM ferro-magnets, antiferro-magnets	1D-3D nanocomp., plasmonic structures, metamater.	superconducting nitrides, oxides, MgB_2	dichalcogenide MOGs	VO_2, Li_xCoO_2, ferromagnets, topological insulators

*SET=single-electron transistors, MTJ=magnetic tunnel junctions, JJ=Josephson junctions, TM=transition metal, TMO=… oxides, MOF=metal organic framework, MOG=2D metal organic graphene analog

Unique Capabilities

A gap exists between Industry and academia; the two entities have different goals and different timelines to achieve these goals. Academic research and development is not bound by the demand of fast investment return; therefore, its project timeline can span between 5-15 years. The academic environment lacks the focus and resources to achieve the goals of Beyond Moore Computing. Industry, however, is deeply concerned with the immediate future – how to reach 7 nm with an eye towards 5 nm; beyond 5 nm presents a large gap in capabilities and information. The National Labs have the resources

to fill that gap and support the above objectives. More importantly, the National Labs can build the infrastructure and platform to help the industry develop beyond Moore technology. Here we detail some of these unique capabilities in support of the desired outcomes.

Facilities

The national labs operate the nation's premier facilities for new materials discovery, nanofabrication and device characterization. The combination of these labs will serve as a grand virtual user facility that industry and academia will use for BMC devices research.

Ames Laboratory

Ames Laboratory focuses on new materials creation, development, and characterization by closely coupling theory, synthesis, and ultrafast spectroscopy of materials performance. The extreme quantum terahertz microscope will enable materials discovery and materials functionality at unprecedented scales of space, time, and energy that are ultrafast, ultra-small, and at very low frequency. The laser-based angle resolved photoemission spectrometer is emerging as an ideal method to probe new quantum materials.

Argonne National Laboratory

Molecular beam epitaxy systems make up a solid research platform to study new technologies for BMC. Several labs house MBE chambers, including seven at Argonne. There, a new semiconductor and oxide/nitride heterostructure synthesis lab is being set up for MOCVD and MBE capabilities for nitrides, carbides, oxides and oxy-chalgogenides. Furthermore, ANL has extensive capabilities for magnetotransport measurements and the characterization of magnetization structure and dynamics.

ANL hosts a state-of-the art 12,000 sq. ft. (soon to be 18,000 sq. ft.) comprehensive cleanroom for device fabrication. ANL also houses the synchrotron based 3-D x-ray nanoprobe imaging technique at CNM-APS, enabling ~30 nm spatial resolution three-dimensional imaging. The Advanced Photon Source (APS) offers a suite of additional capabilities, including microscopic probes of in situ device operation which tracks structural, electronic, and magnetic functionality. The Stanford Synchrotron Radiation Lightsource (SSRL) and the Advanced Light Source (ALS) host a variety of x-ray beamlines for state-of-the-art characterization of magnetic properties, including element specific, time and spatially resolved spectroscopies for powerful studies of spin transport and magnetization dynamics.

Lawrence Livermore National Laboratory

At LLNL, the HPC4Mfg program will be used as an existing and efficient vehicle to enable lab/industry partnerships (expanding lab partners) by identifying the capability at the labs needed by the industry partners. The upcoming Advanced Manufacturing Lab, a 13,000-square-foot facility, will become a modern collaborative hub for developing next-generation materials and manufacturing technologies, and will feature various laboratories and collaboration spaces. Efforts in manufacturing science range from developing new additive manufacturing processes to carbon fiber composites. Size scales span from micrometer and nanometer-sized structures to meter-sized components, and materials sets range from polymers to metals and ceramics. Specific expertise includes: multiscale, multiphysics modeling; tailored synthesis of nanomaterials; material characterization; microfabrication; and custom additive manufacturing techniques.

LLNL also hosts the Center for Micro- and Nano-Technology (CMNT), a 3500-square foot class 10-100 cleanroom facility used for the development of photonics integrated circuits and plasmonic devices. The CMNT houses a wide range of equipment capable of performing all aspects of micro and nano machining, silicon microelectronics, III-V semiconductor optoelectronics, photonics, etc. Other labs provide material characterization and device-testing capabilities, microscopic inspection, packaging, and electrical and optical testing of devices. Extreme UVL capabilities have been developed through a consortium with INTEL and large area optics (for NIF and NASA) with nanometer control have been created using exploiting our exquisite Laser Interference Lithography system, one of the best in U.S. In addition, LLNL also possesses various high-resolution diagnostic tools, such TEM and Dynamic TEM enabling ultra-fast dynamic studies at the nanoscale.

Los Alamos National Laboratory

LANL hosts a unique combination of nanoscale fabrication, characterization and simulation capabilities that can be leveraged to develop electronic and nanophotonic devices. The Center for Integrated Nanotechnologies (CINT) possesses extensive expertise in designing and implementing electronic and nanophotonic architectures based on a variety of materials. CINT has unique capabilities in fabrication of epitaxial complex oxide nanocomposites with full 3D control over composition and structure using pulsed laser deposition (PLD). These can be used to create new materials and devices applicable in architectures based on both memristor and tunneling magnetoresistive components. LANL also uses polymer assisted deposition and dip-pen lithographies to scale fabrication and integration of nanostructures to industrial levels. LANL features a wide range of thermodynamic measurements, scanning probes and broadband (THz to XRay) ultrafast spectroscopies. LANL can also leverage co-located user facilities to provide information on device behavior under filed (NHMFL), irradiation (IBML) and mechanical (static pressure and shockswaves) extremes.

Oak Ridge National Laboratory

ORNL facilities enable device and circuit probing at the die and wafer level, and include resources for wide range temperature characterization (environmental chambers and hot chucks) and prototype device packaging. Collectively these resources enable thorough device characterization and simulation model development that will be critical to new device optimization and integration. ORNL's CNMS nano-fabrication facility is ideally suited for creating new devices and probing never-before-seen materials, building them a single atom at a time. The lab maintains a full suite of conventional lithography, semiconductor processing tools, and characterization equipment within 10,000 sq. ft. of clean room space. This tool set is augmented by nanopatterning and bottom-up material deposition technologies, including a focused-ion/electron dual-beam system capable of on-demand etching and depositing a variety of materials, a JEOL JBX-9300FS electron beam lithography tool, and an Oxford FlexAL atom layer deposition tool. Recently a 3D laser lithography tool that uses nonlinear two-photon absorption processes to selectively initiate the polymerization photoresists was added to enable rapid writing of arbitrarily complex 3D shapes.

Sandia National Laboratory

Sandia National Laboratories (SNL) hosts advanced device fabrication facilities (CINT) where work in memristors, single electron transistors (SET), and optical interconnects are studied. SNL is particularly well-positioned to design, fabricate, and characterize Si-based TFETs and SETs due to our expertise in Si quantum computing gained from the QIST project and the extensive fabrication capabilities at MESA. While the energy consumption of on-chip interconnections is drastically reduced through the use of reduced voltage or non-charge/voltage based post-CMOS devices, off chip interconnects are still limited in bandwidth and energy consumption in an all-electronic paradigm. A number of research efforts both here and abroad are actively pursuing integrating silicon photonics directly with high-value silicon CMOS electronics, such as networking, memory, and computing chips. SNL has hosts a significant effort in silicon and III-V photonics and has been collaborating with industry and academia in these areas for several years. Its 'research foundry' in silicon photonics will expand for BMC devices. This will enable SNL and other labs to study the impact of photonics in HPC systems as both interconnects and computing devices.

Simulation

In order to succeed in selecting the best device and materials classes and optimize their performance for beyond-Moore computing, new comprehensive simulation tools are necessary for at least the three main computing paradigms presented in**Error! Reference source not found.**: Charge, Magnetic and Superconducting. Among, them the creation of the universal charge-based device simulator is considered to be the most crucial.

The national labs house the world's largest computing facilities, such as OLCF at ORNL, with the most combined computing power. These facilities will be brought to bear on the challenge of simulating and predicting how a new device will behave based on first principles.

Argonne National Laboratory

ANL has world-class capabilities and experience in micromagnetic (down to ~ 1nm length scales) modeling of spin textures and dynamics in inhomogeneous structures, as well as experience in developing analytic approaches to magnetization dynamics. In the area of resistive-switching structures, ANL provides experience and expertise using density-functional theory based methods (DFT+U, DFT+SIC) for structure and transport, and also non-equilibrium Green's functions techniques. In addition, ANL presents home-grown numerical models for electron and ionic charge transport in memristic nanostructures.

Lawrence Berkeley National Laboratory

LBNL has developed large scale O(N) scaling density functional theory (DFT) methods and codes, which can be used for *ab initio* device simulations. For example, the linear scaling three-dimensional method (LS3DF) is capable of calculating systems with hundreds of thousands of atoms relevant for a novel device. LBNL also developed a special way to calculate the quantum elastic transport that is compatible with plane wave pseudopotential calculations without the use of a localized basis set. This method/code has been used to simulate tunneling field effect transistor (TFET), and it will allow the study of transport for a million-atom system. Additionally, LBNL created various methods for studying electron-phonon coupling effects. This approach could be used to study current leakage caused by trap state hopping.

Lawrence Livermore National Laboratory

At LLNL, Livermore Computing (LC) is the home for the computational infrastructure that supports the advancement of HPC capabilities. LC delivers multiple petaFLOP/s of compute power and offers massive shared parallel file systems, powerful data analysis platforms, and archival storage capable of storing hundreds of petabytes of data. We develop *ab initio* electronic structure, molecular dynamics, and kinetic Monte Carlo, DFT tools for the study of the structural, electronic, optical and optoelectronic properties of semiconductors and nanostructures, emphasizing the relationships among defects, electronic structure, and device performances. Tools developed at LLNL include Qbox, a Massively Parallel First-Principles Molecular Dynamics code. In addition, the High Performance Computing Innovation Center (HPCIC) exists for outreach by LLNL to U.S. industry.

Los Alamos National Laboratory

LANL offers extensive expertise and HPSC facilities for multi-scale *ab initio* and first-principles material and device simulations, allowing simulation of complex interactions among spin, charge, lattice and orbital degrees of freedom. Toolsets include molecular dynamics simulations for atomistic modeling of materials structure, and DFT codes (VASP, Wien2k, RSPt) for determining electronic structure and magnetic/electric ordering parameters. LANL is also a leader in molecular dynamics (MD) methods for simulating materials and device responses in non-equilibrium states. Recently, LANL developed the TBM^3 package where MD and DFT approaches are combined to model, from first principles, materials and devices that include structural, magnetic and electronic irregularities and interfaces at atomistic to macroscopic scales. Finally, machine learning (ML) approaches are being developed to establish connections between theoretical and experimental results and accelerate optimization of materials and device performance using “co-design” approaches.

Oak Ridge National Laboratory

ORNL has a toolset known as Sentaurus Device which is a multidimensional device simulator for device design and optimization. The simulator supports conventional devices such as MOSFETs but also allows user-defined models for new device types, which will arise from this research. In addition, the tool provides the capability for generating compact models for SPICE simulations from the device simulations, enabling architectural and circuit simulations to be performed in parallel to device prototyping activities. Sentaurus Device will be used by the device designers for both functional verification and for porting a rough model up the hierarchy to the circuits and architectures teams. Early model development, even approximate, will be key to streamlining the device development process to reduce iteration cycle times.

