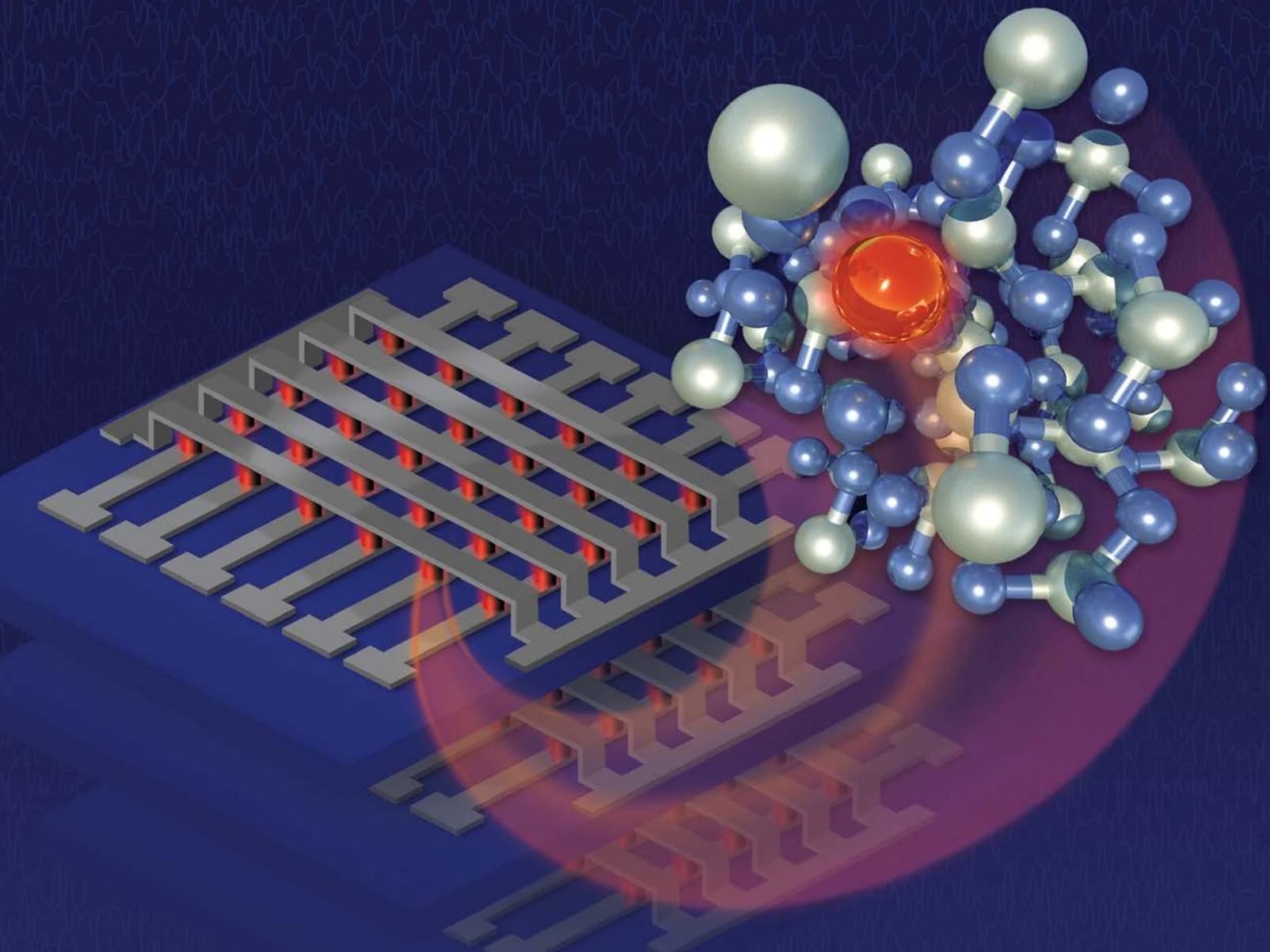


Basic Research Needs for
Microelectronics



*Report of the Office of Science Workshop on
Basic Research Needs for Microelectronics
October 23 – 25, 2018*

Cover image: A simplified schematic of a cross-bar circuit element designed for future low power, non-volatile memory or neuromorphic computing applications. "Row" and "column" metal interconnects form the cross-bar structure, with nanoscale memory elements residing at the intersections. Current research is focused on the design of these materials at the atomic level to enable dense digital and analog memory arrays with performance characteristics well beyond today's circuits.

Image courtesy of Argonne National Laboratory.

DISCLAIMER: This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government.

Basic Research Needs for Microelectronics

REPORT OF DEPARTMENT OF ENERGY OFFICE OF SCIENCE WORKSHOP,
OCTOBER 23-25, 2018

CHAIR:

Cherry Murray, Harvard University

CO-CHAIRS:

Supratik Guha, Argonne National Laboratory and University of Chicago

Dan Reed, University of Utah

TECHNOLOGY LIAISON:

Gil Herrera, Sandia National Laboratories

PANEL LEADS:

Panel 1: Microelectronics for Big Data at Future Facilities: Memory and Storage

Kerstin Kleese van Dam, Brookhaven National Laboratory

Sayeef Salahuddin, University of California at Berkeley

Panel 2: Co-design for High Performance Computing beyond Exascale

James Ang, Pacific Northwest National Laboratory

Thomas Conte, Georgia Institute of Technology

Panel 3: Power Control, Conversion and Detection

Debdeep Jena, Cornell University

Robert Kaplar, Sandia National Laboratories

Panel 4: Crosscutting Themes

Harry Atwater, California Institute of Technology

Rick Stevens, Argonne National Laboratory

PLENARY SPEAKERS:

Dushan Boroyevich, Virginia Tech University

William Chappell, DARPA

Tsu-Jae King Liu, University of California Berkeley

Justin Rattner, Intel (retired)

Michael Witherell, Lawrence Berkeley National Laboratory

REPORT WRITING:

Khurram Afridi, Cornell University
Simon Ang, University of Arkansas
Jon Bock, Sandia National Laboratories
Srabanti Chowdhury, Stanford University
Suman Datta, University of Notre Dame
Keith Evans, Great Lakes Crystal Technologies
Jack Flicker, Sandia National Laboratories
Mark Hollis, Massachusetts Institute of Technology
Noble Johnson, Palo Alto Research Center
Ken Jones, Army Research Laboratory
Peter Kogge, University of Notre Dame
Sriram Krishnamoorthy, Pacific Northwest National Laboratory
Matthew Marinella, Sandia National Laboratories
Todd Monson, Sandia National Laboratories
Sreekant Narumanchi, National Renewable Energy Laboratory
Paul Ohodnicki, National Energy Technology Laboratory
Ramamoorthy Ramesh, University of California at Berkeley
Michael Schuette, Air Force Research Laboratory
John Shalf, Lawrence Berkeley National Laboratory
Shadi Shahedipour-Sandvik, SUNY Polytechnic Institute
Jerry Simmons, Sandia National Laboratories (retired)
Valerie Taylor, Argonne National Laboratory
Tom Theis, IBM (retired)

OFFICE OF SCIENCE LEADS:

Eric Colby, High Energy Physics
Robinson Pino, Advanced Scientific Computing Research
Andy Schwartz, Basic Energy Sciences

SPECIAL ASSISTANCE:

Administrative

Katie Runkles, Basic Energy Sciences

Editorial/Publication

Joseph Harmon, Michele Nelson, Vicki Skonicki, Argonne National Laboratory

Table of Contents

Acronyms	v
Executive Summary.....	1
1. Introduction.....	5
2. Priority Research Directions.....	9
PRD 1 Flip the current paradigm: Define innovative material, device, and architecture requirements driven by applications, algorithms, and software.....	11
Introduction.....	11
Scientific Challenges.....	14
Research Thrusts.....	19
Scientific and Technology Impact.....	20
PRD 2 Revolutionize memory and data storage	23
Introduction.....	23
Scientific Challenges.....	25
Research Thrusts.....	25
Science and Technology Impact.....	29
PRD 3 Reimagine information flow unconstrained by interconnects.....	31
Introduction.....	31
Scientific Challenges.....	32
Research Thrusts.....	35
Scientific and Technology Impact.....	40
PRD 4 Redefine computing by leveraging unexploited physical phenomena	41
Introduction.....	41
Scientific Challenges.....	42
Research Thrusts.....	45
Scientific and Technology Impact.....	47
PRD 5 Reinvent the electricity grid through new materials, devices, and architectures	51
Introduction.....	51
Scientific Challenges.....	53
Research Thrusts.....	55
Scientific and Technology Impact.....	56

3. PANEL REPORTS	59
Panel 1 Microelectronics for Big Data at Future Facilities: Memory and Storage	61
Introduction.....	61
Science Drivers	61
Novel Integration Methods to Enable the “Compute in Storage” Paradigm	63
Heat Dissipation Challenges in Microelectronics	66
A Multi-scale Co-design Framework	67
Materials, Devices, and Architectures for HPC	69
Panel 2 Co-design for High Performance Computing Beyond Exascale	71
Introduction.....	71
Current Status and Recent Advances	72
Scientific Challenges and Opportunities.....	75
Panel 3 Power Conversion, Control, and Detection	83
Introduction.....	83
Current Status and Recent Advances	83
Scientific Challenges and Opportunities.....	84
Panel 4 Crosscutting Themes	95
Introduction.....	95
Component Advances in Co-design Framework.....	95
New Electronic Materials and Phenomena for Information and Energy Transfer.....	96
New Photonic and Optoelectronic Materials and Phenomena for Information and Energy Transfer.....	100
Thermal Energy Management: Materials, Structures, and Architectures	105
Accelerated Co-design of Novel Materials, Device Concepts, and System Architectures for Channel Operation near Quantum Noise/Dissipation Limits	108
Appendix A Preparatory Material for DOE Office of Science Basic Research Needs for Microelectronics Workshop	111
Appendix B Workshop Participants	121
Appendix C Agenda	125

Acronyms

AI	Artificial Intelligence	FEFET	Ferroelectric Field Effect Transistor
AMM	Abstract Machine Model	FeRAM	Ferroelectric RAM
ANN	Artificial Neural Network	FET	Field Effect Transistor
API	Application Programming Interface	FOM	Figure of Merit
ARX	Autoregressive Exogenous Input	FPGA	Field-Programmable Gate Array
ASC	Advanced Simulation and Computing Program		
ASCR	Advanced Scientific Computing Research	GP-GPU	General-Purpose Graphics Processing Unit
ASIC	Application-Specific Integrated Circuit	GPU	Graphics Processing Unit
ATLAS	A Toroidal LHC Apparatus		
		HBM	High Bandwidth Memory
BES	Basic Energy Sciences	HEP	High Energy Physics
		HL-LHC	High-Luminosity Large Hadron Collider
		HMFOM	Huang Material Figure of Merit
CBRAM	Conductive Bridge RAM	HPC	High Performance Computing
CHIPS	Common Heterogeneous Integration and IP Reuse Strategies		
CiS	Computation in Storage	IDRS	International Device Roadmap for Semiconductors
CMOS	Complementary Metal Oxide Semiconductor	ILP	Integer Linear Programming
CMS	Compact Muon Solenoid	I/O	Input/Output
CNT	Carbon Nanotube	IoT	Internet of Things
CPU	Central Processing Unit	IT	Information Technology
		ITRS	International Technology Roadmap for Semiconductors
DARPA	Defense Advanced Research Projects Agency		
DDR	Double Data Rate	LAMMPS	Large-Scale Atomic/Molecular Massively Parallel Simulator
DFT	Density Functional Theory	LCF	Leadership Computing Facility
DOE	Department of Energy	LDOS	Local Density of States
DRAM	Dynamic RAM	LHC	Large Hadron Collider
		LPT	Large-Power Transformer
EB	Exabyte		
EDA	Electronic Design Automation		
ENIAC	Electronic Numerical Integrator and Computer		
EPB	Energy per Bit		
ERI	Electronics Resurgence Initiative		
EUV	Extreme Ultraviolet		

MCF	Moore Co-design Framework	RAM	Random Access Memory
McPAT	Multicore Power, Area, and Timing	RC	Resistive Capacitive
MD	Molecular Dynamics	R&D	Research and Development
MISO	Multiple Input Single Output	RRAM	Resistive RAM
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor		
MRAM	Magnetic RAM	SC	DOE Office of Science
		SERDES	Serializer/Deserializer
		SoC	System on a Chip
NCFET	Negative Capacitance Field Effect Transistor	SPP	Surface Plasmon Polariton
NEMS	Nanoelectromechanical System	SRC	Semiconductor Research Corp.
NERSC	National Energy Research Scientific Computing Center	SRAM	Static RAM
NVSim	Nonvolatile Memory Simulator	SST	Solid-State Transformer
NW	Nanowire	STT-MRAM	Spin-Transfer Torque Magnetic RAM
		TCAD	Technology Computer-Aided Design
PB	Petabyte	TFET	Tunnel Field Effect Transistor
PCH	Platform Controller Hub	TMD	Transition Metal Dichalcogenide
PCIe	Peripheral Component Interconnect Express	TSV	Through Silicon Via
PCM	Phase Change Memory		
PCS	Power Conversion System	UWBG	Ultra-Wide Bandgap
PRBS	Pseudorandom Binary Sequence	VASP	Vienna Ab Initio Simulation Package
PRD	Priority Research Direction		
		WBG	Wide Bandgap
		WDM	Wavelength Division Multiplexing

Executive Summary

Moore's Law — realized via a combination of device physics advances, technology investments, and economic returns— allowed the number of transistors on a chip to double roughly every two years for over five decades. During those five decades, the cost of a unit of computing dropped by eight orders of magnitude. Those declines and the associated computing advances have dramatically affected every aspect of society, from science and technology through business and health to national security.

This virtuous cycle of computing advances and Moore's Law driven innovation is now threatening to end, leaving conventional computing at a critical crossroads. New approaches to miniaturize components and to move information between different parts of a computer chip at high speeds and low energies are needed. We also have vast, unmet needs for cheap, fast, and dense memory storage that today's technology cannot meet.

Unless major innovations in microelectronics occur, advances in computing similar to those experienced over the past five decades will not be possible, with profound implications for U.S. national security, energy efficiency, and economic competitiveness. Equally worrisome, at this critical juncture, the need for further computing advances has never been greater. New computing workloads related to data analytics and artificial intelligence are ill-matched to the 70-year-old von Neumann computing model that we continue to use today, where power limitations restrain data movement, storage, and analysis.

Today, further computing advances are limited by the economics of chip fabrication and the physical limits on transistors, electrical interconnects, and memory elements at the nanoscale. If we are to evolve new generations of computing systems over the next decades that are both faster and more energy efficient, a complete reorganization of the science and technology underlying computing is needed.

Within the Department of Energy's (DOE's) mission, which includes a pivotal and historical role in the evolution of computing, the needs are critical as well. Advanced computing and simulation underpin all aspects of DOE missions in energy, the environment, and national security, requiring energy-efficient computing beyond the exascale range. Edge computing, low-power computing technologies, and computers optimized for artificial intelligence are key to next-generation scientific facilities for high-energy physics research, as well as neutron and x-ray facilities. Finally, microelectronics will also play a major role in reshaping the U.S. electricity grid: from its current state to one that is cleaner, more efficient, cyber-secure, and resilient to widespread events, both natural and manmade. Realizing this vision will require advancements in both system architecture and power electronics.

To enable continued advances in computing technologies, a fundamental rethinking is needed of the science behind the materials, synthesis and placement technologies, architectures, and algorithms. This cannot be modular and linear, as it has been in the past. Rather, these advances must be developed collectively, in a spirit of **co-design**, where each scientific discipline informs and engages the other to achieve orders of magnitude improvements in system-level performance.

To explore these challenges, the Office of Science convened a Basic Research Needs Workshop for Microelectronics in October 2018 and charged workshop participants to conduct a thorough assessment of the scientific issues associated with advanced microelectronics technologies for applications relevant to the DOE mission. The workshop examined research relevant to the extension of complementary metal oxide semiconductor (CMOS) and beyond CMOS technologies; however, topics of direct relevance to quantum information science and quantum computing were outside the scope of this workshop.

The workshop participants included 77 panelists, and an equivalent number of observers. Roughly half the panelists had expertise in the physical sciences, and half were computer scientists and computer engineers. Academic and national laboratory researchers each contributed about 40% of the participants, with industry participants constituting the remaining 20%.

The workshop participants identified five priority research directions that should form the basis for future DOE research in microelectronics. These priority research directions are highlighted below, along with a summary of the underlying critical challenges.

1. Flip the current paradigm: Define innovative material, device, and architecture requirements driven by applications, algorithms, and software

Key Questions: *How can we optimize and integrate across physical, logical, and communication and control hierarchies? How will system-level optimization enable directed materials/device discovery and innovation?*

Materials properties, microelectronic devices, architectures, and algorithms must be understood and designed from the atomistic to the systems level to address the critical technical challenges facing DOE in its missions of science, energy, and national security. The outcome of an “end-to-end co-design framework” will reshape high performance computing, data analytics, the electricity grid, and other computing and power intensive applications.

2. Revolutionize memory and data storage

Key Questions: *How do we link physics, materials, architectures, and algorithms to overcome current physical limits on access and retention times for memory and storage? What innovations will minimize data movement and reduce energy consumption by orders of magnitude?*

Memory technologies are critically important in all aspects of data acquisition, analysis, and storage, and have the potential to perform efficient computations within, or proximally close to, the memory element. We face fundamental tradeoffs between fast memory access, capacity, and data retention time, as well as key challenges in energy usage and heat dissipation. Meeting these challenges will require coordinated breakthroughs in materials, devices, computer architecture, and algorithms.

3. Reimagine information flow unconstrained by interconnects

Key Questions: *How can we minimize data movement while maximizing information transfer? What novel electronic/optical states of matter can be discovered and manipulated to design non-traditional interconnects at the atomic, micro, and macro scales?*

A co-design approach to developing novel interconnect architectures will enable seamless integration of large-scale, real-time computation with communications and sensing to dramatically improve data transfer rates, connectivity, and reconfigurability.

4. Redefine computing by leveraging unexploited physical phenomena

Key Questions: *What unexplored materials, phenomena, or alternative computing models could perform computation far more efficiently than today’s technology? How will these new systems be modeled and programmed?*

The capabilities of the prevailing model of computation, the von Neumann model, are increasingly constrained by the energy inefficiency of established device, interconnect, and architectural approaches. Understanding and using new computing models based on unexploited phenomena require a co-design approach spanning architectures and algorithms to physics, materials science, and new devices.

5. Reinvent the electricity grid through new materials, devices, and architectures

Key Question: *Using a co-design approach, how do we create novel devices based on new materials to enable revolutionary breakthroughs in the performance, reliability, and security of power conversion systems?*

Revolutionary advances in power electronics for the electricity grid will require the design, synthesis, understanding, processing, and integration of advanced semiconductors and magnetic and dielectric materials. Novel device, circuit, and thermal transport concepts will be developed to exploit the unique physical properties of these materials. Such energy-efficient power conversion systems are necessary to replace the century-old electricity grid with one appropriate for the 21st century. They could also be applicable to electric transportation and used in extreme environments such as accelerators and power generation facilities.

This page intentionally left blank.

1. Introduction

The cost of computing has declined approximately eight orders of magnitude over the past fifty years, enabling the digital revolution that has influenced almost every facet of human life, including energy, public health, national security, and business. All of science and engineering now depends upon advanced computing, from the design of molecules to the furthering of our understanding of the universe. This remarkable progress has been driven by the miniaturization of integrated circuit technology. The transistors on computer chips, based on the element silicon, have roughly halved in size every two years (this is known as Moore's Law), becoming faster, smaller, more energy efficient, and cheaper over the past five decades. Throughout these technological advances, the computing approach, the von Neumann model, has largely remained unchanged since its introduction 75 years ago. This computing model consists of an instruction and processing unit connected to a memory that contains both instructions and data. The delay in moving instructions and data to and from the memory for processing is called the "von Neumann bottleneck" and is increasingly a limit on the performance of computer systems for both scientific simulation and data intensive computing. This combination of silicon microelectronics technology and the von Neumann computing model forms the basis for almost all computing appliances in the world today.

Two recent developments have brought computing to a crossroads that will require extraordinary attention from the science and technology community. The first is the recent end of Moore's Law: the recognition that chip technologies, in their current manifestation, have miniaturized to the point where economics, physics, processing chemistries, and materials limitations prevent them from being made smaller or faster. As a result, we can no longer depend on the doubling in performance and density (at the same cost) in each generation of chip that we came to expect over the past few decades.

The second is the explosive growth of data and the emergence of artificial intelligence (AI) that aims at identifying patterns and making inferences from the large masses of exponentially growing data. This analysis is ill-suited to the sequential nature of the current von Neumann computing model that, while effective for high-precision calculations, consumes excessive energy and is wasteful in AI applications. New computing models and architectures beyond the von Neumann approach are needed.

The sophistication of today's silicon-based computing technology is embodied in exascale computing. The world's first exascale machines will be installed in the Department of Energy's (DOE's) national laboratories (Argonne, Lawrence Livermore, and Oak Ridge) within the next few years. Each of these will consume upwards of 40 megawatts of power. What lies beyond?

Continued progress is no longer a matter of incrementalism, but will require radical rethinking of the science and technology underlying computing. Transistors, the switches that process data, have neared economic and physical limitations. Devices and materials for storing data are not fast, cheap, or dense enough to keep up with today's needs. Electrical wires (called "interconnects"), which transport data between different parts of a computer chip, will not efficiently transport tomorrow's data rate loads that the von Neumann computing architecture demands.

There is a clear and pressing need and the opportunity to reshape computing as we know it, seeking to make it orders of magnitude more energy efficient and powerful than it is today. One possibility is computers that can be as powerful as an exascale machine, or can analyze massive amounts of data, yet consume watts instead of megawatts of power. Another possibility is computers that consume milliwatts of power, are centimeter sized, and perform powerful analytics on-the-fly as part of an internet-of-things (IoT) sensor network. *Realizing this vision calls for deep rethinking and broad investigation into the underlying science that shapes and drives computing, including computer science, computing architectures, physics, chemistry, and materials science.*

We need new, energy-efficient computer architectures that supplement the von Neumann approach. We need a deeper understanding of the physics underlying information transfer, processing, and storage to identify new ways of using electrical, optical, magnetic, and thermal excitations at nanometer scales to design the efficient computing hardware of the future. Equally importantly, to build the information processing engines of the future, we need to identify new materials and discover new ways of synthesizing, processing, characterizing, and configuring them at atomic length scales—all of which are simply beyond today’s capabilities. As we have learned from previous research, success requires the principle of **co-design**—the recognition that to achieve orders of magnitude improvement in system level performance, these different aspects of scientific investigation need to inform and guide each other synergistically.

In parallel with the need to develop a new energy-efficient computing paradigm, a pressing need exists to revolutionize the manner in which electrical power is generated, transmitted, and consumed. For example, the U.S. electrical grid must be transformed from its current state to one that is more efficient, is cyber-secure, can handle distributed sources of clean energy that may be intermittent, and is bi-directional and resilient to widespread events both natural and manmade. By 2030, 80% of all electricity will be handled by power conversion equipment, a substantial increase from today. However, to realize this vision for the future electrical grid, numerous advancements are needed in system architecture and power electronic devices. Much of the underlying fundamental science shares common ground with that needed for microelectronics for computing.

Power electronics for the grid can benefit from the same miniaturization strategy that has benefited computing—a Moore’s Law for power electronics. For instance, a 1000 kVA sub-station of the future could be suitcase-sized instead of room-sized, similar to the size reduction from the 1940s-era Electronic Numerical Integrator and Computer (ENIAC) to today’s desktop computers. To build these ultra-compact and efficient power converters of the future, we need new dielectric materials (for capacitors and transistors), new magnetic materials (for inductors), and advances in semiconductor materials and devices that can handle high voltages, frequencies, and currents. Further, as the size of power converters decreases, new science for thermal management is needed to handle and manipulate high power densities at device and even atomic length scales. Such strategies are also a need for electronics for computing.

Likewise, at the power system level, lessons learned from microelectronics can advance the state of the art. To develop an electrical grid that is both cyber-secure and resilient, an extensive ecosystem of sensors and their subsequent information streams must be balanced and optimized. System design tools that can efficiently lay out and model integrated power, communications, and controls on the network are needed. Such tools have been pioneered in the microelectronics industry, as computing has become more complex and chip performance has pushed the envelope of what is possible.

Developing the enabling foundational science for next-generation microelectronics and power electronics will also require us to develop ways to characterize properties of materials, devices, and systems that we are unable to measure today—at scales and resolutions that are currently beyond our grasp. How do very dissimilar materials in a device element behave as a function of time, for instance, when they are located within a few atomic jumps of one another, and are subject to highly non-equilibrium conditions that can be brought about by, say, a voltage pulse? Today we can speculate about these things, but accurate measurement under realistic conditions remains difficult. Such characterization techniques include, for instance, *in-operando* imaging/spectroscopy methods at high spatial and depth resolution and at temporal scales down to the ultra-fast (picoseconds) regime that can relate the performance of devices to materials phenomena at the atomic level. They will leverage the rapid progress in scientific capabilities being made available today, such as those available at the Department of Energy’s Office of Basic Energy Sciences (DOE-BES) user facilities across the nation. From a systems perspective they also include, for example, characterizing the performance of a system or application in terms of the behavior of the devices and circuits in a subsystem. This would enable rapid evaluation and testing of new devices in performance-critical components of a processor. Another desired capability that is beyond our current systems modeling technology is the ability to rapidly generate a hierarchical simulation framework that can accommodate a variety of heterogeneous computing elements and use that framework for testing and evaluating alternative approaches to systems design.

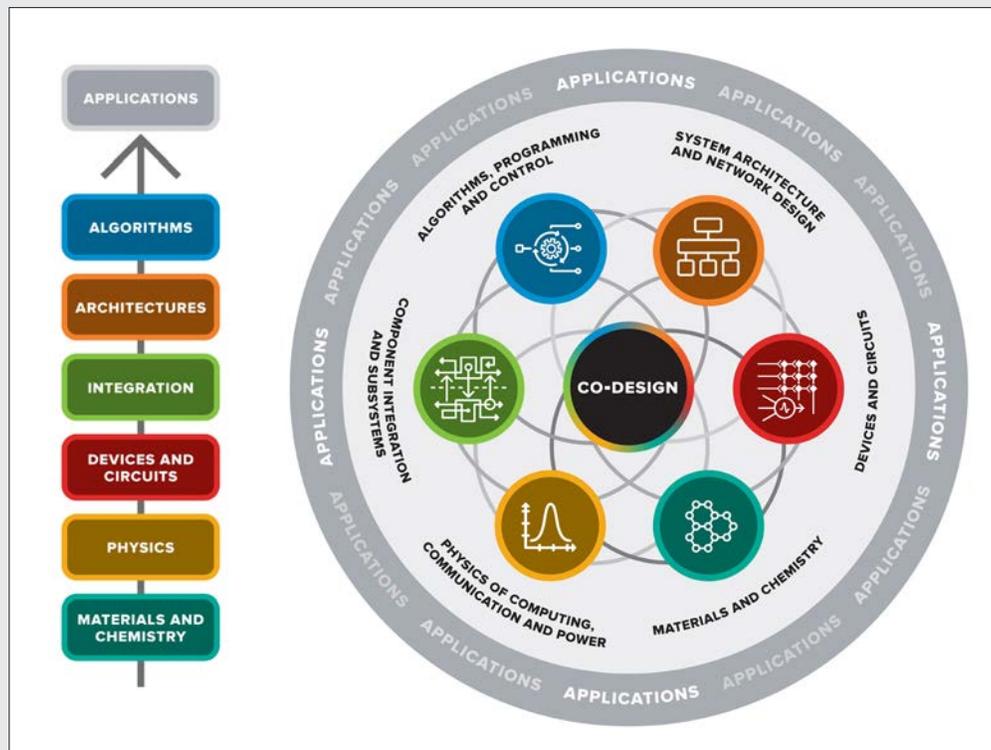
The consequences to the nation are enormous. DOE’s high-performance computing applications will extend our understanding of nature and human systems. New applications related to artificial intelligence and distributed computing such as those used in sensor networks for the IoT are poised to dramatically expand the worldwide semiconductor market, which is currently ~\$480 billion per year. The energy consumption in all of microelectronics is likely to top 20% of worldwide primary energy use. Large-scale scientific facilities, such as those related to high energy physics or the DOE’s various advanced x-ray and neutron source facilities, are undergoing major improvements in their scientific capabilities, which, in turn, place extensive needs for data and computing resources.

Addressing these challenges motivated the Basic Research Needs Workshop, which was held October 23-25, 2018, with 77 subject matter experts as panelists and 70 observers. Approximately 40% of the panelists were from academia, 40% from national laboratories, and 20% from industry. Roughly half of them had a computation systems/architecture background, and roughly half had a background in the physical sciences. The workshop participants were charged with identifying critical scientific challenges, fundamental research opportunities, and priority research directions that require further study as a foundation for future advances in microelectronics for computing, communications, and sensing, with particular emphasis on areas that are aligned with the missions and needs of the DOE Offices of Advanced Scientific Computing Research (ASCR), Basic Energy Sciences (BES), and High Energy Physics (HEP). This document summarizes the key findings of this workshop, including the identification of five priority research directions that are essential for progress in the fields of microelectronics and power electronics.*

* NOTE: Workshop participants were instructed to consider both extensions to CMOS and beyond-CMOS research directions; however, research directly targeting quantum information science and quantum computing were outside the scope of this workshop.

CO-DESIGN: THE FUTURE OF MICROELECTRONICS

The design, development, and manufacturing of present-day microelectronics technologies integrate contributions from many disciplines in a many-layered stack (left side of figure). Specialists in each layer are focused on models or abstractions that are built on, but largely independent of, the models and abstractions of other layers. Thus, there has not been a strong need for scientists and engineers working within each layer to intimately understand the challenges faced by those working below and above. Instead, information has been typically passed from layer to layer, up the stack. While extraordinarily successful for many decades, this model needs to change going forward.



As Moore's Law nears its end, a fundamental rethinking of the science behind the materials, devices, synthesis and fabrication technologies, architectures, and algorithms is needed to enable continued advances in computing, communication, and sensing technologies. This cannot be sequential, as it has been in the past. Rather, these advances must be conceived and developed collectively, in a spirit of **co-design**, where each scientific discipline informs and engages the others, with multi-directional information flow, to achieve orders of magnitude improvements in system-level performance (right side of figure). Materials scientists, chemists, device physicists and engineers, circuit designers, and micro-architects, on up to language, algorithm, and even application designers must work across the traditional layers of abstraction. This collaboration will be difficult because this independence of the various layers has allowed deep disciplinary expertise and enhanced our collective ability to reason about and build highly complex systems. Among the greatest challenges will be to establish a multi-disciplinary culture and language of information exchange while maintaining expertise and innovation in each of the necessary disciplines. In success, the result will be parallel but intimately networked efforts to create radically new capabilities that would not have resulted from the historical linear development process. These "innovation loops" will be driven by application, and the optimization of all loops will likely vary with application class.

2. Priority Research Directions

The workshop discussion identified five Priority Research Directions (PRDs) that define the basic research needs for microelectronics. Each PRD is discussed in depth with the associated research thrusts in this chapter. As background, Chapter 3 of the report provides an in-depth assessment of the current status of relevant research in the field of microelectronics.

LIST OF PRIORITY RESEARCH DIRECTIONS

- 1 Flip the current paradigm: Define innovative material, device, and architecture requirements driven by applications, algorithms, and software.**

Key Questions: *How can we optimize and integrate across physical, logical, and communication and control hierarchies? How will system-level optimization enable directed materials/device discovery and innovation?*

- 2 Revolutionize memory and data storage.**

Key Questions: *How do we link physics, materials, architectures, and algorithms to overcome current physical limits on access and retention times for memory and storage? What innovations will minimize data movement and reduce energy consumption by orders of magnitude?*

- 3 Reimagine information flow unconstrained by interconnects.**

Key Questions: *How can we minimize data movement while maximizing information transfer? What novel electronic/optical states of matter can be discovered and manipulated to design non-traditional interconnects at the atomic, micro, and macro scales?*

- 4 Redefine computing by leveraging unexploited physical phenomena.**

Key Questions: *What unexplored materials, phenomena, or alternative computing models could perform computation far more efficiently than today's technology? How will these new systems be modeled and programmed?*

- 5 Reinvent the electricity grid through new materials, devices, and architectures.**

Key Question: *Using a co-design approach, how do we create novel devices based on new materials to enable revolutionary breakthroughs in the performance, reliability, and security of power conversion systems?*

This page intentionally left blank.

PRD 1 Flip the current paradigm: Define innovative material, device, and architecture requirements driven by applications, algorithms, and software

INTRODUCTION

Half a decade after the first integrated circuit was invented, Gordon Moore observed that integrated circuit density was doubling every year and predicted this trend would continue. He made a similar observation about the reduction of the cost per transistor in an integrated circuit. His prediction proved to be accurate, and by the 1970s, industry evolved the observation into a goal that drove progress. By the 1980s, the technical challenges of continuing to shrink the size of transistors on an integrated circuit were becoming hugely expensive for companies acting on their own. In response, the microelectronics industry in conjunction with the U.S. government collaborated to sponsor pre-competitive academic research via the Semiconductor Research Consortium.¹ By the early 1990s industry, in the U.S. and internationally, collaborated to create technology roadmaps as a means to communicate to their suppliers and the research community the specific technical parameters required to maintain dimensional scaling. In turn, semiconductor materials and device researchers used these roadmaps to help guide their work. Similarly, computer architects used the roadmaps to predict what processor and memory performance would be available and made design decisions based upon what the semiconductor industry planned to deliver. The computer industry's interdependences were hierarchical, driven by advances in microelectronics performance.

By the early 2000s microelectronics scaling (i.e., Moore's law) was showing signs of economic and physical limitations.² Frequency scaling ended, voltage and dimensional scaling slowed, and manufacturing costs accelerated to the point that a new fabrication facility costs on the order of \$5B to build. At this point, many integrated circuit manufacturers either stopped scaling or sold their fabs and became fabless semiconductor companies. Those that continued to fabricate chips drew upon materials and device research advances at an accelerated pace. By 2015, the number of different materials required to make a state-of-the-art integrated circuit had tripled relative to 1995. The silicon metal-oxide-semiconductor field-effect transistor (MOSFET) was redesigned with a hafnium oxide-based dielectric (replacing silicon dioxide) and a metal gate electrode (instead of polycrystalline silicon). New device structures (for example, the fin field effect transistor) were introduced. The relationship among semiconductor process engineers, materials researchers, and device designers evolved to become a more cooperative partnership to address the technical challenges of continued scaling.

The demise of frequency scaling meant that future high-performance processors could not advance solely by increasing the chip clock speed. At this point, parallelism was extended to the chip level, and multicore processors became the norm. In turn, this advance created challenges for high-performance computer designers, who had already adopted node-level parallelism, as they now had to design both complex chip-to-chip interfaces and on-chip interfaces. This also created challenges for the system software and application developers to exploit both chip- and node-level parallelism. With respect to high-performance computer designers, this required increased interactions with integrated circuit companies so that high-performance computer architects could take advantage of multicore processors and potentially influence new designs. With respect to system software and application developers, this required increased interactions with high-performance computer architects to utilize the multicore processors efficiently.

The increasing complexity of advanced computer architectures required a concomitant increase in collaborations between computer architects and application developers. New benchmarks were developed that better reflect the resource demands of increasingly complex applications. Several abstract machine models (AMM) were developed that highlighted architectural aspects that were important or relevant to performance and code structure.³ The AMMs also served as a communication aid between application developers and computer architects during the design process.⁴

The aforementioned examples of improved communication and collaboration demonstrate the rising importance of “co-design,” where each of the technical abstraction layers in modern computer system design, from fundamental materials research through applications (Figure 1), inform and engage other abstraction layers. In each case, these co-design activities largely occur between adjacent technology abstraction layers (e.g., between materials and devices or computer architects and software designers). Because of the maturity of the field and the ability to accurately simulate or assess system-level as well as device-level performance, target performance metrics have been clear in traditional silicon microprocessor technology within each abstraction.