Sandia National Laboratories

SNL has developed several computing tools specifically aimed at simulating and optimizing beyond-Moore devices. The first tool is a universal continuum-based device simulator, Charon, for conventional semiconductor and transition metal oxide devices. It takes into account multiple charge species transport, radiation, self-heating and other effects important for memristors and phase-change devices. The second tool called CBR3D is a novel and efficient code for fully quantum-mechanical transport simulation, geometry and doping optimization of FinFETs/MuGFETs, TFETs, and potentially SETs and other novel devices.

Outcomes and Metrics for Success

Several main outcomes of the logic and communications effort have been outlined as goals for this working group.

1. **Discovery of new devices**. We propose to create a universal framework for the development of new devices that may harness or drive the discovery of new materials. Based on co-design principles, similar to the ones envisioned by the Materials Genome Initiative, this unifying framework will emphasize rapid feedback from the experimental validation to theoretical prediction and vice versa to ultimately enable real-time refinement of theory and fabrication to achieve desired performance parameters. Simultaneously, to investigate the most promising technologies that aren't currently covered by systematic studies at academia and industry, we introduce:
   - Single-electronics initiative – to explore energy efficient technologies capable of exceeding the imminent limit of CMOS logic density and approach a few-atom device sizes.
     - Annual budget $5-10M (year 1-5)
     - Outcomes: comprehensive feasibility study, small circuitry (e.g. an adder) demonstrations with ~5nm gate devices
   - Memristor (resistive logic) initiative – to investigate the utility and competitiveness of two-terminal device logic and non-von Neumann architectures for general purpose computing.
     - Annual budget $2-3M (Year 1-5)
     - Outcome: comprehensive feasibility study, small circuit demonstration
   - Magnetic/spintronics initiative – to investigate the utility of novel oxide and topological materials and heterostructures for developing electric-field controlled magnetic memories and logic elements operating close to the aJ/switch limit.
     - Annual budget $3-5M (Year 1-5)
     - Outcome: 3D structural control of oxide/topological/metal nanocomposites, feasibility study, small circuit demonstration
   - Optoelectronics initiative – nanophotonic and plasmonics for ultrafast manipulation and non-perturbing interconnects.

2. **The Need for Electric-Field Control of Magnetization.** Over the last decade, extensive research has gone into trying to control magnetization with an electric field. From an application point of view, the rationale is intuitive: whenever a current is applied to create a magnetic field or spin transfer torque, a significant amount of energy is dissipated in the current carrying wire itself simply in the form of Joule heating. It has been pointed out by multiple researchers that the actual dissipation in magnets is extremely small - usually of the order of a few 10s of kT, whereas the dissipation in the wire is many millions of kT. Simply from this perspective, it would be valuable to replace current carrying wires with electric fields. Thus, the central goal of this proposal is to control and manipulate magnetism with an electric field wherein a

fundamental unit of operation (for example, writing a state) takes energy ≤1 attoJoule (10-18 J). This corresponds to approximately 10-6 J/cm^2 in terms of energy density.

3. **Creation of Universal "beyond-Moore" Charge Device Simulator** that utilizes excellent HPC resources at DOE and allows:
    - Accurate prediction of the I-V, switching time, heat generation, leaking, and power consumption of charge-based nanodevices.
    - Parameter space exploration (materials, geometry, doping) for device optimization.

    Specific Universal beyond-Moore Charge Device Simulator outcomes with timeline:

    i. The Simulator should be able to capture most physical effects relevant to charge-based device performances:
        a. Accurate materials description (e.g. *ab initio* electronic structure, alignment, discrete dopants, impurities and atomic level roughness) – Year 1-3
        b. Accurate quantum transport treatment (solid and efficient NEGF kernel) – Year 1-3
        c. E-e interaction including especially the Coulomb blockade – Year 1-3
        d. E-phonon interaction for thermal fluctuation and heat generation – Year 3-7
        e. Time dependent simulation for switching speed – Year 3-10
        f. Magnetic/Spin effects – Year 3-10
        g. Ability to treat ion/ion vacancy transport for resistive switching – Year 2-5

    ii. The Simulator should be able to do all the above in a reasonable time such that different (and multiple) device geometries and materials can be optimized; methods and codes for both HPC and more accessible computers will be developed for different purposes. The software should have a user-friendly interface that allows effortless use by all device physicists and designers - including experimentalists, materials and circuit researchers both inside and outside of the national laboratories.

    Annual budget: $3-4M (Year 1-3), $4-5M (Year 3-5), $3-4M (Year 5-10)

4. **Novel characterization techniques** that provide access to process parameters and device performance on the fundamental scales of atomic and electronic motion under *in operando* conditions.
5. A **virtual user facility/center** that serves as a clearing house for new logic device ideas. The facility can direct new device ideas to the appropriate lab level user facilities/nanocenters for material and device design. This user center will consist of both simulation and device fabrication. The goal is to allocate resources from across the labs in order to ensure that the entire co-design framework is implemented seamlessly for a user who proposes a material and device idea. As an example, an idea for a new device based on a newly discovered material might be simulated with *ab initio* codes as detailed above, and then designs derived from the results could be sent to nanocenters such as CNMS for experimental fabrication, followed by

characterization. Characterization, simulation, and fabrication can work together in a feedback loop to refine the design. Finally, advanced device characterization of near-final devices could be carried out in some of the labs' advanced device testing facilities.

Annual budget: $1M for support and virtual center lab integration services.

6. **Scalability**. At the lab level, demonstrate the ability to scale new device fabrication on the order of 10,000 devices. These demonstrations serve as an example to industry that new device creation and production are feasible.

Milestones

Years 1-2

1. Provide a preliminary joint experimental-computational assessment of the potential speed, power, and interconnect characteristics of i) single-electronics, ii) resistive logic, iii) spintronics, iv) superconducting logic, and v) optoelectronics initiatives.
2. In conjunction with the Materials effort, create industry and academic partnerships needed to more quickly advance required understanding.
3. Build a foundation for the Universal *ab initio* "beyond-Moore" Device Simulator

Years 3-5

1. Demonstrate a small single-electron transistor (SET) *circuit* operated at room temperature.
2. Demonstrate (proof of principle) a low-power spintronics gate that can be integrated into circuitry.
3. Demonstrate (proof of principle) an all solid state Mottronics FET with a path toward sub-60 mV threshold slope at low voltages (< 0.5 V).
4. **Year 5:** Calibrate, verify and demonstrate the predictive power of Universal "beyond-Moore" Charge Device Simulator on a variety of charge-based devices.
5. **Year 5:** Down-select to several of the most viable pathways for scalable, energy-efficient BMC devices.

Working Group 4: Materials

Beyond Moore Computing relies on the development of a new ecosystem of software, architecture, devices, manufacturing and materials. Fundamental discoveries and transformative innovation are necessary not just in each of these, but in a synergy that cuts across these areas. Below, we detail fundamental materials research that can underpin BMC, and its relation to other areas in BMC, particularly new devices. We also describe how the unique resources of the DOE lab complex can be leveraged to accelerate materials discovery and innovation.

Motivation

To achieve the goal of Beyond Moore Computing (BMC), fundamental materials research must be undertaken to enable the development of novel devices that allow for the continued miniaturization of computers with increased computing power and vastly improved energy efficiency. Currently, the MOSFET transistors and interconnects are the major consumers of energy and also set limits on the data processing speeds. A targeted and well-focused fundamental research effort to develop materials with functionalities aimed at reducing energy consumption underpins the development of novel BMC devices. Within that context, major research initiatives are necessary to:

- Develop responsive materials for use in devices that exploit the current technology of electrostatically controlled (field-effect) gating with a goal to augment or replace Si-based CMOS transistors.
- Discover and develop responsive materials systems that enable entirely new switching technologies and alternative control paradigms, such as superconducting, topological or spin-based logic.
- Discover and develop new materials and structures that minimize dissipation in interconnects between a variety of gating technologies.

Through a combination of these initiatives, our goal is to discover and develop materials that enable creation of devices that will reduce the energy consumption from current value of 10nJ/operation to 1aJ/operation. Specific challenges in each initiative are outlined below and the devices and materials are summarized in Table 2.

Table 2: Examples of device types and examples of relevant materials

MATERIALS CLASSES	ATTRIBUTES	DEVICE TECHNOLOGY ENABLED	OPPORTUNITIES	CHALLENGES
TDMCs Black Phosphorus III-V alloy nanostructures, monochalcogenides	Tunable band gap, quantum confinement	Improved FET	Self-assembly, synthetically-tunable; solution deposition, improve speed, reduce power, 3D integration	Integration, defects, doping, low mobility in TDMCs
Carbon nanotubes, Quantum dots, MOGs	Path to low-capacitance, high tunneling resistance SET	Novel gating technologies, single electron transistor, ferroelectric gates, tunneling FET	Low energy, high conductivity (CNs)	Low ON current, new logic, speed, amplification, integration
Graphene	High mobility, high conductivity, quantum confinement	Improved FET, interconnects	Low energy, high speed	Integration, no or small bandgap
TM ferromagnets, magnetic semiconductors, YIG, hexaferrites, magnetoresistive materials, Heusler alloys, antiferromagnets	Persistent magnetization, spin-charge coupling, strongly nonlinear, spin-orbit coupling	Spin-based logic and memories, spin-based (or magnonic) interconnects, all-spin logic, analog or neuromorhpic computing	Low energy, non-volatile, well-established pathway towards CMOS integration	Spin-to-charge interconversion, detection (read-out), amplification,
Multiferroics	Magnetic-ferroelectric-strain coupling	Spintronics, novel gating, electric-field control of magnetism	Low energy	Low coupling magnetism-ferroelectricity, interconversion

MATERIALS CLASSES	ATTRIBUTES	DEVICE TECHNOLOGY ENABLED	OPPORTUNITIES	CHALLENGES
Superconducting nitrides, MgB_2, cubprates, pnictides	Superconducting	Superconducting switching, superconducting interconnects	Low energy	Need cooling, scaling, interconversion
BFO, PZT, $SrBi_2Ta_2O_9$	Ferroelectricity	Ferroelectric gating	Low power, non-volatile (in principle) memory	Volatility, defects, endurance
Transition metal oxides (VO_2, V_2O_3, Li_xCO_2,…) perovskites	Electronic correlations, strong lattice-charge-spin coupling	Mottronics	Improved SS/high on current	Defects, high charge density, turn 'off'
Topological insulators (TI); Bi_2Se_3, Bi_2Te_3, MOGs	Dissipationless (topologically protected) surface states, spin-charge coupling	Spin-based logic and interconnects, Majorana switches, environmentally robust computing	Massless carrier/low energy, environmentally robust	Basic properties, room temperature
Intercalation Li-MOx	Tunable through reversible ionic intercalation	Non-volatile redox transistor (NVRT)	Very low voltage, non-volatile, analog	Speed
Metamaterials, dielectrics, nonlinear optical materials	Low-loss photon propagation, photon-plasmon coupling	Photonic switching, interconnects	Low energy dissipation, speed	Interconversion, device density
Organics, MOFs, coordination polymers	Abundant, light, flexible, can be printed, long spin lifetime, multiple charge states	single-electron devices, spintronics, spin-based logic and interconnects	Low power, cost	Conversion efficiencies, aging, low mobilities, speed

Materials perspectives

The directed materials research can be roughly broken into three scenarios, similar to those of other areas in the BMC endeavor. The first scenario involves near-term research to improve existing technologies, in particular, improving existing switching technologies based on electrostatic gated logic. Here, pathways are fairly clear, materials sets are identified, and characterization and computational tools relatively mature. In the second scenario, materials sets can enable disruptive switching and device technologies. Potential material sets have been discovered and some characterization and computational tools exist, but the research is still to a large extent at a fundamental stage. Materials characterization and research are not mature, particularly with respect to interfaces and heterogeneous systems, and computational tools need further development, especially with respect to scaling of the system size. Examples of this scenario are superconducting or organic materials. Finally, the third scenario involves materials sets that can enable completely new paradigms of computing but the materials are just being discovered or are conjectured and computational and characterization tools may be too immature to characterize the basic material properties. Examples here are topological materials, such as topological insulators or Weyl semi-metals, or correlated materials.