For instance, a materials scientist designing a hypothetical future “drop-in” replacement for a silicon CMOS (complementary metal oxide semiconductor) transistor had clear guidelines for this replacement that were derived through input from the higher technology abstractions at the software, architecture, and circuit layers. These methods have proven effective for conventional silicon microelectronics but are insufficient to meet future technology needs and to continue delivering increases in computing performance in a post-Moore’s law era.

To achieve orders of magnitude improvements in system-level performance in the post-Moore’s Law era, co-design must be employed throughout the technology abstraction hierarchy, not limited to adjacent interactions. While holistic co-design approaches are being defined, there is a significant opportunity to realize fundamentally different models of computing, such as non-von Neumann architectures, new devices, novel materials, and new algorithms. A top-down hierarchical design view of specific non-von Neumann architectures should dictate the performance benchmarks of new materials and devices, and vice versa. A bottom-up view of how specific devices and their performance will impact performance at the system level does not exist today for most non-von Neumann approaches at a satisfactory level of detail and depth. In addition, to continue to enhance future computer system performance with time, we need to redesign the innovation process, replacing a hierarchical approach with a collaborative, co-design methodology.

A redesign of the innovation process calls for a holistic co-design framework in which the entire computing ecosystem is co-optimized. This includes application and algorithm requirements, the system software stack, chip /systems-level architecture, circuit designs, device physics, and materials integration. This notional framework is illustrated in Figure 2. We must replace the hierarchical approach driven by Moore’s Law and scaling of silicon CMOS device hardware with co-design collaborations among software developers, computer architects, circuit designers, device physicists, materials scientists, and chemists to guide their R&D strategy. Below, we provide illustrative examples of how researchers working on various elements of computing technology can engage in holistic co-design:



Figure 1. Traditional technology abstraction layers hierarchy

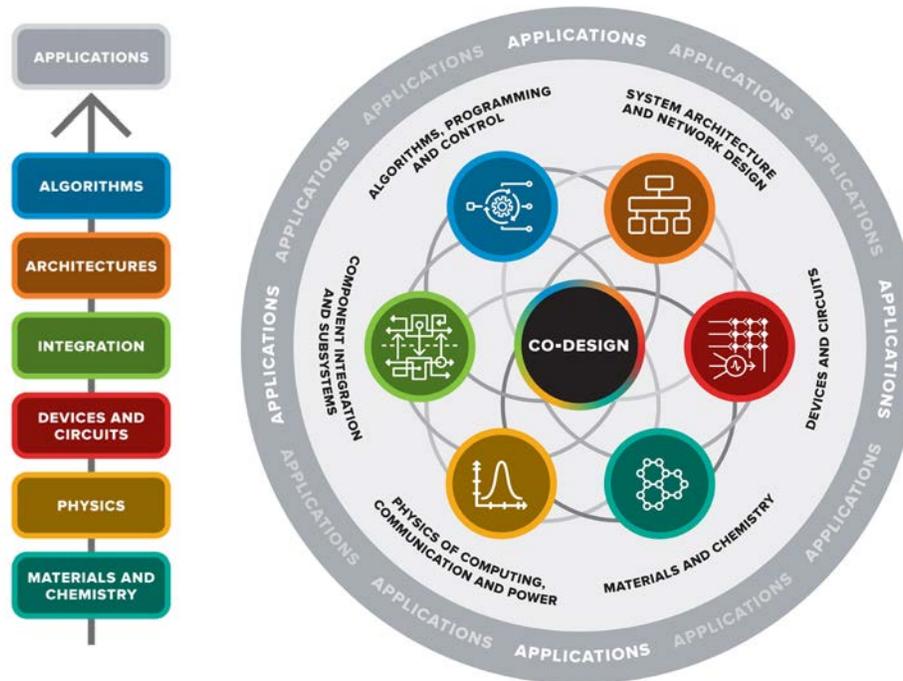


Figure 2. Co-design framework: From the traditional hierarchy of abstraction levels (left) to a holistic system framework (right)

- **Materials Scientists and Chemists.** These individuals can conduct R&D to enable a more agile synthesis flow and can collaborate with computer and system architects to develop a more nimble architectural design process. They can also create new materials models that will facilitate device, circuit, and manufacturing process design. These materials science and chemistry capabilities include computational and combinatorial materials discovery and device simulation, materials synthesis and characterization, process integration, and transition of laboratory-scale materials synthesis to industrial-scale manufacturing production. There are also key collaborations with chemical and process engineers at microelectronics foundries that are relevant to large-scale production.
- **Device Designers.** These individuals can collaborate with materials scientists to create new devices to replace existing transistor, memory, interconnect, and other fundamental devices. They also can collaborate with circuit designers and process engineers to assure that circuits can exploit the new device technology, and that the devices are manufacturable at reasonable cost. Architects can engage to make sure they understand the advantages and limitations of the new device technology, and provide feedback on opportunities for improvement. Also, software developers can engage to make sure they understand how the architects may use the new device technology, and provide feedback in terms of efficient software development.
- **Computer and System Architecture Designers.** These individuals can help develop a multi-lateral co-design framework that will support methods for synthesizing new architectures to optimize performance and power usage in the context of new materials/devices, application profiles, and system software. Optimization of computer and system architectures for a given set of materials and devices is a fundamental step needed to optimize and possibly automate the overall co-design process. The computing community recognizes the need to systematically develop new systems and architectures that can rapidly incorporate the advances of new-material enhanced devices for computing or memory and translate these into realized system performance for applications.
- **Power Grid Engineers.** While this PRD is focused on computing, the holistic co-design approach to fostering multi-disciplinary collaborations is also relevant to co-design from the future power grid to high power devices to fundamental materials science. Whereas power grid engineers will be engaged in PRD 5, their collaborations with device physicists and materials scientists are equally relevant. These individuals can develop fundamental understanding of how to dynamically control grid behavior (via power electronics

controls or adaptive grid topologies) through a variety of grid situations (bi-directional power flow, dynamics, cyber-attack, natural disaster, etc.) through the use of real-time analysis of communications, power flow, and other information streams. These individuals can also develop a framework to utilize behind-the-meter information from (i) “internet of things” connected devices for bulk-grid operations, (ii) electronic design automation (EDA)-like tools that can autonomously optimize grid design or operations through the use of compact models of lower hierarchical layers (e.g., power converters and generators), and (iii) use of a co-design hierarchy that can predict system- and circuit-level operational benefits from material/device design parameters as well as inform the direction of materials/device research based on system-design criteria.

- **Application and Algorithm Developers.** These individuals can connect microelectronics capabilities to application goals, including Department of Energy (DOE) goals for scientific discovery and power grid systems engineering. This co-design framework will enable application and algorithm developers to evaluate technology alternatives in the context of system and application performance, and to extend the experimental systems to provide accurate bounds on the uncertainty of the model predictions.

For many decades, much of the rapid progress in computing has come from exponentially compounding miniaturization of well-established silicon CMOS devices, supported by successive introduction of new or improved materials and integration processes. In the traditional, hierarchical model depicted on the left-hand side of Figure 2, materials scientists worked primarily with device physicists and engineers, confident that advances in materials and devices would translate to improved capabilities of circuits and subsystems.⁵ During these many decades, software developers benefited from a stable, enduring AMM. This approach has reached economic and technological limits as new materials, devices, algorithms, computing models, and architectures are required simultaneously if we are to further increase the performance of computation, subject to unit volume, power consumption, and R&D and fab investment.

The holistic co-design framework depicted on the right-hand side of Figure 2 is a multi-lateral approach to system optimization, where application and algorithm requirements can help define innovations from the systems and architecture perspective, which, in turn, can lead to the definition of target metrics for new circuits, devices, and materials. The anticipated benefit from this co-design approach is orders of magnitude improvements in energy-efficient computing performance. The guidelines and metrics resulting from this approach can be used by microelectronics scientists in research on new materials and devices. In turn, the new capabilities derived from new materials and physics process innovations can inform systems and software architects, offering them additional design space for new innovations and capabilities.

Cost has always been a driver in the evolution of microelectronics technology. Leading-edge semiconductor fabrication facilities now cost over \$5B. As a result, only three companies world-wide have invested the capital required to develop leading-edge technology. Therefore, the microelectronics community must consider the cost of manufacturing and other cost drivers (e.g., the cost of creating and maintaining EDA tools) when identifying areas for potential R&D investment.

The holistic exchange of information among co-design elements to enable global performance improvements extends beyond microelectronics and computing. One example relevant to DOE is the design of next-generation electrical power grids. (See also the discussion in PRD 5.) To optimize the electric grid, co-design is needed of a wide variety of flows (e.g., communications and power flows) and technology capabilities (e.g., compact models for power systems, devices, and materials). To ensure global system optimization, tools must be developed to allow system designers to seamlessly and quantitatively understand design tradeoffs across all levels.

SCIENTIFIC CHALLENGES

The well-established silicon CMOS technology is no longer scaling according to Moore’s law and automatically providing increased computational performance for each generation of chip design, but as new materials, devices, computing models, and chip and computer architectures and algorithms are invented the ability to increase application performance may be possible through the holistic utilization of co-design. The technical challenge of this PRD is to develop new holistic co-design capabilities to support the development of computer and system architectures and associated designs for microelectronics components and application-specific integrated circuits driven by the requirements for future high performance computing and smart grid applications.

One key microelectronics goal is to continue the rate of increase of application performance, normalized by power consumption and chip fabrication costs. For the electric power grid community, a key goal is to make the future smart grid adaptive to the bidirectional flow of power and information. These computing and power grid goals will need to be addressed by engineering the optimization of multi-disciplinary, holistic design systems while considering economics and power consumption.

In both systems engineering scenarios, a co-design framework that integrates physical elements, logical elements, controls, and software elements is needed. This framework will enable scientists and engineers to develop unified EDA simulation tools for materials, devices, circuits, system architecture, software, and applications that will achieve improved performance, efficiency, and resilience, as well as reduced design and development time for mission-critical systems.

To date, DOE has pursued co-design activities with a primary focus on collaborations involving application-high performance computing (HPC) architecture, rather than multilateral collaborations of the type described earlier. For example, an application developer may have a performance profile of the application code that indicates how much time is spent in various functions as represented in the source code. Modeling can translate these functional operations into energy/delay profiles at the abstraction level of the source code on given hardware. Translating these high-level application/source code metrics into lower level materials/device metrics would require at least two additional co-design collaborations among the microelectronics technology elements in Figure 2 that do not exist today. This PRD emphasizes new R&D support and opportunities for DOE to pursue performance modeling, characterization, and simulation of new devices and circuits targeting new or enhanced computing architectures that can deliver significant capabilities to future computing systems. These mid-level performance modeling and simulation co-design tools (from devices to processors) could also be leveraged by power grid or sensor/detector hardware R&D teams to accelerate microelectronics innovation in those areas as well.

Co-design collaboration requires a model that relates the application functional/energy profile to the given architecture in a way that apportions the time/power behavior to the micro-architectural details and device limits of the specific system. Such models may integrate fine-grained power and performance measurements at the circuit/subsystem level with fine-grained architectural simulators, assessments of the combined structure with application profile traces, and mapping from the source code level to that of the physical architecture and circuits. Such a co-design loop would involve collaborations among all of the technology elements in Figure 2.

Co-design collaboration also relates performance characteristics for the materials and device physics to circuit and system architecture performance for applications. A goal of this PRD would be estimating how changes in a materials parameter would affect change in an application metric and *vice versa*, assuming an effective mapping between the application and the materials. Computationally, this goal could be accomplished by estimating the partial derivatives of the application performance metrics in terms of the materials parameters. A major science challenge for this PRD is enabling this multi-lateral flow of information and sensitivities through a collection of integrated adjoint simulators. Below, we provide several co-design examples, each involving three to four elements, to further illustrate the needed co-design collaborations.

Co-design Example: Applications – Algorithms – System Software – Computer and System Architectures

When developing purpose-designed hardware and system architectures, it is important to include the requirements and needs of DOE's applications, algorithms, and system software users and developers. Application drivers would expand from DOE's Office of Science traditional modeling and simulation applications to artificial intelligence (AI)-enabled science applications that integrate large-scale simulations, data-driven predictive models, experiments, and theory. It is very likely that a re-engineered, "co-designed" software environment may be necessary to efficiently map DOE applications and algorithms to the successive generations of purpose-designed hardware and system architectures. Cost is a persistent consideration, including weighing the cost to design/build purpose-designed hardware versus the cost of application development for new generations of commodity hardware.

The co-design loop (involving applications, algorithms, system software, and purpose-designed hardware) will likely require a number of design iterations to yield a purpose-designed hardware that explicitly integrates requirements from users and developers of DOE applications, algorithms, and system software stacks. The advanced computing and smart grid design space will require the exploration of many such co-design collaborations.

An associated science challenge is to develop an “adjoint” version of an end-to-end application tailored to device simulators. Although this approach could estimate sensitivity of applications performance to materials properties, it would not account for the inevitable changes in system architectures that would likely occur given advances in materials and devices. Rather, one must concurrently optimize system architectures, subject to power, reliability, performance, and cost constraints, with optimization points for new materials and new devices. This would require advances in automating design choices such as some form of generative models for architectures. Optimization points would permit a kind of searching for new architectures that would also include data analytics and scientific machine learning applications.

For illustrative purposes, below are some examples of utilization of hierarchical co-design that reaches beyond the traditional model.

Co-design Example: *Computer & System Architectures – Circuits – Low Voltage Devices and Enabling Materials – Chemistry and Processes*

Computer and system architectures include many possible design points, including (i) conventional von Neumann architecture compute nodes that enable the use of existing software code bases, (ii) purpose-designed machine learning and data analytics architectures that are of interest in large-scale sensing networks, experiments, and AI-enabled science, and (iii) heterogeneous application drivers that include HPC and combinations of (i) and (ii). Central to any co-design process is identifying materials and devices, along with specific performance considerations, that result in energy-efficient system architectures, while emphasizing the dependencies among system considerations and the physics and materials. This approach highlights the need for a science-based approach to the co-design of materials for energy-efficient devices, the chemistry to synthesize and produce the needed materials, and new device and circuit designs that can support the target design computer and system architecture.

Co-design Example: *Real-Time Control Applications/Algorithms – Real-Time System Software – Distributed Computing and Communication Integrated into Smart Grid System Architectures*

The current paradigm for the electrical grid was developed in the late 1800s and is predicated on the unidirectional flow of power from large, centralized synchronous generators to the decentralized loads of customers. Two main developments have shifted the traditional paradigm of the electrical grid. First, the continuing exponential decrease in the price of stochastically distributed energy resources (e.g., wind and solar) has resulted in bi-directional power flows within the distribution system, the lowering of system inertia due to displacement of traditional synchronous generation, and more complex transient behavior due to stochastic changes in power generation output. Second, as illustrated in Figure 3, the explosion of microelectronic and power electronics devices has drastically increased the number of electronic loads on the electrical system, increasing harmonics and changing traditional load behaviors. These changes

require complex control schemes that operate on shorter time scales than have been traditionally used and may require significant alterations to the grid structure itself to fully optimize system performance, reliability,

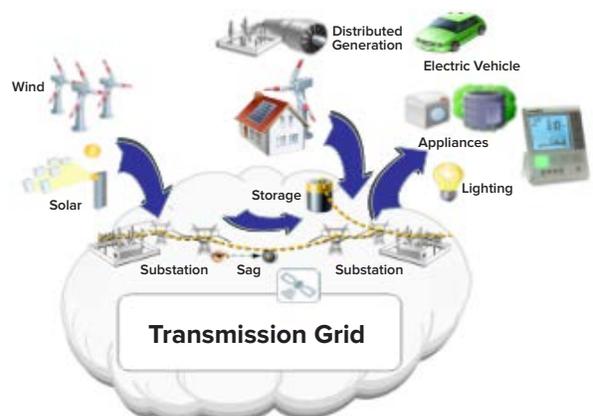


Figure 3. Schematic of future smart grid system architecture with stochastically distributed energy generation and new microelectronic and power electronic loads. Courtesy of Henry Huang, Pacific Northwest National Laboratory. Images from NIST Smart Grid Framework and Roadmap for Interoperability Standards (Release 3.0), NIST Special Publication 1108r3, September 2014.

and resiliency. The introduction of solid-state power conversion devices would support the opportunity to change grid design and control. More details of the future power grid are provided in PRD 5.

A co-design framework would allow grid designers to optimize the grid's systemic behavior, either through controls or future planning. To enable controls that can respond to the dynamic conditions inherent in a grid with a high percentage of renewable energy sources, real-time data must be combined with models (e.g., from AI and machine learning) for the power electronic converters to change operating behaviors or dynamically re-adjust the grid architecture during transients or abnormal operating events.

Additionally, grid planning tools are needed to characterize the behavior of distributed energy resources. Analogous to EDA in microelectronics, grid planners looking to design the system require compact models of generators, power converters, and protection elements that can be co-designed with the global system to ensure robust operation under a variety of operating conditions.

Co-design Example: *Smart Grid System Architectures – Circuits – Devices – Chemistries – High Power Electronics Materials*

Wide bandgap materials have received significant attention over the past 15 years for power electronics applications. These materials enable development of semiconductor devices that could provide higher voltage operation and higher frequency operation. Both of these advantages manifest directly in circuit topologies and systems. High-voltage hold-off allows development of higher voltage-rated devices, which simplify circuits and reduce costs for electric grid applications. This enables cost-effective production of grid-level power electronics circuits and can enable new high voltage components on a future smart electrical grid (e.g., cost-effective, high-voltage DC converters) that would allow continent-spanning grids to function effectively. Operating devices at higher frequencies would reduce the use of expensive, heavy, passive components (i.e., capacitors and inductors). In turn, this could reduce the size and weight of the corresponding circuits, which would enable new technologies in consumer electronics, mobile platforms, and the electrical grid by reducing the costs for transporting and installing power electronics systems.

Understanding what system-level benefits are derived from materials-level decisions (and vice versa) requires a complete co-design environment that couples all elements, from system to circuit to device to materials. Again, a more detailed discussion of the future power grid is in PRD 5.

Co-design Example: *Smart Sensors and Experimental Diagnostics – Materials – Devices and Circuits – Component Integration – Algorithms, Programming, and Control*

The DOE experimental facilities have a long history of developing custom- and purpose-designed instrumentation. The concept of “flipping the paradigm” is not actually relevant for this community of scientists and engineers, as a co-design methodology is typically used to develop custom instrumentation, diagnostics, and sensors to support a specific application — capturing experimental measurements. Such custom instrumentation is traditionally bound to DOE experimental facilities, and new challenges exist for how to design/optimize a system for analysis of streaming data from the Large Synoptic Survey Telescope, Dark Energy Science Collaboration, European X-Ray Free-Electron Laser Facility, as well as coherent synchrotron, nuclear physics, and high-energy physics events.

This co-design example also has an opportunity to leverage the significant investments that the commercial community is making in both the Internet of Things (IoT) and machine learning/deep learning (see sidebar). A key opportunity for the IoT involves the use of embedding computing at the edge to integrate processing capabilities into sensors that support local processing and filtering of measurements. Some of this processing could include low-power, inference engines that are trained by deep neural networks that run on HPC or cloud computing servers. This example of microelectronics co-design will leverage industry focus on low-power node analytics. The resulting smart sensors can enable edge computing and sensor networks that are distributed “in the field,” enabling new types of experiments that are no longer facility-bound. Of course, the opportunity to co-design hardware architectures for new inference engines and for machine- and deep-learning neural network accelerators is a priority for many companies.

SEMICONDUCTOR HARDWARE FOR IOT APPLICATIONS

There has been a surge of recent interest in semiconductor hardware for IoT applications involving artificial intelligence. Most such applications involve low-power computing at the edge for inference applications. These activities have relevance to sensor networks, image sensors, and autonomous vehicular transport. Figure 5 highlights the rapid growth of funding for semiconductor hardware for AI. There are roughly 70 companies in the U.S. today involved in building processors for AI applications at the edge. Almost all of these applications use existing silicon CMOS processes and capabilities, and the innovations primarily involve circuit design approaches for silicon CMOS, such as power staging, sub-threshold operation of transistors, and use of memory. The challenges have been in efficient computing and the ability to use maximal amounts of onboard memory in ultra-energy efficient ways to stay within power limitations for edge applications. While significant progress has been made, we remain limited in the capabilities of these edge computing appliances—for instance, most of these applications are limited to inference and classification at the node, while the training and model selection are done on the cloud or remotely. The technology limitations also hinder the size of the neural networks that can be run, affecting the granularity of the inference.

Clearly, there are major needs for fundamental innovation. This is an instance of the need for application-driven algorithm-system-hardware co-design that includes the development of new memories, power delivery, and routing approaches, new ways for interconnectivity, and new ways of computing within memory that overcome the limitations of today’s in-memory schemes. Overall, we need to transcend the limitations of silicon CMOS described above if we are to fulfil the promise of low-power computing at the edge. In particular, such applications have significant impact for the large-scale experimentation with the DOE user facilities and high-energy physics applications where processing at the edge with AI is paramount to reduce the amount of data that needs to be transmitted long distance. These can be ideal testing grounds for such technologies since they can withstand some flexibility in cost and possibly power. In the end, however, they have to be part of standard future foundry and design processes to be practical.

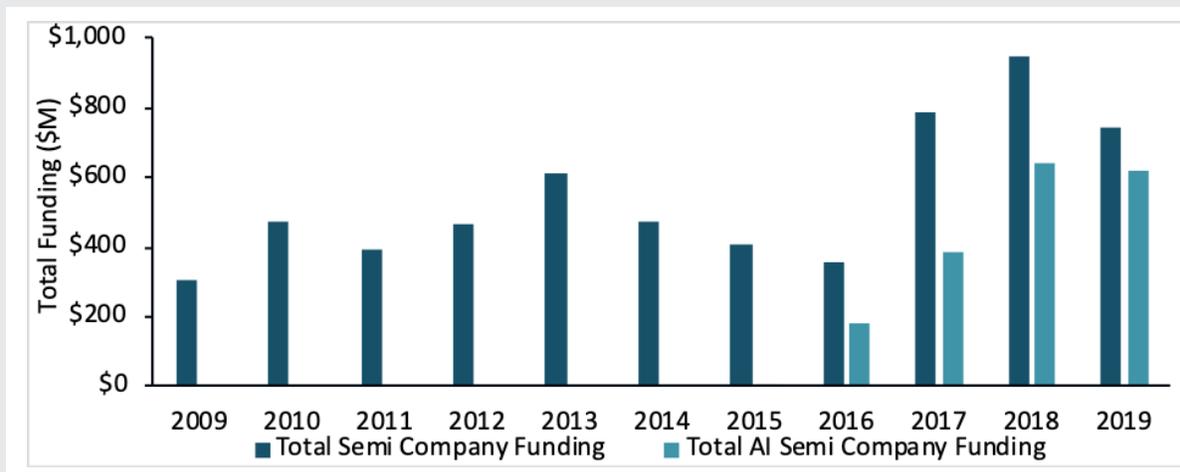


Figure 5. Chart showing the increase in equity financing by semiconductor companies involved in hardware for artificial intelligence. Courtesy of Rudy Burger, Woodside Capital. Conveyed through Jeff Bier, Embedded Vision Alliance. Data for 2019 only covers March 2019. Note how growth is already comparable to 2018 investment.

RESEARCH THRUSTS

Co-design among the different elements of Figure 2 presents an important opportunity to pursue different approaches that can result in orders-of-magnitude performance improvements. The requisite research includes identifying capabilities and mechanisms for communication and collaborations among the traditional abstraction elements. For example, proxy applications help non-application developers understand the technical challenges and bottlenecks that exist in DOE applications. The computer and system architecture community has an analogous communication vehicle in AMMs, a simplified representation of the computer architecture that can be used by software developers. Figure 4 illustrates the role of proxy applications and AMMs in the co-design interactions between applications and node and system architectures.⁶

Co-design capabilities analogous to proxy applications and AMMs are needed to facilitate communication and collaborations among specific technology domain experts and domain experts in other technology abstraction elements. For example, materials scientists can develop and define simplified or abstracted materials models that can be used by device physicists to guide their device designs. These device physicists can, in turn, develop and define abstracted device models that can be used to communicate device design tradeoffs to either materials scientists or circuit designers. The research to develop a portfolio of co-design capabilities such as these will facilitate co-design collaborations and discussions among subject matter experts on a technology element with counterparts on other elements.

Thrust 1. Identify Key Capabilities and Metrics per Element

For each abstraction element in Figure 2, it is important to identify the capabilities and key performance metrics, especially those that can impact or are impacted by the other elements. A starting point for this thrust is to consider the capabilities and metrics that impact other elements of the traditional stack and to broaden to consider other elements. For example, for the application element, some capabilities include major functions, and metrics include execution time models, power requirements, memory footprint, dataflow between functions, etc. Models and simulators can be developed for these capabilities and metrics to relate to algorithms, such that different methods for implementing the given functions can be explored. The methods could include different numerical solvers, surrogate models, and machine learning methods. For the device and circuits element, capabilities could include various output functions, performance characteristics and limitations, voltage requirements, and manufacturing cost. For this thrust, it is important to be very broad in the capabilities and metrics for the different elements.

This thrust may involve further element-specific research to further define the key capabilities. One example is the exploration of bi-stable (or even multi-stable) states in functional materials, which can be manipulated with applied voltages as small as 100 mV or lower. All else being equal, energy dissipation in digital switching is proportional to the square of the voltage. Hence, successful development of such low-voltage devices could increase the energy efficiency of computing systems by one or more orders of magnitude. Success will require a precise definition of the electronic structure as well as knowledge of how the relevant phenomena (magnetic, ferroelectric, charge correlation, optical, and chemical) can be manipulated at such energy scales. It will also require a systematic (computational optimization and/or high-throughput combinatorial experimental search) discovery of new materials that enable such performance. Significant research is needed to establish fundamental limits of the energy/length/time scales of these phenomena, as these will directly impact device- and circuit-level attributes such as power consumption, latency, and speed. In many cases, establishing these limits will require the use of state-of-the-art atomic-level simulations and the state-of-the-art experimental probes available at DOE user facilities.

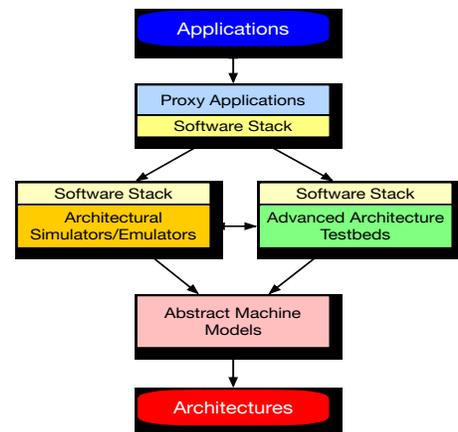


Figure 4. Schematic of the DOE HPC communities' traditional view of co-design, showing the role of proxy applications, testbeds, simulators, and abstract machine models.

Thrust 2. Develop All-to-All Relationships among Abstraction Elements

A co-design effort that spans all of the traditional abstraction elements of Figure 2 must develop relationships among all the elements. One needs to extend the capabilities and metrics to models, possibly simulators or parameterized models, that can be used to communicate co-design optimization opportunities. This thrust addresses the use of parameterized models for device physics that can be used to communicate co-design collaboration opportunities with materials scientists or software developers. It also addresses application and algorithm needs such as the ease of writing and porting software to a given architecture to determine the performance characteristics of devices and materials. If the paradigm has truly been flipped, the underlying application and algorithm needs will be directly supported in purpose-designed architectures and the supporting system software stack. The intent is to shift application development effort from porting legacy code to enhancing applications with new capabilities, such as scientific machine learning. Addressing these issues requires an end-to-end design exercise where new science impact in device physics and materials is guided by, and related to, system level needs by considering the various dependencies related to computing, memory, communication, workload, footprint, cost to manufacture, or code and energy consumption.

As an example, spin-based logic switches offer the potential for extremely low energy switching. However, they are also slow; consequently, they cannot be used to directly replace silicon transistors. One could, however, explore architectural and circuit innovations for highly distributed low-power sensor networks that could utilize spin-based devices. In such a model, what are the requirements on the devices and the materials in terms of power and performance?

The development of vector matrix multipliers using non-silicon hardware is a second example. A third example is that of architectural designs for digital or analog (or mixed) circuits that take advantage of in-memory computing approaches and three-dimensional chip architectures of the future. How might these applications guide discovery of new materials for memory, signal transmission, switching, and thermal management to microelectronics researchers?

Thrust 3. Provide Hardware Design and EDA Tools and an Analysis Framework

Research is needed to develop a new generation of open-source hardware design and EDA tools that are readily adaptable to a variety of application spaces, both for HPC and the electrical grid. The primary goal of these tools would be to incorporate the results from the first two research thrusts on capabilities and metrics and relationships, such that one can explore the impact of changes at one element with respect to other elements. To encourage and support open innovation, the hardware design tools should target open system-on-chip ecosystems. For this research thrust, lessons learned from the open source software community are relevant.

SCIENTIFIC AND TECHNOLOGY IMPACT

This vision for co-design leads to the development of architecture concepts that are defined to meet application requirements. These architecture concepts, in turn, yield requirements for co-design collaborations with circuit and device designers and, ultimately, materials scientists, physicists, and chemists. Application and algorithm developers no longer just respond to changes in new architectures. Instead, the costs of software development (i.e., development of applications, algorithms, and software stacks) are factored into the design choices that computer and system architects make, because these software developers have multi-lateral input into the design of these hardware architectures. In addition, the costs of hardware development are also included.

With the end of Moore's Law, a significant number of microelectronics materials scientists, device physicists, and circuit designers will no longer be driven by a scaling-dominated research viewpoint. Instead, using a holistic co-design approach, they will collaborate with computing application, system software, computer architects, and software developers and system architects for power grid communication and control. As a result, technology advances will not be paced by the arrival of the next reduced transistor feature size; rather, new generations of processor designs will be paced by the multi-lateral development of new designs for microelectronics and computing component design. This change will also drive the development of new EDA tools for accelerating the development of microelectronics designs.

More importantly, development of the proposed microelectronics co-design framework will significantly enhance the DOE research capability to attack future challenges faced by the nation and enable new research programs that are otherwise impossible to do. With the end of Moore's Law, to continue to improve performance, reduce power, and improve application capabilities, we must increase the level of innovation and the velocity in incorporating new advances in materials and devices into computing and the power grid.

REFERENCES

1. Semiconductor Research Corp., *SRC 1984 Annual Report*, <https://www.src.org/about/corporate-annual/1984.pdf> (1984).
2. C. Mack, *The Multiple Lives of Moore's Law*, IEEE Spectrum, <https://ieeexplore.ieee.org/document/7065415> (April 2015).
3. J.A. Ang, R.F. Barrett, R.E. Benner, D. Burke, C. Chan, J. Cook, C.S. Daley, D. Donofrio, S.D. Hammond, K.S. Hemmert, R.J. Hoekstra, K. Ibrahim, S.M. Kelly, H. Le, V.J. Leung, G. Michelogiannakis, D.R. Resnick, A.F. Rodrigues, J. Shalf, D. Stark, D. Unat, N.J. Wright, and G.R. Voskuilen, *Abstract Machine Models and Proxy Architectures for Exascale Computing v2.0*, SAND2016-6049, <https://cfwebprod.sandia.gov/cfdocs/CompResearch/docs/CALAMMv2.0.pdf> (June 2016).
4. R. Bair, J. Cook, D. Donofrio, J. Kuehn, and S. Moore, *Hardware Evaluation Outreach: Application Development Challenges Now and for the Exascale Era*, SAND2019-4136R, https://cfwebprod.sandia.gov/cfdocs/CompResearch/docs/AMM_ProgCentric_2019_04.pdf (April 2019).
5. F.B. McCormick, J. Shalf, A. Mitchell, A.L. Lentine, and M. Marinella, *DOE Big Idea Summit III: Solving the Information Technology Energy Challenge Beyond Moore's Law: A New Path to Scaling*, SAND2018-2328R, <https://www.osti.gov/biblio/1426401-doe-big-idea-summit-iii-solving-information-technology-challenge-beyond-moore-law-new-path-scaling> (April 2016).
6. J.A. Ang, T.T. Hoang, S.M. Kelly, A. McPherson, and R. Neely, *Advanced Simulation and Computing: Co-design Strategy*, SAND2015- 9821R, <https://www.osti.gov/biblio/1226118> (November 2015).

This page intentionally left blank.

PRD 2 Revolutionize memory and data storage

INTRODUCTION

Memory technologies are critically important in all aspects of data acquisition, analysis, and storage, and have the potential to perform efficient computations within, or proximally close to, the memory element. There is a rapidly increasing demand for cheap and fast memory due to the rise in data-intensive workloads worldwide. Within DOE, the anticipated expansion of its scientific capabilities will also require extensive data processing capabilities. For example, the ~10-year projected data handling needs for the Basic Energy Sciences (BES) light sources are expected to grow from less than 10 petabytes (PB) today to nearly 700 PB, and the disk storage needs of the Compact Muon Solenoid (CMS) at the Large Hadron Collider are expected to rise from less than 5 PB today to over 3 exabytes (EB) (see Figure 1). Figure 2 presents the projected near-term rise of storage requirements from 100 PB today to over 5 EB in the coming 10 years for the ATLAS (A Toroidal LHC ApparatuS) experiment at the Large Hadron Collider.

At the same time, we face fundamental tradeoffs among memory access latency, capacity, bandwidth, cycling endurance, and data retention time, as well as key challenges in energy consumption and power dissipation. Today's computational needs increasingly demand a closely packed and on-demand exchange of data between compute and memory elements—leading to concepts of compute-in-memory or memory-in-compute paradigms. However, the historical delineation between compute and memory blocks has led to physical and architectural designs that pose a significant barrier to implementing such novel computing models. Meeting these challenges will require coordinated breakthroughs in materials, device design, computer architecture, and algorithms.

How efficiently data can be processed in today's computing systems depends on the locality of data. Cache memories provide the fastest access time, and are accessed and reused frequently.¹ However, their relatively small size leads to what is known as the von-Neumann bottleneck. For example, in a typical machine learning task, during the training phase, one needs to continuously update the weight matrix for the data. This updating requires the machine to store the initial value. However, for almost all practical problems, the static random-access memory (SRAM) array, which is the on-die cache memory, is not large enough to hold both the data over which the learning is performed and the weight matrix. This means that the processing is performed on a small portion of the data. Subsequently, the processed data are sent back to the dynamic random-access memory (DRAM) and saved. Next, a new batch of data is brought back into the SRAM, and a new processing cycle begins. This back and forth between the central processing unit (CPU) and the DRAM, which is not on the same die, comes with significant penalty of latency and power dissipation incurred from moving data back and forth through the communication bus. As a result, this is one of the most significant technological barriers in designing computing systems for data centric applications, such as those related to artificial intelligence.

Shuttling data across the various levels of memory hierarchy leads to enormous energy dissipation. This becomes a critical roadblock for large scale simulations and big data applications. For example, future high energy physics experiments such as the High-Luminosity Large Hadron Collider (HL-LHC), may require data processing rates of 10 Pb/s at or in proximity to the detector. If a nominal energy dissipation of 1 pJ/bit is assumed to be the energy cost of communication between the computing blocks and the memory systems, the required power for just data transport becomes ~90 kW, which is unacceptable under circumstances where a fixed amount of energy is available and anything used on memory and processing is not available to the detector.