Within the realm of field-effect devices and electrostatic gating, the most promising route to low-power is discovering materials where fast manipulation of electronic states is possible with weaker fields. A clear goal is to control device behavior at <100 mV. In doing so, we must also address the challenge of maintaining thermal stability while allowing for low-energy manipulation. This includes the development of materials that will enable transistors with < 60 meV sub-threshold swing while still delivering significant ON current for most applications. There are several very promising examples of materials that can be utilized to achieve this goal, such as carbon nanotubes, graphene, MoS_2 and materials displaying an interaction-driven Mott metal-insulator transition. In addition, design and discovery of small band gap, high-mobility semiconductors may provide additional promising candidate materials.

Fundamental materials research that enables the development of entirely new switching technologies (for example, using spin-based logic) has the potential to create transformational changes in the computational paradigm while drastically reducing power consumption beyond what is possible with electrostatic gating. The performance potential is extremely promising, but fundamental open questions about alternative switching schemes and suitable materials must be answered through basic research before establishing a roadmap towards aJ performance. For example, materials that are candidates for use in spin-based logic, such as topological materials and skyrmions, are at the forefront of basic research; moving beyond proof-of-principle demonstrations is required in order to realize their unusual properties in applications.

The discovery of new materials that are responsive to stimuli that couple to spin, orbital, and/or lattice degrees-of-freedom (such as ferroelectrics, multiferroics, or topological quantum matter) will be key to the development of alternative switching technologies. Furthermore, fundamental research that underpins the development of new types of interconnects linking transistors within microprocessors is critical to reducing energy consumption and increasing processing speed. There are several promising routes that include high carrier mobility materials such as graphene, carbon nanotubes and topological

insulators, dissipationless superconducting interconnects, non-Ohmic spin transport, plasmonics and photonics. Ultimately, the combination of logic and interconnects that comprise these new computational approaches will further the growth and economic impact of spintronics, photonics and plasmonics. Independent of the particular initiatives, there are fundamental problems that cut across all possible materials that need to be addressed.

There is a need to synthesize promising new materials and integrate them into manufacturable architectures including, for example, inhomogeneous and geometrically complex assemblies, metal contacts, and encapsulated structures. Advanced characterization techniques are required to understand the energy scales and lifetimes of quantum states, the addressability and manipulation of these states, and the transport and response at nm length scales and ps timescales in inhomogeneous, nanoscale structures (including interfaces). Finally, simulation and theory are critical to guide both synthesis and device development and will be validated by advanced characterization data, closing the co-design loop between synthesis, characterization, device design and performance, and theory and modeling. The goal of operating devices at <100 mV puts stringent requirements on acceptable levels of inhomogeneity in the structure and composition of materials and devices. These requirements underpin the need to (i) develop versatile approaches to atomically precise scalable materials synthesis and device engineering, (ii) advance non-intrusive characterization techniques, and (iii) elevate accuracy of computational modeling methods to well above the current state-of-the-art. The solution to these problems will require broad collaborative efforts between scientists at the forefront of different aspects of quantum materials science.

Research Agenda

In addition to fundamental research to improve existing or enable transformative technologies and devices, materials research and development is needed to reduce energy consumption and enhance the speed of communications through various types of interconnects. An overview of the materials research is shown in Table 2. We will thus pursue a staged approach that combines research and development in all materials classes to enable near term success, as well as to ultimately identify the best candidate materials in the longer term. In the following, we describe in more detail the materials challenges for the different device technologies outlined Table 2.

New materials for traditional gating logic and interconnects

Improvements of existing gating technologies

For traditional device technology based on the field effect transistor (FET), desirable properties are high charge-carrier mobility for fast operation, a high on/off ratio for effective switching, and high conductivity and low off-state conductance for low energy consumption during operation. Promising pseudo-2D materials and layered 2D materials that combine high mobility, tunable bandgaps, and high on-off ratios show great potential for FET applications. The interplay among traditional electric, optical and new spin and valley properties of charge carriers in these layered 2D materials can be explored for future devices. Some of these materials are relevant for both switching devices and interconnects.

Graphene has exceptionally high carrier mobility at room temperature for devices encapsulated in BN dielectric layers. However, because pristine graphene lacks a bandgap, graphene-FETs have low on/off switching ratios. Bandgaps in graphene can be tuned up to 0.2 eV using either nanostructuring, chemical functionalization or by applying a high electric field to bilayer graphene, but these methods add complexity and diminish carrier mobility.

In contrast, 2D transition-metal dichalcogenides (TMDCs) (such as MoS_2) possess moderate carrier mobility and sizable bandgaps around 1–2 eV (depending on the thickness, strain level, and chemical composition) and are promising materials for FET and optoelectronic devices. However, the carrier mobility of TMDCs to date does not yet exceed that of Si (~500 $cm^2/V/s$). The valley index of charge carriers in TMDC is another property that has been explored beyond FET, leading to potential "valleytronic" devices. It exploits the confinement of electrons or holes in distinct conduction-band minima or valence-band maxima at the same energies but different positions in momentum space. In the future, more layered 2D materials will be explored for these applications, including monochalcogenides such as GaS.

Black phosphorus (BP) bridges the gap between graphene and TMDCs (see Figure 5). Independent of the number of layers, the band gap remains direct, in contrast to the indirect gap in TMDCs. Importantly, the size of the bandgap for BP also strongly depends on the number of layers, due to quantum confinement of the charge carriers in the out-of-plane direction. This effect is stronger in BP than TMDCs and provides excellent tunability. Interestingly, as seen in Figure 6, the field effect performance of BP FETs and other properties also bridge the gap between graphene (very high mobility and poor current on/off ratio) and TMDCs (low mobility and excellent on/off ratio).

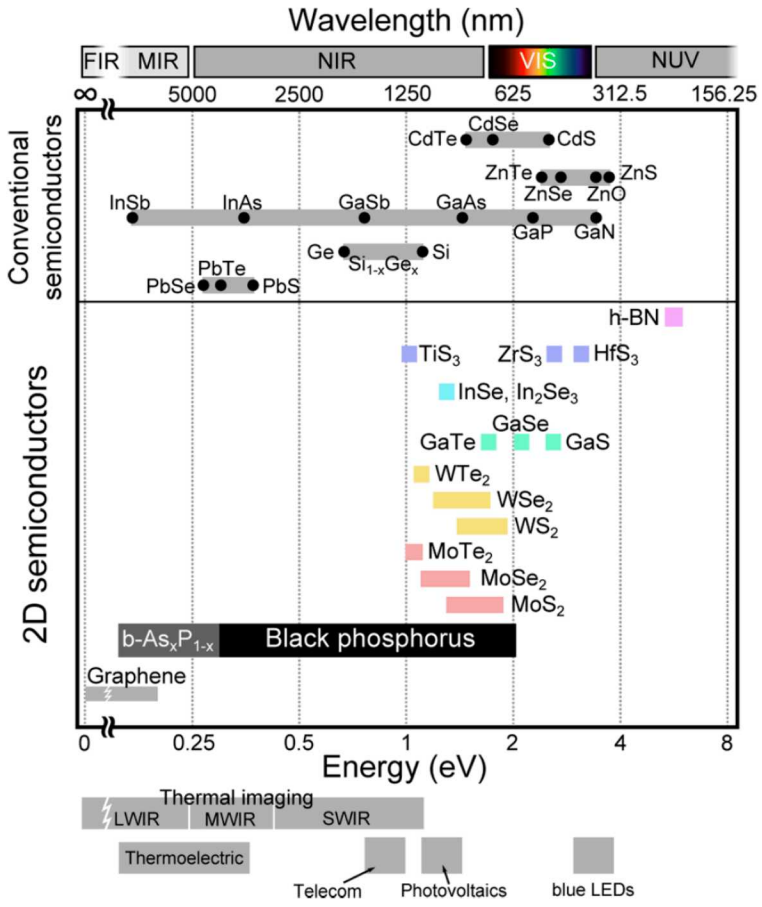


Figure 5: Comparison among the bandgap values for different 2D semiconductor materials. (A. Castellanos-Gomez, J. Phys. Chem. Lett. 6, 4280-4291 (2015))

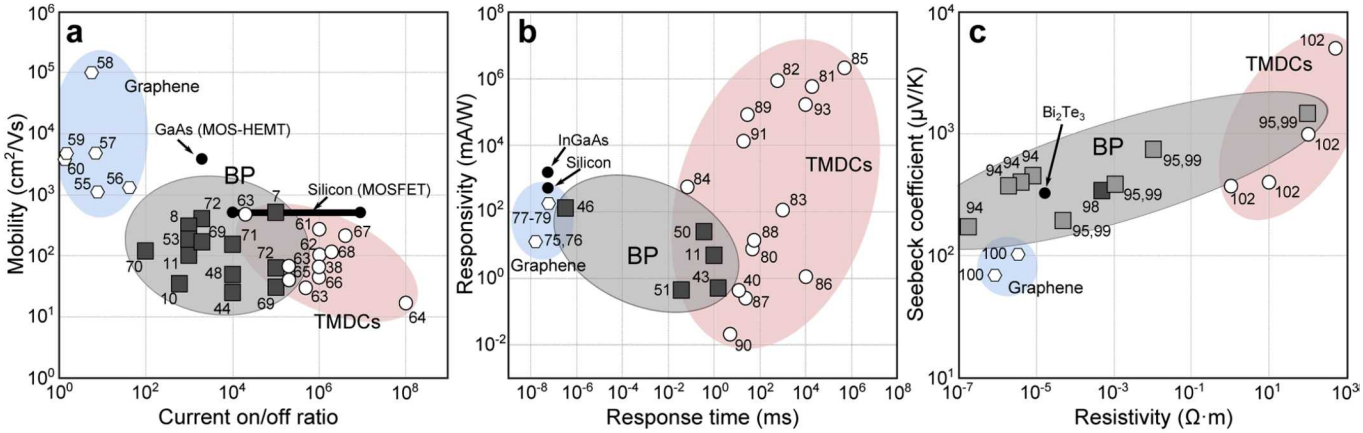


Figure 6: Comparison among the performances of 2D-based nanodevices. (The numbers are the references in A. Castellanos-Gomez, J. Phys. Chem. Lett. 6, 4280-4291 (2015))

Transformational device technologies

Spin-based logic constitutes a new paradigm that stipulates the control and manipulation of spin degrees of freedom instead of movement of the electrons (charge). Because spin current is in principle dissipationless, using spin as the logic element promises a dramatic improvement in energy consumption and speed.

General issues that need to be addressed include: (i) generating and maintaining spin coherence and propagation; (ii) developing methods to manipulate spin and local magnetic moments using spin transfer torque, spin-orbit coupling, spin-Hall effect and magnetoelastic/electric couplings; and (iii) sensitive (in space and time) spin detection techniques. Magnetodynamics is inherently strongly nonlinear in magnon density, and the nonlinearities can be explored to manipulate spin currents, especially in low-damping materials. Magnetostatic interactions are also long range and can be explored to entangle information in spin super current in easy-plane magnetic systems and Bose-Einstein condensates of magnons, for example.

Basic materials sets of interest include: (i) transition metal ferromagnets, including binary and ternary alloys (e.g., Heusler alloys), in particular for improved spin polarization and reduced damping; (ii) low-damping magnetic insulators such as YIG, and hexaferrites, for exploring magnetic nonlinearities; (iii) antiferromagnets, both metallic and insulating, for high-frequency and spin transfer torque applications; and (iv) materials with strong spin-orbit coupling, including topological insulators and quasi-2D materials such as dichalcogenides.

Effects that may be utilized include: (i) spin and charge manipulation and conversion (including spin textures, e.g. skyrmions); (ii) (inverse) Edelstein effect for efficient conversion between spin and charge currents; (iii) spin-transfer torque at ferromagnet (FM)/TI interfaces for writing/storing information; (iv) tunneling magnetoresistance for reading; (v) FM/TI for giant spin batteries and quantized charge sources; and (vi) spin Seebeck power conversions for thermal energy recycling.

Biological and chemical computing technologies use organic materials that are diverse, abundant, easy to modify, and can be tuned for specific chemical interactions. In addition, organic materials offer flexible processing routes, including vapor, low-temperature liquid/solution, photolithography, solid, roll-to-roll, and “printing.” They are also adaptable to various substrates and geometries; flexible electronics based on organic materials is a rapidly growing field. Novel conducting coordination polymers and metal organic frameworks (MOFs) that support multiple charge states can enable room temperature operation of SETs and other types of single-electronics. Organic materials can also support multi-functional modalities, including spintronics or spin-based information technology. Issues that need to be resolved include efficiencies (typically conversion efficiencies in organic materials are low), aging, and low mobilities.

Plasmonic and nanophotonic materials: These materials exploit plasmons and their couplings to light and other degrees of freedom to enable logic operations at very low energy dissipation. The main issues here lie in developing materials and materials structures with plasmon resonances that can be controlled and manipulated, both in terms of frequency response as well as losses, and also understanding how to efficiently couple plasmons to photonic or electronic degrees of freedom. Traditional plasmonic materials

included low-loss metals such as Ag or Au incorporated into dielectrics such as Al_2O_3 for surface plasmon use; however, our focus will be on the integration of optical nonlinearities into nanostructured materials to design controllable plasmon resonances and couplings.

Superconducting switches and interconnects are based on low-temperature superconductors with engineered interfaces (Josephson junctions) and tailored magnetic properties. This technology can potentially enable very large scale integration for high frequency and low power computation (800 GHz, 2-5 mV, 10^{-19}J/switch, zero/low dissipation interconnects). Fundamental issues to address include: (i) understanding and control of electronic and magnetic defects; (ii) preservation of superconducting and magnetic properties upon multilayer integration and scaling; (iii) controlling (increasing) the maximum current density (de-pairing current); and (iv) minimizing the effects of dissipation, in particular that due to vortex motion and the influence of magnetic fields. Specific materials of interest are cuprates, iron pnictides, MgB_2 and superconducting nitrides.

Materials for Single Electronics: single electronics is based on quantum-mechanical phenomena including the Coulomb blockade, where the tunneling of discrete charge leads to macroscopically observable conductance changes. Achieving Coulomb blockade requires a charging energy, $e^2/2C$, sufficiently high for thermal stability, and tunneling resistance that exceeds the von Klitzing constant, h/e^2. Thus, the main requirement is to produce **low-capacitance structures** of 1aF or less with **high tunneling resistance** of 25kΩ or more. Although room-temperature operation of SETs has been demonstrated experimentally by several groups, novel low atomic density materials such as conducting metal organic frameworks/graphene-analogs could enable highly parallel SET fabrication schemes with atomic precision for optimum performance.

Materials for strongly transformative computing technologies

Topological insulators and topological semimetals: A new, very promising class of materials for spin-based logic applications is topological insulators and semimetals. Topologically protected edge states that have unusual spin-momentum locking, Weyl points that are spin monopoles in the momentum space, as well as Majorana zero modes seem uniquely suited for spintronics. Topological matter may also provide pathways to multi-functional elements (e.g. combined logic and storage) for non-von Neumann architectures. Although there is promise for a route towards energy efficient computing using topological materials, the potential energy savings are presently not well understood. It should be emphasized that significant fundamental research is necessary to fully understand the potential and pathways for functionalizing these materials. In addition to processing and integration issues that may be especially challenging for topological matter, other basic issues that need to be addressed include: (i) improvement in surface conductivity versus bulk insulating properties; (ii) maintaining topologically protected states in complex heterostructures where, for example, FM metals can quench surface metallic state; (iii) localized and controlled placement of (non)magnetic impurities for atomic-scale spin/charge manipulation; (iv) optimization of TI performance including spin-orbit coupling and band inversion, controlling spin/charge scattering rates, and controlling spin coherence for storage and transport; (v) coherent optical/THz/X-ray excitation and detection for ultrafast spin and charge currents manipulation; and (vi) superconductor/TI junction devices for discovery and applications of Majorana fermions.

This class of materials includes time-reversal TIs, such as Bi_2Se_3, crystalline TIs such as $Pb_xSn_{1-x}Se/Pb_xSn_{1-x}Te$, Kondo TIs such as SmB_6, Weyl semimetals and other materials with Dirac fermions and Fermi arc states. Understanding the properties and behavior of TI materials at interfaces is critical. A few systems of particular importance are TI/ferromagnet/High-T_C superconductor heterostructures, TI states produced at complex oxide interfaces (such as LAO/STO), and quantum spin Hall insulators or 2D topological insulators (such as MBE grown InAs/GaSb).

Mottronic switching technologies are based on interaction driven transitions of a material state. They exploit intrinsic large interaction energy scales to rapidly accomplish the state change by means of a small external stimulus field. The paradigmatic example is the Mott transistor, which is based on the correlation driven Mott metal-insulator transitions (MITs). Materials that exhibit a Mott MIT are attractive candidates for low power logic devices because a small change in carrier density can trigger an MIT with steeper subthreshold slope compared to conventional band-type semiconductors. To realize a high-performance Mott FET, however, requires significant optimization of materials quality, dielectric and metal interfaces, and improved theoretical understanding of the device relevant physics, including (i) what are the relevant quasiparticles and how do we manipulate, control, and detect them at energies < 100 meV; (ii) how can composition, doping, and defects be controlled to energies within 100 meV; iii) can we find effective low voltage switching mechanisms besides E-field; (iv) can we understand density of states evolution with carrier density; and (vi) can we exert complete control of the Mott/dielectric and Mott/metal interfaces? These challenges can be addressed by integrating state-of-the-art correlated materials growth, characterization, modeling, and innovative device design concepts to demonstrate proof-of-principle, high performance Mott FETs that would motivate further significant industry investment. Materials of interest here, at least initially, include vanadium oxides such as VO_2 and V_2O_3, as well as perovskite oxides.

Advance enabling capabilities

Implementation of the new physical principles and new materials systems in device technologies requires parallel development of enabling capabilities, primarily versatile synthesis, non-intrusive in situ multi-modal characterization, and reliable multi-scale modeling.

- As the areal/volume density of devices increases, so does the density of interfaces between materials that form device components. Future (BMC) synthesis approaches should be able to control not only the quality of functional materials as such but also the behavior of interfaces between them, including electronic structure, chemical composition, characteristic spatial parameters, and defect distribution.
- Further development of characterization and visualization tools will enable in situ observations of device assembly and operation. In order to minimize interference, including damage, induced by the act of measurement, sparse low-dose spectroscopy and microscopy capabilities and novel metrology tools need to be developed. Automated on-the-fly data analysis, together with machine learning algorithms, can become a powerful method for early detection of runaway processes during the synthesis/operation stages. When coupled with automated feedback, these characterization tools will allow for quantitative analysis of acceptable materials variability ranges and corresponding external conditions.

- Finally, the accuracy of computational materials modeling methods and their ability to access multiple time and length scales need to be advanced in order to guide the development of synthesis and data analysis approaches as well as to capture the effects of inhomogeneties on the behavior of the overall device.

Outcomes and Metrics for Success

Materials research and development underpins the creation of new devices with new functionalities and reduced energy consumption. Consequently, the expected outcomes and metrics for success in the materials research and development space are tied to successful delivery of materials that enable post-Moore computing with reduced energy consumption. A key metric for success is therefore the development of a materials structure that will enable a device that operates at 1 aJ per operation at GHz frequencies or ns switching times. The energy consumption per operation is also tied to the applied fields necessary to operate a device. A second metric is then the successful proof-of-concept demonstration of device operation at less than 100 mV or 10 mT fields. Central to success is also the realization of scaled-down devices. Therefore, materials research and transfers to device development groups must be carefully coordinated so that the impact of contacts, dielectric interfaces, and dimensions/scaling on device operation can be jointly addressed and mitigated by materials and device teams. In a broader sense, success will be determined by a successful hand-off and subsequent demonstration of workable devices that can operate at 1 aJ/operation at GHz frequencies and are stable at operating temperatures.

Unique Capabilities

The DOE National Laboratory complex provides a powerful combination of cross-cutting subject-matter expertise. In addition, existing and forthcoming world-leading experimental and computational capabilities from the complex will address the fundamental physics and materials science problems that need to be resolved and understood in order to realize and accelerate BMC. The available resources include: state-of-the art facilities for single-crystal growth and the discovery of new materials; deposition and growth techniques, including in situ growth combined with time- and space-resolved spectroscopy and electro-optical measurements at the light sources, as well as ultra-sensitive solid-state NMR; and the unique capabilities provided by the SNS to map lattice and spin dynamics. The electron microscopy and nano centers provide 3D/4D multi-modal electron tomography and Lorentz imaging, as well as probe microscopies and other unique characterization techniques, such as multi-modal X-ray microscopy and ultrafast spectroscopy. The synchrotron light sources offer capabilities ranging from in situ tracking of time-resolved structural, electronic, and magnetic properties to tools for understanding the synthesis of new materials in real-time.