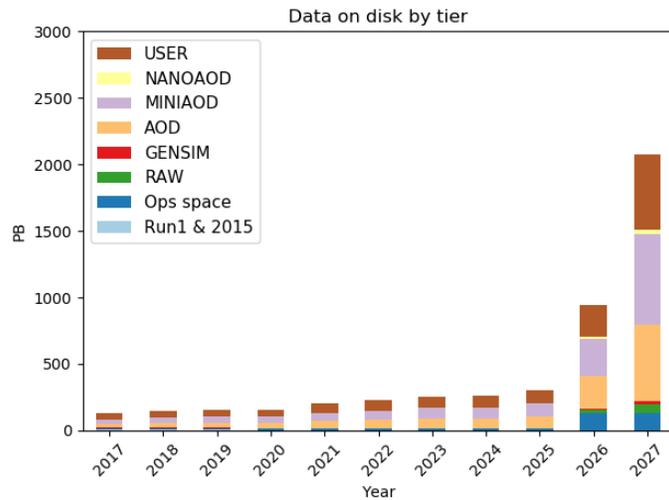
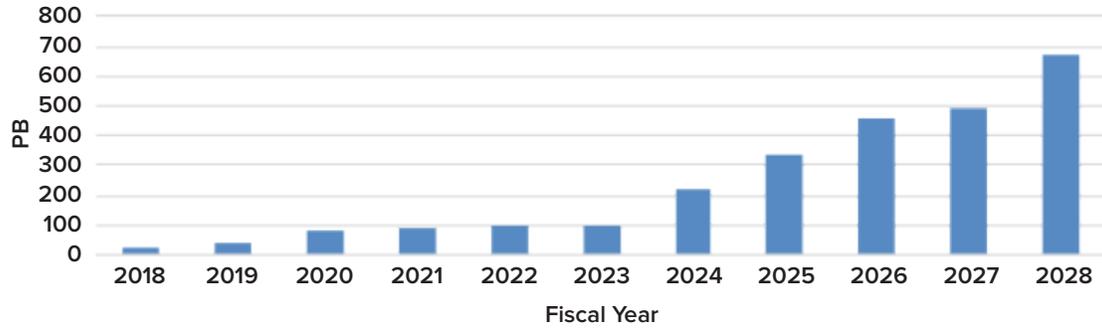


Figure 1. (a) Projected aggregate data generation rates at BES light sources (aggregated together). From data call for *DOE-BES User Facilities Data Management and Analysis Resource Needs in Advanced Scientific Computing Research*. (b) Projected data storage needs from CMS experimental facility at Large Hadron Collider. Courtesy of David Lange, Princeton University and CERN.

Significant research is underway to improve existing memory technologies. However, each alternative suffers from its own limitations.^{2,3} The critical need for memory systems that allow significant improvement in data processing capability cannot be stressed enough. This document outlines the basic research challenges that need to be addressed to develop more effective memory technologies.

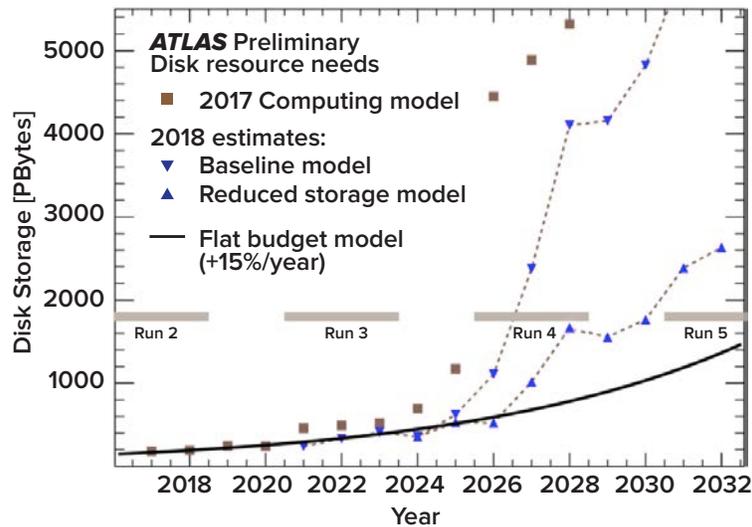


Figure 2. Anticipated data storage requirements for ATLAS facility at Large Hadron Collider (from CERN Public Record, ATLAS Upgrade Project).

SCIENTIFIC CHALLENGES

Challenges in the realization of efficient memory systems come from fundamental physics considerations. Today, extremely large-density memory systems can be fabricated. Commercially available three-dimensional NAND Flash technology offers a staggering number of storage bits within a small volume. Other memory technologies, such as magnetic and conductive bridge random-access memory (MRAM and CBRAM), phase change memory (PCM), and ferroelectric field effect transistors (FEFET), are at various stages of research development or early production. Recent research shows that DNA-based memory systems might be possible, leading to densities larger than what is achievable today. However, all these memory systems are governed by a set of common limitations—to store memory with large retention times, the energy cost of “writing” the memory is relatively high. Another challenge is that the memory storage must be sufficiently decoupled from the environment to minimize errors; and this then leads to slow access times.

A trade-off exists among energy, speed, error rate, and retention in the memory technologies that are being explored today. In addition, an efficient memory system must compute inside or in close proximity to memory. The above trade-off and the need for computing functions within memory highlight the need for exploring new physics-based approaches and materials that look beyond current approaches, or seek to advance them along paths that are not already being pursued today. Furthermore, “in-memory computing” brings with it an algorithmic challenge, requiring the principle of co-design in its resolution.

In sum, a complete rethinking of the entire memory hierarchy, including material synthesis, device design and fabrication, integration with logic, interconnection in three dimensions, and new algorithms to guide and take advantage of such systems, will be necessary to overcome the challenges and usher in a new era for memory systems.

RESEARCH THRUSTS

Thrust 1. Explore novel materials and physics that can overcome the cost-density-speed tradeoffs

Whether in a HPC system or at the edge where data are generated by experimental devices, memory and storage advances are required to meet DOE challenges. DOE program requirements include HPC systems with extreme capacity, reliability/retention, high bandwidth, and low latency, all in the face of power constraints. Different programs require different combinations of these factors. For instance, long-term archival storage requires extreme capacity and retention but can generally tolerate low latency. We expect these advances will require disruptive innovations in materials and physics that can overcome the current tradeoffs in memory and storage devices.

Given that data are generated in HPC systems or experimental facilities, these locations impose varying environmental constraints, such as high radiation environments at experimental devices and constrained cooling environments. The DOE facilities include many terrestrial locations from sea level to high altitude, as well as aerospace and extraterrestrial locations. These locations come with different radiation environments, such as rates of thermal, epithermal, and fast neutrons on earth as well as high energy protons in space. Technological advances need to address these constraints where they are expected to be applied and fielded.

DOE needs, as well as those of the computing community at large, include new memory and storage elements that minimize the cost-density-speed tradeoffs. Today, storage elements such as tape and hard disk drives are inexpensive and dense (>1 Tb per in.²), but slow (millisecond access time); while memory elements such as SRAM are comparatively low-density (~ 1 Gb per in.²) and expensive but fast (nanosecond access time) (Figure 3). To meet DOE mission requirements, the underlying science and technology for future memory devices need to circumvent this tradeoff. For instance, could one make a memory that approaches the cost, energy, and density scaling of tape, but attains read/write speeds that approach those of SRAM, while—at the same time—retaining non-volatility, endurance, and other characteristics?

This challenge calls for the exploration of new physical approaches and phenomena, and their exploitation in new materials with unique characteristics that expand much beyond the limited set of materials that are currently explored for memory technologies. It also calls for developing memories that push the extremes of the nanoscale, with storage involving hundreds to single atoms (instead of tens of thousands as is the case today), while developing ways to synthesize, place, and connect these nanoscale devices and utilizing them in complex, tightly integrated devices.

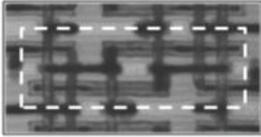
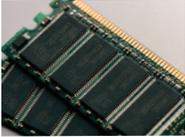
SRAM	DRAM	Flash	Hard Disk	Tape
				
>\$300/GB	>\$7/GB	>\$1/GB	>\$0.03/GB	>\$0.01/GB
10 nano-sec	50 nano-sec	micro-sec	milli-sec	sec
~1 Gb/sq. in. equiv	~100 Gb/sq. in. equiv	>1 Tb/sq. in.	>1 Tb/sq. in.	>1 Tb/sq. in.

Figure 3. Memory and storage devices typically used in systems today. Note variation in density, access times, and cost. The SRAM image courtesy of Majhi Prashant, Intel. Other images reproduced from <https://unsplash.com/photos/ING1Uf1Fc30> (DRAM), <https://unsplash.com/photos/hvBHile6dMw> (Flash), https://commons.wikimedia.org/wiki/Hard_disk (Hard Disk), and <https://commons.wikimedia.org/wiki/File:3592Tape.JPG> (Tape).

Thrust 2. Explore fundamental physics that governs energy efficiency in logic

A memory system cannot be efficient without highly efficient computing hardware. The algorithms that use the memory system are run (or at least predominantly run) on a computing block.^{4,5} Energy efficiency in computing, however, is facing its own fundamental challenges. Traditional voltage scaling of the silicon CMOS transistor has essentially stopped. This limitation is due to the fact that the supply voltage in transistors is currently limited by the Boltzmann distribution of electrons. As a result, there is a fundamental minimum limit on the voltage required to turn on a transistor from the off condition.

Extensive research in the academic community in the last decade has explored new physics in the operation of transistors in order to remove this limitation, such as tunneling⁶ and correlated phenomena,⁷ exploitation of the negative capacitance phenomena,^{8,9} and use of new materials such as carbon nanotubes, graphene, and two-dimensional semiconductors. However, for such novel schemes, materials quality, integration with silicon, device scalability, device-to-device variation, and in some cases, architectural incompatibility pose significant challenges. Breaking the barriers of energy dissipation in today's CMOS transistors requires new approaches that rely on novel physical concepts; however, such approaches should be firmly grounded in terms of scalable material systems and a re-thinking and re-architecting of the traditional CPU-memory hierarchy.

Thrust 3. Explore novel synthesis, integration, and architecture to enable a three-dimensional (3D) integrated logic-memory paradigm

One approach to attain high throughput in memory systems is to pursue the merged logic-memory paradigm, where logic and memory blocks work in unison, overcoming the von-Neumann bottleneck. One way of embodying this paradigm is via 3D integration of logic and memory to enable a truly monolithic, high-density, high-throughput logic-memory technology (Figure 4). Today's memory technologies were developed in isolation from the logic devices. As a result, most often their synthesis is not compatible with high performance processors. Currently, significant effort is being devoted to "embeddable" memory systems. However, none of the existing embeddable technologies can be integrated at very high density. Additionally, for a large memory, addressing the end memory locations requires significant access time.

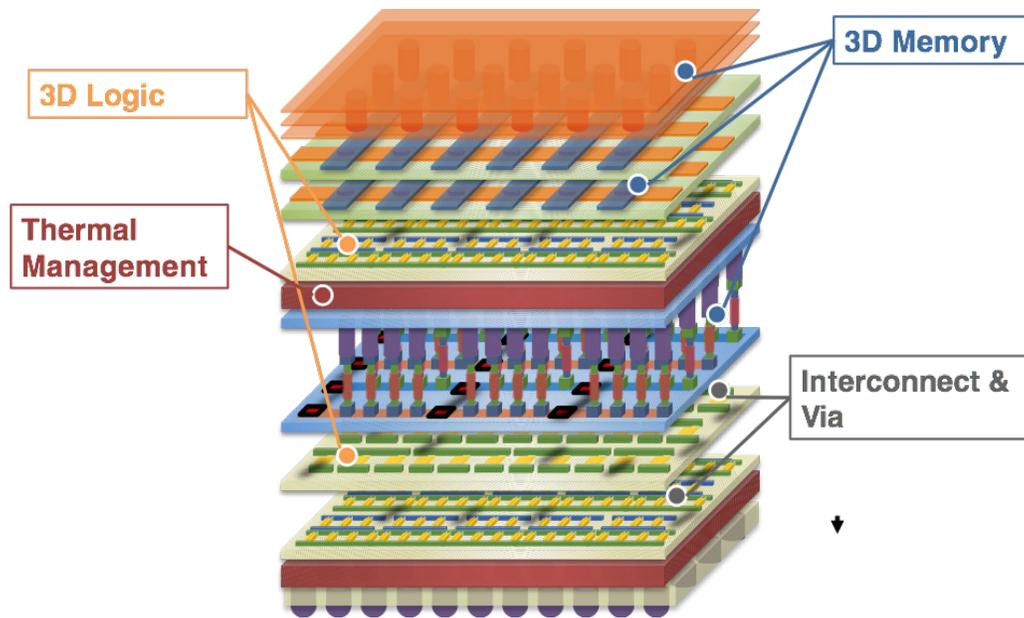


Figure 4. Schematic of 3D architecture where memory and processing layers are stacked on top of one another to closely integrate these two functions and minimize wiring distances. Adapted from M.M. Sabry Aly et al., *Computer*, 48 (2015) 24-33.

Another approach is to integrate two different technologies such as logic and memory in the package, the so-called “2.5D/3D die stacking” technologies. However, to truly overcome the aforementioned challenges, the ultimate goal is to develop the ability to disperse memory and logic blocks, as desired and as required by algorithms, in 3D with varying degrees of granularity. Further, as we move toward systems with numerous such distributed functions, we will need to understand how to set up, monitor, and control a very heterogeneous set of computational threads that simply is not like anything we have today, even with graphics processing units (GPUs).³ This capability requires a holistic approach and a complete rethinking of the materials, devices, and interconnect technologies, as well as abstraction and programming. In addition, finding the optimum placement of the inter-leaved logic and memory elements in 3D without creating unintended routing congestion is a non-convex optimization (“NP hard problem”). Therefore, breakthrough concepts will be needed to ensure an efficient way of arranging devices in 3D.

Magnetism has long been the basis for information storage devices, but recent materials discoveries have raised the possibility of magnetic devices for digital logic. With the digital state of each device represented by a persistent magnetic polarization, the state of a computation would not need to be saved before the power is turned off. This capability could be of immediate value in power-starved systems dependent on intermittent power sources. In the longer term, it could profoundly change computer architecture by exploring magnetism, including voltage-controlled magnetism, as a way of integrating memory with logic, provided the energy efficiency of magnetic polarization switching is made competitive with that of electronic switching.

Thrust 4. Explore neuromorphic devices and multi-level to analog memory

Artificial neural networks (ANNs) with analog memory elements as synaptic weights are being aggressively explored as energy-efficient architectures for execution of machine learning algorithms. However, attaining their full potential may require the introduction of new materials and devices that more compactly and efficiently implement key network functions.¹⁰ For example, much effort has gone into the development and demonstration of analog memory devices to store the weights of synaptic connections, but the material properties and resulting device characteristics are still far from ideal for this application.¹¹ Some researchers are modifying algorithms to better match the device characteristics, while others are pursuing materials and device approaches that may be better suited to the well-established algorithms. Most investigations into materials examined today for ANNs have been opportunistic, using materials such as oxide dielectrics that have been developed and made available for other purposes. New ultra-efficient ANN hardware represents a major opportunity for computing; however, discovery research for new materials and phenomena is needed for a new generation of analog memories and selector switches with true analog characteristics and controllability.

Thrust 5. Develop methods for atom-scale and 3D nanofabrication (also relevant to PRD 3)

As technology scales, we are now encountering two limitations in semiconductor fabrication methodologies: (i) microelectronics nanofabrication has been historically developed for planar, two-dimensional structures, and (ii) stochastic or other physical limitations inherent within many semiconductor fabrication methods (ion implantation, for example) limit us in our ability to reproducibly, routinely, and efficiently pattern below ~5-10 nm at scale. Because of limitation (i), we do not have the ability to pattern and create heterogeneous structures in a monolithic and efficient manner for 3D architectures. Limitation (ii) will restrict us from creating ultra-dense structures beyond what we are able to achieve today within existing materials and nanofabrication paradigms.

These limitations highlight the need to develop an entire family of new nanofabrication techniques that will enable us to build dense three-dimensional computing architectures of the future with the heterogeneous material environment within them enabling the memory, processing, data communication, and thermal management functions needed at a cost that is economical to deploy. This effort will involve developing new deterministic placement or statistical but robust self-assembling methods in three dimensions that are scalable, and creating engineered heterogeneities via synthetic chemical functionalization means that may be combined with patterning techniques. (Organic functionalization of surfaces in atomic layer deposition for front-end-of-line applications in silicon microelectronics is an early example that is being explored today.) This effort will also involve the discovery science of new materials, morphologies, and approaches that can be multifunctional in accommodating communication, computing, memory, and thermal transport needs. Biological systems are an example of the latter, for instance, where fluids aid in both delivering power and removing heat. Developing these techniques will require close interaction among experiment, theory, and simulations.

Thrust 6. Analyze and characterize tradeoffs among space, time, energy, and precision in devices, algorithms, and applications

Traditional digital computing applications are often designed as deterministic procedures executed on perfect hardware to produce results to a user-desired precision. However, over the past few years computing has been increasingly driven by other less traditional applications: streaming-data analysis of high-volume, high-velocity data streams on HPC and high throughput computing services, as well as edge devices; data-driven graph analysis and machine learning applications; and in-memory computing for data reduction from large volumes of data. These applications have been designed to make use of traditional computing architectures and will not effectively utilize future compute-memory designs. Given the expected increase in memory hierarchy depth, the overarching consideration related to minimization of energy, and the need to accommodate new algorithm models, future memory technologies and designs can expose a far greater range of architectural design points in terms of the performance, energy requirements, and associated error rates.

We anticipate that technology-algorithm mapping will lead to new classes of algorithms co-designed with the hardware. Indeed, with the fast rise of complex workflows in science, these will present their own type of challenge, given the heterogeneity in their computational and data requirements across the workflow at any given point in time, quite different from traditional HPC applications. Now more than ever there is a pressing need to understand the fundamental memory access and compute patterns in DOE applications, beyond specific existing implementations, particularly given the advent and preponderance of data-driven computing in science.

We are no longer confined to consideration of performance through locality and memory sizes. Indeed, the number of degrees of freedom for this challenging optimization problem now includes heterogeneity in the architecture and technologies for memory and storage, energy, reliability, and acceptable bounds for approximate computing. Hence, the memory-compute patterns that form the basis of co-design need to be well understood and encapsulated in tools and “compact apps” available for co-design. The complexity of optimization of data movement along the design criteria and in coordination with the needed compute capabilities will lead to application development that is programming model and runtime aware enough to enable dynamic optimization of heterogeneous architectures. Existing tools and techniques typically focus on individual abstraction layers (e.g., memory access patterns in a kernel) or metrics (e.g., cache misses). These tools and approaches need to be rethought to enable design-space exploration involving multiple algorithmic and application-level objectives and constraints. Furthermore, with the expected addition of completely new hardware, such as DNA-based storage and quantum co-processors and machine learning/neuromorphic accelerators, we need to further review both metrics and measurement methods. The scale of this challenge

necessitates the development of novel tools and approaches to enable this innovation across a broad set of heterogeneous hardware and software.

Thrust 7. Build programming models, algorithms, and data structures to mitigate the enforcement of memory consistency

Algorithms, data structures, programming environments, and entire software stacks have been built to address the challenges associated with efficient data access and movement of data between components of the memory hierarchy and the processing units. Indeed, assumptions about the memory hierarchy are so fundamental to current parallel software and system architectures that we classify them as distributed memory, shared memory,¹² or partitioned global address space models. The key design choices underpinning this classification via the memory model are described via the notions of memory coherence and consistency. While stronger consistency guarantees may facilitate programmability, they imply ever greater data coherence, traffic, and synchronization requirements and the communication bandwidth and power to perform the required checks. Research is needed to understand the suitability of current models and programming abstractions for new kinds of memory and the platforms in which they are configured.¹³

Paradigm-changing programming models and software solutions are needed to complement novel memory architectures in order to minimize application development effort and ensure that these new systems are effectively exploited. Research is needed to rethink the on-chip support for shared memory programming paradigms, including detailed analyses of current bottlenecks, development of techniques that further scale existing approaches, and creation of novel strategies to avoid or eliminate these performance bottlenecks.^{14,15} Programming environments, supporting tools, and best practices are needed that enable a developer to create applications with minimal coherency requirements. In addition, novel programming environments and automated solutions must be developed to enable this fundamentally different approach to parallel programming.

SCIENCE AND TECHNOLOGY IMPACT

The research actions recommended in this PRD will deliver a number of critical advances that will significantly improve our understanding and practice in the way memory is used in computing, and that will overcome the serious barriers that we face today for memory technologies. Success would revolutionize the field of memory devices, as well as enable key application areas in the DOE mission. From a scientific perspective, success would enable us to explore the physical limits of microstructure and state control in materials, as well as to discover new materials and novel physics and determine their role in exploring the limits of energy efficiency, density, and performance of memory.

The development of new physics-based approaches and materials would help circumvent the cost-density-speed tradeoffs that constrain today's memory and storage devices. Atom-scale nanofabrication will enable us to create ultra-dense and low-power processing elements that intimately integrate memory with processing, meeting the needs for future memory-hungry workloads. Three-dimensional integration, artificial neural networks, and neuromorphic approaches, as well as paradigm-changing memory models and software solutions, are needed to overcome the current memory bottleneck in computing.

Atom-scale nanofabrication and the synthesis and manipulation of heterogeneous materials with atomic precision can impact many other scientific fields. While our current ability to control materials down to the 5-10 nm level robustly and across scale is no doubt impressive, being able to do so at the ~ 1 nm level opens up entirely new arenas of science at the nanoscale. The exploration of novel physics and new materials discovery for energy-efficient switching and memory, as well as the approaches for artificial neural networks and neuromorphic computing, will lead to a deeper understanding of the relationships among processing, memory, and communications in information processing and the extent to which these functions can begin overlapping.

The energy consumption in today's supercomputers is dominated by the energy costs for storing and moving data. By creating ultra-low energy, ultra-dense storage, tightly integrated with computational capabilities, the research envisaged under this PRD could have a remarkable impact on how we architect computing systems of the future. In addition to requiring less power to store data, this will additionally collapse today's deep storage hierarchies by tightly integrating storage with the computing function, thus reducing or eliminating the need to move data.

Besides enabling more-efficient supercomputers, this research will enable compact, but highly computationally powerful edge-computing devices tightly coupled within sensor and experimental measurement networks for building new classes of powerful data acquisition systems. In particular, these systems will facilitate groundbreaking new scientific experiments and applications within DOE's large-scale scientific facilities.

As discussed earlier, new exascale computing systems and advanced experimental facilities envisaged for the future, such as the Advanced Photon Source Upgrade at Argonne and the HL-LHC, will create data at unprecedented scales. In addition to the volume of data created, expected to be in the exabyte range for both Exascale Computing and the HL-LHC, the rate at which data are created will be a critical component. For instance, HL-LHC expects to generate 1 Pb/s of raw data. These data volumes are too large to be stored in their entirety, and need to be reduced *in-situ* via automated, on-the-fly scientific analysis. This is where breakthroughs in edge computing, and associated with it, the research envisaged by this PRD, will prove to be highly impactful.

Increasingly powerful sensors will enable sensor networks of the future to collect extreme volumes of data. Examples range from the electrical power grid to subsurface sensor networks. Here, the ability to process ultra-high data volumes at the point of creation is essential. Delivering high throughput data processing and some data retention within a very low energy envelope will be key to transforming these types of sensor networks, from limited long-term data collection tools to experimental and operational environments that support real-time analytics.

REFERENCES

1. H.-S.P. Wong and S. Salahuddin, Memory leads the way to better computing, *Nature Nanotechnology*, 10(3) (2015) 191.
2. S. Sayeef, K. Ni, and S. Datta, The era of hyper-scaling in electronics, *Nature Electronics*, 1(8) (2018) 442.
3. J.S. Vetter et al., *Extreme Heterogeneity 2018: Productive Computational Science in the Era of Extreme Heterogeneity Report for DOE ASCR Basic Research Needs Workshop on Extreme Heterogeneity, January 23–25, 2018*, DOI: 10.2172/1473756 (2018).
4. T.N. Theis and H.-S.P. Wong, The end of Moore's Law: A new beginning for information technology, *Computing in Science and Engineering*, 19(2) (2017) 41-50.
5. T.N. Theis and P.M. Solomon, In quest of the "next switch": Prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor, *Proceedings of IEEE*, 98 (2010) 2005-2014.
6. B.M. Borg, K.A. Dick, B. Ganjipour, M.-E. Pistol, L.-E. Wernersson, and C. Thelander, InAs/GaSb heterostructure nanowires for tunnel field-effect transistors, *Nano Letters*, 10(10) (2010) 4080-4085.
7. S. Manipatruni, D.E. Nikonov, C.-C. Lin, T.A. Gosavi, H. Liu, B. Prasad, Y.-L. Huang, E. Bonturim, R. Ramesh, and I.A. Young, Scalable energy-efficient magnetoelectric spin-orbit logic, *Nature*, 565 (2019) 35-42.
8. J.C. Wong and S. Salahuddin, Negative capacitance transistors, *Proceedings of IEEE*, 107(1) (2019) 49-62. DOI: 10.1109/JPROC.2018.2884518.
9. A.K. Yadav, K.X. Nguyen, Z. Hong, P. García-Fernández, P. Aguado-Puente, C.T. Nelson, S. Das, B. Prasad, D. Kwon, S. Cheema, A.I. Khan, C. Hu, J. Íñiguez, J. Junquera, L.-Q. Chen, D.A. Muller, R. Ramesh, and S. Salahuddin, Spatially resolved steady state negative capacitance, *Nature*, 565 (2019) 468-471.
10. W. Haensch, T. Gokmen, and R. Puri, The next generation of deep learning hardware: Analog computing, *Proceedings of IEEE*, 107(1) (2019) 108-122. DOI: 10.1109/JPROC.2018.2871057.
11. T. Goken, M. Rasch, and W. Haensch, Training LSTM networks with resistive cross-point devices, arXiv:1806.00166 (2018).
12. S.V. Adve and K. Gharachorloo, Shared memory consistency models: A tutorial, *IEEE Computer*, 29(12) (1996) 66-76.
13. B. Chapman, The challenge of providing a high-level programming model for high-performance computing, in *High Performance Computing: Paradigm and Infrastructure*, New York: Wiley Publishers, pp. 21-50 (2006).
14. R. Komuravelli, S.V. Adve, and C.-T. Chou, Revisiting the complexity of hardware cache coherence and some implications, *ACM Trans. Architecture and Code Optimization (TACO)*, 11(4) (2015) Article No. 37.
15. A. Shriraman, H. Zhao, and S. Dwarkadas, An application-tailored approach to hardware cache coherency, *Computer*, 46(10) (2013) 40-47.

PRD 3 Reimagine information flow unconstrained by interconnects

INTRODUCTION

Today, ultra-large-scale integrated circuits contain close to one hundred million transistors on a millimeter square sliver of silicon. Over thirty miles of interconnect wires run on each chip strewn over ten or more levels of dielectrics to shuttle electrical data and clock signals around the chip for high-throughput information processing. With the anticipated logic gate latency approaching one picosecond at the 10-nm technology node, the interconnect-induced delay poses the critical bottleneck for information transfer above a critical length.

The interconnect bottleneck, which limits the data bandwidth, extends beyond the chip and is exacerbated as the volume of data to be ingested, processed, and analyzed grows super exponentially. By some estimates, today we produce the same amount of data in minutes that we produced over the past hundred years. By 2025, we will produce the same in less than ten seconds. Ingesting, processing, and analyzing such abundant data with high throughput requires unhindered data movement through the interconnect fabric between the compute cores and the physical memory.

For example, experimental high energy physics (HEP), which enhances our basic understanding of fundamental particles, will benefit significantly from extending the boundaries of data intensive computing. It involves acquiring massive amounts of data by detectors, moving the data to a HPC cluster, and processing the data to identify patterns, perform clustering analysis, and reconstruct critical events. The interconnect bottleneck restricts the flow and interactive analysis of data, and the problem will get worse as the HL-LHC comes online and contributes to the data explosion. Other data-intensive research activities face similar challenges.

As noted in PRD 2 and 4, neuromorphic computing has emerged as a complementary architecture to traditional von Neumann systems. The brain-inspired neuromorphic architectures are characterized by extreme connectivity and parallelism, requiring not only co-located memory and processing units but local feedback for information processing. There is a need for alternative interconnects that go beyond traditional electrical “wires” and serve as conduits for propagation of novel information “tokens,” such as asynchronous spikes, collective oscillations, chemical messengers, mechanical strain, and photonic excitations. Radical innovation in emerging devices may provide pathways toward achieving such non-Von Neumann capabilities that approach the biological brain in terms of compute performance and energy efficiency.

PRD 3 seeks revolutionary scientific ideas and new materials that can lead to interconnect technologies that go well beyond the traditional paradigm and enable efficient data flow in future computing systems across many length scales (Figure 1).

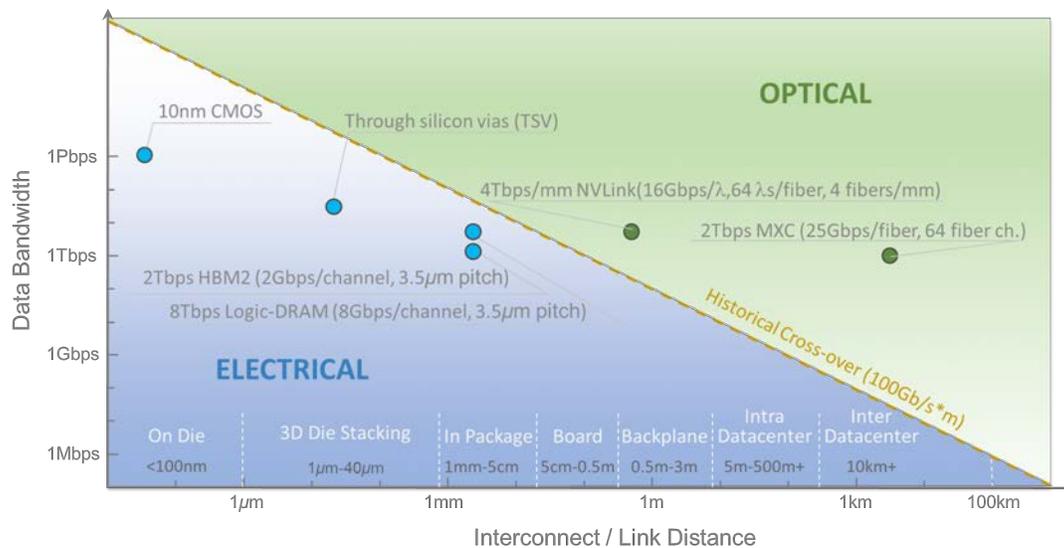


Figure 1. Data bandwidth as a function of link distance in today's interconnect paradigm. Electrical-to-optical interconnect cross-over occurs when bandwidth density crosses 100 gigabits per second per meter of link distance (Gb/s/m). PRD 3 seeks disruptive approaches not only to increase the respective electrical and optical bandwidth densities to their ultimate limits, but also to blur the historical cross-over boundary using novel collective states of condensed matter. HBM = high bandwidth memory.

SCIENTIFIC CHALLENGES

What novel electronic, optical and new states of matter can be discovered and manipulated to design and implement non-traditional interconnects at the atomic, micro, and macro scales?

On-die interconnect (<100 nm): Local and semi-global interconnect scaling is rapidly becoming a major roadblock in transistor technology at advanced nodes (Figure 2).¹ Local interconnect latency, rather than transistor switching speed, is quickly becoming the critical bottleneck as the gate delay in transistors reduces to less than one picosecond at the 10 nm technology node. New electronic conduction mechanisms that exhibit much lower mean free path and resistivity product are promising candidates to realize extremely scaled vias and local interconnect. A new figure of merit for future-generation interconnect materials at extremes of scaling is the inverse of the product of the mean free path for back scattering (λ) and the resistivity (ρ) at scaled dimensions.

Three-dimensional (3D) die stacking (1-100 µm): As discussed in PRD 2, recent years have witnessed a shift from traditional technology scaling in the two-dimensional plane to integration of high-density memory and logic elements using 3D die integration. One approach for this is “die-stacking,” i.e., partitioning the processor into several blocks where each block can be a chiplet fabricated on a different technology node, and aggregating those blocks or chiplets across multiple planes that are stacked atop one another. For example, Intel Foveros is a silicon stacking technique that allows different chips to be connected by through-silicon-via (TSVs) such that the I/O (input/output), the logic cores, and the DRAM can be fabricated as separate chiplets and then stacked together. In this instance, Intel considers the I/O chiplet, the chiplet at the very bottom of the stack, as a sort of “active interposer”, that can route data between the logic and the memory chiplets on top.² Additionally, this technology also brings new opportunities to integrate heterogeneous components (sensor arrays, mixed signal analog circuits, non-volatile memories, etc.) in a single chip.

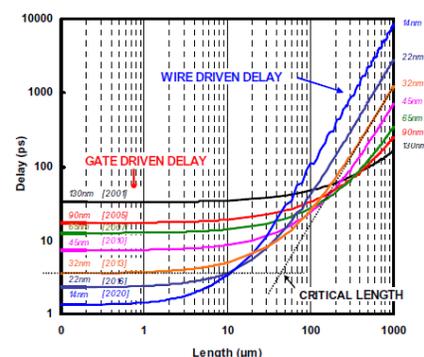


Figure 2. Logic gate delay as a function of wire distance. At less than 1 µm, the transistor gate delay dominates, limited by the intrinsic switching speed and parasitic resistance and capacitance. Above that, the resistive-capacitive (RC) delay associated with local and semi-global wires dominates. Reproduced from M. Sellier et al., 9th International Symposium on Quality Electronic Design. Copyright (2008) IEEE.

In TSV-based stacked 3D integrated circuits, forming vertical structures to electrically connect a multitude of vertically stacked layers involves high-aspect-ratio etching, silicon wafer thinning, and die bonding. Defects that occur during these extra processes cause performance and yield loss, limiting the number of heterogeneous layers that can be stacked in a 3D integrated circuit technology. Additionally, thermal management is a second critical issue in 3D volumes with multiple active dies (each generating heat), resulting in a thermal dissipation bottleneck. There is an opportunity for fresh thinking here—in discovering new materials and interconnects that leapfrog over the limitations of TSVs, in actively removing heat from dense 3D integrated structures via nanoscale thermal engineering, and in developing methods for materials processing and nanofabrication to create multi-level stacked and monolithic 3D information processing volumes.

In-package (chip-to-chip) (1 mm-5 cm) interconnects: Chip-to-chip interconnect bottlenecks arising from high volume data movement between memory blocks and computing units restrict system performance. Optical interconnects can support 1 Tb/s data bandwidths via use of wavelength division multiplexing (WDM) channels, albeit with the complex integration challenges of new materials such as compound semiconductors, large device sizes, and thermal stability issues, relegating the optical interconnect usage to link distances over 10 cm. Revolutionary ideas, beyond traditional electrical and optical interconnects, that can address the so-called “last centimeter barrier” are needed to support high bandwidth data transfer between the chips within the package. In the 2.5D approach, the dies are placed side-by-side on top of a silicon interposer, which incorporates TSVs. The interposer acts as the bridge between the chips and a board, which, in turn, provides more I/O and bandwidth in packages. Disruptive approaches that can dramatically lower energy per bit (<100 fJ/bit) and increase inter-die bandwidth per unit edge length (>500 Gb/s/mm of edge) are required to find alternatives to conventional copper-based wiring on silicon interposers with embedded millimeter-wave transmission waveguides.

Photonic interconnects offer potential for much higher bandwidth and lower latency, as interconnect lengths span the 1-mm to 5-cm regime between the processors and memory. Chip-based dielectric waveguides embedded within the package are required to achieve acceptably low materials losses, minimize optical crosstalk in densely nested waveguide arrays, and attain insertion losses of <0.3 dB per waveguide/device transition. Moreover, today’s silicon photonics and silicon nitride waveguide technologies are composed of essentially passive components, whereas future large-scale photonic interconnects will require a means of generating on-chip gain with high power efficiency. These interconnects would benefit from the availability of ultra-compact, broadband modulators with sub-femto-joule switching energies. This would drive the discovery of new material building blocks for active modulator components that are capable of generating ultralow power optical nonlinearities and electro-optical modulation with sub-picojoule per bit switching energies. Future ultracompact large-scale photonic networks will also benefit from new scientific concepts and phenomena for optical isolation of sources and detectors to enable both transmit and receive modes of operation. The requirement for optical isolation implies that portions of the waveguide network would ideally be composed of non-reciprocal optical media. Emerging photonic science concepts for violating Lorentz reciprocity in optical media include excitation of magnetic fields, temporal modulation of transmitted and scattered signals, and development of new optically nonlinear media. Waveguide networks composed of topologically protected optical channels also have potential for optical isolation and improvement of the signal-to-noise ratio of photonic interconnect networks.