The Lab complex also provides expertise and the Leadership computational facilities for fundamental materials theory and modeling. The National High Magnetic Field Laboratory (Florida State/LANL) has unique high-field and NMR capabilities. Sandia has a clean room fab for silicon and MEMS fabrication as well as testing, reliability, and failure analysis capabilities. However, additional investments will be needed to enhance deposition, integration, and patterning capabilities for new heterogeneous materials sets (e.g., magnetic materials, perovskite oxides).

Milestones

Years 1-2

1. Provide a preliminary joint experimental-computational assessment of the potential speed, power, and interconnect characteristics and options available for the various identified BMC materials options.
2. In conjunction with the Communications and Logic Devices effort, create industry and academic partnerships needed to more quickly advance required understanding.

Years 3-5

1. Demonstrate a workable low-power cryogenic memory (note IARPA C3 project).
2. Demonstrate a high-mobility organic conductor that can be integrated into circuitry.
3. Demonstrate a low-power organic FET (ref: Chemical Sciences and Society Summit 2012).
4. Demonstrate (proof of principle) a low-power spintronics gate that can be integrated into circuitry.
5. Demonstrate (proof of principle) an all solid state Mottronics FET with a path toward sub-60 mev threshold slope at low voltages (< 0.5 V).
6. Demonstrate (proof of principle) a patterned controlled topological surface state for conduction
7. Demonstrate a controlled bandgap and doping of 2D chalcogenides.
8. Demonstrate a workable high mobility interconnect.

Year 5

1. Down-select to three potential pathways for scalable, energy-efficient BMC.

Interactions with other Efforts

The Materials effort will stimulate the work of the Devices team by generating materials systems that demonstrate new physical phenomena which could be utilized in logic and/or memory technologies, and by providing recommendations regarding device architecture, physical parameters and operation protocols. In turn, the Device team will provide quantitative analysis of device overall behavior, identify performance bottlenecks and set improvement targets. These targets will then be used to guide design of the materials structure and properties, as well as direct efforts on advancing synthesis procedures and control of defects.

Close interaction between the Materials and Manufacturing efforts will guide the development of robust metrology tools and processing controls (e.g., heat management) for device manufacturing, and will facilitate the design, discovery and development of compatible materials systems amenable to scalable production. Coupling between the Device, Materials and Manufacturing efforts will enable the establishment of acceptable ranges of compositional, structural and performance variability, which in turn, will guide the development of optimal strategies for deterministic and/or imprecise computing by the Software and Hardware efforts

Working Group 5: Advanced Manufacturing

Introduction

Beyond Moore Computing is a unified effort by material, device, fabrication, architecture, and software experts to design the next generation of compute devices necessary to remain competitive in the global economy, and promote ongoing growth of an application space that assumes continued improvement in compute power efficiency. Manufacturing & Fabrication focuses on delivering viable strategies to realizing working, reliable apparatuses based on these new designs that can be cost effectively manufactured at scale, and exploring new fabrication pathways to bring the next manufacturing paradigm into focus. In this section, we propose ambitious research strategies that build upon current manufacturing and scientific expertise, lay out key requirements for scalable fabrication, and describe interaction between DOE facilities, working groups, academia, and industry to prototype Beyond Moore computational infrastructure.

Motivation

Gordon Moore famously noticed that continual improvement in manufacturing of integrated circuits had resulted in a paradigm where "the complexity for minimum component costs [is increasing] at a rate of two per year." At the same time, the cost and complexity of the core fabrication methods of integrated circuits, such as photolithography, deposition, and etch have also risen exponentially. As feature length-scales approach 7 nm (~20 Si atoms), the threat to continuation of Moore's Law is two-fold: (1) shrinking macroscopic processes to molecular/atomic scales is running into the physical limits of matter itself; and (2) practically, the cost of the next generation tools is becoming prohibitive.

Continuous improvement in the quality of manufacturing has historically been the driving force behind the exponential increase in computing efficiency, rooted mainly in increasing the number of logical operations possible in a given power footprint. By contrast, the primary role for manufacturing in transitioning to a Beyond Moore Computing (BMC) paradigm is path-finding in how to achieve scale for the materials, devices, and architectures that themselves will provide the continued improvement in energy per operation, and risk mitigation in pursuing only those which can achieve scale. The Manufacturing & Fabrication working group also seeks to develop capabilities that enable the discovery of new devices, circuits, and architectures by pushing fabrication tools, and the associated manufacturing materials and processes, to the molecular/atomic limit. Herein a subtle distinction arises with respect to developing materials which are a part of the fabrication process itself; to avoid the confusion with the efforts of the Materials working group, these materials are consolidated under the term *process* for the rest of this document.

Top-down meets bottom-up

DOE Labs are currently engaged in a broad set of metrology and fabrication activities aimed at improving the current so-called top-down manufacturing paradigm, based on shrinking macroscopic photolithography, etch and deposition processes to the point they reach fundamental limits at the atomic/molecular scale. Critically, there are conceptual gaps in understanding these macroscopic processes at the atomic/molecular scales; from how radiation-matter interactions work (e.g. photon/soft-

matter for EUV photolithography), to identifying useful metrology at hard-soft interfaces during processing, (e.g. the etch damage to the molecules at the edge of a resist mask).

In the near term, the biggest impact to manufacturing will likely originate from developments at the device and materials level. A key contribution of BMC manufacturing will be to determine how new technologies can be translated into large-scale manufacturing, thus helping forecasting the likelihood of realizing their potential. Significant impact will originate from fabrication workflows that incorporate coherent design of tech trees to optimally utilize current manufacturing capabilities. This approach expands current top-down manufacturing capabilities, and adds elasticity in integration of new components to existing algorithms and architectures, as summarized by Scenario I in Figure 1. Furthermore, this manufacturing approach strategically positions this multifaceted collaborative effort to meet challenges in Scenario II and III (Figure 1).

There are extensive benefits in simultaneously pursuing fabrication approaches completely outside of the current fabrication paradigm, in the realm of devices or circuits built bottom-up starting at the molecular/atomic scale. In the near-term, by offering substantially higher precision in fabrication and metrology, non-production-scale tools can provide critical 'look-ahead' risk mitigation in determining which manufacturing processes, material improvements, or devices are the most fruitful to pursue. In the long-term, exploring unconventional processing introduces opportunities to discover new pathways within the current top-down paradigm, and to discover pathways which may form the backbone for future manufacturing paradigms.

Finally, the development of new capabilities, from both the top-down and the bottom-up perspectives, is complementary. Broadly, these new capabilities will generate new insights within the other working groups. We propose a fully integrated, cross-cutting manufacturing approach capable of rapidly testing material, device, and architecture prototypes, and simultaneously developing methodology for industry to package new devices into existing computational infrastructure. This approach also naturally evolves into a close collaboration with Circuits and Algorithms efforts in manufacturing the first completely nontraditional computing testbeds.

Research Agenda

Multi-modal visualization and process simulator

Multi-modal visualization presents a major opportunity in understanding macroscopic processing at the molecular/ atomic scale, which is likely to be critical in achieving scalability and throughput. Unsolved problems include complete three-dimensional characterization of inhomogeneous non-crystalline material (such as developed photoresist), understanding radiation- and particle- matter interactions at the atomic/molecular scales, and nanometer-scale correlation of process variations at every step of fabrication, over a wafer, with device and circuit performance. Novel approaches to data collection, processing and storage will be used to feed large-scale dynamic statistical simulations aimed at understanding the underlying physics. In turn, real world physics insights uncovered in this process will drive decisions and instrument settings, with the ultimate goal of accelerating the development of new manufacturing methods, process flows, and process simulators.

Real-time feedback and control

Understanding the chemical and structural composition of a wafer, or a chip, before and after a processing step, is a traditional open-loop debugging approach. The creation of fabrication instruments that monitor key processes *in-situ* opens the door to process-optimizing closed-loop feedback in real time. Analyzing multimodal experimental data in real-time will provide rapid information exchange between experiment and simulation that act as both, a guide and a validation steps; which ultimately distills to predictive capabilities in the manufacturing process. To unlock this *in-situ* multimodal monitoring capability in real time, significant emphasis must be placed on forming a tight loop between dynamic streaming of experimental data, and model-driven interpretation of the data back to the instrument controller. As these techniques mature, theory-in-the-loop will significantly expedite the quality control process, and other important facets necessary to transition from scientific test-beds to high volume manufacturing.

Advanced manufacturing

Advanced manufacturing techniques, outside the current paradigm, offer the possibility of an advanced low-cost discovery platform, and for an accelerated path to integration. Atomic-precision manufacturing offers a risk-mitigation opportunity by eliminating shortcomings in conventional fabrication in evaluating materials, devices, and circuits, along with identification of what fabrication process improvements offer the biggest payoffs for a given technology. Examples of existing technology which could have future impact in making idealized prototypes includes atomically precise patterning of a monolayer of hydrogen, acting as a resist, on silicon <100> by using scanning probe microscopy; as well as the induction of electrochemical reactions at nano-scale in solid-vapor and solid-liquid interfaces, amorphous to crystalline phase transformations with atomic layer precision, and the directed motion of a specific single dopant atom within a crystal lattice by focused electron and/or ion beams.

Molecular scale approaches that rely on self-assembly provide opportunities for scalable manufacturing of meso-scale functional structures that depend on collective, multi-atomic processes. While individual success stories can be illustrated for a multitude of samples and techniques, robust techniques that produce high yield, scalable devices, accounting for structural and chemical variability, have not been demonstrated. Non-traditional manufacturing paths, such as additive manufacturing techniques, could provide an opportunity for directing self-assembly. Additive manufacturing also presents a way to accelerate integration between widely different device and material types, including, but not limited to, self-assembled structures.

Variation and Defect Tolerant Manufacturing

Direct simulation feedback and robust process control will continue to enable and broaden deterministic manufacturing. However, a continuous stream of new materials, device architectures, and software from other Working Groups will require more tolerant, probabilistic approaches to process variation and defects, while maintaining commercially viable yields. This notion must propagate throughout all stages of development, from basic device design, to architecture integration, to *in-situ* repair, to software adaptation, etc. We have identified the following strategies to aid in addressing defect contribution to device performance, and ways of ameliorating them:

- Probabilistic analysis of defect propagation (location and type) through the manufacturing process to allow software based error correction
- *In-situ* device testing and repair (similar to existing methods for improving memory yield) for logic circuits
- Identification of computational problems compatible with non-deterministic device performance.

Outcomes and Metrics for Success

The metric of success for the fabrication and metrology technologies the Manufacturing & Fabrication group seeks to develop, in the short-term, is how transferable new methodologies are to current-day practices, particularly in silicon based production. Longer term solutions need to be co-developed, such that there is a clear path for passing new technologies off to partners in the supply chain.

Ultimately, the goal of the other working groups is to identify which combination of technologies promise to achieve 10 fJ per equivalent operation, while the manufacturing group seeks to predict the probability that these technologies can achieve manufacturable scale .

Interaction with academia and industry

The research activities of the Manufacturing & Fabrication Working Group are in the spotlight to address tooling needs for metrology and fabrication, material preparation and characterization; and finally device manufacturing and testing. A multifaceted interaction with industry is critical, as they will be the ones delivering solutions based on the discoveries the BMC effort accelerates. Leveraging existing partnerships, we will pursue early engagement to align project goals towards market-viable solutions. Simultaneously, we will continue developing relationships with academia, accelerated leveraging existing infrastructure in SRC or similar organizations, to accelerate the most promising advances from university researchers. Persistent contact with both academic and industrial partners will be critical in pushing the viability of new processes, tools, and approaches. Longer-term, disruptive technologies are likely to create widespread supply-chain issues, which notoriously inhibit the adoption of new technologies, and too need to be addressed by the Advanced Manufacturing group.

One possible path for future academic interactions could be to direct them through existing infrastructure, such as SRC. Other options will also be explored.

Interactions with other Efforts

Manufacturing & Fabrication effort will interact closely with all of the efforts, as Materials, Devices, Architecture and Software dictate form and function of a BMC device. The key challenge for Beyond Moore manufacturing is timely coaction with all the Working Groups. Risk-assessment, roadblock workarounds, and prediction of dead ends to manufacturability need to be addressed at every level of the framework. This includes, for example, the integration of new materials into existing process flows, reliable device fabrication, assessing the manufacturability of circuits, and the integration of heterogeneous accelerator circuits into architectures.

Unique Capabilities

Critical facility-level capabilities are available at several national labs, typically inaccessible to industry or academia. Chief amongst these is supercomputing facilities, including half of the top 10 performing supercomputers as of Nov. 2016. Five DOE Labs also host Nanoscale Science Research Centers, which house complete sets of tools necessary for nanoscale fabrication and characterization, and combine them with signature capabilities and areas of expertise, some of which will be described below. These capabilities often also leverage truly unique facilities at individual labs- Ion Beam Laboratory (SNL) for ion implantation, the spallation neutron source (ORNL) for materials characterization, the Advanced Light Source (ALS) for bright tunable ultraviolet, extreme ultraviolet, and soft x-ray light, the National Synchrotron Light Source II (NSLS-II) for very bright mid-range x-rays, and the Advanced Photon Source (APS) for very bright hard x-rays.

Argonne National Laboratory

Argonne National Laboratory's APS leverages the brightest storage ring-generated hard x-ray beams in the western Hemisphere to deliver an unmatched ability to spatially resolve 3D structure, chemistry, strain, magnetism, and oxidation states of nanostructures and buried interfaces which are critical to the functioning of nanoscale devices. Real-time functioning of devices are probed with operando studies, and the mechanics of synthesis and fabrication are probed with existing wide-range of in-situ manufacturing capabilities, including oxide MBE, PLD, ALD, ALE, MOCVD, and sputter deposition. Spatial resolution down to ~20nm can be achieved in direct 3D imaging at the Hard X-ray Microscopy beamline, and phase imaging using the coherent scattered x-rays provide highly sensitive measurements of sub-Ångstrom atomic displacements. Through the pulsed nature of the X-ray source, it is also possible to track transient processes and dynamically driven changes in properties with a time resolution of ~ 100 ps.

Lawrence Berkeley National Laboratory

A signature multi-level capability which extends the traditional top-down fabrication paradigm has been nurtured at Lawrence Berkeley National Laboratory. Leveraging the ALS, a synchrotron providing one of the brightest tunable sources of extreme ultraviolet light in North America, LBNL, in a 20 year partnership with industry, has developed the tools, materials, and methods to implement extreme ultraviolet (EUV) photolithography. Significant investments in cryogenic and laser-assisted etching, used to transfer the pattern into silicon, have been critical in establishing processes based on EUV.

Oak Ridge National Laboratory

Oak Ridge National Laboratory has made significant investments in attaining full control of atomic arrangement and bonding in three dimensions. The last two decades witnessed substantial industrial, academic, and government research efforts directed towards this goal through various lithographies and scanning probe based methods. These technologies emphasize 2D surface structures, with some limited 3D capability. Recently, at ORNL a range of focused electron and ion based methods have demonstrated compelling alternative pathways to achieving atomically precise manufacturing of 3D structures in solids, liquids, and at interfaces. Electron and ion microscopies offer a platform that can simultaneously observe dynamic and static structures at the nano and atomic scales, and also induce structural rearrangements and chemical transformation. Currently, energetic beams have been used to create free-standing

nanoscale 3D structures, fabrication potential with liquid precursors, epitaxial crystallization of amorphous oxides with atomic layer precision, as well as visualization and control of individual dopant motion within a 3D crystal lattice.

Pacific Northwest National Laboratory

Pacific Northwest National Laboratory developed instrumentation capabilities and expertise enabling multi-modal in situ characterization of solid and liquids phases and solid/liquid interfaces. These include nuclear magnetic resonance spectroscopy suite, sparse low-dose electron microscopy sampling, extensive expertise in design and in situ characterization and control of electrochemical devices, and a range of mathematical and computational methods facilitating capture and analysis of experimental data. While originally developed for the national security applications, these methods proved to be useful in interpretation of, for example, electron microscopy data.

Sandia National Laboratories

Sandia National Laboratories has signature capabilities aimed at integration of materials, devices, and circuits into multi-functional platforms. Chief amongst these is a limited production scale fab for silicon and III-V semiconductors. This has enabled taking research-scale efforts to manufacturing-scale in MEMS and photonics. Conversely, the combination of the fab with sophisticated metrology tools has enabled the investigation of manufacturing yield issues in many research-level problems, which turn around to feed a deeper understanding of the fabrication processes themselves.

Generally, these unique capabilities are complemented by leading supercomputing facilities. These supercomputing resources are leveraged for everything from atomistic simulations to understand and tailor materials and processes, machine learning for design optimization, and real-time analytics of large-scale image data sets. In the future, such computing resources may prove critical in the control systems which enable the next generation of manufacturing.

Other Important Work

- Manufacturing support: Materials, Architectures, etc.
- Competitive landscape analysis (China)

Conclusion/Overall Plan: Distillation/Rollup of Plans

- Restate the problems and the framework solution
- Overall vision and outcomes: Roadmap with milestones at 1, 3, 5, 10, and 15 years
    - Year 1
        - Detailed roadmap
        - OGA & Industry engagement
        - Small "team-building" projects in efforts
        - ROM cost
    - Year 3
        - World-first demonstrations
            - Novel materials?
            - Device physics and devices?
            - Failure analysis, models = pieces of the Framework value prop ➔ demonstrate capability of generating data for a single prioritization recommendation/downselect example?
        - Example of Industry support
        - ROM cost
    - Year 5
        - Demonstration of full Framework (mark 0.9 version)
        - Several draft scenarios to 20 fJ/instr
        - Industry engagement
        - Capability of generating data for several prioritization recommendation/downselect examples?
        - Multiple Industry partnerships started
        - ROM cost
    - Year 10
        - Viable, defendable, comprehensive path to 20fJ/instr across all Framework levels
        - Industry transition/engagement starting
        - Start of transition from pre-competitive to competitive
        - ROM cost
    - Year 15
        - BMC devices in pre-production that will address energy crisis
        - Still have strong partnerships with Industry where BMC BI adds value?
        - ROM cost
- Motivational charge to the government: start now and be serious!

Conclusion

The 50-year sustained growth of computing capability at relatively low consumer cost in terms of both dollars and **energy** has masked the urgency of the present situation. Analyses of the growth of computing consumption versus performance scaling indicate that global electricity consumption by the IT sector may need to grow from its current 3-4% fraction to over 30% in the next two decades to support the progress we've come to expect. This sector represents the fastest growing consumer of energy, and uncontrolled, this demand would have significant implications on the U.S. energy landscape. In one example of projected IT energy growth, Cisco reports that data center traffic (a useful metric for energy demand) is projected to grow at a compound annual rate of 25% from 2014-19. With no improvement in computing efficiency (i.e., the end of Moore's Law), one would expect this growth to be directly reflected in increased energy demand going from 91 billion kilowatt-hours in 2013 to 252 billion kilowatt-hours in 2018 out of a total US consumption of over 4000 billion kWhrs. Simply *meeting* this increased demand would require 60 new 500-megawatt power plants. The energy requirement is likely to be exacerbated in the next decade with the end of conventional Moore's Law technology scaling.

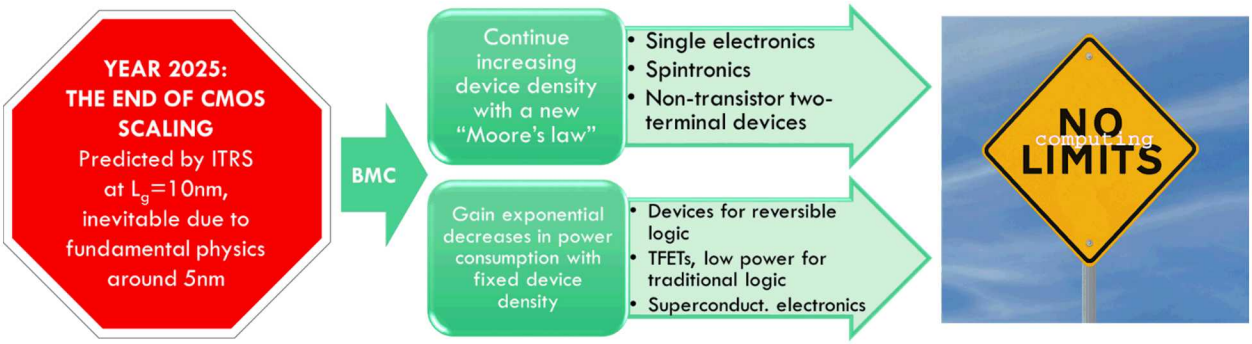
Economic and **national security** timelines are also urgent, and disruptions in the pace of our progress in computing can have dramatic consequences: financial management, advanced manufacturing, online sales, global supply chain management are all demanding greater information processing fidelity at higher speeds, and the consequences of missteps are increasing (witness recent layoffs and increased consolidation of the entire ecosystem). A similar situation exists in our military and intelligence missions against terrorism, cyber-warfare, and modern ground and air warfare, where the pace of strategic and tactical decision-making is rapidly accelerating.

DOE has a unique opportunity to apply unique DOE capabilities to new Public-Private programs for basic/applied research to accelerate the development of energy efficient IT beyond the end of current roadmaps, as well as to maintain an advanced manufacturing base in the economically critical semiconductor space. These programs will enable the government to coordinate current and future BMC investments more effectively, as well as to leverage significant industry investments. Pieces of this may already exist, in programs within multiple DOE and NSF program offices, within IARPA and NSA, and at NIST and DARPA. A large number of independent university projects exist in this area, supported by DOE and NSF funding, and the Semiconductor Research Corporation (SRC) uses DARPA, NSF, and industry funding.