Ultimate scaling of photonic interconnects will also demand new tools and phenomena to enable extreme subwavelength light confinement as well as low optical losses. One approach could feature nanophotonic motifs and excitations that exploit polaritonic modes, including plasmon, exciton, and phonon polariton modes, which are opening new opportunities for extreme light confinement. Recently, polaritons in two-dimensional materials, such as plasmons in graphene and dielectric modes in transitional metal dichalcogenides, have demonstrated remarkable potential for reduction in wavelength and mode volume relative to their free space counterparts. At the same time, novel electro-optical, phase change, and valley-tronic phenomena in two-dimensional materials can enable significant modulation of optical permittivities in monolayer and multilayer films. The ability to exploit the potential of reduced dimension materials for implementation of active interconnect components may open a completely new avenue for chip-to-chip high-speed data communication.

Backplane and intra- and inter-data center (1 m–100 km):

Large-scale data centers and HPC systems consume megawatts of power. For example, Aurora, the exascale machine anticipated to be operative at Argonne National Laboratory in 2021, will consume 43 MW of power. Conventional electronic interconnects will not meet the future intra- and inter-data center interconnect requirements because of bandwidth-density and power-consumption limitations. Low-power, high-speed optical interconnects are emerging as alternatives for data center and HPC communications. They offer opportunities for continued energy-efficiency and bandwidth-density improvements. Figure 3 shows the floor plan of a conceptual peta-scale compute node that exploits photonic communication links. Photonic links have several advantages related to reduced energy dissipation and higher bandwidth. First, in photonic interconnects, there is no RC charging of electrical lines. Second, because signal propagation lengths are much longer, photonics dramatically reduce the complexity of electronic circuitry, such as clock and data recovery circuits, line coders, and serialization and deserialization circuits. Third, WDM systems featuring broadband optical amplifiers and low dispersion fibers can enable extremely high bandwidth optical links.

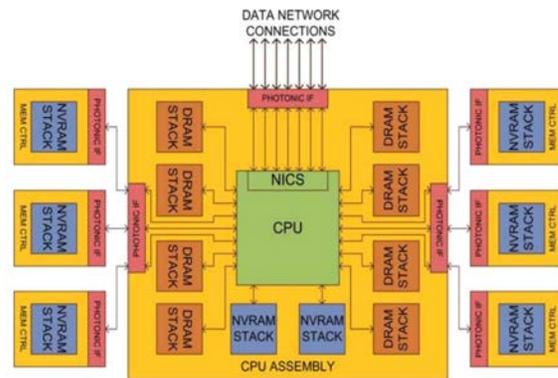


Figure 3. Peta-scale compute node exploiting photonic interconnect links for intra- and inter-node data traffic. Reproduced with permission, François Bodin, *HPC Today*, <http://www.hpctoday.com/state-of-the-art/non-volatile-memory-nvm-technology-for-exascale-computing-a-position-paper> (2014).

Photonic interconnects need to interface seamlessly with electronics, since processors and memories are largely anticipated to remain electrical in the future. Thus, circuitry is needed to interconnect the processor I/O with the photonic link on-board. This circuitry has to drive the photonic interconnects and requires proper matching between the electronic and photonic worlds. It includes compact and efficient drivers, modulators, detectors, trans-impedance amplifiers, and clock data recovery circuits for instance. There is an opportunity going forward for new advances with emerging materials, such as polymers, graphene, 2D layered materials, and topological materials that may lead to hybrid interconnect schemes with ultrahigh density and speeds.

How can we minimize data movement while maximizing information transfer?

A typical work flow for HEP data acquisition and analysis involves detector signal recording, event reconstruction, and data analysis. In the event reconstruction step, the physics of broad interest such as the trajectories of charged particles and particle hypotheses is extracted from raw instrument signals. Event reconstruction is performed on full raw data sets and is expensive in terms of compute time and energy. Data analysis typically involves processing the reconstructed data, calculating the statistical summaries, and identifying the most relevant summaries. Analysis is an iterative process where low latency and interactivity are keys to making progress. There is an opportunity to explore alternative approaches to in-memory data processing to enable interactive analysis to meet the needs of the next generation of particle colliders. For example, recent years have witnessed much progress in in-memory computing architectures based on resistive and phase-change cross-bar memory for highly parallel and reconfigurable information processing. Such computing architectures show promise to accelerate streaming vector matrix math.

Data rates vary over time due to the characteristic dynamics in edge computing applications, such as those needed in sensor networks. Within DOE such needs are anticipated to arise in the large-scale experimental facilities within light and neutron sources and in the context of HEP experiments. Adaptivity and flexibility are critical at the levels of logic gate components and the interconnects to enable scaling of performance and power consumption during runtime. If a lower (or higher) data rate is present on the link, the performance (e.g., bandwidth) of the adaptive component can be reduced (or increased), and therefore, power can be saved (or expended). One way to achieve such adaptability is by switching a part of the electronic or optical link components on or off as needed. Disruptive innovations such as dynamically reconfigurable materials (e.g., metal-insulator phase transition materials or super cut-off transistors with sub-Boltzmann switching characteristics) could make next-generation interconnects adaptive and flexible.

Need for a co-design approach for interconnect architectures

The rapid growth of abundant data challenges us to completely re-think hardware architecture from a communication-centric, rather than purely computation-centric view. Technology trends clearly highlight the importance of interconnect-conscious design (see Figure 2), since communication latency has not scaled as fast as the logic gate or the memory array access delays in advanced nodes. The creation of future on- and off-chip interconnect networks that offer high performance, energy efficiency, thermally resilient, and security by using heterogeneous integration of diverse technologies (e.g., electronic, photonic, or plasmonic) will require a holistic design platform for careful exploration of the design space.

Comprehensive simulation platforms should be established to understand the critical interplay between applications, system architecture, and hybrid interconnect fabrics that rely on novel phenomena and emerging device concepts. The creation of a tightly coupled experimental and simulation platform will have a significant influence on the design of future-generation integrated microsystems, and will also foster new research directions for heterogeneous compute platforms (using CPUs, GPUs, field-programmable gate arrays, accelerators, stacked DRAM, and cross-point phase change memories) that can support mission-critical compute workloads.

RESEARCH THRUSTS

Thrust 1. On-chip interconnects

The first research thrust is to develop on-chip interconnects that can reduce the product of mean free path and resistivity by exploiting new physical phenomena and new materials (e.g., topological materials, plasmonic media, magnons, room-temperature ultra-high conductivity materials, superconductors, and artificial axons) as well as interfacial and material design to mitigate electrical and thermal transport losses. At the mesoscopic scale, the Landauer-Buttiker transport formalism holds, and the product of the mean free path and the resistivity simply reduces to the inverse of the number of transmission modes available within a Fermi window. One approach to address this could be by creating new materials with anisotropic Fermi surfaces that provide a large density of modes along the transport direction, or by exploring new interconnect materials with topologically protected surface and edge states that reduce carrier scattering. These innovations are required to overcome the resistance and capacitance increase associated with local, semi-global, and global interconnect scaling in integrated circuits at advanced technology nodes.

The effective resistance of tungsten-based local wires and copper-based semi-global and global wires will become a critical bottleneck for interconnect-dominated chip performance at small dimensions because the large mean free path in these materials will lead to higher sidewall scattering and the need for an amorphous diffusion barrier (typically with higher resistivity). According to the Landauer theory of electrical resistance of a mesoscopic conductor, the conductance is a function of both the scattering properties of the conductor and the number of available transmission channels.³

The inverse of the product of mean free path and resistivity is proportional to the number of available transmission channels, $M(E)$, where E is the energy level of the carriers responsible for transport within the available bias-dependent Fermi window. Figure 4 orders the known metals in terms of their resistivity and the product of mean free path and resistivity. Today's most advanced integrated circuits in mass production are already moving from tungsten-based local contacts to cobalt, while research demonstrations are under way to replace copper-based semi-global wires with ruthenium. The challenge is to design and synthesize new materials beyond conventional metals (shown in Figure 4)³ with intricate anisotropic Fermi surfaces that maximize the density of states along the current transport direction and, hence, the available number of transmission channels. The anisotropy of the Fermi surface in reciprocal space is related to the periodicity and symmetry of the crystalline lattice of the constituent lattice.

Recent advances in the low temperature deposition and the ability to do conformal and selective growth of materials by developing new variants of atomic layer deposition and physical vapor deposition have revolutionized current microelectronics. Similarly, development of new synthesis techniques will offer new opportunities to synthesize materials with tailored Fermi surfaces. Further, the development of angle-resolved photoemission spectroscopy has provided the condensed matter community with a direct probe to resolve the

Fermi surfaces of crystals, ranging from conventional metals to high temperature superconductors, topological insulators, and Weyl semi-metals.

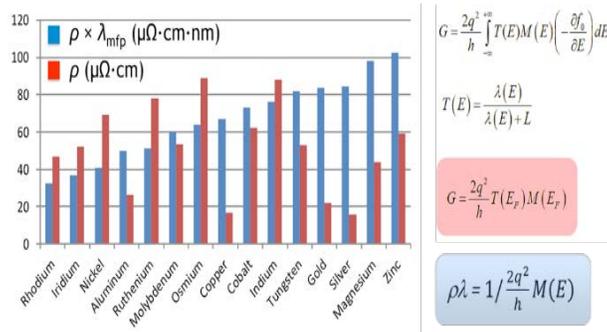


Figure 4. Material candidates for interconnect applications ordered in terms of the resistivity and mean free path product and the resistivity alone. Data in figure taken from D. Gall., *J. Appl. Phys.*, **119 (2016) 85101.**

The design, exploration, and identification of new interconnect materials with novel crystal structures will offer new opportunities to synthesize materials with tailored Fermi surfaces. Further, the recent development of angle-resolved photoemission spectroscopy has provided the condensed matter community with a direct probe to resolve the Fermi surfaces of crystals, ranging from conventional metals to high temperature superconductors, topological insulators, and Weyl semi-metals.

The design, exploration, and identification of new interconnect materials with novel crystal structures and symmetries have the potential to resolve the latency bottleneck for on-chip data movement. Accelerated discovery of these materials would benefit from computationally efficient algorithms for predicting the transport properties of new materials, as well as new approaches to high-throughput materials synthesis and experimental analysis.

Thrust 2. Interconnects within 3-D integrated stacks

The second research thrust is to use charge and electromagnetic excitations in new ways to create non-traditional interconnect structures for 3D interconnects that go well beyond the data movement limits of traditional TSVs. Future interconnects would also ideally be multifunctional and multipurpose: carrying data, as well as source power and energy, and providing a thermally conductive infrastructure for thermal management. A bio-inspired approach to 3D-interconnect fabric that functions concurrently as signal propagation medium, homeostatic feedback unit, and thermal management structure could have a transformative impact on future high-performance computing.

As one example aimed at circumventing the limitation of TSV-based 3D integrated circuits, inductive- or capacitive-coupling approaches can serve as the conduit through which signals are transmitted vertically to enable communication between the stacked layers (Figure 5). Efficient inductive/capacitive-coupling links will require technology scaling, discovery of novel magnetic materials, and the use of resonant coupling techniques. In the case of near-field magnetic resonant coupling, there are significant opportunities to improve the complex conjugate matching between the inductive coils and maximize the electrical energy transfer by means of meta-material-based matching elements. The challenge before us is to find novel link architectures or to identify disruptive pathways to realize alternative coupling links to inductive/capacitive interaction (including mechanical and chemical interaction) that would make such links realizable for more energy-efficient, interlayer communication bandwidths.

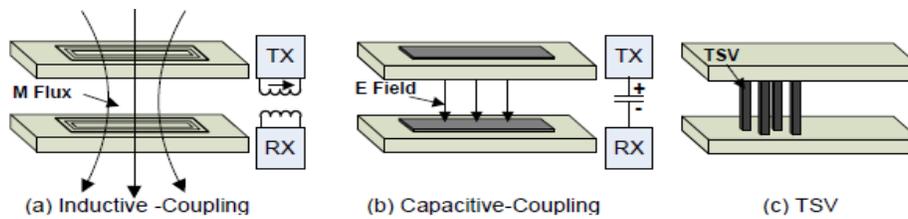


Figure 5. Conceptual schematic of the inductive, capacitive, and TSV-based physical links for data communication between the stacked layers in a 3D integrated circuit (TX refers to transmit and RX refers to receive circuits). Reproduced with permission, J. Ouyang et al., “Evaluation of using inductive/capacitive-coupling vertical interconnects in 3D network-on-chip,” IEEE/ACM International Conference on Computer-Aided Design. Copyright (2010) IEEE.

Thrust 3. Chip-to-chip interconnect

Conventional chip-to-chip communication approaches (other than 3D die stacking) are increasingly incapable of delivering the high-bandwidth density requirements for future-generation computing systems. Recent advances in 2.5D packaging as well as silicon interposer-based electrical and optical interconnects have supported high bandwidth communication between chips, but still fall short of achieving 1 Pb/s/mm of bandwidth density with energy efficiency less than 1 pJ/bit over a link distance of 10 mm and beyond. This is the figure-of-merit the future generation of interconnect (electrical or optical) has to meet or exceed to be competitive.

Brute force scaling has resulted in dense electrical wires on silicon interposers that allow higher aggregate communication bandwidth between the chips. As an example, with a copper stripline differential pair of 20- μm pitch for interconnects on a silicon interposer and channel data rate of 10 Gb/s, the bandwidth density achievable today is 500 Gb/s/mm over a link distance of 4 cm with 5 pJ/bit.

Integration of nanophotonics to form optical interposers has also been investigated, as nanophotonics can enable significantly higher bandwidth density using fine-pitch silicon photonic waveguides and WDM. For a waveguide pitch of 10 μm with eight WDM channels (each supporting 10 Gb/s), the bandwidth density achievable today using nanophotonics integration on an interposer is over 8 Tb/s/mm. However, for a given photonic interconnect design, the energy per bit consumed in electrical-to-optical conversion and vice versa, along with laser energy efficiency, sets a minimum distance beyond which the utilization of optical interconnect becomes feasible.

Recent advances in large silicon interposer and novel micro-alignment techniques supporting close proximity placement of multiple heterogeneous chips on the same substrate allow re-thinking of the current partition between electrical and optical communication for short and intermediate chip-to-chip communications. Research on emerging phenomena, such as surface plasmons—collective excitations of electrons on the surfaces of artificial materials—has enabled novel plasmonic devices⁴ that act as sub-terahertz waveguides, combining the density advantage of electrical conductors and the non-interference attribute of optical waveguides.⁵ Furthermore, plasmonic modulators have been developed for integration with low-loss dielectric waveguides in silicon photonics technology, and plasmonic modulator designs have enabled atto-joule energy per bit dissipation in subwavelength-scale device form factors (Figure 6).⁶

Non-reciprocal elements are critical components in photonic communication systems since they enable unidirectional transmission of optical signals and blocking of the propagation of modes in the other direction, thereby avoiding interference and ensuring isolation of optical sources and detectors. The research thrust here is to come up with disruptive concepts for a hybrid electronic-photonic waveguide networks with non-reciprocal materials, phenomena, and devices that can enable optical isolation.⁷ Emerging photonic science concepts for violating reciprocity in optical media include excitations of magnetic fields, temporal modulation of transmitted and scattered signals, and development of new optically nonlinear media.

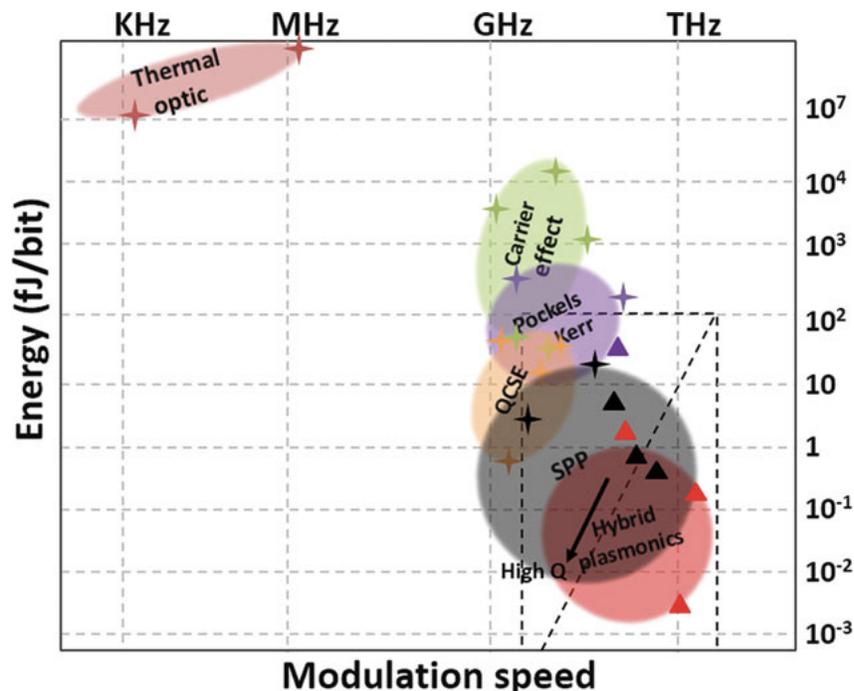


Figure 6. Energy per bit as a function of modulation speed for electro-optic modulators with various mechanisms. From K. Liu et al., *Laser Photonics*, 9(2) (2015) 172–194.

Thrust 3a. Metastructures and collective phenomena for inter-chip communications

The primary advantage of chip-to-chip communication via electrical interconnects lies in their efficiency in transmitting baseband electrical signals. Electrical interconnects are preferred for short-distance communication when the energy cost of data transfer dominates the bandwidth of communication, whereas optical interconnects are superior for long-haul communications, when the speed of communication dominates all other factors. This research thrust will address several questions. Is it possible to envisage design of a novel communication medium that prioritizes different factors (i.e., energy and bandwidth) at different times? Is it possible to realize an interconnect system that combines quasi-electrical with quasi-optical characteristics? Demonstration of surface plasmon polariton (SPP) waveguides illustrates that sub-terahertz wavelength optical modes can be supported on surfaces of metals and artificially sculpted metallodielectric materials or metamaterials, where effective permeabilities and effective permittivities that are different from the permittivities and permeabilities of natural materials can be realized by design of resonant structures and layered heterostructures.⁸ The proof-of-concept demonstration of SPP interconnects and components that are compatible with truly sub-wavelength planar CMOS encourages us to find more revolutionary concepts arising from collective phenomena in novel materials and geometries to establish a new paradigm in inter-chip communications.

Thrust 3b. Novel interconnects for chip-to-chip communication and in-network data processing

Optical interconnects are currently the only compelling alternative to electrical interconnects for next-generation chip-to-chip communication, as the link distance exceeds one meter. Achieving massive data transfer rates requires WDM of many wavelengths, requiring many parallel optical sources that, if generated using discrete lasers, would consume far too much power and size for integration. The aim of this research thrust is to find new approaches that would enable ultra-dense WDM and the ability to access multiple wavelengths without using multiple discrete lasers. One example, for instance, could be frequency combs operating at telecom wavelengths and created by parametric generation from a single laser source.⁹ This technology would enable light sources that consist of a large number of evenly spaced lines as an alternative for WDM. The key ingredients—Kerr nonlinearity, dispersion management, and amplitude shaping and modulation—need to be achieved in integrated platforms in a compact form factor.

Significant energy dissipation and time delay are required to convert optical signals to electrical signals in a network router, to decode them to set the path, and then to convert them back to optical signals that are routed through the new path. Hybrid electrical/optical routing capabilities have been demonstrated to establish optical

routing paths.¹⁰ More advanced capabilities with all optical routing are required to dramatically reduce the latency of communication. If it were possible for optical logic to decode the routing instructions and change the optical path, this could significantly reduce the latency and potentially the energy of optical routing in a data center or local area network. Research is needed to identify novel materials, structures, and devices that could perform optical logic such as decoding instructions, identifying paths that are available, and switching the optical data stream to a new path to implement dynamically reconfigurable low latency and energy-efficient all-optical interconnect networks for intra-and inter-data center communication. While all optical interconnects and optical logic need some kind of local optical memory, which has been largely elusive, there are opportunities to implement “memory-less” optical interconnect networks transporting continuous streaming data.

Thrust 4. Nanoscale thermal management (also relevant to PRD 2 and PRD 5)

A major consequence of 3D integration of multiple device layers that are interconnected with fine-grained and dense vertical interconnects is the need for effective thermal management. Heat is mostly carried by phonons in semiconductors. Thus, the manipulation of heat-carrying phonons that propagate and scatter at the nanoscale promises to yield beneficial thermal transport properties. When the system size reaches the nanoscale, heat conduction is affected by two physical effects: phonon confinement, as discussed above, and enhancement of boundary scattering. Reducing semiconductor materials to sizes comparable to the characteristic lengths of phonons has unveiled new physical mechanisms and engineering capabilities for thermal energy management and conversion systems.¹¹ However, yet to be established are a detailed understanding of phonon properties and characterization of scattering mechanisms at confined systems. This research thrust is to develop theoretical methods that are able to characterize phonon properties of a non-periodic system, and that can take into account the effect of the presence of heterogeneous materials, surfaces, interfaces, defects, or local strain in a non-empirical manner, without employing any fitting parameters. This research thrust must be accompanied by advances in thermal spectroscopies and imaging techniques that can image local temperature fields and heat-carrying terahertz phonons with nanoscale resolution^{12,13} in complex nano-integrated architectures.

The performance of photonic components for optical interconnects is strongly temperature dependent, requiring careful thermal management. It is not uncommon for a temperature drift of 10 degrees to drive a photonic integrated circuit outside of its operational tolerances. Consequently, there is a need for active temperature stabilization or homeostatic control in integrated electronic-photonic integrated circuits. The thermal management solutions should be implemented in the context of the heterogeneous integration scheme within the framework of a tightly coupled co-design and co-optimization.

For example, in the case of stacked 3D integration, this problem is exacerbated by the fact that thermal energy evolved in the electronic integrated circuit travels through the optical chip on its way to the external heat sink. The relatively high thermal impedance of the interconnects between the electronic and photonic integrated circuits creates a significant temperature gradient across the vertical “stack” of the integrated device, and reduces the overall performance. Thermal transport at the nanoscale in dense integrated circuits is fundamentally different from that at the macroscale and is determined by the distribution of phonon mean free paths and phonon dispersion in a material, the dimensions of the heat sources, the separation between them and the distance/heterogeneity of the path over which the heat is transported.

Recent experimental and numerical studies expose the inadequacy of Fourier’s law in nanomaterials, even when the characteristic dimensions are longer than the phonon mean free path.¹⁴ The ultra-high heat flux arising from heat sources with high-temperature gradient and super-low cross-sectional area, as well as the collective interactions of the closely spaced heat sources (Figure 7), induces remarkable deviations of nanoscale thermal energy transport from the traditional Fick’s Law. Several non-Fourier heat conduction theories have been developed to describe the heat transport in nanomaterials, including phonon dynamics, phonon hydrodynamics, thermomass theory, and extended nonlocal theories. There is a need for

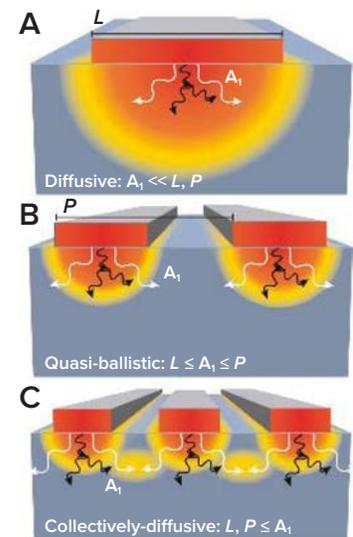


Figure 7. Nanoscale thermal transport in closely spaced thermal sources; when both length (L) and spacing (P) of heat sources shrink, both long and short mean free path phonons interact as if they originate from a single, larger heat source. From K. M. Hoogeboom-Pot et al., *PNAS*, 112 (2015) 4846-4851.

fresh thinking here applied towards the design of non-Fourier material for thermal transport in the context of tomorrow's integrated circuits with dense and heterogeneous features.

SCIENTIFIC AND TECHNOLOGY IMPACT

Translation of the scientific and engineering advances envisioned in this PRD to commercial hardware will help invigorate the global microelectronics and optoelectronics industry and thus advance the capabilities of computing systems. The benefits will then flow to ASCR, BES, and HEP facilities, as well as to other high-priority mission critical DOE projects. At the basic science level, the pursuit of this research will lead to discovery of new degrees of freedom that allow us not only to manipulate single-particle-based electronic and phonon flow, but also control the transmission of electronic/optical/plasmonic collective modes. At the technology level, breakthrough and advancement in the bandwidth, energy efficiency, reliability, and adaptivity of novel interconnect architecture will allow unconstrained movement and processing of abundant data.

REFERENCES

1. M. Sellier, J.-M. Portal, B. Borot, S. Colquhoun, R. Ferrant, F. Boeuf, A. Farcy, Predictive delay evaluation on emerging CMOS technologies: A simulation framework, *9th International Symposium on Quality Electronic Design*, DOI: 10.1109/ISQED.2008.4479784 (2008).
2. Ian Curtress, Intel's interconnected futures: Combining chiplets, EMIB, and foveras, <https://www.anandtech.com/show/14211/intels-interconnected-future-chipslets-emib-foveros> (2019).
3. R. Landauer, Spatial variation of currents and fields due to localized scatterers in metallic conduction, *IBM Journal of Research and Development*, 1 (1957) 223–231.
4. H.A. Atwater, The promise of plasmonics, *Scientific American* 296(4) (2007) 56-62.
5. Y. Fang and M. Sun, Nanoplasmonic waveguides: Towards applications in integrated nanophotonic circuits, *Light: Science & Applications*, 4 (2015) e294.
6. K. Liu, C.R. Ye, S. Khan, and V.J. Sorger, Review and perspective on ultrafast wavelength-size electro-optic modulators, *Laser Photonics*, 9(2) (2015) 172–194.
7. D.L. Sounas and A. Alù, Non-reciprocal photonics based on time modulation, *Nature Photonics*, 11 (2017) 774–783.
8. J.B. Pendry, L. Martín-Moreno, and F.J. Garcia-Vidal, Mimicking surface plasmons with structured surfaces, *Science*, 305 (2004) 847–848.
9. A. Hugi, G. Villares, S. Blaser, H.C. Liu, and J. Faist, Mid-infrared frequency comb based on a quantum cascade laser, *Nature*, 492 (2012) 229–233.
10. A.V. Krishnamoorthy, The intimate integration of photonics and electronics, *Advances in Information Optics and Photonics*, SPIE, Chapter 28 (2008).
11. S. Kwon, M.C. Wingert, J. Zheng, J. Xiang, and R. Chen, Thermal transport in Si and Ge nanostructures in the confinement regime, *Nanoscale*, 8(27) (2016) 13,155-13,167.
12. F. Menges, H. Riel, A. Stemmer, and B. Gotsmann, Quantitative thermometry of nanoscale hot spots, *Nano Lett.*, 12(2) (2012) 596–601.
13. F. Menges, P. Mensch, H. Schmid, H. Riel, A. Stemmer, and B. Gotsmann, Temperature mapping of operating nanoscale devices by scanning probe thermometry, *Nature Commun.*, 7 (2016) Article number 10874.
14. D.G. Cahill, W.K. Ford, K.E. Goodson, G.D. Mahan, A. Majumdar, H.J. Maris, R. Merlin, and S.R. Phillpot, Nanoscale thermal transport, *J. Appl. Phys.*, 93 (2003) 793.

PRD 4 Redefine computing by leveraging unexploited physical phenomena

INTRODUCTION

We are entering a period of great upheaval in the architecture of computing machinery. In 1936, Alan Turing proposed a model for carrying out computations a small step at a time, on a small “bit” of data, with a separation between the data being processed and the discrete and low level instructions that guided the computation. In 1945, John von Neumann wrote a document, the EDVAC Report,¹ that became the cornerstone on which the basic architecture of all modern computing systems has since been built. His model of computing was aimed at the solution of problems in arithmetic with extreme precision, and again separated the data from the sequence of instructions. His big innovation beyond the Turing model was that the commands that dictate each step could be placed in a memory themselves, akin to a step-by-step “recipe” that could be changed just like data.

For many decades, this von Neumann architecture has been a constant in computing. Speed and energy efficiency have improved exponentially, enabled primarily by relentless iterative miniaturization of devices and circuits and by relatively straightforward elaboration of the basic architecture through increasing parallelism. In the intervening years, there have been many forks in the road, but the tried-and-tested von Neumann architecture remains well suited to the arithmetic operations for which it was designed. However, continued extension of the von Neumann architecture is proving increasingly difficult: the field effect transistor—the dominant workhorse for digital computing—is now approaching some fundamental physical constraints to its further improvement. Additionally, data movement rates are limiting fast access to memory, and the lack of affordable, high bandwidth, and fast memory devices is constraining data intensive computing. Thus, the historical rate of gain in the performance and energy efficiency of high-performance computers has slowed. One response has been the exploration and development of fundamentally new devices for switching and memory, which can take digital computing beyond the limitations of the long-established device technologies. Another approach, rapidly gathering momentum at this time, is the exploration and development of fundamentally new architectures for computing.

While processors based on von Neumann architecture are well suited to sequential processing, they employ a fixed hardware structure with sequential instruction sets and fixed bus widths. Though they offer considerable flexibility in general purpose computing and ease of programming, they can be wasteful of throughput and energy for many important computing problems. For such problems, application-specific integrated circuits, digital signal processors, and field-programmable gate arrays routinely deliver one- to three-orders-of-magnitude improvements in computational performance and energy efficiency. This is because the associated design and programming models result in better mapping of the operations of the algorithms to the physical resources used to perform the operations. System architecture may, therefore, be moving toward heterogeneous systems consisting of a great many such specialized units executing specific operations. However, there is risk of complicating and encumbering the programming model. Perhaps more important, the number of different useful algorithms is large, and there seems to be little commonality of optimal hardware resources, even among algorithms seen from a software perspective as closely related.² The challenge, then, is to identify the most important and broadly applicable operations and algorithms and explore and develop optimal architectures for their execution. A prime example is the recent emergence of deep learning algorithms³ as the preferred approach to solve many long standing and important problems in pattern recognition.

The inefficiency of the von Neumann architecture for execution of these algorithms is driving wide-ranging experimentation in computer architecture. Thus, artificial neural networks (ANNs) are being aggressively explored as energy-efficient architectures for execution of machine learning algorithms and could become the most profound development in computer architecture in many decades. Most of this development is based on digital circuits implemented with long-established device technologies. However, attaining the full potential of ANNs may require the introduction of new devices that more compactly and efficiently implement key functions.⁴

The resulting systems may combine analog and digital devices and circuits. For example, much effort has gone in to the development and demonstration of analog memory devices to store the weights of synaptic connections, although the material properties and resulting device characteristics are still not ideal for this application.⁵

Beyond these ongoing developments, it is time to broadly reconsider the way we compute, and to reexamine the approaches that were eclipsed by the success of the von Neumann model. Many of these approaches focus on leveraging novel physical phenomena as dynamical systems that can express computation. Nonlinear optical systems can be engineered to behave as digital switches or as analog spatial light modulators for the solution of difficult optimization problems. DNA computing is a nascent field with much potential for the solution of problems with great combinatorial complexity. New devices that act as digital switches, yet retain their current state when the supply voltage is removed, may enable architectures that intimately merge computation and memory, reducing the communication penalty inherent in the von Neumann separation of these functions. Many other examples could be cited, and perhaps many are yet to be discovered.*

With diminishing returns from continued elaboration of the established devices and architecture, with growing excitement around machine learning and ANNs, *now* is the time to explore these broader possibilities. Various approaches must be identified and weighed for their capabilities. Formal models and physical demonstrations should be developed for the most promising approaches. The upside is large: we could extend the frontiers of science and engineering by solving problems unsolvable using traditional computational techniques, either due to energy constraints or algorithmic limitations or exponential growth in hardware requirements with the problem size. And new algorithms—new ways of solving hard problems—may be inspired by this work.

SCIENTIFIC CHALLENGES

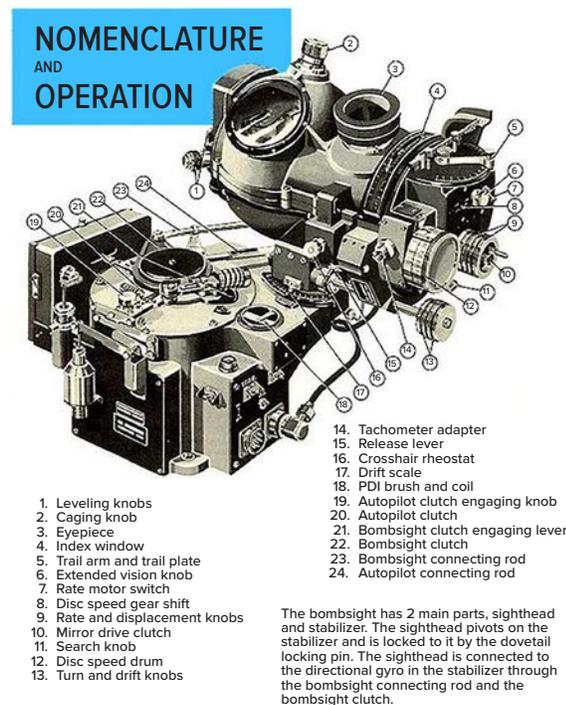


Figure 1. Example of an analog computer: the Norden bombsight was a highly sophisticated optical/mechanical analog computer used by the U.S. Army and Air Force during World War II, the Korean War, and the Vietnam War to aid bomber pilots. From https://en.wikipedia.org/wiki/Analog_computer.

We have come almost full circle. At the dawn of the computer era there was no broad appreciation of the advantages of digital devices and circuits, and analog approaches were pursued and developed by some of the greatest minds of the time. Now, as we broaden our horizons to include the possibility of radically new architectures and new physical systems for implementing those architectures, we look again to analog and hybrid analog-digital approaches. It is therefore instructive to recall some of the reasons why digital computing utterly eclipsed the early analog approaches. First, the available digital devices—the vacuum tube and then the bipolar transistor—were more compact than the analog devices of the time and lent themselves to continued miniaturization, and thus ease of manufacture and low cost per device. More important in the longer run, digital devices augmented by digital error correction delivered numerical precision and reproducibility limited only by the available physical resources—despite the presence of noise and the unavoidable variability of components. Also, since any function could be implemented as a digital function, the digital approach to computing turned out to be, in an important sense, universally applicable. Thus, many of our key research challenges involve finding the right trade-offs between the potential energy efficiency and performance of analog systems and the now obvious benefits of digital systems.

The challenges ahead revolve around answering several questions.

* Although qubit/gate-based quantum computing is yet another non von Neumann approach, it is receiving considerable attention elsewhere and thus is not the focus of this PRD.

How can we discover, explore, and understand physical systems that are particularly useful for computation?

Any physical system can be seen as computing its own evolution in time. Equilibrium and steady-state dissipative systems are uninteresting in this respect, but many dissipative systems exhibit fascinating dynamical behavior as they evolve over time from a non-equilibrium perturbed state to an equilibrium or quasi-equilibrium steady state. A subset of those is known to exhibit dynamics that can be mapped to the solution of important computational problems. For example, the dynamics of systems described by the Ising model correspond to the solution of many interesting optimization problems such as the traveling salesman problem or the graph-coloring problem. Driven systems of coupled oscillators are also promising in this respect. The most desirable systems will have broad applicability and flexible programmability, but can guidelines for the discovery of such systems be formulated?

For optimum energy efficiency and speed of operation, the system and the devices that compose it should be scalable to very small dimensions. For adoption, the systems must be manufacturable at low cost. With integrated manufacturing processes such as those used in the microelectronics industry, the cost per device tends to fall inversely with the number of devices integrated onto each substrate—so scalability to small dimensions is a powerful key to affordable systems. Of course, for any given set of manufacturing processes, a point is reached where further reduction in device dimensions leads to increasing yield loss and increasing cost per device. An economically viable fabrication process must balance integration complexity (enabled by device and circuit scalability) with yield. Recent years have seen marked advances in the demonstrated scalability of nanomagnetic devices, non-linear photonic devices, nanomechanical devices, and more. Any new system should be examined with scalability and manufacturability in mind.

What levels of precision, accuracy, reproducibility, and endurance are needed for various proposed applications?

Because machine learning algorithms, for example, deal with data from the noisy analog world, it is now commonly argued that these algorithms do not require digital precision, accuracy, and reproducibility. This may be true for many applications, but there are likely to be exceptions. Consider, for example, a self-driving car, where the algorithm guides actions that may have life or death consequences. In today's commercially available "neural network" chips, the implementation is entirely digital. Given a *digital* record of the input from the car's various sensors, the operation of the system can be reproduced with a high degree of certainty. The operations that led to an accident can, therefore, be replicated and ultimately understood. If the digital chip is replaced with an ANN in which synaptic strengths are encoded in analog memory, some reproducibility will be lost in return for improved energy efficiency. To what degree will this trade-off be acceptable, and for which applications? By extension, similar trade-offs are to be expected for any new architecture which encompasses analog devices and functions, and the appropriate trade-off will depend on the application.

What are the appropriate formal models and measures of complexity for these novel approaches to computing?

The guiding complexity models for modern computing are built on the Turing model. We anticipate that future computing in non-von Neumann architectures will be in one of two classes: (1) those that are Turing complete or (2) those that are used as accelerators for specific computations in conjunction with a von Neumann system. In the abstract, the foundation for time complexity modeling of modern systems will be applicable to class (1) computing approaches. However, significant theoretical work will be needed to bridge from the non-Von Neumann computational paradigms of class (2) approaches to the Turing model of class (1) to connect the new approaches to the long established and deep body of theoretical results for modern computing. The theoretical computer science community will need to be engaged in order to perform this work.

A separate aspect of complexity modeling is energy complexity.[†] Historically, it has been difficult to model the energy complexity of an algorithm since the energy usage depends on the hardware on which it runs. The field of non-equilibrium thermodynamics has expanded in recent years.⁶ This theory has the potential to provide a basis for energy complexity modeling of both von Neumann and non-von Neumann systems. With this modeling, energy complexity comparisons will become possible. This requires incentivizing a collaboration between the theoretical physics community working on open systems thermodynamics and the theoretical computer science community.

Are these new approaches to computing compatible with functional composability, particularly in heterogeneous systems?

Composability is a significant and powerful property of the modern computing paradigm. It is the property whereby modules that perform well-known functions can be composed into an ensemble to perform a more complex function. Two conditions must be satisfied for a system to allow such ensembles: (1) the modules, if they store state, must store state based on their past or current inputs alone, and (2) the composition of two modules in series must produce a linear composite of one module operating on the results of the prior. Not all systems share this property. Chaotic analog systems based on non-linear dynamics, for example, are of interest for, and should not be excluded from consideration of novel compute approaches. By definition, they are not composable with each other. Composability in general will require finding known boundaries between modules that provide for predictable, negotiated interfaces.

How should data be encoded?

Should data be encoded in discrete (digital) or continuous (analog) formats? As mentioned above, ongoing research and development of ANNs encompasses both approaches, with different groups championing one approach or the other. For encoding the strength of synaptic connections, the right answer depends on the desired trade-off between the accuracy, precision, and reproducibility of digital systems and the energy efficiency and performance offered by analog approaches. Similar trade-offs must be determined for any novel architecture that embraces analog data encoding. Sparse representations and processing of sparse data streams are intertwined and will require co-design of encodings with near memory accelerators.

Need for broad cross-disciplinary co-design

Present day computing integrates contributions from many disciplines and skills in a many-layered stack. Specialists in each layer are focused on models or abstractions that are built on, but largely independent of, the models and abstractions of layers below. Thus there has been little need for specialists within each layer to understand and communicate with those working below and above.

Inherent in finding novel ways to compute is the requirement that disciplines collaborate. (See Figure 2.) For the first time in many decades, materials scientists, device physicists and engineers, circuit designers, microarchitects, etc., on up to language and algorithm designers must work across the traditional layers of abstraction. This collaborative co-design will be difficult because the independence of the various layers has enhanced our collective ability to reason about and thus build highly complex systems.

For example, the pursuit of better electronic devices has traditionally motivated much materials research, and conversely, advances in materials research have often enabled improved devices. Materials and device researchers generally know how to work together. Now, truly new devices with the potential to address the current power-performance bottleneck are emerging, but such devices will not precisely duplicate the characteristics of the long-established devices. As a direct consequence, circuit designers are becoming much more engaged in early-stage materials and device research. Today, research on ANNs is driving interdisciplinary communication across the entire stack of computing disciplines illustrated in Figure 2. This is because the full potential of ANNs may only be attained with the introduction of new devices that more compactly and efficiently implement key network functions.⁷ Because the properties of known materials and characteristics of known devices are not ideal for this purpose, system architecture and algorithms directly guide materials and device research, and *vice versa*. Exploration of similar trade-offs will be key to the development of other new architectures, as discussed above. More generally, ongoing research illustrates many possibilities for profound

[†] Note that a measure of power complexity is enabled by models of energy complexity combined with time complexity since power is the rate at which energy is consumed.

advances in the capability of information technology hardware through advances in materials. Diverse examples are summarized in the sidebar at the end of this PRD.

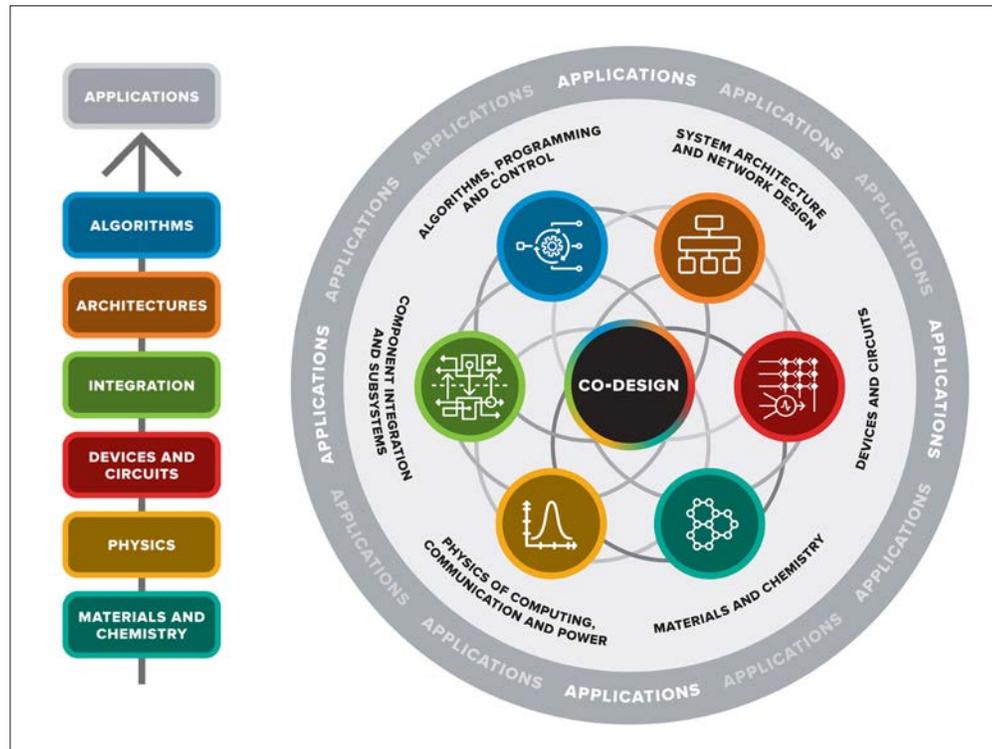


Figure 2. Levels of computing hierarchy. Computing is currently based on a layered stack of abstractions with each layer largely independent of those above and below. Emerging devices, such as the negative capacitance field effect transistor, bring “new physics” into the device operation. While similar to conventional field effect transistors, they will not be “drop in” replacements. New materials and chemistry, physics, and devices and circuits must be co-developed. These and other emerging devices may have characteristics that are well-suited to new analog or hybrid digital-analog computer architectures. In that case, all levels in the computing hierarchy must be co-developed.

RESEARCH THRUSTS

The overarching goal of the research thrusts given below is to leverage novel physical processes to perform useful computation. Both natural and man-made devices are ubiquitous, carrying out functions for computing, sensing, energy generation, force transduction, bio-regulatory operations such as protein folding, etc. The repertoire is very rich, with devices operating on electronic, mechanical, magnetic, optical, and bio-chemical principles. Examples include current CMOS transistors, bio-molecular machines such as mitochondria and ribosomes, ionic and memristive devices, spintronics, photonics, superconducting Josephson junctions, carbon nanotubes, nano/micro-electro-mechanical systems, CMOS compatible silicon micro-mirrors, DNA, systems of neurons, etc.

Over the 20th century, many, if not all, of the man-made devices have undergone tremendous scaling down in size, driven by economic imperatives (e.g., reducing manufacturing cost per device) and the pursuit of computational performance metrics (e.g., system clock speeds of gigahertz and lower switching energy of a digital transistor). Voltage scaling of the CMOS transistor has reduced system energy consumption by many orders of magnitude, although we remain far from limits set by thermal voltage fluctuations and acceptable error rates, and we have barely begun to explore reversible computing, which could, in principle, reduce energy dissipation in computing below the Landauer limit—the minimal energy loss associated with erasure of information (see below for further discussion). Importantly, as new low-voltage devices are introduced and device dimensions are further reduced, thermodynamic fluctuations will make devices increasingly stochastic in operation. The current paradigm for computing is inadequate for harnessing such stochasticity, even when it is in fact needed at the algorithmic level (e.g., randomized algorithms).

This suite of existing devices can be coarsely categorized as either (1) artificial devices, such as CMOS transistors, whose design is guided purely by computational and functional goals with thermodynamic constraints seen as ultimate limits, or (2) systems that evolve under strong thermodynamic constraints or collective interactions but with limited or no consideration for specific computational goals. Examples of the latter range from natural systems, such as mitochondria or systems of cellular neurons to artificial systems, such as coupled oscillators, for solving hard optimization problems. To date, no systematic survey of the computational potential of these class-2 devices has been performed.

Thrust 1. Optimization machines

Optimization problems are among the most computationally expensive problems challenging today's computers. Annealing has long been known as a natural process that appears to solve optimization in linear time. Relatively recently, new approaches to annealing have emerged, including quantum processes^{8,9} and modified Boltzmann machines implemented in hybrid analog/digital electronics.¹⁰ Other classes of devices, such as coupled oscillators¹¹ and evolutionary biosystems, are well known to have computational potential for optimization. There are likely many more such potential optimization machines that remain to be discovered. A targeted research thrust in the study of both existing and potentially new optimizers has yet to be determined and is urgently needed. A specific need is a systematic examination of how such systems can be synthesized and interconnected for the collective response sought. This direction opens up opportunities for designing and examining new classes of materials that are beyond conventional semiconductors and can be tailored for the response sought—related to electronic phase transitions; magnetic response; electronic, ionic, or electromagnetic excitations; or devices that rely upon positive feedback effects unlike field effect transistors, for instance. It also opens up the need for designing and synthesizing new ways of connecting these collective devices for signal transduction and power delivery—ways that can incorporate features such as self-configurability and adaptability.

Thrust 2. Computational models

In 1961, Rolf Landauer of IBM proposed a fundamental connection between information loss in computing and energy dissipation (increase in physical entropy).¹² According to Landauer's principle, each bit of information erased at temperature T results in a minimum of $kT \ln(2)$ of energy dissipation. At the simplest level, this follows directly from the second law of thermodynamics and the statistical-mechanical understanding of entropy as unknown information. Whenever some previously known (i.e., correlated) information is lost (e.g., to the environment) and is subsequently thermalized, with its prior correlations having become inaccessible, the entropy has, at that point, by definition, increased, and free energy has decreased.

The field of non-equilibrium thermodynamics has expanded in recent years and has given us much in the way of theoretical foundation for reasoning about the energy bounds of computational systems.⁵ The research thrust here is to combine the well-established field of complexity analysis based on the Turing model with the new results from non-equilibrium thermodynamics to create a new foundation to measure the complexity of approaches that will emerge in this research area.

Thrust 3. Partitioning computation between non-von Neumann and von Neumann architectures

One way to view non-von Neumann computational approaches is as a general replacement for modern techniques of computation. This would be a "total takeover" by these new computational devices. Although it should not be discounted as a possibility, the more likely scenario is that non-von Neumann and von Neumann architectures co-exist in a larger hybrid system. For solving a given problem, then, these hybrid systems will require algorithmic partitioning, potentially dynamically, between what is efficient in von Neumann and what is efficient in the new approaches. This requires a research thrust with extensive development of new languages, algorithms, and software tools. Much of this parallels the current push for specialized accelerators (e.g., sparse vector-matrix operation acceleration) added to existing compute platforms. Thus, we anticipate that the community created to reach those goals will also be interested in solving the more difficult problem of partitioning in non-von Neumann / von Neumann hybrid systems.

SCIENTIFIC AND TECHNOLOGY IMPACT

The performance of today's computing systems is overwhelmingly power constrained – limited by economically acceptable bounds on power and cooling across nearly all applications. The most important technological impact of the proposed research would be breaking these bounds. Each and every advance in energy efficiency will deliver broad benefits.

New digital devices for logic and memory can extend the supply voltage scaling limits of the conventional field effect transistor and thereby break the barriers of speed and efficiency in existing systems. Recent measurements on experimental devices and circuits⁷ suggest that roughly an order of magnitude improvement in energy efficiency with no sacrifice in performance can be achieved in the next few years. Further large improvements are physically possible,¹³ and the benefits of these digital device breakthroughs would flow to nearly all established applications of information technology (see sidebar). New architectures, aimed at algorithms and applications for which the von Neumann architecture is ill-suited, can multiply these benefits by additional orders of magnitude.

Without a doubt, such broad and profound improvements in energy efficiency would contribute to sustained US leadership in information technology. Computing for scientific discovery would be invigorated and accelerated. Furthermore, in thinking broadly and deeply about the intimate connections between computer architecture, algorithms, and energy efficiency, we may discover and develop new ways of reasoning about computation. We may invent better algorithms and architectures for tasks that are already commercially important. For example, many observers already predict ever broadening applications of machine learning as algorithms and architecture co-evolve. Perhaps more important, we may find new algorithms and architectures to solve previously intractable problems in areas such as optimization or real-time unsupervised learning. Once these problems become tractable, important commercial applications are likely to follow. Examples might include breakthroughs in intelligent autonomous systems (e.g., self-driving cars), portable devices freed from the constraints of battery capacity, internet-of-things devices that are self-powered and thus implantable in infrastructure, etc. In general, at every point in the current state-of-the-art where energy efficiency limits the applicability of computation to solve a problem, the research and development of non-von Neumann architectures, if successful, can overcome these limitations.

New devices for memory: Emerging random access memory (RAM) devices include spin-transfer torque magnetic RAM (STT-MRAM), ferroelectric RAM (FeRAM), conductive bridge RAM (CBRAM), resistive RAM (RRAM), and phase change memory (PCM).¹⁴ These very distinct devices, each based on a different class of materials, all share some highly desirable attributes. In particular, they can all be fabricated at relatively low process temperatures, enabling integration of memory devices directly above blocks of logic. This fine-grained integration of memory and logic is seen as a key to the implementation of energy-efficient logic-in-memory and memory-in-logic architectures. At the same time, each of these devices offers advantages and disadvantages compared to others and will advance only with further advances in the properties of its requisite materials. The write energies versus the memory cell area for a number of these technologies are shown in Figure 3. A challenge is to be able to reduce write energies to $\ll 100$ fJ with memory cell sizes that contain 100s rather than 10,000s of atoms.

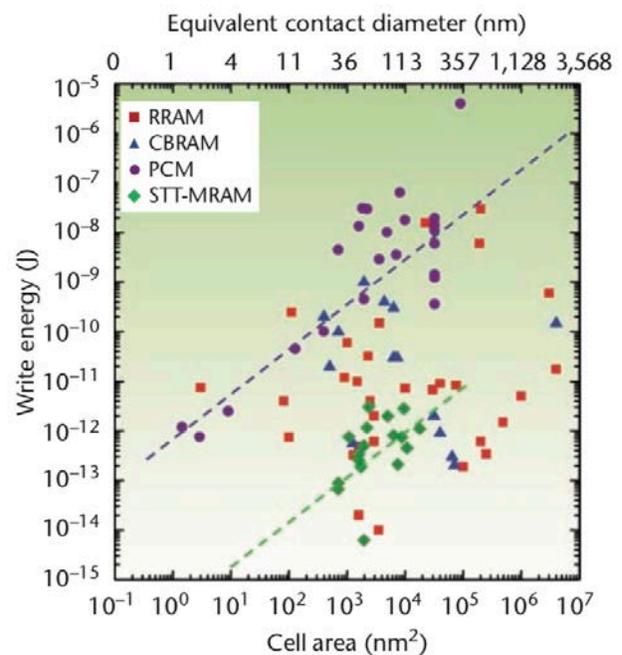


Figure 3. Programming energy versus memory cell area from published data for leading emerging nonvolatile memory technologies. Both STT-MRAM and PCM require a critical current density to switch the memory state; thus, the programming energy is proportional (purple and green dashed lines) to the memory cell area. Conduction in CBRAM and RRAM is filamentary; therefore, the programming energy is independent of the memory cell area. Data from <https://nano.stanford.edu/stanford-memory-trends>. Image from T.N. Theis and H.-S. Wong, *Computing in Science and Engineering*, 19 (2016) 41-50.

ONGOING RESEARCH ILLUSTRATES POSSIBILITIES FOR PROFOUND ADVANCES IN COMPUTING

Low-voltage, low-power devices: Many studies over the past fifteen years have investigated new classes of transistor-like devices that switch by various physical principles which are fundamentally different from the operating principle of the conventional field effect transistor, and can thus transcend some of the field effect transistor's (FET's) fundamental limits.¹² A rough taxonomy of such an exploratory device is illustrated in Figure 4. Among them, the negative capacitance field effect transistor (NCFET)¹⁵ is one example that is being pursued actively, enabled by a breakthrough discovery in thin-film ferroelectric materials.¹⁶ Other compelling low-voltage, low-power device concepts are less developed, perhaps awaiting other advances in materials science.

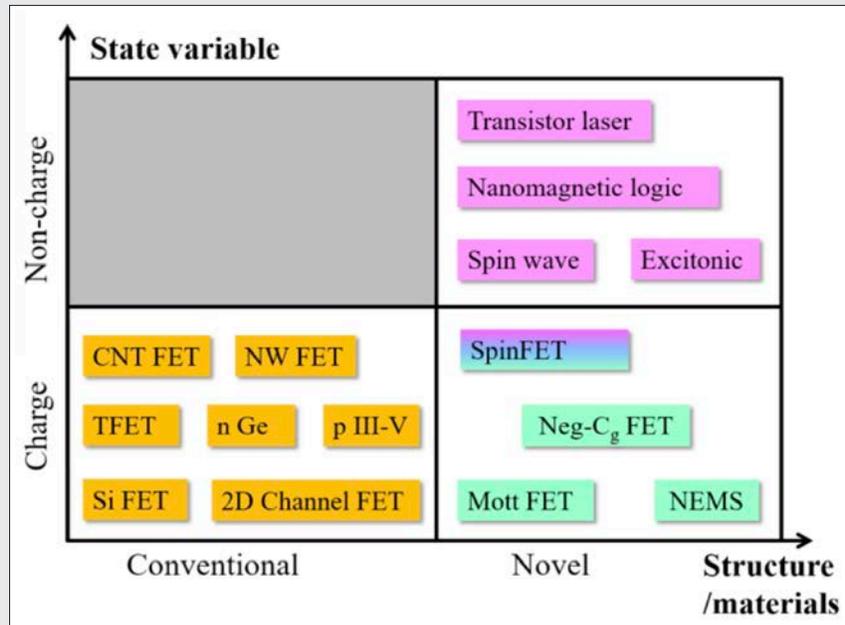


Figure 4. Taxonomy of some promising device concepts explored in recent years. Reproduced with permission, *International Roadmap for Devices and Systems*. Copyright (2017) IEEE. FET = field effect transistor, CNT = carbon nanotube, NW = nanowire, TFET = tunnel FET, n Ge = n-channel germanium, NEMS = nanoelectromechanical system.

New devices combining the functions of memory and logic: Magnetism has long been the basis for information storage devices, but recent materials discoveries have raised the possibility of magnetic devices for digital logic. With the digital state of each device represented by a persistent magnetic polarization, there would be no need to save the state of a computation before the power is turned off. This capability could be of immediate value in power-starved systems dependent on intermittent power sources. In the longer term, it could profoundly change computer architecture. First, however, the energy efficiency of magnetic polarization switching must be made competitive with that of electronic switching. One exciting research frontier is the study and demonstration of mechanisms for voltage-controlled magnetism in ferromagnetic and antiferromagnetic materials.^{17,18}

Nanophotonics and non-linear optical materials: Frontier research in nanophotonics is focused on demonstration and development of tiny non-linear optical devices based on the large non-linear optical coefficients obtainable with 2D materials integrated in nanometer-scale optical resonators.¹⁹ This research is enabling new technological opportunities in optical communications, including energy-efficient all-optical gates and 100 THz all-optical modulators. Such devices may allow computational functions to be distributed in optical networks for smart routing and management of data flow.

Neuromorphic devices: Artificial neural networks are being aggressively explored as energy-efficient architectures for execution of machine learning algorithms. However, attaining their full potential may require the introduction of new devices that more compactly and efficiently implement key network functions.⁴ For example, much effort has gone into the development and demonstration of analog memory devices to store the weights of synaptic connections, but the material properties and resulting device characteristics are still far from ideal for

this application.⁵ Some researchers are modifying algorithms to better match the device characteristics, while others pursue materials and device approaches that may be better suited to the well-established algorithms. Another example is provided by oscillatory dynamical systems – arrays of oscillators coupled by resistive and capacitive circuits that give rise to interesting and potentially useful phase and frequency dynamics, which can be controlled and mapped to the solution of computationally hard problems like graph coloring. Recent demonstrations involve insulator-to-metal transition devices and spin-torque oscillators, but the general approach can be translated to other hardware substrates, such as micromechanical and optical systems.

In general, the computational performance of a device will depend on the architecture of the system in which it is embedded. For example, some devices that are poorly matched to the demands of digital logic may have distinct advantages in the implementation of neural network architectures.

REFERENCES

1. John von Neumann, *First Draft of a Report on EDVAC*, University of Pennsylvania (1945).
2. V.C. Cabezas and P. Stanley-Marbell, Parallelism and data movement characterization of contemporary application classes, *Proc. 23rd Ann. ACM Symp. Parallelism in Algorithms and Architectures*, pp. 95–104 (2011).
3. Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature*, 521 (2016) 436–444.
4. W. Haensch, T. Gokmen, and R. Puri, The next generation of deep learning hardware: Analog computing, *Proceedings of the IEEE*, 107(1) (2019) 108-122. DOI: 10.1109/JPROC.2018.2871057.
5. T. Goken, M. Rasch, and W. Haensch, Training LSTM networks with resistive cross-point devices, arXiv:1806.00166 (2018).
6. C. Jarzynski, Nonequilibrium equality for free energy differences, *Phys. Rev. Lett.*, 78 (1997) 2690.
7. D. Kwon et al., Response speed of negative capacitance FinFETs, 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, pp. 49-50 (2018). DOI: 10.1109/VLSIT.2018.8510626.
8. T. Kadowaki and H. Nishimori, Quantum annealing in the transverse Ising model, *Phys. Rev. E*, 58 (1998) 5355. DOI:10.1103/PhysRevE.58.5355.
9. A.B. Finilla, M.A. Gomez, C. Sebenik, and J.D. Doll, Quantum annealing: A new method for minimizing multidimensional functions, *Chem. Phys. Lett.*, 219 (1994) 343.
10. S. Kumar, J.P. Strachan and R.S. Williams, Chaotic dynamics in nanoscale NbO₂ Mott memristors for analogue computing, *Nature*, 548(7667) (2017) 318-321. DOI: 10.1038/nature23307.
11. A. Raychowdhury, A. Parihar, G.H. Smith, V. Narayanan, G. Csaba, M. Jerry, W. Porod, and Suman Data, Computing with networks of oscillatory dynamical systems, *Proceedings of IEEE*, 107(1) (2019) 73-89. DOI: 10.1109/JPROC.2018.2878854.
12. R. Landauer, Minimal energy requirements in communication, *Science*, 272 (1996) 1914.
13. T.N. Theis and P.M. Solomon, In quest of the “next switch”: Prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor, *Proceedings of IEEE*, 98(12) (2010) 2005-2014. DOI: 10.1109/JPROC.2010.2066531.
14. T.N. Theis and H.-S.P. Wong, The end of Moore’s Law: A new beginning for information technology, *Computing Sci. Engin.*, 19(2) (2017) 41-50.
15. S. Salahuddin and S. Datta, Use of negative capacitance to provide voltage amplification for low power nanoscale devices, *Nano Lett.*, 8 (2008) 405-410.
16. S. Mueller, J. Mueller, A. Singh, S. Riedel, J. Sundqvist, U. Schroeder, and T. Mikolajick, Incipient ferroelectricity in Al-doped HfO₂ thin films, *Adv. Function. Mater.*, 22(11) (2012) 2412-2417.
17. S. Fusil, V. Garcia, A. Barthélemy, and M. Bibes, Magnetoelectric devices for spintronics, *Ann. Rev. Mater. Research*, 44 (2014) 91-116.
18. C. Song, B. Cuia, F. Li, X. Zhou, and F. Pan, Recent progress in voltage control of magnetism: Materials, mechanisms, and performance, *Progress Mater. Sci.*, 87 (2017) 33-82.
19. A. Autere, H. Jussila, Y. Dai, Y. Wang, H. Lipsanen, and Z. Sun, Nonlinear optics with 2D layered materials, *Adv. Mater.*, 30 (2018) 1705963.

This page intentionally left blank.

PRD 5 Reinvent the electricity grid through new materials, devices, and architectures

INTRODUCTION

Our national and economic security depend on continued leadership in science and technology. Just as Moore's Law has drastically driven down the price of computing, the electrification of numerous industries has enabled new functionality, increased performance, and decreased cost. However, while the U.S. has been rapidly moving toward further electrification in several market segments, the delivery system for electricity – the grid – has been relatively static in the level of its technology. The aging U.S. power grid, fundamentally unchanged for over a hundred years, is presently unable to meet the demands associated with a growing and more “plugged-in” population.

A resilient and reliable electric grid is of utmost importance to national security. To emphasize this, the blackout that affected the most households in North American history, the Northeast Blackout of 2003, caused 50 million people to be without power for 2 days, resulted in 11 deaths, and cost an estimated \$6 billion.¹ Today, an enduring blackout on this scale would likely be significantly more catastrophic, both in terms of human lives lost and dollars spent to restore the grid, due to society's increased reliance on electricity-enabled basic services, especially telecommunications.² Additionally, as more renewable generation is added to the grid, a much greater degree of flexibility will be required. The traditional “hub-and-spoke” model of large, fixed sources supplying distributed loads will no longer apply — rather, sources will be not only geographically distributed, but also intermittent in time (Figure 1).³ This new grid architecture will require advanced power electronics to control the flow of power from sources not only to loads, but also to extensive energy storage capacity. Indeed, it is estimated that by 2030, 80% of power in the grid will flow through power electronics.⁴

Additionally, as computing utilizes a larger and larger fraction of the total available electricity (e.g., due to expanding numbers of server farms), the problem of efficiently and economically converting medium-voltage AC power from a distribution feeder to low DC voltages (1-5 V) for powering microprocessors has become a significant economic issue. While several power distribution architectures may potentially be deployed, existing technology accomplishes this through four to six rectification and step-down conversion stages, each of which is between 85% and 99% efficient. The combined efficiency of all the stages is thus typically below 80%, resulting in significant wasted energy and exacerbating the expense and challenge of heat management. Furthermore, multiple conversion stages result in a mass of bulky power conversion hardware. The development of new power circuit topologies, capable of high-ratio step-down power conversion in as few stages as possible (including package-level and/or chip-level power distribution), is needed to enable efficient conversion. Similarly, vehicle electrification is a key element in curbing carbon emissions as well as maintaining an economic advantage in an increasingly competitive global marketplace. Finally, today's power management technology is not sufficiently robust to support the stringent power requirements for applications such as next-generation scientific facilities, the “smart” grid, the IoT, and beyond-Moore computing centers.^{5,6}

The key to enabling a high-performance smart electrical grid that can meet the power requirements and deliver resiliency for the future is one based on *widespread power conversion systems (PCSs) enabled by advanced power electronics*. A prototypical example is the solid-state transformer, which not only allows for cost and volume reduction, but also enables dramatically expanded functionality. However, this solid-state PCS has thus far been unavailable at the power levels required for application in the grid.

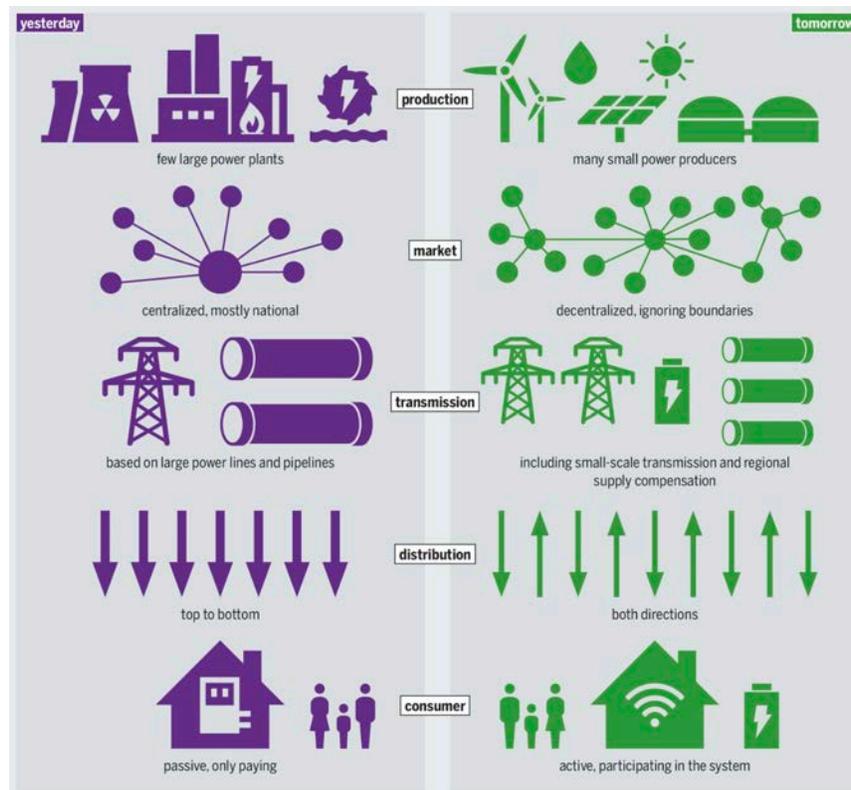


Figure 1. Characteristics of the future smart grid (right) compared to the traditional electrical distribution system.
From https://en.wikipedia.org/wiki/Smart_grid.

A new generation of ultra-wide-bandgap (UWBG) semiconductors for power electronics is required to enable PCSs that are higher power, more efficient, smaller, and cheaper. These, in turn, will enable a smart grid that is higher performing and more resilient than what exists today.

Today's mature semiconductor technologies have benefitted from past fundamental investments in substrates, epitaxial growth, defect science, and processing technology, and similar investments are now required to advance the next generation of power semiconductors. Today, the laboratory-scale availability of suitable substrates, coupled with new epitaxial growth capabilities (e.g., the ability to grow at very high temperatures), has resulted in initial laboratory demonstrations of device-quality epitaxial material and preliminary device demonstrations. We are now at the stage where expanded basic research is needed to fully understand these materials so that UWBG-based power electronics can be realized. Discovering new power semiconductor materials and bringing them to a state of practicality will require development and utilization of modeling at many levels, advances in growth and defect science, and development and application of new and advanced characterization techniques. Further, before efficient and compact power electronics based on high-switching-frequency UWBG power devices can be realized, new magnetic and dielectric materials that are capable of operating at high frequency, power, and temperature, as well as unconventional integration and thermal management techniques, must be similarly studied and brought to fruition.

UWBG semiconductors have the potential to achieve levels of performance for high-power switching electronics that are substantially superior to those of currently available conventional (e.g., silicon) and even state-of-the-art wide-bandgap (e.g., silicon carbide and gallium nitride) semiconductors.⁷ As discussed in the Panel 3 report, many of the figures of merit for device performance scale super-linearly with bandgap, with the consequence that UWBG semiconductors potentially offer compelling advantages over their narrower bandgap counterparts for power conversion applications. A primary advantage is a huge increase in power density (Figure 2), leading to orders-of-magnitude decreases in power converter weight and volume.

SCIENTIFIC CHALLENGES

For all known UWBG semiconductors, and likely for any new materials in this category that may be discovered in the future, several key scientific challenges exist. First among these is the ability to synthesize substrates of suitable quality, purity, and size to enable epitaxy of device-quality layers. Some low-defect-density UWBG substrates are available on the market today, such as aluminum nitride (AlN) (Figure 3), gallium oxide (Ga_2O_3), and diamond (Figure 4), but their sizes are inadequate for commercial production. Difficulties in increasing the size of high-quality UWBG substrates arise mainly from the need to employ vapor-phase processes for their growth (except for Ga_2O_3) and the lack of large-area, high-quality seed crystals (for materials where liquid-phase approaches are available, the seed crystals can be self-grown). In some cases, approaches such as stitching together smaller substrates and the use of non-native seeds have been tried but suffered from defect generation at the stitched interfaces and across the wafer, respectively. Thus, these techniques are not good long-term solutions, and new approaches to the growth of high-quality, large-area substrates are required. Specific problems that must be avoided are dislocations, impurities and other point defects, and wafer bow.

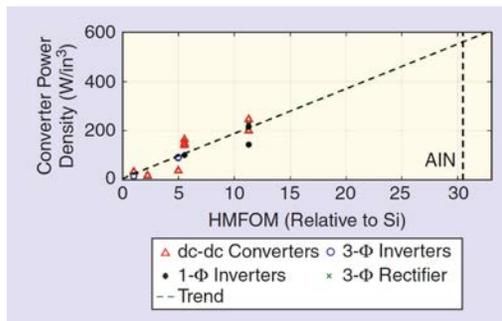


Figure 2. The volumetric power density as a function of Huang material figure of merit (HMFOM) for several power converters, extrapolated out to the HMFOM for AlN. From Robert J. Kaplar et al., *IEEE Power Electronics Magazine*, 4(1) (2017) 36-42.



Figure 3. Array of 25-mm AlN boules used to make substrates for epitaxial growth. Although substrates up to 2-in. dia are commercially available, larger substrates are needed for economic feasibility. Courtesy of Greg Mills, HexaTech.



Figure 4. Diamond homoepitaxial growth. Courtesy of Bob Nemanich and Franz Koeck, Arizona State University.

Challenges also exist in the epitaxial growth of UWBG semiconductors. The much higher bond strengths in UWBG semiconductors compared to conventional semiconductors and even wide bandgap semiconductors will require exploration at extreme growth conditions, such as very high temperatures and pressures. The growth of quaternary UWBG alloys such as BaInGaN is desirable because of their potential for creating lattice-matched heterostructures. But such growth is particularly challenging due to the different growth conditions required for the binary constituents, and progress will require fundamental new insights into growth kinetics. Extreme growth conditions will also fundamentally impact the incorporation and properties of impurity dopants, as well as unintentional compensating impurities and native point defects.