The Department of Energy (DOE) is particularly well positioned to help address the developing threats, with program offices and national labs that serve the nation in all three of these problem areas (**energy**, **economy**, and **national security**), and a wealth of unique capabilities and depth of useful expertise. DOE Labs already support some of the ongoing work listed above, but largely in a diffuse, "individual relationship" manner. A broader initiative would provide easier access to the wealth of subsidized capability at the Labs, as well as break down some stovepipes and remove duplicate efforts. The DOE also has a leadership role in the National Strategic Computing Initiative, is a primary driver for the "top of the computing pyramid" (HPC), and has successful experience with energy efficiency initiatives. The bulk of the impact of a successful initiative in this area would be achieved by increasing the efficiencies and

manufacturing base for mobile, consumer, and "commodity" computing rather than "bleeding–edge" HPC, but advances from HPC investments within DOE will drive computing hardware and software progress forward, such that leading edge HPC capability becoming a commodity is accelerated. Additionally, the science performed on, and enabled by, HPC at the labs benefits industrial breakthroughs and spin-off support technologies.

The essence of the Beyond Moore Computing challenge is that, even if the significant near-term technological and economic difficulties for the continued increase of logic element density are solved by the industry, CMOS devices as they are currently understood and manufactured can be pushed no further below 5 nm size scale, meaning that the increases in computational power that have come with every new process node will effectively end in the very near future.



There are two principle paths for continuing the significant increases in computational power for future generation electronics [DMG15]. One is to continue **increasing the density of logic elements by extending Moore's law with non-CMOS devices** such as single-electron transistors, spin-based transistors or by utilizing (non-transistor) two terminal devices such as memristors or phase-change devices. The other path is to attempt to achieve **exponential decreases in power consumption** due to improvements in materials, devices, circuits, architectures, and/or algorithms while **keeping the device density fixed**. Each of these two paths presents daunting technological, economic competitiveness, and energy challenges, and will require coordinated and coupled advances in materials and device technology, in scalable circuit integration, manufacturing, and packaging technologies, and in novel system architectures and programming models.

In order to succeed in implementing the corresponding recommendations from the four efforts (Algorithms and Software Environments, Hardware and Circuit Architecture, Communications and Logic Devices, and Materials), the team of 8 DOE national labs is proposing a Multiscale Co-Design Framework that will enable both top-down coupling of application and architectural requirements to circuits and devices, as well as bottom-up coupling of materials and device physics constraints to algorithms and architectures. Portions of this framework exist in today's industrial toolkit, but they are narrowly focused on the current silicon CMOS technology stack. It is extremely challenging for the industry to tightly couple activities across such different technical disciplines and paths. Additionally, the time and expense to develop strong coupling is poorly matched to the rapid pace of industry development cycles.

The structure of the Multiscale Co-Design Framework that we envision is shown in Appendix B. The framework will couple modelling at all of the key technology levels, validated with experimental data, parameter extraction, and physical demonstrations. This approach will exploit a large base of multiscale, multiphysics modelling and HPC at the Labs, as well as the large unique base of experimental, fabrication, and metrology capabilities maintained across the nation within these facilities. Indeed, small-scale versions of this approach are already underway at the Labs (and universities). The framework will also leverage the broad and deep expertise of Labs' scientists, and the "shared culture" of the DOE Labs to overcome communication hurdles encountered in working across the technical disciplines but within the layers of the framework. Such a framework will provide a means to identify the most promising Beyond Moore Computing ideas for application impact and to avoid some of the "blind alleys" (material incompatibilities, unmanufacturable devices, circuit resiliency, etc.). It will also enhance communication and coordination across disciplines and organizations in industry, academia, and government. This co-design framework can eventually deploy powerful optimization and sensitivity/variance analysis codes to evaluate design tradeoffs across many levels and provide a means to develop "educated bets," quantify potential gains, and measure progress in this multi-disciplinary R&D environment. Only with co-design covering this broad space and consideration of manufacturing challenges can we expect to make complementary progress in all areas and bring about real change to the IT energy outlook.

In addition to containing the growth of IT related energy demand, the output of this work will provide a path to sustaining exponential growth in computing capabilities to enable new scientific discoveries, maintain U.S. competitiveness in all segments of the computing market (from IoT, to datacenters, to supercomputing), and thus guarantee U.S. economic competitiveness and national security.

To meet the goal of broad societal impact, we must not only ensure the transition of basic research to high volume manufacturing but more fundamentally shape basic research from the start, with an eye to manufacturability. This will be achieved through the development of a multi-lab ecosystem serving as a unified facility that can evaluate and demonstrate the manufacturing and energy savings feasibility of next generation technology options. Technologies will be rigorously evaluated for potential benefits on energy, implications on architecture, and programming paradigms. The most promising technologies will be evaluated for issues around high volume manufacturing followed by ramp-up demonstration and their ability to deliver on the energy requirements.

References

[DJL14] P. D. Düben, J. Joven, A. Lingamneni, et al., On the use of inexact, pruned hardware in atmospheric modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 372, no. 2018, 2014.

[BHM16] A. H. Baker, D. M. Hammerling, S. A. Mickleson, H. Xu, M. B. Stolpe, P. Naveau, B. Sanderson, I. Ebert-Uphoff, S. Samarasinghe, F. De Simone, F. Carbone, C. N. Gencarelli, J. M. Dennis, J. E. Kay, and P. Lindstrom. Evaluating lossy data compression on climate simulation data within a large ensemble. Geoscientific Model Development Discussions, 2016:1–38, 2016.

[BXM14] A. H. Baker, H. Xu, J. M. Dennis, M. N. Levy, D. Nychka, S. A. Mickelson, J. Edwards, M. Vertenstein, and A. Wegener. A methodology for evaluating the impact of data compression on climate simulation data. In Proceedings of the 23rd International Symposium on High-performance Parallel and Distributed Computing, HPDC '14, pages 203–214, New York, NY, USA, 2014. ACM.

[MIT16] Sparsh Mittal. 2016. A Survey of Techniques for Approximate Computing. *ACM Comput. Surv.* 48, 4, Article 62 (March 2016), 33 pages. http://dx.doi.org/10.1145/2893356

[AAH13] Armin Alaghi and John P. Hayes. 2013. Survey of Stochastic Computing. *ACM Trans. Embed. Comput. Syst.* 12, 2s, Article 92 (May 2013), 19 pages. http://dx.doi.org/10.1145/2465787.2465794

[ACK06] B. E. S. Akgul, L. N. Chakrapani, P. Korkmaz and K. V. Palem, "Probabilistic CMOS Technology: A Survey and Future Directions," 2006 IFIP International Conference on Very Large Scale Integration, Nice, 2006, pp. 1-6.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4107595&isnumber=4107582

[JHO13] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," *2013 18th IEEE European Test Symposium (ETS)*, Avignon, 2013, pp. 1-6.
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6569370&isnumber=6569335

[KPL13] Krishna Palem and Avinash Lingamneni. 2013. Ten Years of Building Broken Chips: The Physics and Engineering of Inexact Computing. *ACM Trans. Embed. Comput. Syst.* 12, 2s, Article 87 (May 2013), 23 pages. http://dx.doi.org/10.1145/2465787.2465789

[WIG06] Avi Wigderson. 2006. The power and weakness of randomness in computation. In *Proceedings of the 7th Latin American conference on Theoretical Informatics* (LATIN'06), José R. Correa, Alejandro Hevia, and Marcos Kiwi (Eds.). Springer-Verlag, Berlin, Heidelberg, 28-29.
http://dx.doi.org/10.1007/11682462_7

[DMG15] Denis Mamaluy, Xujiao Gao, "Fundamental Downscaling Limit of Field Effect Transistors", Appl. Phys. Lett., 106, 193503 (2015). http://dx.doi.org/10.1063/1.4919871

Appendix A: Workshop Agenda and Breakout Group Topics

Solving the Information Technology Challenge Beyond Moore's Law

DOE Big Idea National Lab Meeting

July 27-28, 2016

Albuquerque, NM

Topics addressed in the workshop:

1. Identify an inventory of each lab's key contributions (expertise, capabilities, facilities, funded projects, etc.)
2. Identify and define compelling aggregate (multi-lab) Value Propositions for the Big Idea
3. Draft timelines and propose critical projects for review by key DOE program offices

July 27, 2016 – CINT Chaco

8:00 AM	Open Workshop: Dave Sandison, *SNL*
8:15 AM	Big Idea Recap & Feedback: Rick McCormick, *SNL* & John Shalf, *LBNL*
9:15 AM	NSCI Overview/National Landscape: Bill Harrod, *SC/ASCR*
9:45 AM	Advanced Manufacturing for the End of Moore's Law and Beyond: Mark Johnson, *EERE/AMO*
10:15 AM	Break
10:30 AM	LANL Capabilities and Interests: Toni Taylor & John Sarrao
11:15 AM	PNNL Capabilities and Interests: John Johnson
12:00 PM	Hosted Lunch
1:30 PM	ORNL Capabilities and Interests: Bobby Sumpter
2:15 PM	ANL Capabilities and Interests: Olle Heinonen
3:00 PM	Break
3:15 PM	LLNL Capabilities and Interests: Peg Folta & Matt Horsley
4:00 PM	LBNL Capabilities and Interests: Patrick Naulleau
4:45 PM	SNL Capabilities and Interests: Matt Marinella & Si Hammond
5:30 PM	Adjourn
7:00 PM	Hosted Dinner at El Pinto

July 28, 2016 – CINT Chaco

8:00 AM	Breakout Kickoff: Multiscale Codesign (MSCD) Framework Background/Expertise: Toni Taylor, *LANL*	
8:15 AM	Group 1: Implications for Algorithm and Software Environments	Room 1026
	Group 2: Hardware and Circuit Architectures	Room 1151
	Group 3: Communication and Logic Devices	Room 1041
	Group 4: Materials	Room 1024
10:45 AM	Break	
11:00 AM	Breakout Group 1 Debrief	
11:30 AM	Breakout Group 2 Debrief	
12:00 PM	Hosted Lunch	
1:00 PM	Breakout Group 3 Debrief	
1:30 PM	Breakout Group 4 Debrief	
2:00 PM	Break	
2:15 PM	Discuss/Distill Value Propositions from Breakout Working Groups	
3:00 PM	Summarize and Finalize Key Conclusions to Include in Report	
3:45 PM	Develop Impact Milestones	
4:30 PM	Review Actions Items and Plan Next Meeting	
5:15 PM	Wrap-up	
5:30 PM	Adjourn	

Breakout Group Instructions:

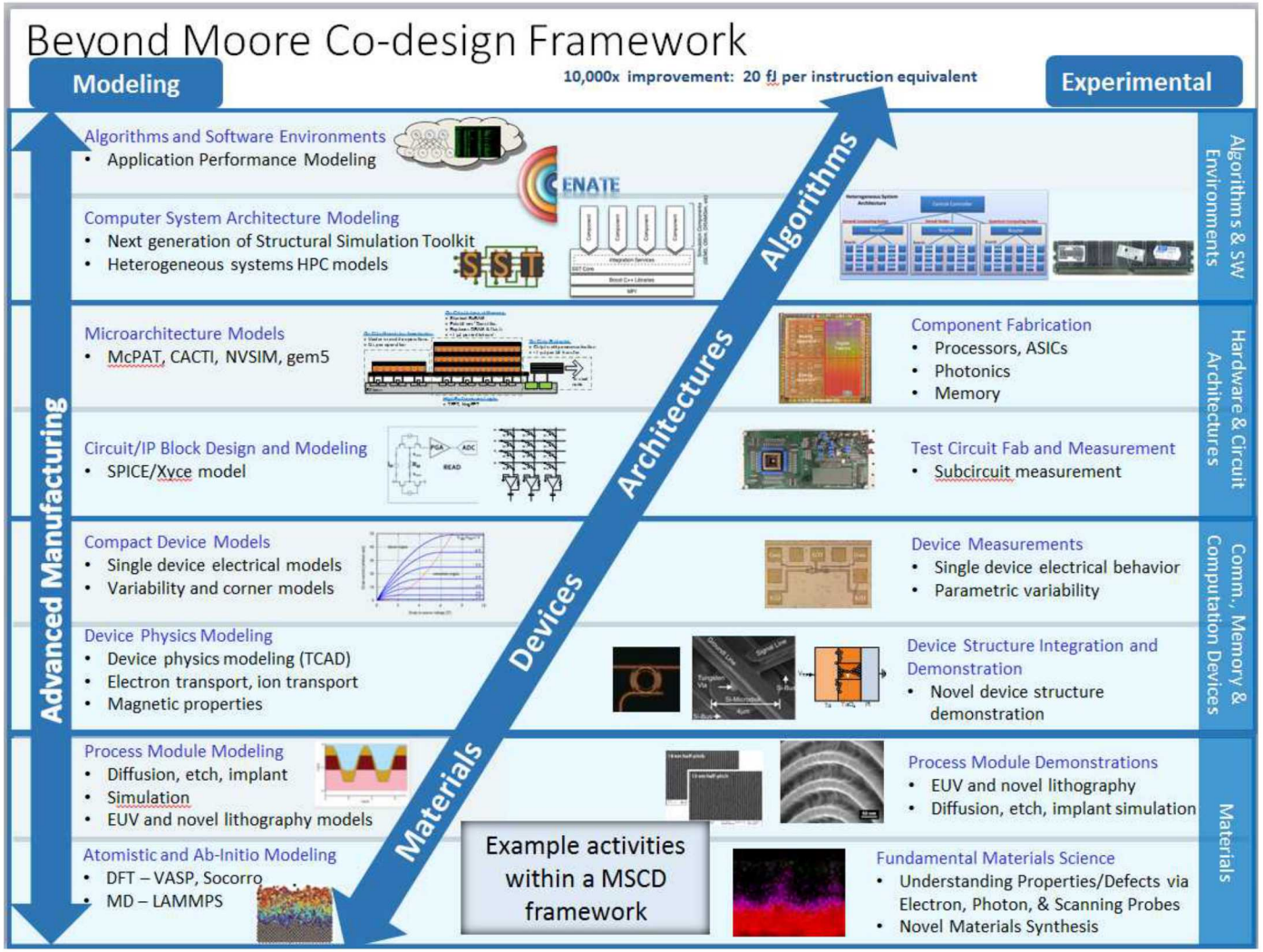
1. POCs will give an overview of their area and introduce team
2. Facilitators will conduct the breakout sessions, keeping time and communicating the goals
3. Scribes will record discussion

Breakout Group Deliverables:

1. For the first hour, discuss and inventory each lab's key contributions (expertise, capabilities, facilities, funded projects, etc.)
2. During the second hour, identify and define compelling aggregate (multi-lab) value propositions
3. During the last thirty minutes, draft an exemplar project(s) with rough order of magnitude schedule and costs

Group 1: Implications for Algorithm and Software Environments	Room: 1026
• What key existing capabilities and programs from each lab can be leveraged? Example: SST, Exascale project, CENATE. • Which DOE program areas are we targeting and what are their interests? What application targets should be modeled in the initial framework? Broadly, we have defined this as Scientific Computing; what more specific directions and codes do we want to target? • How extensively can existing modeling codes be leveraged to develop the modeling framework? • What new software ideas are needed? • What does a successful project look like? (key challenges, capabilities, & risks) • What are the current readiness, anticipated schedule, and budget for this work? • Identify any potential gaps in capabilities and strategy. • Summary statement of how labs will organize themselves to address the topical area.	
Group 2: Hardware and Circuit Architectures	**Room: 1151**
• What existing research projects, expertise and capital equipment from each lab can be leveraged? • What key Beyond Moore architecture/component approaches can be demonstrated within the framework? • How do the hardware and circuit models integrate with the Software/Algorithm models? • What is the path to continuous manufacturable technology advancement? (New Moore's Law) • What does a successful project look like? (key challenges, capabilities, & risks) • What are the current readiness, anticipated schedule, and budget for this work? • Identify any potential gaps in capabilities and strategy. • Summary statement of how labs will organize themselves to address the topical area.	
Group 3: Communication and Logic Devices	**Room: 1041**
• What existing device research, expertise and capital equipment from each lab can be leveraged? Examples: MESA, Berkeley Molecular Foundry, CNMS. • What are the key logic devices we want to model and demonstrate in the framework? • What are the key communication devices we want to model and demonstrate in the framework? • What are the key device manufacturability issues that must be addressed? • What does a successful project look like? (key challenges, capabilities, & risks) • What are the current readiness, anticipated schedule, and budget for this work? • Identify any potential gaps in capabilities and strategy. • Summary statement of how labs will organize themselves to address the topical area.	
Group 4: Materials	**Room: 1024**
• What existing simulation capabilities, projects and expertise from each lab can be leveraged? Examples: The Material Project, Which BES facilities can be leveraged? • What key new electronic and photonic materials should be investigated for low energy computing? • What are the key materials manufacturability issues that must be addressed? • What does a successful project look like? (key challenges, capabilities, & risks) • What are the current readiness, anticipated schedule, and budget for this work? • Identify any potential gaps in capabilities and strategy. • Summary statement of how labs will organize themselves to address the topical area.	

Appendix B: Beyond Moore Co-design Framework



Appendix C: Existing Lab Capabilities for Software Technology

Capabilities by Lab

Lab	Capabilities
Argonne	• Compiler and optimizations for FPGA • Inexact/approximate/lossy computing (fundamental, SZ compressor) • Optimization tool (auto-tuning, Hovland) • Messaging layers (MPI) • Compiler (LLVM) • Numerical libraries (PETSc) • OS (Argo) • I/O (file systems) • Resilience (error detection)
Berkeley	• Domain-specific languages & compilers (stencils) • Auto-tuning, performance modeling (roof-line) • Global-address space programming • Light-weight communication • Checkpoint restart • Motif-specific frameworks (AMREX – mesh refinement, etc.) • SuperLU
LANL	• Domain specific language (Scout) • Abstraction libraries (FleCSI) • Mini-apps (SNAP, COMD, PENNANT, CLAMR, NuT, etc.) • Scientific computing on Hadoop, etc. • Development and usage of LEGION • Open MPI expertise • Burst-buffers and IO • General LLVM expertise • *Quantum computing • Visualization tools (Paraview)
LLNL	• Working with large capacity memories (cluster CATALYST, attached high performance PCIe flash) • Emulator capabilities (FPGA) • OS-level caching of persistent memory etc. • Checkpoint restart (SCR) • Data structure portability/layout (RAJA) • Rose (source2source translator) • HYPRE package • CHAI • Performance tools • Power measurement/adaptation • SAMRAI (adaptive mesh refinement) • Proxy apps • Hetero computing • *Neuromorphic computing capabilities • Viualization tools (Visit) • Debugging at scale (STAT)

Lab	Capabilities
ORNL	• ASPEN (coarse-grained performance modeling tool) • OpenARC (accelerator research compiler) • NVL-C (extension to C for programming NV memory) • Open ACC2FPGA compiler (reconfigurable computing) • XSIM (modeling scaling of parallel apps) • *QC and NC • UCX (messaging)
SNL	• N2A (algorithms to architecture compiler) • SST • O/S and runtimes • Kokkos • DARMA • QThreads • Profiling tools • Portals • Scientific libraries • Trilinos • *FPGA tools • Mantevo mini-app suite (multi-lab) • Power-API • UQ and SA (DaKOTA) • *QC and NC • Multi-vendor computer architecture test beds • CBR3D (quantum transport beyond-Moore device simulator), Charon TCAD
PNNL	• One-sided communication (COMEX/ARMCI) • Performance modeling • CENATE • Power analysis • Data and graph analytics

Experiences with Recent Disruptions

Moving forward, it is valuable to learn from our recent experiences with disruptions in architectures, such as the transition to GPUs and many-core processors. The recent move to GPUs started in 1999 when NVIDIA invented the discrete GPU. Over a few years, several early adopters found general-purpose uses for this graphics hardware (J.D. Owens, D. Luebke et al., "A Survey of General-Purpose Computation on Graphics Hardware," Proc. Eurographics 2005, State of the Art Reports, 2005, pp. 21-51). In 2009, NVIDIA launched the CUDA programming language that transformed the programming of these often previously esoteric devices. In 2009, NVIDIA added much improved double precision floating point support and ECC to its memory systems, reflecting the growing importance of GPUs in scientific computation. Also, in the same year, organizations collaborated to propose OpenCL, an open programming language for heterogeneous systems. In 2010, GPUs powered two of the world's fastest systems: Tsubame2 and Tihane-1A. In 2012, ~35 GPU-based systems were on the TOP500 including four

of the top 10 systems. For programming models, in 2011, the first OpenACC compilers began to appear on systems, and in 2015 the first OpenMP offload compilers arrived.

The move to GPUs was aided by the fact that these units allowed for the gradual modification of a single application kernel at a time. In most cases, GPUs were added as an expansion card to commodity servers, and the application's most intensive kernels were rewritten in CUDA or converted to use library calls for specific, critical functions. The GPU's architecture was dramatically different than that of a CPU: GPUs provide throughput-optimized cores, where each core is very simple when compared to a CPU core. Also, the memory systems were different. If GPU memory accesses are coalesced, the memory system can provide significantly more bandwidth relative to a CPU based architecture.

In order to exploit the underlying design of the GPU, applications had to expose massive amounts of concurrency (possibly 10,000 threads or more), minimize synchronization, schedule data into scratchpad memory, and arrange global memory accesses so that they did create hotspots (contention). Meanwhile, the devices and logic of the GPU were similar if not exactly the same as that for a CPU.

The labs, in partnership with industry, consortia, standards bodies, and universities, have played a major role in designing and validating new programming models, language constructs, and software APIs for massively parallel architectures.

Understanding Disruptions

	Example	Features
Algorithm		Computational Complexity, Data complexity
Programming model	MPI, OpenMP	Concurrency, data movement, data placement
Language	C, Fortran, OpenCL, CUDA	
API		
Runtime and Operating system	Linux, MPI runtime, OpenMP runtime, OCR	
Architecture	Processing in memory	Heterogeneity, partitioning, data movement, data placement
ISA		
Microarchitecture		
FU	Approximate/inexact/lossy	Rethink numerical methods, algorithms, compilers, runtimes
Device	FinFETs, single electronics, memristors, superconductive electronics, etc	Rethink software according to the architectural changes