Fundamental scientific questions related to impurity doping and point defects are universal amongst all known UWBG semiconductors, and thermodynamics dictates that as-yet undiscovered UWBG materials will face the same challenges. Impurity dopant ionization energies generally increase with bandgap, and as the ionization energy increases, the fraction of carriers that are thermally excited to the conduction and valence bands decreases exponentially. This makes it very difficult to achieve the free carrier concentrations that are required for viable devices. Additionally, UWBG materials are susceptible to self-trapping of carriers, particularly for holes. Thus, even if suitable shallow acceptors can be identified, this tendency for carrier localization will suppress carrier mobility. Finally, dopants in UWBG semiconductors tend to be compensated by impurities, native defects, or defect complexes, and new insights into the relationship between growth conditions and defect formation will be required to achieve controllable doping. Fundamentally new approaches to doping that circumvent the use of thermally activated impurities may be needed – for example, the exploitation of the polarization properties of UWBG crystals to create free carriers.

Another scientific challenge for UWBG semiconductors is related to the very high electric fields at which the devices based on them will operate. These high fields will drive physical phenomena much farther from equilibrium than occurs in conventional and even today's wide bandgap semiconductors. Exploration of such regimes that have not previously been studied is expected to lead to the discovery of new phenomena and concepts related to carrier dynamics and transport (including breakdown phenomena), impurity scattering, electron-phonon coupling, photon emission and absorption, photoelectron emission, and thermal conduction.

Finally, as the bandgap of the semiconductor increases, the number of suitable semiconductor/insulator pairings with band offsets sufficiently large to achieve carrier confinement decreases significantly, creating a problem for the realization of field-effect devices. Moreover, surface and interface polarization charges may significantly influence the properties, performance, and requirements of passivation, isolation, and confinement by the layers. The breakdown fields of UWBG semiconductors (5-15 MV/cm) may be comparable to those of popularly employed dielectrics, such as Al_2O_3 , SiO_2 , and SiN_x . A key performance requirement for insulators in power electronic devices is that they have much higher breakdown fields than the active semiconductor with which they are integrated. Thus, tailoring the relative breakdown strengths of UWBG semiconductors and compatible insulators is an important and significant challenge. This difficulty is compounded because the bandgaps and dielectric constants for insulators — which are both desired to be large — tend to be inversely related. Overall, identifying semiconducting and/or insulating materials for carrier confinement, as well as understanding the electronic structure of their interfaces (including defect states) with UWBG semiconductors, is a universal and fundamental scientific challenge.

UWBG semiconductors will also face significant scientific challenges associated with their thermal properties. The temperature distribution and thermal fields in future UWBG power devices will play a critical role in their performance, efficiency, and reliability. Typically, methods to manage the thermal fields in electronics are left to packaging engineers after the device architecture has been developed. Future systems will require a full electro-thermo-mechanical co-design approach that can implement thermal solutions directly in the device architecture to improve the thermal dissipation of UWBG devices. Critical parameters that must be measured or accurately modeled include process- and temperature-dependent thermophysical properties, as well as interfacial thermal resistance at device contacts. For low-thermal-conductivity UWBG materials (e.g., Ga_2O_3 and alloyed nitrides), it is expected that thermal energy will need to be efficiently extracted from interfacial thermal contacts to these materials into high-thermal-conductivity substrates, or through the electrical contacts on the devices.

For high-thermal-conductivity UWBG materials (e.g., diamond), thermal mismatch in terms of the phonon density of states will also create additional challenges for heat dissipation across interfaces. Overall, thermal interfaces will require special attention as they can form bottlenecks for heat dissipation. The key will be discovery science of electrical and thermal contacts between semiconductors and metals, as well as interfaces between different semiconductor layers in UWBG heterostructures and heterogeneously integrated devices. At present, appropriate models are lacking an electro-thermo-mechanical co-design methodology that can yield appropriate property-processing relationships for the full behavior at interfacial contacts to UWBG devices. Finally, methods for high-resolution characterization of electric fields and thermal gradients in devices will be necessary to better understand device performance and to link these topics to fundamental failure and reliability mechanisms. Understanding the physics of failure through multiscale modeling will be critical for the development of high-performance, high-reliability UWBG devices.

The increasing demand for high power density and reduced footprint requires higher switching frequency for power electronic systems, which is a primary motivator for the development of UWBG semiconductors. At higher switching frequency, the size of the energy storage inductors and capacitors decreases, and as such the power density of the power electronic system increases. However, in addition to the semiconductors, development of a wide range of novel materials is required to enable this capability. Of particular importance are magnetics and dielectrics, and as such, high-frequency and high-power magnetic and dielectric materials must be concurrently developed with UWBG semiconductors. This is necessary so that the advantageous properties of the UWBG semiconductors will be fully realized at the system level. Similarly, the performance of UWBG-based power converters will be limited not only by semiconductor, magnetic, and dielectric components themselves, but also by the packaging of these components and by their integration with other circuit components. For example, very fast switching causes parasitic inductance to play a prominent role in the performance of the converter — the high slew rates of voltage and current produced by fast switching of UWBG power devices are expected to significantly degrade circuit performance unless precautions are taken to mitigate these effects. Accordingly, parasitic inductance must be minimized for optimum high-speed switching of UWBG devices, and new design and fabrication techniques are required to minimize its impact. Thermal management is also a large concern, and novel packaging materials, integration methods, and thermal management techniques (including transport across complex multi-material interfaces at the package and system levels) will be essential to overcome many of the performance limitations anticipated for next-generation UWBG-based power electronic systems.

RESEARCH THRUSTS

Thrust 1. Investigate materials science (growth, doping, defects, and transport) in UWBG semiconductors

The universal scientific challenges for UWBG semiconductors presented in the previous section, including the science of bulk and epitaxial growth, doping and defect physics, high-field phenomena such as carrier velocity saturation and avalanche breakdown, and interface and confinement physics all need to be addressed experimentally and theoretically. Equally important is the creation of a framework to predict the existence of as-yet undiscovered UWBG materials, which will involve the prediction of the band structure of new crystals and the ability to ascertain whether theoretically predicted materials can be practically grown and fashioned into devices. Such a predictive capability will likely require a computational electro-thermo-mechanical co-design framework spanning various levels (atomistic, device, and system) as well as an efficient means to couple predictions to experimental evaluation.

Thrust 2. Integrate UWBG semiconductors with other materials

Regarding the processing and device-level integration of UWBG semiconductors with other materials, several research efforts are needed. As discussed earlier, dielectric passivation and isolation are essential for the fabrication of UWBG power devices, and significant challenges are anticipated in this area for UWBG semiconductors. Hence, detailed studies are necessary to demonstrate suitable dielectrics for UWBG materials that satisfy key requirements such as low interface state density, high breakdown fields relative to the UWBG semiconductors, sufficiently large band offsets with the semiconductors, and the ability to serve as effective diffusion barriers for metals. Further, advances in metal-semiconductor interface physics are required, especially related to the formation of ohmic contacts, because deposited metals tend to form high Schottky barriers on UWBG semiconductors. In addition, new approaches are needed to form low-resistance ohmic contacts to enable the UWBG devices to be integrated with other elements in the power conversion system.

Thrust 3. Develop inductor magnetics and capacitor dielectrics for UWBG semiconductor circuits

High-frequency, high-power magnetics will be required to take full advantage of the advances made in UWBG semiconductors. The next generation of power conversion electronics will require revolutionary changes in soft magnetic materials and magnetic component design. Inductors will no longer be fabricated as discrete, hand-wound devices that must be soldered in place, and instead will be highly integrated with advanced semiconductors and capacitors. This will require advanced magnetic materials that can be embedded into various substrates and are amenable to micro-processing techniques. As power electronics are pushed to ever-higher operating frequencies, eddy currents must be significantly decreased or eliminated. Insulating magnetic materials will be required, but magnetic saturation must not be sacrificed, so that power density can be maintained or increased. Better understanding and control of domain wall motion and its analog in composite magnetic materials will be crucial to keep losses low and performance high. Further, advanced characterization techniques will be required to study magnetization reversal at the nanoscale in all classes of materials.

Similarly, as next-generation power electronics push temperature, frequency, and voltage to higher levels, dielectric materials for capacitors will face difficult challenges related to high-temperature operation and lifetime (to be clear, these dielectrics comprise power circuit elements and serve a very different purpose than do the dielectrics discussed above in relation to field-effect device architectures). Both high-temperature operation and low equivalent series resistance, which are necessary to avoid self-heating due to high-frequency operation and ripple currents, will be required. These needs push past the limits of today's electrolytic capacitors as well as common polymer dielectric capacitors, such as polypropylene and polyethylene. Better understanding is needed of the self-healing behavior of high-temperature dielectrics, as well as new materials and techniques for increasing the probability of self-healing in high-temperature polymers.⁸ Additionally, while results have been promising for novel ceramic dielectrics that retain high dielectric constants at high temperature,⁹ many challenges concerning processing and lifetime remain. An understanding of the degradation effects in new capacitor materials is needed, as is material and circuit co-design that potentially allows self-repair of dielectric materials during circuit operation.

Thrust 4. Develop materials and architectures for thermal management (also relevant to PRD 2 and PRD 3)

The integration of thermal dissipation techniques within the UWBG semiconductor devices (electro-thermal-mechanical co-design), with particular emphasis on interfaces, is likewise extremely important for the development of compact, power-dense electronic components. Advances outside of the switching device in packages and power modules, as well as their interfaces with the device, are also needed. The development of high-temperature, reliable, high-performance thermal interface materials, methods of forming interfaces with low thermal resistance, and robust interconnects will be critical enablers. New materials that can facilitate intentional manipulation of heat flow will be very helpful in terms of enabling heterogeneous integration of multiple components with different temperature capabilities within a small volume (e.g., a 250°C-rated UWBG switch placed close to a 125°C-rated capacitor). New materials for encapsulants, baseplates, and heat exchangers will also be necessary. Multi-scale models to predict defect initiation and propagation and to help develop predictions of lifetime will also be required, as will multi-physics models to facilitate electro-thermo-mechanical co-design of packages and modules. Finally, also needed are integration methods to mitigate electrical parasitics, such as monolithically integrated gate drive electronics, embedded circuit elements, and 3D interconnection schemes.

SCIENTIFIC AND TECHNOLOGY IMPACT

Advances in power electronics have historically come from several directions. These include improved components, such as semiconductor switches based on new materials and new device structures that enable higher frequency, voltage, and temperature operation. Likewise, innovative circuit topologies and the associated control strategies required to achieve high performance while utilizing imperfect components are important. Finally, better system architectures for power conversion are critical. Future advances in power electronics will continue to come from innovations at the material, component, circuit, and system levels, and there is need for basic research at all of these levels. Fundamental research at the materials level will result in new understanding of the growth of UWBG semiconductors. Additionally, new physics concerning the behavior of these materials under conditions of extreme electric field, temperature, charge density, and radiation will be revealed. The ability will be gained to integrate these novel semiconductors with dielectrics and metals, as will the understanding and control of interfaces between these materials. Fundamental understanding of magnetic and dielectric materials that can handle high power and frequency will also be developed. New understanding of thermal transport phenomena, both within bulk UWBG materials and across interfaces of these materials not only with each other (i.e., heterostructures) but also with dielectrics and metals, will also be uncovered. Finally, related devices, such as new detectors and sensors, ultraviolet photonics, cold cathodes, and single-photon emitters, will likely be developed, with this effort leading to new physical understanding regarding these types of devices.

The applications that are driving innovation in solid-state power conversion and control (i.e., power electronics) include transportation electrification (ground, marine, and air), renewable energy generation, energy storage, grid modernization (such as solid-state transformers and DC distribution), and electronic loads (such as light-emitting diode lighting and data centers). All of these applications demand power electronics to be more efficient, more reliable, smaller, and less expensive. Successful deployment of power electronics in these applications can have a tremendous impact on our society. For example, if road transportation (which accounts for 23% of U.S. total energy consumption, compared to 38% for electrical power generation¹⁰) is fully electrified, it will increase the electrical energy consumed in the U.S. by more than 60%. This presents a tremendous opportunity for innovation and discovery in microelectronics to support the development of a sustainable electrical energy future. Another example is the electric grid, as mentioned at the outset, which must be adaptive, resilient, and reliable. A key enabler for this is the substation-scale solid-state transformer, which will be able to not only replace today's traditional transformers, but also enable additional functionality, including phase and frequency decoupling, reactive power control, power quality management, natural integration of DC sources/loads (e.g., photovoltaics combined with energy storage), and integration of variable frequency sources (e.g., wind). It is anticipated that employing UWBG semiconductor materials and devices, along with new magnetic and dielectric materials coupled through an electro-thermo-mechanical co-design approach, will allow compact and efficient solid-state transformers that can be integrated into a "substation in a suitcase", the key component enabling the next-generation flexible and resilient electric grid.¹¹

REFERENCES

1. U.S.-Canada Power System Outage Task Force, *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, <https://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf> (2004).
2. J.S. Foster, E. Gjeldre, W.R. Graham, R.J. Hermann, H.M. Kluepfel, R.L. Lawson, G.K. Soper, L.L. Wood, and J.B. Woodard, *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack*, http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf (2005).
3. Wikipedia, Smart grid, https://en.wikipedia.org/wiki/Smart_grid, accessed June 2019.
4. L. M. Tolbert, B. Gzpineci, J.B. Campbell, G. Muralidharan, D.T. Rzy, A.S. Sabau, H. Zhang, W. Zhang, X. Yu, H.F. Huq, and H. Liu, *Power Electronics for Distributed Energy Systems and Transmission and Distribution Applications*, <https://info.ornl.gov/sites/publications/Files/Pub57485.pdf>, p.1-1 (2005).
5. National Academies of Sciences, Engineering, and Medicine, *Enhancing the Resilience of the Nation's Electricity System*, National Academies Press, Washington, DC, Chapter 4, <https://doi.org/10.17226/24836> (2017).
6. *Grid Modernization Multi-Year Program Plan*, U.S. Department of Energy, <https://www.energy.gov/sites/prod/files/2016/01/f28/Grid%20Modernization%20Multi-Year%20Program%20Plan.pdf> (2015).
7. R.J. Kaplar, J.C. Neely, D.L. Huber, and L.J. Rashkin, Generation-after-next power electronics, *IEEE Power Electron. Mag.*, 4(1) (2017) 36-42.
8. M. Rabuffi and G. Plcci, Status quo and future prospects for metallized polypropylene energy storage capacitors, *IEEE Trans. Plasma Sci.*, 30 (2002) 1939-1942.
9. A Zeb and S. J. Milne, High temperature dielectric ceramics: A review of temperature-stable high-permittivity perovskites, *J. Mater. Sci.: Mater. Electron.*, 26(12) (2015) 9243-9255.
10. U.S. Department of Transportation, Bureau of Transportation Statistics, *Transportation Statistics Annual Report*, Chapter 7, <https://doi.org/10.21949/1502596> (2018).
11. D. Boroyevich, CPES Research: SSPS – *Building Blocks for the Future Electronic Power Grid*, U.S. Deptment of Energy Solid-State Power Substation Roadmapping Workshop, North Charleston, SC, June 27, 2017, <https://www.energy.gov/sites/prod/files/2017/09/f36/%5B7%5D%20VT%20-%20Dushan%20Boroyevich.pdf> (2017).

This page intentionally left blank.

3. Panel Reports

The Basic Research Needs Workshop for Microelectronics was structured around four panels, including a panel focused on crosscutting themes:

PANEL 1: MICROELECTRONICS FOR BIG DATA AT FUTURE FACILITIES: MEMORY AND STORAGE

PANEL 2: CO-DESIGN FOR HIGH PERFORMANCE COMPUTING BEYOND EXASCALE

PANEL 3: POWER CONVERSION, CONTROL, AND DETECTION

PANEL 4: CROSSCUTTING THEMES

These reports formed the basis for identifying the five PRDs described in Chapter 2.

This page intentionally left blank.

Panel 1 Microelectronics for Big Data at Future Facilities: Memory and Storage

INTRODUCTION

The DOE Office of Science manages, or is closely involved in research at, several large-scale facilities that are unique and critical to progress in the sciences. These facilities involve thousands of researchers worldwide, whose work impacts a variety of fields. Examples include the five DOE Basic Energy Sciences (BES) light (X-ray) sources, which impact fields such as physics, materials science, and biology. Another example is the DOE High Energy Physics (HEP) researchers using the ATLAS (A Toroidal LHC Apparatus) and CMS (Compact Muon Solenoid) particle detectors in the Large Hadron Collider (LHC). As these and other facilities are continually updated with sensing and measurement capabilities that push the limits of spatial, temporal, and energy resolutions, the demands on data storage and data processing multiply. For instance, by 2028, the data generation in the five light sources will be in the exabyte (EB) range, and data (on disk) needs for the ATLAS and CMS detectors at the LHC will be ~6 EB. These represent increases by factors of 20-24 from today, and it is anticipated that these data handling needs will further multiply over the next few decades. Our current technology is inadequate for meeting these data storage, data movement, and on-the-fly processing challenges.

In light of these challenges, Panel 1 met to discuss the microelectronics research needs for the future that would be required for data intensive and edge computing. The panel identified five principal areas of research: memory and storage, computation, communication, heterogeneous integration, and new tools and methods to facilitate this research. These needs are complex and interconnected, and include the roles of data storage densities and latency, as well as energy costs in transporting, computing, and storing data. It was clear from the discussions in Panel 1 that a major leap forward would need the discovery of new materials, new ways of fabricating materials in three dimensions at high spatial resolutions, the use of novel physics approaches, and the design of new types of devices. The description of needs and the recommendations from the panel are described in detail below.

SCIENCE DRIVERS

Current Status and Recent Advances

This panel began its discussion with the introduction of several science driver descriptions specific to DOE science and technology needs, and with characteristics that may not necessarily be found in industry.

High-luminosity LHC data and computing challenges

The LHC experiment at CERN is known for its extreme data volumes, which are currently hardware filtered at acquisition time to less than 10% of their original size. The trigger rate is currently 1 kHz, with detected pileups of 20 for a luminosity of 30 (fb⁻¹).* Over the next 7 years, the LHC will undergo significant upgrades, introducing new technologies to increase the luminosity available for experiments. The new high-luminosity LHC (HL-LHC) will have a luminosity of 3000 fb⁻¹ (Figure 1). Consequently, there will be higher data rates, ~1 petabyte (PB)/s, and more complex events: 5-10 kHz trigger rate and 150-200 pileups.

* The inverse femtobarn (fb⁻¹) is a measurement of particle-collision events per femtobarn (a measure of area; one femto barn equals 10⁻⁴³ m²). One fb⁻¹ corresponds to approximately 10¹² proton-proton collisions.

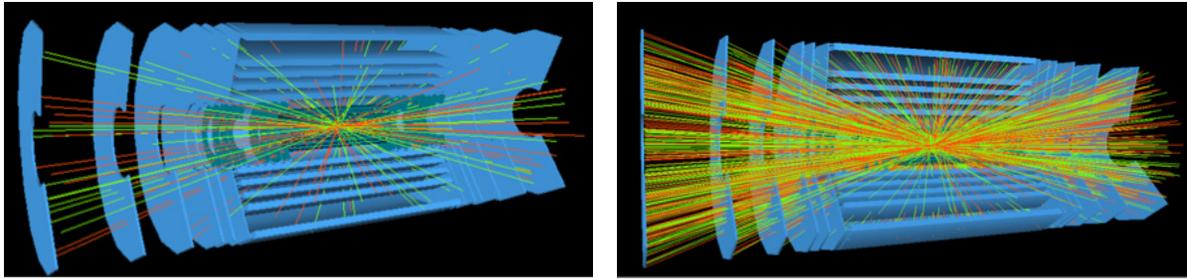


Figure 1. LHC events now (left) and for (right) HL-LHC. From Peter Vankov, *EPJ Web of Conferences*, 28 (2012) 12069.

Therefore, new data acquisition technologies are needed to capture the necessary event data and reduce it to its critical elements in real time. Due to the complexity of the data, software data reduction techniques will need to be introduced. This calls for powerful computing close to the experiment and will bring a number of significant challenges. HL-LHC poses specific requirements for the electronics, such as the ability to operate within high radiation environments, equivalent to a nuclear reactor. Furthermore, power and cooling are major issues within the detector electronics; these limits were set when the detectors were built and can only be marginally changed. What is available in terms of power and cooling for computing is, therefore, extremely limited. At the same time, maintenance and repair are difficult due to the tight construction and design of detectors. Hence, the experiments seek edge computing and networking capabilities that can run reliably with extremely low energy and cooling footprints, with a small form factor, and within a high radiation environment.

Data challenges at BES light source facilities

The use of microelectronic components is a key element within the overall process pipeline of the BES Light Source Facilities that connect the experimental activity eventually to the collection, curation, and analysis of the data. A particular point was made during the presentation that in many cases it is not sufficient to develop the “multipurpose next generation CPU/GPU (central processing unit/graphics processing unit) or memory” in a vacuum. Instead, detection electronics, computing, data storage, and data transfer have to be matched to each other to process scientific results most efficiently. The block diagram in Figure 2 can be generally applied to HEP experiments, light and electron sources, cryo-electron microscopy, ultrafast electron diffraction, and in some cases, neutron sources. It would be helpful to create centers where experiments from different scientific domains and computing are co-located to encourage solutions that are developed in an integrated manner, rather than rely upon single-point solutions.

Two science use cases were discussed by the panel, and a number of key challenges that future microelectronic solutions would need were identified. More rigorous studies are needed to determine when edge computing devices are more effective to use than large-scale central computing resources. In cases where edge computing is the appropriate solution, and the panel members thought that there would be a significant number, these data and computing devices may have to be able to withstand hostile environmental conditions. Another key pattern identified in data capture and analysis for large-scale experimental facilities was the crucial need for high bandwidth access to storage, as well as random access to small amounts of data during analysis.

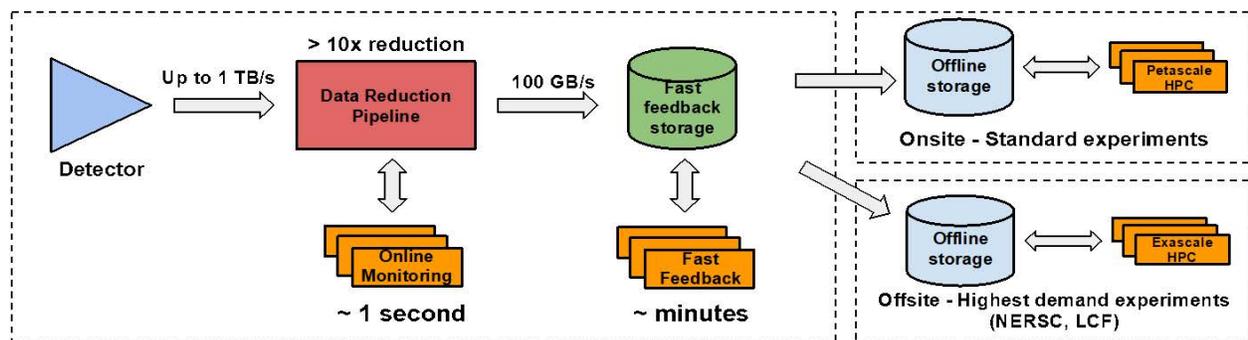


Figure 2. Standard data flow for large-scale experiment from the detector to the initial data analysis compute system (NERSC = National Energy Research Scientific Computing Center, HPC = high performance computing, LCF = Leadership Computing Facility). Courtesy of Norbert Holtkamp and Melinda Lee, SLAC National Accelerator Laboratory.

Scientific Challenges and Opportunities

Data today are as ubiquitous as they are voluminous, and to unlock their value, they need to be stored in an accessible way and analyzed, often many times in different settings. This means that data have to be moved over and over, an inefficient and time-consuming process. Because of the demanding scientific requirements of future research, it is more desirable to limit data movement by finding solutions that allow us to embed computing actions at the source of the data. There is also a need to re-configure networks on the fly to efficiently utilize data bandwidth, as well as the compute capacity available along the data transfer path. Primary research directions for this vision identified by Panel 1 are to create:

- Boolean and non-Boolean edge computing fabrics with high computation density embedded in memory with extreme energy efficiency and resilience
- Novel reconfigurable transduction elements that can enable a heterogeneous network
- Protocols to enable reconfigurable computing within the network fabric
- New algorithms that can leverage data analytics networks on the fly

The panel felt that now would be the time to expand this type of research, as we have a better understanding of the relationship between data and information density, as well as improved heterogeneous integration techniques. If successful, these developments could facilitate a range of advances, including coupling between various states of matter (e.g., electrical to optical), pushing the boundaries of information throughput and energy efficiency for spatio-temporal data processing while restricting data movement and enabling a new class of data-abundant scientific experiments.

NOVEL INTEGRATION METHODS TO ENABLE THE “COMPUTE IN STORAGE” PARADIGM

Current Status and Recent Advances

Scaling of storage device capacity has outpaced bandwidth scaling in the past and is likely to do so for the future. The gap between storage capacity and bandwidth renders efficient utilization of high-density storage devices challenging. The main limitation for device bandwidth scaling is the high energy cost of data movement, which has led to the idea of moving computation closer to storage. To enable this new “computation-within-storage” (CiS) paradigm, we seek key contributions in the following research areas.

Materials and devices

Existing complementary metal oxide semiconductor (CMOS)-based technologies are optimized for either storage or logic operation, rendering the integration of both challenging. We seek new materials and devices to integrate logic and memory on the same substrate to provide high performance and energy efficiency. Of particular interest are technologies that improve device characteristics on multiple levels, improving the tradeoff between retention energy speed and error rate.

Packaging technologies

We seek new approaches that enable heterogeneous storage stacks and multi-chip modules, including memory, nonvolatile random access memory (RAM), flash, logic, I/O (input/output), and sensors. Other areas of interest are intra-chip interconnection technologies that enable low power communication between chips at <0.01 pJ/bit and novel techniques for efficient heat dissipation of integrated multi-layered systems.

Microarchitecture

We seek new logic design architectures for CiS and memory systems. This includes new architectures (conventional as well as non-von Neumann) that exploit the unique characteristics of storage devices, such as asymmetric read/write performance, high latency, and limited endurance.

Programming models and systems

We seek new approaches to integrate compute-in-memory devices into existing systems efficiently. This includes new programming models and languages to specify CiS kernels and operating system enhancements and techniques to share data efficiently between the host and CiS devices.

An important topic discussed by the panel was the future trajectory for storage devices. The panel felt that while there is an industry roadmap for storage devices, whether it can actually be achieved is unclear. A key design feature in the coming decade is the increased use of complex 3D structures in storage solutions. At present, the mechanical stability of such solutions is insufficient to deliver reliable components. The panel concluded that new materials and engineering methods would be needed to create such devices at scale. Futuristic approaches such as storage based on biological elements were discussed. The panel discussed the viability of DNA-based storage systems and the challenges of integrating such future technology. A key insight from the discussion was the definition of the physical boundaries that science is encountering when trying to provide extreme amounts of storage, while providing fast access at the same time.

Metasurface integration

By using metallic and dielectric nanostructures precisely sculpted into two- and three-dimensional nanoarchitectures, the fields of microelectronics and photonics have experienced a remarkable evolution in the last decades. Electrons and photons can now be manipulated, confined, and processed in ways that are impossible to achieve with conventional materials and geometries. More recently, the introduction of optical metasurfaces has further revolutionized the ways in which electromagnetic waves can be controlled, and thus, the prospect of planar, lightweight, and ultra-compact optical devices is becoming a reality.

Metasurfaces consist of planar arrays of sub-wavelength structures that locally modify the properties of the electromagnetic waves (Figure 3). Since they are fabricated by the same techniques as used to manufacture microelectronics circuits, they should offer seamless integration with electronic components. Monolithic fabrication of metasurfaces that are integrated within electronic circuitry may offer advantages related to data-intensive applications, advanced sensing, and hardware security. For instance, squeezing light to nanoscale dimensions can enable dense optical integrated circuits, overcoming fundamental challenges related to bandwidth and energy dissipation.

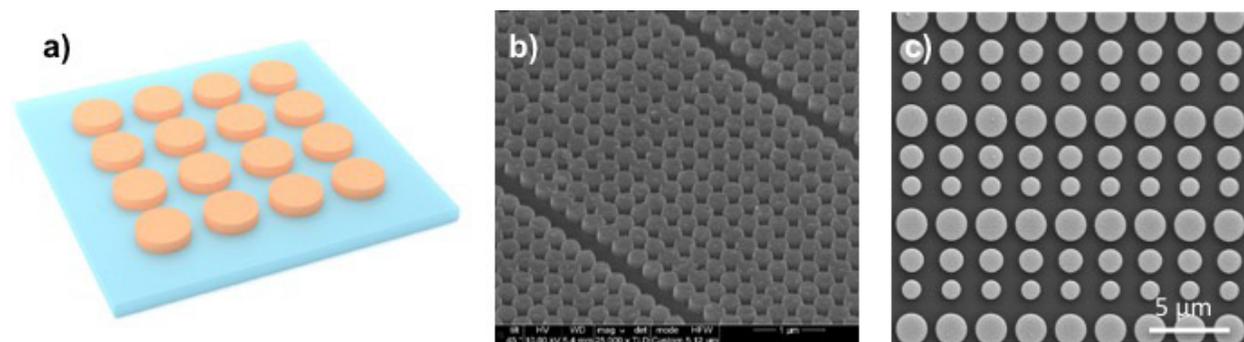


Figure 3. (a) Schematic representation of a metasurface: planar array of sub-wavelength structures (disks) that can modify the local properties of electromagnetic waves. These metasurfaces can be fabricated by using dielectric structures, such as the TiO_2 nano-resonators shown in (b), or metallic ones, such as the plasmonic antennas made of Au shown in (c). (a,b) Courtesy of Daniel Lopez, Argonne National Laboratory. (c) From T. Roy et al., *Applied Physics Letters Photonics*, 3 (2018) 021302.

Scientific Challenges and Opportunities

Memory and storage

There are fundamental physics limits on the retention energy speed and error rate in today's memory and storage technologies. These require tradeoffs between storage density and the achievable access speeds and retention times. To overcome these challenges, we need to consider solutions that go beyond the traditional digital memory paradigm, looking at new materials (including non-traditional semiconductors and insulators, as well as biological materials), their integration with more traditional devices, and their effective uses. One specific research challenge is the need for high speed access to large volumes of data for its analysis. Key

physical and engineering limitations need to be overcome to allow an effective, tight integration of logic and interconnect functionalities with the memory. A DOE requirement for certain specific applications is the ability for these new materials and devices to perform in harsh conditions, such as at low temperatures or high radiation environments. To address these challenges, the panel suggested four research areas:

- New states of matter that can overcome the current tradeoffs
- Emerging physics at the ultrafast time scales
- Novel synthesis methods that enable extremely high density and energy efficient communication for random access
- Novel integration methods to enable a 3D integrated logic-memory paradigm

Successful research in these areas would create a fundamental understanding of the metastable states of matter and their time and energy dynamics, as well as insights into their interaction lifetimes at ultra-small scales. These, in turn, could be the foundation for new materials and devices that could revolutionize storage technology.

Such new memory technology could have a significant impact on scientific research by enabling novel ultra-high data throughput and processing from leading experimental facilities, distributed sensor networks, and leadership class computing facilities. Real-time feedback and control in experimental physics and other complex systems would become much more feasible. Finally, extremely energy-efficient computing systems could be built that would enable much more cost-effective, data-intensive computing.

Metasurfaces

Panel 1 discussed that a key challenge was the capability to integrate optics with CMOS electronics to demonstrate monolithically fabricated novel sensors and detectors. Currently, metasurfaces are very sensitive to light wavelength used, and broadband metasurfaces are needed in order to have a real impact on optical systems. New materials and time-varying metasurfaces could complement CMOS timing applications.

New physical effects to enable 100 aJ per single computation

Computing today is not limited by our ability to build large enough systems, but by our ability to provide the energy required to operate our ideal systems. Therefore, energy efficiency has to be a key driving issue in microelectronics innovation. Panel 1 decided to set a specific goal for this research direction of 100 aJ per single computation. To achieve this, we need to overcome a number of key scientific challenges such as the ability to harness new materials that enable device structures for low voltage switching. At the same time, we need to explore new approaches (such as neuromorphic, analog, and adiabatic/reversible computing) that would enable low energy processing. Such hybrid compute paradigms bring their own scientific challenges, for example, those which combine traditional and neuromorphic computing. To address these challenges, Panel 1 suggested the following research directions:

- Explore and develop new materials and associated device structures aimed at specific new computing paradigms. Two examples are: memory materials optimized for artificial neural networks and new materials and switching concepts for digital logic.
- Co-design algorithms and hardware implementations.

It is expected that such research would help discover new phenomena that instantiate computation, as well as would lead to new understanding of the physics in known materials and devices. At the same time, the research would create new techniques for *in-situ* measurement and the correlation of electrical, thermal, and physical properties at atomic resolution. Achieving 100 aJ/single computation represents an improvement of 10,000x over current HPC systems, and would enable edge processing of massive datasets in real time for large facilities.

Synthesis, nanofabrication, and characterization

The panel concluded that new methods to synthesize new materials for heterogeneous integration were needed. Future architectures with very dense heterogeneous integration of dissimilar materials will require addition/removal of materials with extremely high local selectivity and atomic precision. Revolutionary nanofabrication capabilities are needed to enable the ability to pattern and fabricate heterogeneous systems in three dimensions down to sub-nanometer spatial resolution. Accelerated materials discovery will require new theoretical and experimental methods – though these should also address issues of manufacturability upfront. It is also important to develop, in particular, *in-situ* (or *in-operando*) characterization methods for microelectronic materials with nanometer or sub-nanometer spatial resolution and ultra-fast time resolution characterization methods for microelectronic materials to study growth and degradation processes under device-relevant conditions.

HEAT DISSIPATION CHALLENGES IN MICROELECTRONICS

Current Status and Recent Advances

The materials structure in storage and computing devices has a significant impact on the thermal conductivity and heat dissipation of the structure. Today's transistors could run much faster, but are limited by heat dissipation (Figure 4). It is thus important to understand and accurately model nanoscale heat transport physics, and to actively control heat transport at the nanoscale (nanophononics) in very small, tightly packed micro-electronic structures where dissimilar materials are integrated on the scale of nanometers. Investigating phonon transport in complex heterogeneous architectures has to incorporate issues of confinement, surfaces and interfaces, strain, coupling with electrons, spin and plasmons – these are not very well understood today. Today's theoretical modeling capabilities are not detailed enough and also do not account sufficiently for materials defects.

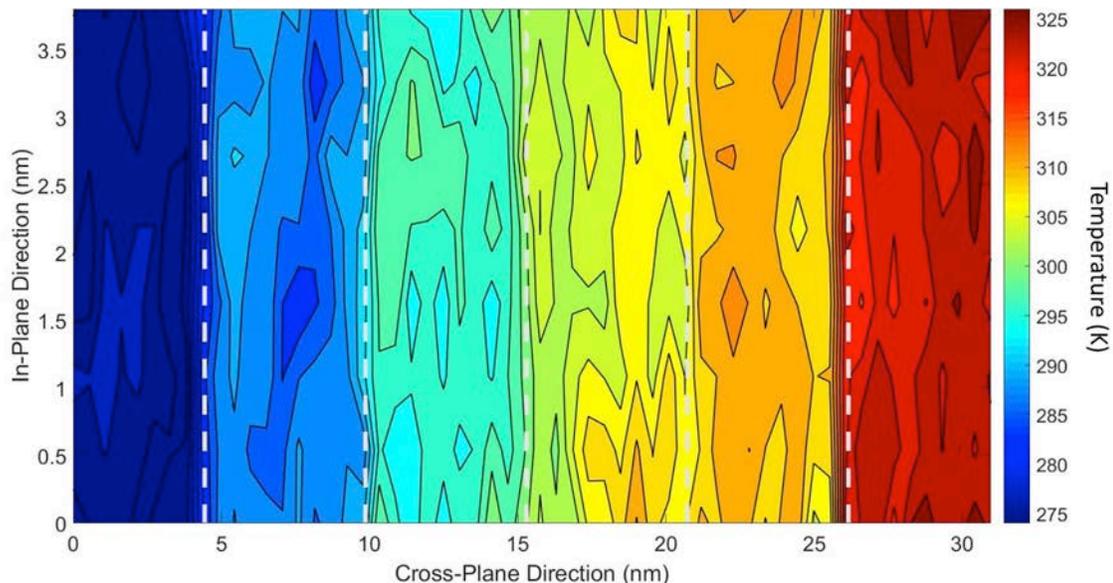


Figure 4. Irregular temperature contours in a silicon-germanium nanostructure under the influence of a stationary cross-plane thermal gradient. The white dashed lines represent the interfaces between the silicon and the germanium layers. Courtesy of Q. Moore and S. Neogi, University of Colorado at Boulder.

More detailed experimental observation and modeling would be extremely useful for the design of effective and efficient devices. Experimental techniques are needed to map electron and phonon transport pathways at high resolutions, specifically, to be able to identify defects and their effect on the overall architecture. Understanding these phenomena better would be a first step to designing storage devices that are not constrained in their performance, form factor, and energy efficiency by heat dissipation.

Synchrotron x-ray techniques and emerging ultra-fast electron microscopy techniques can contribute significant support to the above research needs for developing next-generation microelectronics. The wide variety of functional sample environments that are compatible with x-ray investigative techniques has long been a strength of these methods. Additionally, the spectrum of techniques available at a modern user facilities spans the gamut from spectroscopy, which can reveal atomic coordination and chemical states, to diffraction and imaging, which provide insight into the atomic and electronic structure of novel materials or devices. The ability to image structure and the deformation of ordered materials provides invaluable information on the strain within a material, the development and evolution of structural defects over time, and the eventual failure or diminishment of a phenomenon. At lower x-ray energies, imaging and photon correlation experiments can provide a time-resolved measurement of electronic and surface structure. As the requirements on the performance and efficiency of functional materials grow, nanoscale imaging, provided by these x-ray techniques, is an invaluable tool to characterize high-density devices under *in-operando* conditions. Similarly, ultra-fast electron microscopy diffraction and imaging equipment and techniques are rapidly advancing, with improvements in both stroboscopic and single-shot imaging methods. It is anticipated that these emerging techniques will provide new insights into understanding transient effects in emerging electronic materials.

The emergence of new x-ray and electron microscopy imaging techniques, particularly, those that involve 3D imaging and transient phenomena, in turn, further drive the need for fast on-the-fly analysis and data storage. For instance, the panel noted that imaging a 1-mm 3D material cube at ~ 1 nm resolution could generate up to 250 PB of raw data.

Scientific Challenges and Opportunities

The panel noted that experiments, advances in theory, and detailed numerical modeling are needed in the design of new devices, specifically to study phenomena such as heat dissipation, radiation, and power. These would precipitate a new rational design approach, changing current approaches from engineering to quantitative research.

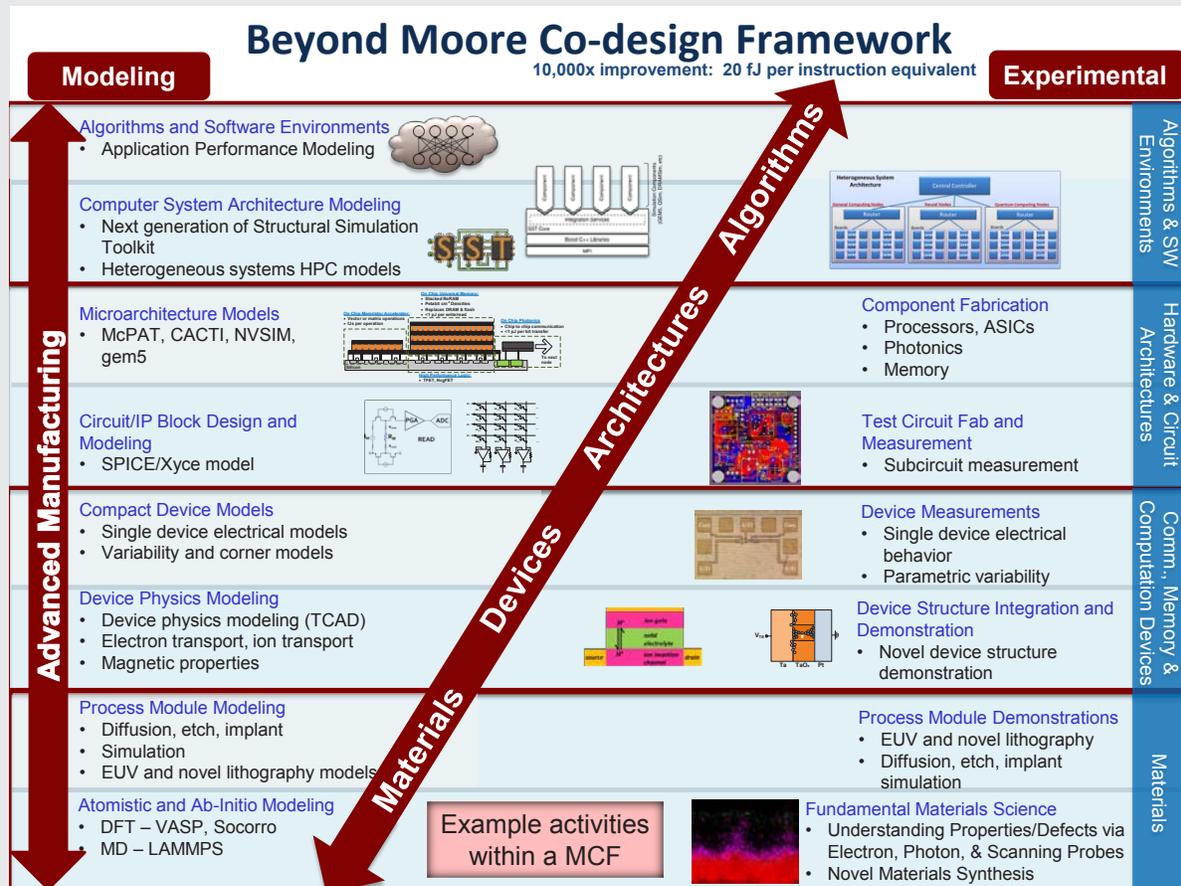
A MULTI-SCALE CO-DESIGN FRAMEWORK

Current Status and Recent Advances

Today's architectures and data intensive applications suffer from the so-called "von Neumann bottleneck," limiting the amount of data that can be loaded from memory into the processor per clock cycle. This bottleneck has been continuously growing, with ever faster processors and increasing memory hierarchies through which data have to move. On the short term, the von Neumann bottleneck can be addressed via heterogeneous integration of high density memory near the logic, thus minimizing the data transport pathways. This is already occurring with technologies such as high bandwidth memory GPUs and Intel/Micron 3D Xpoint architectures. Ultimately, computing must be done directly on data while in memory to avoid data movement costs, and this architecture will benefit from the exploration and exploitation of radically new approaches and physics. New memory, computation, and communication devices combined with new computing paradigms will enable an expansive set of new opportunities: neuromorphic, in-memory, analog, and adiabatic computing. It will be necessary to methodically explore and down select promising pathways. A major challenge in this venture is achieving new low energy regimes, well below 100 fJ of energy per mathematical computation. Today, we have a disconnect between fundamental materials work and its connection to the computational architecture level, and it is crucial to consider the challenges that arise at every integration stage. Development of the new devices needed will require establishment of the connections among the different design levels through a strong co-design framework.

BEYOND MOORE CO-DESIGN FRAMEWORK

10,000x improvement: 20 fJ per instruction equivalent



Comprehensive view of the many computational and experimental techniques needed in the rational design of new microelectronics. Shows the need for strong integration across many levels during the design phase. McPAT = multicore power, area, and timing modeling; NVSim = nonvolatile memory simulator; TCAD = technology computer-aided design; EUV = extreme ultraviolet lithography; DFT = density functional theory; VASP = Vienna ab initio simulator package; MD = molecular dynamics; LAMMPS = large-scale atomic/molecular massively parallel simulator; MCF = Moore co-design framework; ASIC = application-specific integrated circuit. Courtesy of Matt Marinella, Sandia National Laboratories.

Scientific Challenges and Opportunities

Heterogeneous integration

One of the key challenges in creating new microelectronic devices is the move from a novel material to a complex, integrated device or device component that retains the desired properties of the underpinning materials or combines them in effective ways to create the desired properties. For this, we need effective methods to synthesize new materials for heterogeneous integration and to add or remove components selectively down to sub-nanometer spatial resolution. To address this key challenge, Panel 1 identified the following principal research areas:

- Discover new chemistries and methods of materials synthesis
- Design and probe new interfaces and materials for optimal interconnection
- Discover new approaches to integrate desired materials and functionalities into complex 3D structures

Developing the experimental and production capabilities to create new materials with sub-nanometer specificity would enable the community to create much more efficient high-performance interconnects; achieve desired efficiencies for power, thermal, and cooling; and create new sensing and communication technologies.

New methods and tools to facilitate co-design

The challenges discussed at the Workshop demonstrated the need to work very closely across a wide spectrum of areas from physics and devices to architecture. These challenges open up many new dimensions in an already complex search space. To address these challenges, Panel 1 identified the following principal research directions:

- Model and simulate power, performance, and reliability (must capture the impact of new materials and devices)
- Develop an automated, generalized co-design process, including measurement, modeling, and simulation
- Perform workload characterization
- Investigate properties from the atomic scale to those occurring *in operando*
- Investigate scalability and manufacturability of promising new devices and components

If successful, this research would greatly accelerate the cycle from materials to architectures.

MATERIALS, DEVICES, AND ARCHITECTURES FOR HPC

Current Status and Recent Advances

The often-predicted and gradual “end of Moore’s law” is setting the stage for exciting new developments in information technology as the focus of R&D shifts from miniaturization of long-established devices to the coordinated introduction of new devices, new integration technologies, and new architectures for computing.¹ The impetus for this shift in research investment goes back to 2003-2005, when the increase in microprocessor clock frequencies, constrained by considerations of heat removal and power density, began to slow down.² To overcome this power versus performance bottleneck, exploratory research was initiated on devices that operate via physical principles different from those of conventional field effect transistor (FET) and architectures that are distinct from the long-established von Neumann architecture. Among new devices for digital logic, the tunnel FET, the spin transistor, and the negative capacitance FET (NCFET) are worth mentioning. None of these devices has been introduced as commercial products or introduced within processor chips for evaluation, though the NCFET is being actively investigated in industrial R&D today. New devices for digital memory are more advanced, with versions of magnetic RAM (MRAM), resistive RAM (RRAM), ferroelectric RAM (FeRAM), and phase change memory (PCM) in various stages of commercialization.

At the same time, we may be witnessing the emergence of the *second* major architecture for computing, driven by the explosive growth of machine learning algorithms and applications. While the first major architecture, the von Neumann architecture, is optimized for arithmetic, the second will be optimized for pattern matching and discovery. Machine learning algorithms currently leverage established hardware platforms (central processing units, graphical processing units, or field programmable gate arrays) or more specialized designs, but all are based on conventional digital logic and memory devices. However, power and performance are predicted to improve by orders of magnitude if devices, architecture, and algorithms are co-designed and co-developed.³ Artificial neural networks could address complex pattern recognition and problems related to finding correlations much more efficiently than conventional hardware.

Scientific Challenges and Opportunities

With these considerations in mind, we propose three main research challenges for the coming decades:

- Continue to invent, explore, and develop post-CMOS devices for digital logic and CPU/GPU and related architectures.
- Co-optimize materials, devices, architectures and algorithms for artificial neural networks (ANNs). (Specific objectives might include improved memory materials for analog representation of synaptic weights or compact devices to implement the thresholding function at network nodes.)
- Seek new classes of broadly useful algorithms and architectures for their efficient execution that look beyond both the CPU and the ANN.

REFERENCES

1. T.N. Theis and H.-S.P. Wong, The end of Moore's Law: A new beginning for information technology, *Computing in Science and Engineering*, March/April, pp. 41-50 (2017).
2. T.N. Theis and P.M. Solomon, In quest of the "next switch": Prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor, *Proceedings of IEEE*, 98 (2010) 2005-2014.
3. T. Gokmen and Y. Vlasov, Acceleration of deep neural network training with resistive cross-point devices: Design considerations, *Frontiers in Neuroscience*, 10 (2016) 333.

Panel 2 Co-design for High Performance Computing Beyond Exascale

INTRODUCTION

For over 25 years, DOE's HPC community has leveraged the commodity computing ecosystem. Moore's Law technology advances and the clock speed multiplier of Dennard scaling gave rise to the "killer micros" that led to the transition from custom vector supercomputers to large-scale parallel systems. In this timeframe, DOE supercomputers were built from the integration of commodity computing components. DOE practiced multi-disciplinary co-design among different software development activities: system software, tools, algorithms, and applications. A focus for these multi-disciplinary collaborations was improving the scalability performance of DOE's HPC application portfolio with capabilities such as message passing communication libraries and computational domain partitioning tools. The predictability of performance improvements for general-purpose CPU processor technologies provided DOE with a solid foundation for the evolution and co-design of new HPC applications with the software stack. While the DOE HPC user and application community did have some influence over commodity hardware architecture designers (e.g., for x86 vector extensions and cache configurations), commodity processor designs necessarily were biased to the needs of commercial software companies like Microsoft. In the Cray vector supercomputer era, DOE's HPC co-design loop was very interactive – it allowed Cray's computer architects to prioritize among options as they developed their custom hardware designs.

The slowdown of Dennard scaling was marked by the appearance of the first commercially available dual-core processors in 2005 since microprocessor frequencies could not continue to increase at the same pace as before. While the core counts in CPUs have steadily increased, as long as compute nodes were based on the integration of general-purpose, multi-core CPUs and commodity dynamic DRAM (DRAM), DOE's HPC needs were met by continuing the strategy of leveraging commodity computing technology. DOE's computer science investments continued to focus on realizing the compute performance of multicore processors and to address changes driven by reduced memory bandwidth and capacity, as well as reduced interconnect bandwidth, on a per core basis. Overall, the application software and system software stacks were remarkably stable over this timeframe.

A major disruption in DOE's computing model occurred in 2008 when the first large-scale, heterogeneous compute node architecture was introduced by the Roadrunner system at Los Alamos National Laboratory. Supported by the Advanced Simulation and Computing Program (ASC) of the National Nuclear Security Administration, Roadrunner included two dual-core AMD Opteron CPU processors and four IBM Cell eDP GPU video game processors per compute node. With nearly 3.5K compute nodes, the Roadrunner system was the first petaflop HPC system. While IBM ended the Cell processor, HPC systems around the world began to use heterogeneous compute nodes with x86 CPUs and general-purpose GPUs (GP-GPUs). NVIDIA GP-GPUs were deployed on Oak Ridge National Laboratory's Titan supercomputer in 2012. GP-GPUs provided a dramatic increase in energy-efficient processor performance, but the heterogeneous mix of CPUs and GP-GPUs in these "advanced architecture" compute node designs was very disruptive to the application development community. The complexity of these node architectures has driven much of the subsequent DOE Advanced Scientific Computing Research (ASCR) and ASC investments in applied math, along with computer science investments in programming models and tools.

The DOE's HPC community is still leveraging commodity computing. The recent DOE ASCR workshop on extreme heterogeneity provides an extensive overview of the software challenges in this space.¹ A key research topic is emerging: do we need to remain dependent on the commodity commercial ecosystem? If we cannot follow this path, can the DOE and U.S. government at least influence the designs of future commodity computing ecosystem components? Can DOE once again close the loop in co-design and either directly or collaboratively develop custom architectures that can reduce the software development burden, while continuing to meet our

energy efficiency performance goals? The Electronics Resurgence Initiative of the Defense Advanced Research Projects Agency (DARPA)² is also making strategic investments to pave the way toward an era with lower barriers to the development of new architecture designs.

CURRENT STATUS AND RECENT ADVANCES

Despite the end of Moore's Law, there are multiple paths forward to continue the advancement of the performance and efficiency of microelectronics.³ Figure 1 outlines the multiple dimensions of opportunity for DOE to push microelectronics forward. New architectures and packaging will enable continued advances over the next 10 years. At present, DOE has a unique opportunity to change the course of development for new devices, new physics that underpins the principles for operation for the devices, and new materials that enable optimal implementation of the devices. The replacement of our current ecosystem will involve advances in all of these dimensions of opportunity.

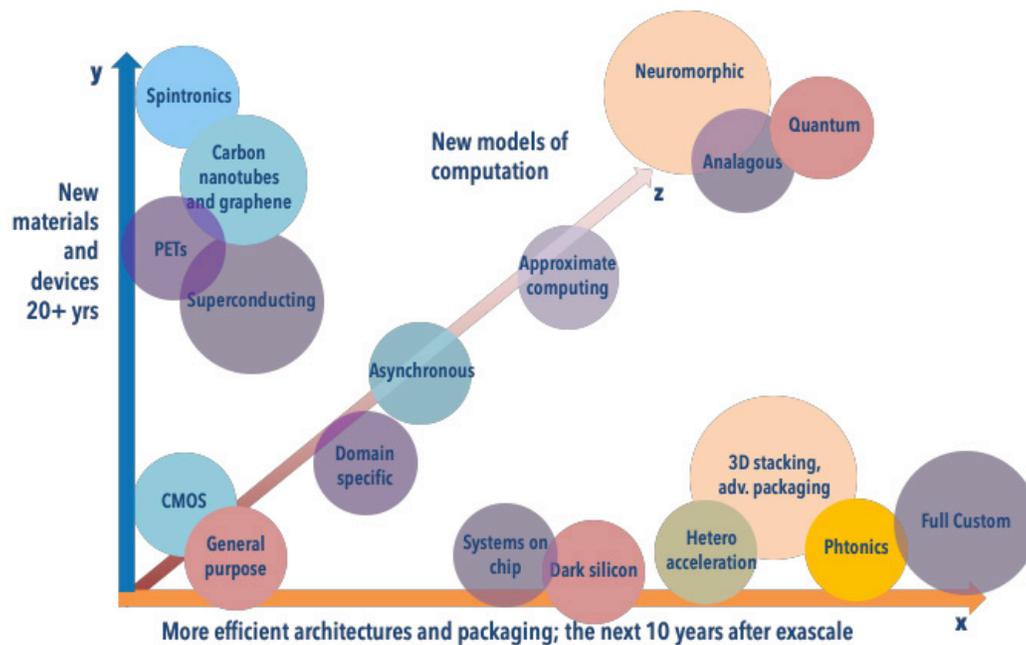


Figure 1. Multiple paths forward to continued performance scaling of microelectronics. An integrated approach that enables exploration of all of these paths together is required, and DOE is the only agency that can provide that level of integration. Courtesy of John Shalf, Lawrence Berkeley National Laboratory.

Microelectronics Materials – Status Quo

The electronics industry has made amazing progress over the past 50 years by using a very limited set of materials systems. Silicon and other group III-V elements have become the primary semiconductor material for integrated circuits because of their favorable chemical properties and physical robustness. However, in conventional FETs, device performance is limited by the voltage swing required to turn them completely on or off. New, more efficient devices would need new fundamental physical principles of operation to overcome the “Boltzmann tyranny,” which limits current devices to 60 mV/decade switching performance. Such an advance will require not only new fundamental physics for switching operations, but also new materials that are able to optimize the properties that enable the new physics.

Moving to a new fundamental principle of operation would enable a steeper slope for turning on/off the switch, which, in turn, would enable lower power and higher performance. However, these emerging devices are held back by inadequate materials such as those that have high contact resistance to metals, chemically interact in ways that make manufacturing impractical, have poor electron or hole mobility, or are mechanically or thermally unstable. The current “artisanal” process for discovery of new materials will not meet the pace of change required for new electronic systems.

Semiconductor CMOS-based electronics rely on the Boltzmann distribution of electron energies, as modified by the applied voltage. With the end of Moore's Law, there is an opportunity to fundamentally change the underlying principles behind logic and memory functions. At the chip level, there is the opportunity to explore non-Boolean architectures such as neural computing. Within the Boolean computing framework, there is the opportunity to keep the binary logic underpinning, yet dramatically reduce the energy consumption per logic operation, for example, by gaining orders-of-magnitude improvement over the ~ 50 pJ/logic operation today, as well as attaining a true merger of memory and logic functions through the creation of a nonvolatile, logic-in-memory architecture that can be dense and power efficient. At the materials physics level, this achievement would need radically new approaches, such as tapping into internal correlational energy scales based on quantum mechanics (exchange interactions in magnets, dipolar interactions in dielectrics, electronic correlations, etc.) as pathways to reduce the external energy consumed to carry out the logic/memory functions. This research direction immediately points to the implementation of new materials that can enable low-power information storage and manipulation. Lowering the energy consumption by several orders of magnitude will require dramatically new materials, along with better understanding of their underpinning physics and chemistry, new synthetic protocols, and new device architectures.

Device Physics – End of Moore's Law

Moore's Law as interpreted in its original form is that the number of transistors per dollar grows at an exponential rate. Historically, this rate has varied, and the doubling of transistors per dollar has ranged from every year to more than two years. Moore's Law is often conflated with Dennard scaling, where the latter is the observation that transistor performance scales with its size. Dennard scaling effectively ended for microprocessors when the power wall was reached, around 2005. Moore's Law may or may not continue in its original form, but Moore's Law for monolithic 2D integrated circuits is nearing its end in the next decade due to a combination of the limits of lithography, processing complexity, and the limits of performance for silicon-based metal-oxide semiconductor FET (MOSFET) performance.

3D Packaging Technology

Evidence is mounting that data movement, not logic characteristics, limit CMOS-based system performance and power. A detailed study of logic gate delay versus distance that data must be moved on a chip⁴ has shown that, for the most widespread CMOS process today of 14 nm, by the time a bit has moved just a few micrometers on a chip ($\sim 0.005\%$ of the way across a typical processor chip), we have exceeded the delay through a typical gate that has generated it. Fixing this problem requires more powerful drivers, which, in turn, introduce more delay at the logic end and draw significantly more power.

Going off-chip is even worse. Moving data from a typical processor chip to a typical memory part requires very complex serializer/deserializer (SERDES) circuits, which typically require a few picojoules per bit. To place this in context, achieving the original exascale goals⁵ of 1 exaflop at 20 MW is equivalent to expending only 20 pJ per average flop. At a few picojoules to just cross a chip interface, reaching this goal means that an executing program can access a 64-b word from off-chip memory at best only very occasionally. This is untenable for modern applications where accesses to sparse irregular data are frequent, and caches are no longer a panacea.

Heterogeneous Node Architectures: Processors and Memory

Starting with the Roadrunner system in 2008, DOE's high-end computer systems have increasingly gone to two or more incompatible designs for the processing cores: one for "general purpose" computing and one optimized for "accelerating" a specific type of computing, such as dense regular numerical calculations. The latter has proven to reach higher performance and to be more energy efficient, but only for a relatively narrow range of applications. A node on Summit at Oak Ridge National Laboratory (the current No. 1 HPC system in the world) has two general purpose chips, each with 22 cores, and six accelerator chips with 80 cores each. As a result, well over 90% of the peak numerical capability is in the accelerators, but virtually all of the management and communication capability is in the general purpose cores. The prior No. 1 HPC system, the Chinese supercomputer TaihuLight, had 256 accelerator cores and 4 general purpose cores on the same chip.

This heterogeneity in processing cores has also been reflected in heterogeneity in memory components and memory-processor interconnect architectures, as different types of processing have different memory needs. In addition, the emergence of radically new memory technologies, especially with nonvolatile characteristics, is forcing even more dramatic changes in memory interfaces and protocol.

Together, this overall heterogeneity has led to greatly increased complexity in programming, particularly in deciding how to break up an application so that different pieces will run on which core type, and how to manage the transfer of data between the memories associated with each core type. Going forward, this heterogeneity will likely accelerate, as new technologies appear in accelerators for specific functions in “mix and match” configurations with systems built from traditional technologies. This situation will be particularly true for technologies that support memory with embedded functionality that by its nature may be incompatible with the “conventional” memory (and its interfaces) used in what is then “conventional” heterogeneous systems.

Heterogeneity in processing is also leading to systems that do not have guaranteed cache coherency, meaning that application codes must take more responsibility for managing copies and synchronizing between threads. Co-design research will be critical for taking new technologies, initially most suitable for some narrow range of applications; architecting accelerators that can integrate with conventional heterogeneous systems without major redesign of the expensive high end chips; and then building software tool chains that allow access to the new capabilities without destroying compatibility with the previous capabilities.

System Software: Operating and Runtime Systems

As HPC system architectures have evolved, operating systems and low-level runtime systems have laid the foundation for portability, providing a stable portable operating system interface across hardware generations. In recent years the operating system and runtime have evolved from monolithic modules providing a fixed set of services for a static execution environment, to a system software stack that supports greater flexibility in resource sharing and isolation among executing entities, dynamic discovery and adaptation, and greater performance portability. However, the ability of the operating system and runtime to provide such a stable evolutionary path has been constrained by the slow evolution in application software and much faster evolution in the architectural landscape.

System Software: Programming Models and Environments and Associated Compilers and Tools

Programming models and environments provide a higher layer of abstraction to which applications interface directly. They consist of application-level interfaces that must evolve with application needs, associated compilers and tools, and lower levels of abstraction that evolve with system architecture to most efficiently implement the primitives exposed to the user. Efficient implementations of existing abstractions and development of new ones has enabled performance portable implementations of applications. With the anticipated increase in architecture complexity and diversity, fundamentally new models co-designed with the system architecture are necessary to ensure performance portability. Yet another change to how we must develop programs in the future is the need to move from a pure batch mode to a mode that, as needed, can stream new data through incrementally. The emerging world of the internet of things will encourage, and possibly force, such alternative thinking.

Finally, the growing use of artificial intelligence and machine learning, when integrated into both system run times and applications, will enable a growing capability for introspection, where the system/application learns as it goes how best to allocate resources and/or schedule computations.

Application State of the Art

Existing software systems and the applications that depend on them have become accustomed to years of steady performance growth from general purpose computing systems. The tapering of Moore’s law improvements in transistor density and performance will have a more profound effect on the programming environment by accelerating the move toward diverse hardware. The near-term approach to performance improvements is to utilize transistors more efficiently by tailoring the architecture to target application.

Overall, the tapering of Moore’s Law and the emergence of new device technologies will lead to a broader range of accelerator or specialization technologies than we have seen in the past three decades (see also the article on the economics of the semiconductor industry by Thompson⁶). Already, the adoption of GP-GPUs has created enormous challenges to programmer productivity for the scientific community, and we have only made incremental progress toward performance-portable programming environments. If our programming paradigm is already challenged by a single kind of accelerator (the GP-GPU), then we are especially ill prepared for a future with many different kinds of heterogeneous accelerators. This situation will require a deep re-think of our programming models in order to develop more productive programming environments for the future.

SCIENTIFIC CHALLENGES AND OPPORTUNITIES

Multidisciplinary Co-design

With the end of Moore's Law, the related scientific challenges span the full spectrum of disciplines embedded within microelectronics, including determination of the fundamental physics of information transport, storage, and communication; the fundamental science of synthesis of atomically perfect materials at length scales approaching a few unit cells; and methods for combining heterogeneous materials with diverse sets of physical properties in a scalable way, culminating in advanced, energy-efficient computing architectures. At a broader scale, these challenges present an unprecedented opportunity to create a "co-design" framework for innovation (Figure 2), which has not happened since the earliest days of electronic computing. Thus, these challenges fall under the purview of several offices within the DOE Office of Science. The key question for DOE HPC is, How will application performance normalized by energy consumed continue to increase? For the power grid community, the key question is, How can we make the future smart grid adaptive to the bidirectional flow of power and information? In both systems engineering scenarios, DOE needs a co-design framework that integrates physical layers, logical layers, and controls. This co-design framework will enable DOE scientists and engineers to develop unified materials, devices, circuits, and system architecture for EDA simulation tools to ensure performance, efficiency, resilience, and reduced design and development time for mission critical systems.

One approach is an intentional (co-design) effort to tie computer/system architectures to application requirements. The definition of these computer/system architectures can, in turn, drive requirements for circuits, devices, and materials. Co-design involves multi-disciplinary, multi-lateral collaboration.

DOE has a unique opportunity to create a multidisciplinary co-design framework for basic/applied research to accelerate the development of energy-efficient information technology (IT) beyond the end of current roadmaps and to maintain an advanced manufacturing base for the economically critical semiconductor space. This framework will allow DOE to leverage significant industry investments with the goal of enabling low-power computing and suitably low-cost smart grid, power electronics, and building electronics.

Solving the daunting energy challenge described earlier will require manufacturing technology advances that permit the continuation of Moore's Law from the device patterning perspective as well as device technology advances going beyond CMOS, system architecture, and programming models to allow the energy benefits of scaling to be realized. Only with co-design covering this broad space and consideration of manufacturing challenges can we expect to make progress in all areas cohesively to 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 that will enable new scientific discoveries and maintain U.S. competitiveness in all segments of the computing market (from internet-of-things, to cloud computing, to high-performance computing).

To meet the goal of broad societal impact, we must ensure transition of basic research to high-volume manufacturing and, even more fundamentally, shape basic research from the start with an eye to manufacturability. This approach will be achieved through the development of a multi-laboratory ecosystem that can be used to evaluate and demonstrate the manufacturing and energy savings feasibility of next-generation technology options. Technologies will be rigorously evaluated for potential benefits on energy and 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 further improvement to deliver on the energy promises. This phase will depend heavily on identifying specific manufacturing/device materials where we can leverage the capabilities of the DOE Materials Project (whose purpose is to provide an online resource for removing guesswork from materials design in various applications)⁷ and current HPC capabilities to accelerate development through modeling and "virtual cycles of learning." Manufacturing feasibility would also include demonstration of whatever patterning technology would be needed to support the various technologies and scaling of those technologies. Delivering on this vision will require integration across the DOE scientific areas shown in Figure 2

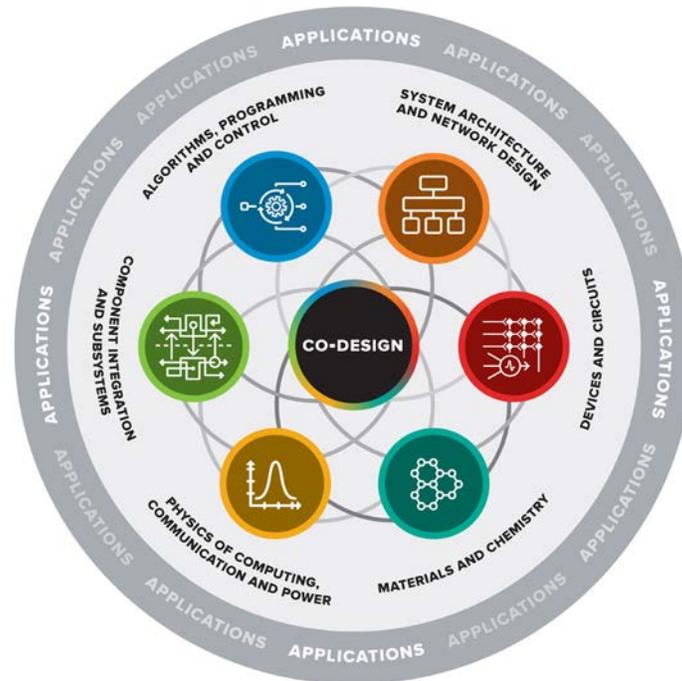


Figure 2. Multiscale co-design framework. Co-design involves multi-disciplinary collaboration that takes into account the interdependencies among materials discovery, device physics, architectures, and the software stack for developing information processing systems of the future. Such systems will address future DOE needs in computing, power grid management, and science facility workloads.

The co-design framework would overcome the following obstacles to advancing microelectronics beyond the end of Moore’s Law:

- The value of new and novel materials or device technologies is not currently understood in a system architecture context.
- Metrics at the application level are not understood in a system context, while metrics at the system context are not currently understood at a device or materials context.
- Scaling of performance will require
 - co-design that spans all layers in Figure 2;
 - multiscale modeling among many technology levels (materials, devices, architectures, system software, algorithms, and applications);
 - linkage between and across levels (e.g., apply device-scale first-principle models to automate search for materials and support algorithm-aware architecture choices; and development of alternative computing models, specialized hardware accelerators, and non-von Neumann architectures).
- Heterogeneity of the system software stack is extreme.

Materials Co-design

The discussions in this panel included many topics that are described in the deep co-design interactions in Figure 3. Materials science thrusts that would form the fundamental science underpinnings of the co-design platform could include the use of a computational materials discovery approach (for example, the Materials Project) coupled with precise materials synthesis to discover specific sets of materials/phenomena. Challenges include new, disruptive interconnect materials for information transport, or new mechanisms for information transport that could eliminate the almost 70% of power dissipation in current CMOS due to flowing current in interconnects. A second broad thrust is to start to explore bi-stable (and preferably multi-stable and analog) states in functional materials, but those that can be manipulated reproducibly and repeatedly with applied voltages in the range of 1-100 mV instead of ~1 V. This thrust will require a precise definition of the electronic

structure as well as understanding of how the relevant correlated phenomena (electromagnetic, ferroelectric, spin, charge, and chemical correlations) can be manipulated at such low energy levels. A significant amount of research is needed to establish the fundamental limits of the energy/length/time scales of these phenomena, since they will directly impact device- and circuit-level attributes such as power consumption, latency, and speed. In many cases, establishing these limits will require the use of state-of-the-art probes available at DOE user facilities.

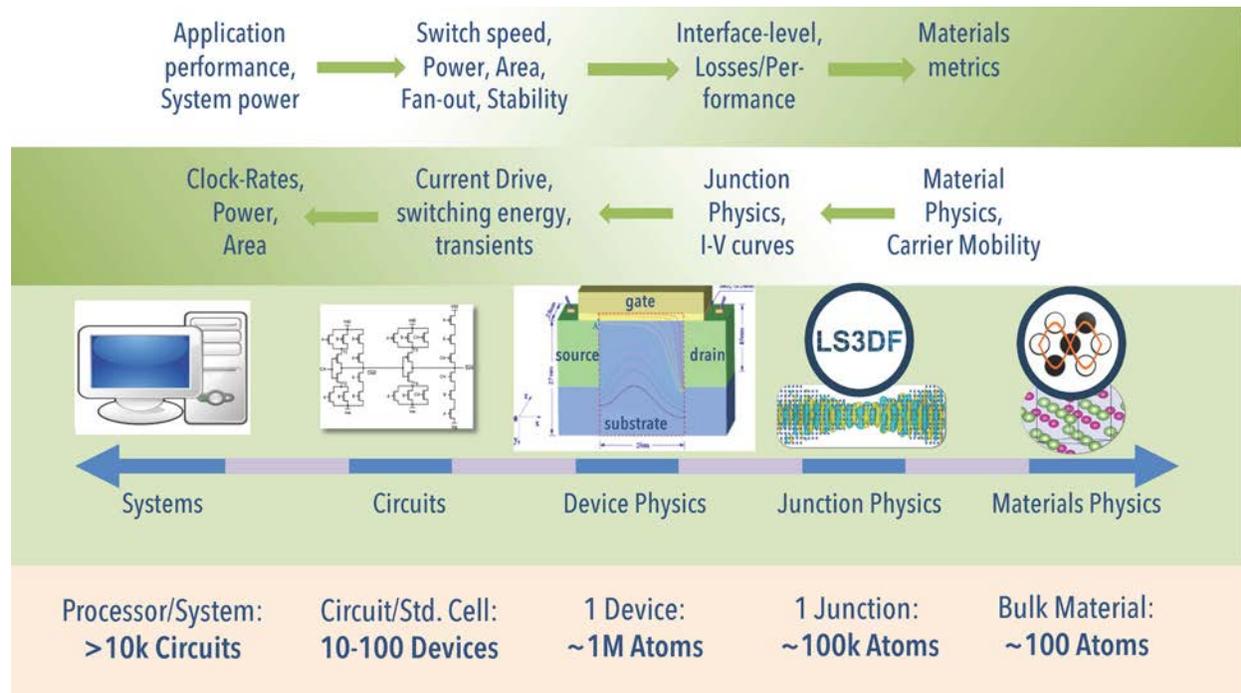


Figure 3. Topics on materials co-design from the atomic scale to processors and circuits. Courtesy of John Shalf, Lawrence Berkeley National Laboratory.

The approach to discovering new materials that could optimize properties for these new devices is “artisanal” in that material discovery and synthesis are extremely labor intensive. While there are numerous promising materials, devices that utilize those materials are inadequately understood. The challenge is describing metrics for improvement that are actionable at the level of materials science. The limitations of current practice slow the process of materials discovery and ultimately hold back the enterprise of rapidly advancing microelectronics. The DOE Materials Project has successfully demonstrated a more “industrialized” approach to automating the search for new and better materials based on well-defined properties, but this approach has been primarily applied to discovering battery materials.⁷ It points a way towards industrializing our process for the discovery of better materials for microelectronics, and would make the pace of discovery much faster and less “artisanal.”

Data Movement and I/O Challenges

Today’s solutions to data movement and I/O challenges are two-fold: pursue a 3D approach and change the way we build interconnections of multiple chips. The 3D approach has two variants. First is to build a stack of CMOS chips with metal-plated holes through each die to allow electrical contact to the dies above and below them. Second is to build 3D arrays of transistors on the same chip. In both approaches, going from a 2D to a 3D chip significantly reduces the distance between logic functions and requires significantly less power than going off-chip. The other alternative, changing the way we build chip-to-chip interconnects, uses a substrate (like a large slice of silicon) that contains wiring between chips. Individual chips then may have very small low-capacitance connections that are joined to the pads on the substrate. Again, these chips often do not require SERDES to go from one die to another, and thus not only take less power but permit far greater numbers of connections. The first approach is being used increasingly today to build 3D stacks of memory, such as “high bandwidth memory”, which are then attached to a microprocessor chip (often a GPU) by the second approach.

Looking forward, a variant of these techniques will become an integral part of co-design efforts that employ different combinations of processor and memory chips, and/or integrate new technology chips that are incompatible with being implemented on a CMOS process but provide new accelerator or memory capabilities. The DARPA Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) program⁸ is taking this one step further to design sets of chips, along with standardized chip-to-chip interfaces, that can use the second substrate approach to allow predesigned “chipllets” to be assembled in a low-cost “on-demand” basis. Further research is needed to take this one step further to include 3D and non-CMOS technologies to be integrated in a co-design fashion.

Copper wire is about as good a conductor as one could expect at room temperature, and the conductance of the thinnest interconnect wires today is limited by electron scattering at the wire surface and the need for cladding layers to prevent copper out-diffusion. Therefore, absent disruptive breakthroughs such as the development of scalable room-temperature superconductors, or new approaches for signal transduction, copper wire-based I/O will continue to be fundamentally limited by the resistance and capacitance of materials. A possible direction for exploration is photonic technologies. For communications (wire replacement), photonics has the benefit of having energy costs that are nearly independent of the distance that data travel, whereas the standard electrical wires have a strong distance-dependent energy cost. Therefore, photonic technology overcomes the fundamental limitation of wire resistance. However, challenges remain in being able to interrogate dense device assemblies at high spatial resolution and handle the complications of harnessing efficient light sources. Improvements in transcoding between the photonic and electronic or spintronic domains may also provide additional benefits for integration density and performance for data movement and I/O. Similar data movement challenges are bottlenecks for DOE experiments, as shown in Figure 4.

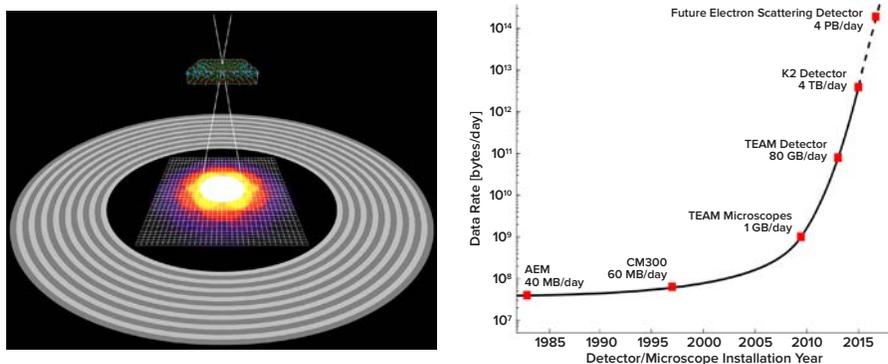


Figure 4. Examples of how I/O data rates for emerging detectors for DOE experiments are overwhelming data interfaces. Integrating processing into the edge is one approach to manage the data movement challenge. Photonics to improve data rates is another option. Courtesy of Peter Denes and David Donofrio, Lawrence Berkeley National Laboratory.

Innovative solutions are needed to reduce the energy cost of data movement. One promising approach to achieving low-loss communication in media that are **not** loss free at room temperature is the use of 2D confined states and quasiparticles, such as magnetic skyrmions. Communication with 2D confined states is a topic for further research.

Programmability, Performance, and Portability

With a stable computing paradigm and slow and predictable hardware evolution, software abstractions have effectively managed programmability while supporting performance and portability. With the anticipated rapid changes in system design, including fundamentally novel computing paradigms and the primacy of data movement costs and energy efficiency, there is an urgent need to co-design algorithms, software abstractions, and the underlying execution models to continue to meet today’s and tomorrow’s DOE application needs. Without such an effort, application developers will need to resort to on-off non-portable optimized implementations or programmable yet inefficient implementations, leading to inefficient utilization of system designs or software development efforts. Importantly, today’s algorithms might not be the most effective on tomorrow’s architectures, requiring a renewed effort in architecture-aware application and algorithm design. As the next-generation computing paradigms are being identified, a timely effort combining expertise in applications, algorithms, parallel software and hardware abstractions, and computing paradigms can address this challenge.

Of utmost importance is the design of algorithms in concert with constraints dictated by materials and system designs: locality, data movement costs, and energy efficiency. Going beyond individual implementations targeting specific architectures, tools and methodologies needs to be done to aid designing and analyzing algorithms in concert with architectural design. This includes research to analyze algorithm-level data movement characteristics, including coherence requirements for communication complexity analysis. There is a need for integrated design of algorithms and software for data-centric systems (computer, memory, and storage). Also needed are holistic end-to-end designs that focus on extreme sparsity and extreme locality. Novel approaches that move away from notions of global addressability and coherence requirements or approaches that can aid the generation of localized implementations from such global views can address the data movement challenges.

Productive languages and frameworks, including domain-specific abstractions, are needed to enable design of architecture-aware algorithms. Simultaneously, to the same goal of algorithm-aware architecture, methods are needed to facilitate early co-design of hardware designs, programming interfaces, and algorithms. Compilers, runtime systems, and associated tools, including code generation and auto-tuning strategies, need to be co-designed with the algorithms to enable performance-portable mapping of these interfaces to diverse architectures.

Non-von Neumann Computing for Scientific Discovery

The capabilities of the prevailing model of computation, the von Neumann model, are increasingly constrained by the energy inefficiency of established hardware and architectures. Understanding and using new architectures based on unexploited physical phenomena require co-design spanning formal models and algorithms to physics, materials science, and new devices (Figure 5).⁹

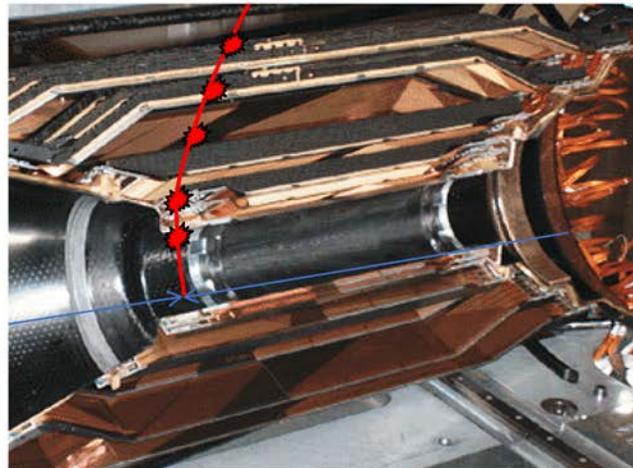
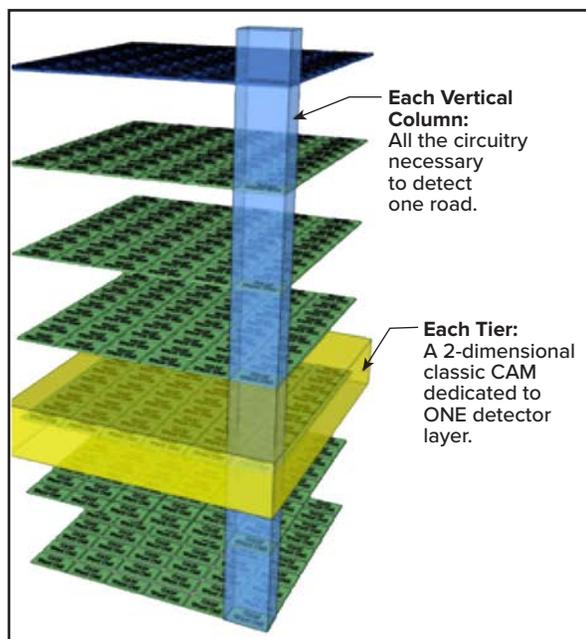


Figure 5. Non-von Neumann spatial computing concept for HEP particle tracking integrated directly into the sensor array. (Left) From T. Liu et al., *Physics Procedia*, 37 (2012) 1973-1982. (Right) Courtesy of Ted Liu, Fermilab.

At the dawn of the computer era, there was no broad appreciation of the advantages of digital devices and circuits, and analog approaches dominated. Now, as we broaden our horizons to include the possibility of radically new architectures and new physical systems for implementing those architectures, we are looking again to analog and hybrid analog-digital approaches. It is, therefore, instructive to recall some of the reasons why digital computing utterly eclipsed the early analog approaches. First, the available digital devices were more compact than the analog devices of the time and lent themselves to continued miniaturization, ease of manufacture, and low cost per device. In addition, digital devices augmented by digital error correction delivered numerical precision and reproducibility limited only by the available physical resources. Also, since any function could be implemented digitally, the digital approach to computing turned out to be, in an important sense,

universally applicable. Thus, many of our key research challenges have involved finding the right trade-offs between the potential energy efficiency and performance of analog systems and the obvious benefits of digital systems. In addition, when such systems are found to be useful for computation, the underlying models of computation need to be developed so that these systems can be evaluated appropriately.

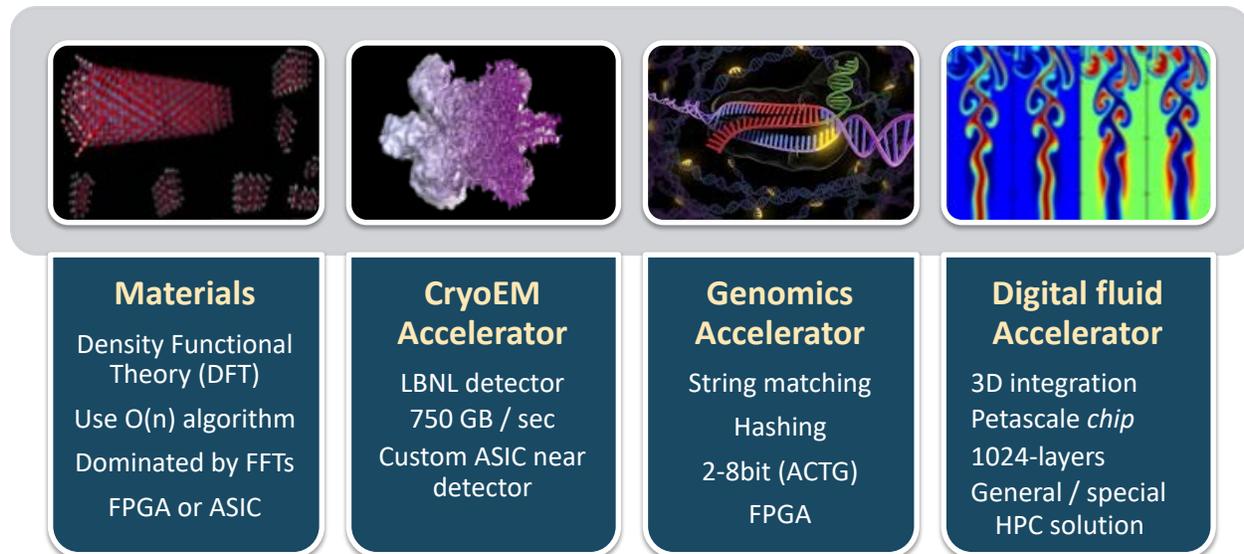


Figure 6. Other opportunities for non-von Neumann processing for scientific discovery (FPGA = field-programmable grid array). Courtesy of John Shalf, Lawrence Berkeley National Laboratory.

The overarching goal of the proposed research is to leverage novel physical processes to measure, process, store, and communicate information. Both natural and man-made devices are carrying out functions for computing, sensing, energy generation, force transduction, bio-regulatory operations such as protein-folding, etc. The repertoire is rich, with devices operating on electronic, mechanical, magnetic, and biochemical principles. Examples include current CMOS transistors, bio-molecular machines such as mitochondria and ribosomes, ionic and memristive devices, spintronics, photonics, superconducting Josephson junctions, carbon nanotubes, nano/micro-electro-mechanical systems, DNA, and systems of neurons. Some of these options are illustrated in Figure 6. Voltage scaling of the CMOS transistor has reduced system energy consumption by many orders of magnitude, although we remain far from limits set by thermal voltage fluctuations and acceptable error rates, i.e., the Landauer limit – the minimal energy loss associated with erasure of information.^{10,11} Importantly, as new low-voltage devices are introduced and device dimensions are further reduced, thermodynamic fluctuations will make devices increasingly stochastic in operation. The current paradigm for computing is inadequate for harnessing such stochasticity, even when it is needed at the algorithmic level (e.g., randomized algorithms).

The suite of existing devices can be coarsely categorized as either (1) artificial ones such as CMOS transistors whose design is guided purely by computational and functional goals or (2) systems that evolve under strong thermodynamic constraints but with limited or no consideration for specific computational goals. Examples of the latter range from natural systems, such as mitochondria or systems of cellular neurons, to artificial systems, such as coupled oscillators for logic circuits. To date, no systematic survey of the computational potential of these class-2 devices has been performed.

An example of class-2 devices is the area of optimizers. A targeted research direction in the study of both existing and potentially new optimizers is urgently needed. A specific need is a systematic examination of how such systems can be synthesized and interconnected for the collective response sought. This area opens up opportunities for designing and examining new classes of materials, beyond conventional semiconductors, that can be tailored for the response sought – related to electronic phase transitions; magnetic response; electronic, ionic, or electromagnetic excitations; or positive feedback effects unlike those from FETs, for instance. This area also opens up the need for designing and synthesizing new ways of connecting these collective devices for signal transduction and power delivery and for incorporating features such as self-configurability and adaptability.

The guiding complexity models for modern computing are built on the Turing model. We anticipate that future computing in non-von Neumann architectures will be in one of two classes: (1) those that are Turing complete or (2) those that are used as accelerators for specific computations in conjunction with a von Neumann system. In the abstract then, the foundation for time complexity modeling of modern HPC systems will be applicable to class (1) computing approaches. However, significant theoretical work will be needed to bridge from the non-Von Neumann computational paradigms of class (2) approaches to the Turing model in order to connect the new approaches to the long established and deep body of theoretical results for modern computing.

If successful, finding and using non-von Neumann systems will yield broad and profound improvements in energy efficiency and compute time. This would contribute to sustained U.S. leadership in information technology. Computing for scientific discovery would be invigorated and accelerated. Furthermore, in co-designing full systems across computer architecture, algorithms, and energy efficiency, we may discover and develop new ways of reasoning about computation.

REFERENCES

1. J.S. Vetter, R. Brightwell, M. Gokhale, P. McCormick, R. Ross, J. Shalf, K. Antypas, D. Donofrio, A. Dubey, T. Humble, C. Schuman, B. Van Essen, S. Yoo, A. Aiken, D. Bernholdt, S. Byna, K. Cameron, F. Cappello, B. Chapman, A. Chien, M. Hall, R. Hartman-Baker, Z. Lan, M. Lang, J. Leidel, S. Li, R. Lucas, J. Mellor-Crummey, P. Peltz, Jr., T. Peterka, M. Strout, and J. Wilke, *Extreme Heterogeneity 2018: DOE ASCR Basic Research Needs Workshop on Extreme Heterogeneity*, Department of Energy, Office of Science, Advanced Scientific Computing Research, doi:10.2172/1473756 (2018).
2. Defense Advanced Research Projects Agency, *DARPA Electronics Resurgence Initiative*, <https://www.darpa.mil/work-with-us/electronics-resurgence-initiative>, accessed May 2019.
3. J.M. Shalf and R. Leland, Computing beyond Moore's Law, *IEEE Computer*, 48(12) (2015) 14-23.
4. M. Sellier, J.-M. Portal, B. Borot, S. Colquhoun, R. Ferrant, F. Boeuf, and A. Farcy, Predictive delay evaluation on emerging CMOS technologies: A simulation framework, 9th Int. Symp. on Quality Electronic Design (ISQED '08), San Jose, CA, pp. 492-497 (2008).
5. P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R.S. Williams, and K. Yelick, *ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems*, Defense Advanced Research Projects Agency Information Processing Techniques Office, <http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf> (2008).
6. Neil Thompson, The economic impact of Moore's Law: Evidence from when it faltered, <http://dx.doi.org/10.2139/ssrn.2899115> (2017).
7. *The Materials Project*, <https://materialsproject.org>, accessed July 2019.
8. A. Olofsson, *Common Heterogeneous Integration and IP Reuse Strategies (CHIPS)*, Defense Advanced Research Projects Agency, <https://www.darpa.mil/program/common-heterogeneous-integration-and-ip-reuse-strategies>, accessed May 2019.
9. T. Liu, J. Hoff, G. Deptuch, and R. Yarema, A new concept of vertically integrated pattern recognition associative memory, *Physics Procedia*, 37 (2012) 1973-1982.
10. R. Landauer, Dissipation and heat generation in the computing process, *IBM Journal of Research and Development*, 5 (1961) 183-191.
11. Ian A. Young and Dmitri E. Nikonov, Principles and trends in quantum nano-electronics and nano-magnetics for beyond-CMOS computing, 47th European Solid-State Device Research Conference (ESSDERC), pp. 1-5 (2017).

This page intentionally left blank.

Panel 3 Power Conversion, Control, and Detection

INTRODUCTION

Panel 3 focused on research needs to enable future electronics for the control and conversion of high-voltage and high-current electrical power. Such power conversion is critically important for applications such as the electrical grid, transportation and machinery, and harsh environments such as those experienced by sensors and in DOE facilities such as accelerators. Discussion centered around topics related to energy efficiency and portability that could be enabled by a new generation of ultra-wide bandgap (UWBG) semiconductor materials, defined to be those with bandgaps greater than that of gallium nitride (3.4 eV). Also discussed were novel device designs that can fully exploit the superior properties of UWBGs for power electronics, sensing, optical devices, and other DOE-relevant applications. Related materials, such as magnetic and dielectric materials capable of high-power and high-frequency operation, were also addressed. Additionally, the group discussed methods to integrate these new materials by a co-design approach that takes maximal advantage of their properties at the system level.

The panel discussion began with a series of short presentations by each of the panelists, coupled with a discussion of how the topics presented related to the plenary presentations earlier in the day. An initial discussion also covered differences between fundamental and applied research, top-down and bottom-up driven research, and the most appropriate way to group the topics addressed by the panel.

CURRENT STATUS AND RECENT ADVANCES

The grid was taken as the prototypical application of interest. Achieving a next-generation grid – with programmability and the ability to rapidly reconfigure itself in response to threats, component failures, etc. – will require new power electronics capable of processing very high voltages and currents.¹⁻⁴ This results in technology needs for UWBG devices, including bipolar devices with conductivity modulation and new high-voltage components in general. There is a need for high-bandwidth and high-speed sensors, thermally aware control and power processing electronics, and new circuit topologies using components beyond traditional metal-oxide varistors, inductors, and capacitors. A “solid-state transformer” may be a key enabling converter in the future. To accomplish this, high-flux and high-density magnetic materials are also needed, as are new winding designs. Defense against electromagnetic pulses, either natural or malicious, is a key concern, as the widespread replacement of traditional transformers with solid-state alternatives could potentially increase vulnerability unless electromagnetic pulse mitigations are developed.⁵ Additionally, many pulsed-power applications cannot be realized with existing semiconductors because of their slow response times and inadequate breakdown fields and current densities. To address the needs of the grid and other applications of interest to DOE, the panel discussed several key topics, including UWBG semiconductors, high-field and high-frequency magnetic and dielectric materials, and electro-thermo-mechanical co-design concepts.

Of primary importance, and underpinning most of the other research needs, is the fundamental physics of emerging materials. Wide-bandgap (WBG) devices such as those based on silicon carbide (SiC) and gallium nitride (GaN) represent today’s state of the art,⁶⁻⁹ and new UWBG semiconductors of the future must surpass the capabilities that these materials provide. For both more mature and less mature materials, 3-5, 5-10, and 10-20 year timeframes are natural dividing points. The panel discussed the UWBG workshop that was convened by Sandia National Laboratories, the Air Force Office of Scientific Research, and the National Science Foundation in April 2016, and the subsequent study and report.¹⁰ One focus of this discussion was materials parameters, and it was noted that many of the commonly used parameters for emerging UWBG semiconductors are either unknown or incorrect. The study attempted to correct this deficiency by tabulating the best-known values (an example of this concerns diamond, where many commonly used parameters are based on old work performed on immature material from approximately 30 years ago). The following prototypical UWBG

semiconductors were discussed extensively by the panel: aluminum gallium nitride (AlGaN), diamond, and gallium oxide (Ga_2O_3). Detailed observations regarding these materials were provided by several of the panelists and are summarized below.

The *AlGaN alloys* have exceptional properties.¹¹ For electronics, they offer (1) direct bandgaps spanning a wide range (3.4 to 6.0 eV), enabling heterostructure devices, (2) high breakdown electric fields (>10 MV/cm for higher Al compositions), (3) high electron mobility for some compositions (>1000 $\text{cm}^2\text{V}^{-1}\text{s}^{-1}$), (4) high saturation velocities ($>10^7$ cm s^{-1}), and (5) readily achievable n-type doping using Si, which has a relatively small donor ionization energy up to $\sim 80\%$ Al composition. All of these are outstanding characteristics for power electronics. However, the AlGaN alloys face three major challenges: (1) difficulty in achieving controlled p-type doping,¹² (2) the lack of readily available large-area single-crystal AlN substrates with the quality necessary for epitaxial growth, and (3) gaps in scientific understanding needed for highly controlled AlGaN epitaxy on such substrates. One method by which the doping challenges may be addressed is polarization-induced doping, which arises when the composition of the material is graded in space and, hence, does not rely upon thermal activation of impurities. AlGaN also suffers from gaps in understanding of high-field phenomena (e.g., breakdown) and carrier confinement. These gaps are common to all the UWBG materials, and suitable experiments and theoretical studies to address them are required.

Diamond, with a bandgap of ~ 5.5 eV, has extreme properties¹³ similar to those of AlGaN, which enable applications such as high-power and high-frequency electronics,^{14,15} radiation detectors, electron emitters for ultra-high-voltage vacuum switches¹⁶ and traveling-wave tube cathodes, and thermionic emitters for energy conversion. Some of diamond's outstanding electronic properties include (1) breakdown electric fields potentially >10 MV/cm, (2) high electron and hole mobilities ($>2,000$ $\text{cm}^2\text{V}^{-1}\text{s}^{-1}$), (3) high saturated drift velocity, and (4) low dielectric constant. Diamond has the highest known thermal conductivity of any material, which is of great importance because in many power electronic applications the device operation is heat-sink limited. Recent materials breakthroughs for diamond include realization of single-crystal substrates¹⁷ with low defect density ($<10^5$ cm^{-2}) and demonstration of n-type doping using phosphorus. For doping densities above $\sim 10^{19}$ cm^{-3} the dopant energy levels spread into the conduction band due to degeneracy, which results in high conductivity (hopping conduction may play a role). Below this doping concentration, standard drift-diffusion transport of the limited number of carriers that are ionized occurs in the conduction band.

The third example of an UWBG semiconductor is Ga_2O_3 .^{18,19} For unipolar power devices, Ga_2O_3 offers some outstanding attributes. First, it has a bandgap of ~ 4.5 eV, leading to a large breakdown electric field of 7-8 MV/cm. Second, it displays good controllability of n-type conductivity over a wide range of n ($\sim 10^{15}$ - 10^{19} cm^{-3}) using Si or Sn doping, and a tunable resistivity spanning the range $\sim 10^{-3}$ - 10^{12} $\Omega\text{-cm}$. However, perhaps the greatest advantage of β -phase Ga_2O_3 is the availability of large-area, affordable, high-quality native substrates. By contrast, the two main drawbacks of Ga_2O_3 are (1) absence to date of reports of successful p-type doping and (2) very poor thermal conductivity. The low thermal conductivity is perhaps the single most serious potential weakness of Ga_2O_3 for power devices, and methods are required to circumvent this problem. We note also that the crystal structure of Ga_2O_3 is quite complex (the β phase is the most common, but several other polymorphs exist) and determines the nature of phenomena such as electric and thermal transport, as well as defect physics (including self-trapping). This complexity and its ramifications on Ga_2O_3 material properties must be fully understood.

SCIENTIFIC CHALLENGES AND OPPORTUNITIES

To mature a technology over a 20-30 year timeframe, a sequence of R&D steps is often needed. The potential path by which this sequence might best occur for power conversion was the focus of much of the discussion. These steps involve the development and maturation of materials, devices, models (at various levels from atomistic to system), and manufacturing processes, as well as eventual market adoption. Several different semiconductors may be required in the future to achieve needed device functions. These steps are often interrelated; for instance, demands of a particular application may necessitate specific device architectures, which, in turn, can drive the selection of certain doping techniques over others. Below we describe in more detail these different areas.

Materials and Devices

Semiconductor materials

The maturity of the UWBG semiconductors is at present quite low. Challenges include achieving high crystalline quality, understanding the physics of growth and doping (Figure 1), and creating well-researched methods (“toolkits”) for the processing of contacts, reduction of defects, etc. For semiconductor materials, there is a need to focus on substrates (Figure 2), including their supply, as well as bulk and epitaxial growth, and to determine how these interplay with the fine details of the substrate morphology.

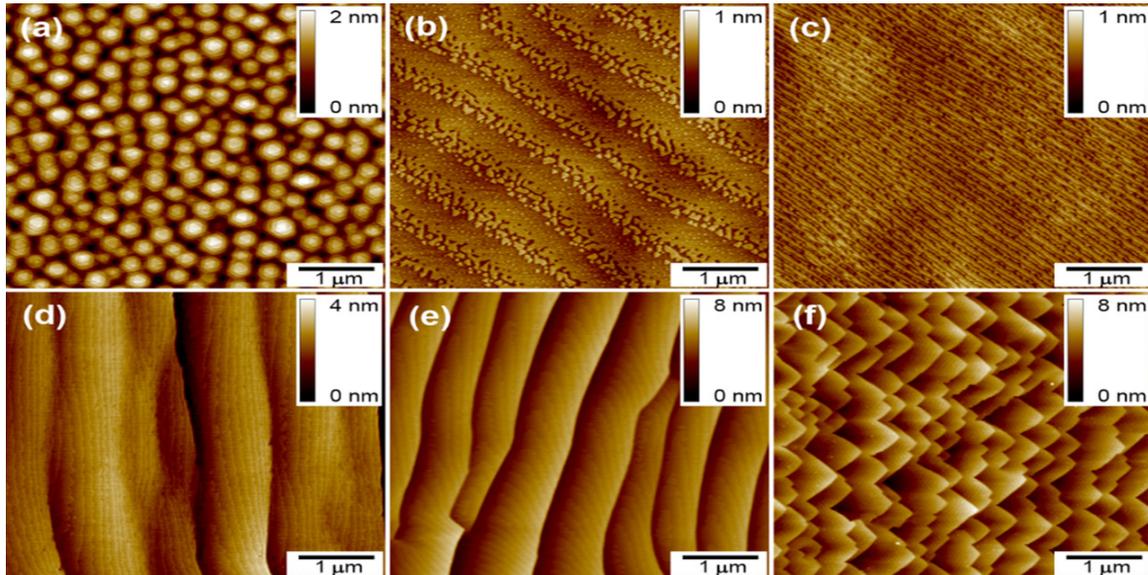


Figure 1. AlGaN surface morphology for different epitaxial growth conditions. Courtesy of Zlatko Sitar and Ramon Collazo, North Carolina State University.



Figure 2. WBG and UWBG substrates. Courtesy of Jacob Leach, Kyma Technologies.

One of the most exciting aspects of UWBG semiconductors is that they present a new realm in which to study the science of the mutual interactions between electrons and holes (both charge and spin), photons, and phonons. UWBG materials were once thought of as insulators, and research challenges include the transport and dynamics of charge carriers in high-field and non-equilibrium regimes. The physics of high-field behavior

in particular, such as transport and breakdown, is largely unknown. This fundamental research problem needs to be addressed. Breakdown has not been carefully measured in the newer UWBG semiconductors, and the ionization parameters are not well known as a function of electric field. Although the purity of the material is believed to be very important, the role of impurities in high-field behavior is poorly understood.

Other topics of interest include defect and impurity centers, and different types of conduction such as hopping. One good example is a diamond PiN diode, where transport is space-charge-limited, like in a vacuum tube. A second example is a deep-depletion MOSFET. Where do the carriers in this structure come from? The time scale for minority carriers to be generated is extremely long – much longer than the timescale over which the device is operated. UWBG materials open the possibility of operating in regimes that have not been considered before. Heterostructures, including unusual dielectrics, may be aligned by the charge-neutrality level, and a single monolayer can change the band alignment by several electron volts. Heterojunctions with energy differences of a few electron volts between the constituent materials are common, and the semiconductors have bandgaps equivalent to those of the dielectrics. In general, these new physical regimes and configurations have the potential to produce phenomena not previously observed, especially as we approach the fundamental physical limits of the materials.

Broadly speaking, an UWBG semiconductor materials toolbox is needed that focuses on fully realizing the properties required for high semiconductor material quality and device performance. Substrate requirements need to be well-defined, and a domestic capability for single-crystal materials growth that can achieve the desired properties is needed.

There are not, at present, many domestic sources of high-quality UWBG wafers. This lack is of great concern for U.S. interests. Development of high-quality n-type wafers for vertical power devices is needed, as well as robust UWBG epitaxy that is uniform and reproducible. Effective surface passivation schemes are also needed, especially for polar materials. The limits of high-temperature operation are currently not imposed by the semiconductors, but rather by the components on the periphery, which also should be the subject of further research. A holistic approach to device reliability is required.

One tantalizing prospect is to develop the ability to predict new materials that have high breakdown strength. The conventional approach for new material development is to formulate a list of candidate materials, then experimentally characterize them in detail. Using this approach, a new material may take 20-30 years to mature. The second approach is to methodically look at new materials using theory and computation first. In this case, one might computationally examine a new material that has many more atoms in the unit cell than is typical of today's state-of-the-art materials. Using this latter approach may cut down on the development time dramatically.

Overall, it will be important for groups to not work in silos for the different materials, as is largely done now – universal principles that are applicable to all UWBG materials should be shared among research groups. For example, computational materials experts using techniques such as density functional theory should be engaged in discussions with experimentalists conducting epitaxial growth. Overarching problems such as designing and growing new materials, identifying new dopants and exploring new methods of doping, and understanding how point and extended defects interact with dopants all require experts in different fields to work together.

Devices that work at extreme voltages and currents

In addition to research on materials, future advances in power conversion, control, and detection will require research on new device architectures to enable operation at extreme voltages and currents. To keep systems at a manageable size and to provide short response times, such devices will also need to operate at higher frequencies than power devices have done previously. Novel device architectures will, of course, require close interplay with materials advances; because of this, the topical division into categories of “materials” and “devices” is somewhat arbitrary.

To evaluate device performance, various figures of merit (FOMs) are often used as a metric to compare measured results to theory. The most well-known is the unipolar FOM (Figure 3), applicable to unipolar devices such as MOSFETs. Subtleties exist when comparing vertical-to-lateral devices, and differently defined FOMs need to be

correctly employed for different situations. Utility-level devices likely require vertical geometry (to achieve the highest breakdown voltages and current densities), good ohmic contacts, and bipolar architectures (to enable higher performance than unipolar devices in terms of FOM).

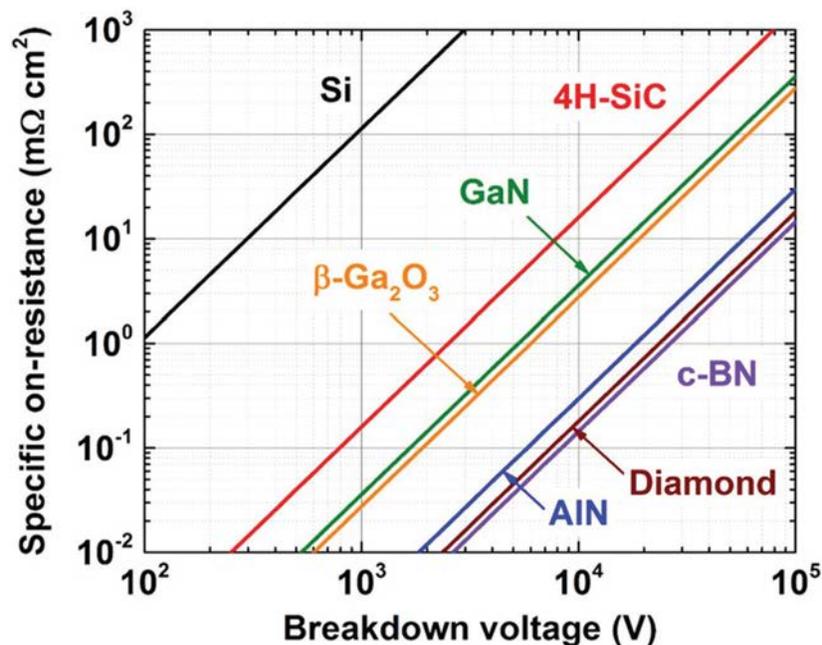


Figure 3. Unipolar figure of merit for WBG and UWBG semiconductors. From J.Y. Tsao et al., *Advan. Electron. Mater.*, 4 (2018) 1600501.

Note that these guidelines and FOMs were established long ago at a time when UWBG semiconductors were not prevalent. Their definitions and applicability must be re-evaluated as new and evolving experimental evidence made available by UWBG semiconductor materials and devices is obtained.

To achieve the theoretical FOM, one needs to be able to make a nearly ideal device, with precise control of defects and doping, yet also keep the carrier mobility high. The most familiar doping approach is impurity doping, but this becomes ineffective for certain UWBG materials because their impurity energy levels are deep within the bandgap, where thermal activation is ineffective. Thus, a key question is how to controllably achieve efficient doping schemes within the emerging UWBG materials.

One example of a novel approach for manipulating the energy levels of impurity dopants is quasi-Fermi level control during epitaxial growth.^{20,21} This can modulate the formation energies of a wide variety of point defects, including the impurity-vacancy complexes that always arise in UWBG semiconductors. Additionally, alternatives to conventional impurity dopants exist, such as modulation doping, but their suitability for power devices is unclear and needs further investigation.

Polarization-induced doping has already been shown to be an effective method with polar materials such as AlGaN,²² and new growth techniques may be required to more effectively utilize this approach for the highest performance. In addition, laterally patterned doping is generally required for vertical devices, and several techniques should be investigated. For selective-area growth, problems growing on etched surfaces and at corners exist. In addition, p-n junctions may be achieved by selective-area growth methods, although challenges exist related to high leakage currents, presumably due to interface disorder and/or impurities. Indeed, this is a fundamental problem, not just a processing problem – how can sharp interfaces be achieved that contain few or no defects?

The growth of thick, low-doped epilayers is critically important for achieving high stand-off voltages in diodes and vertical transistors, and a doping level of less than $\sim 10^{15} \text{cm}^{-3}$ is likely required to exceed the performance of power devices based on conventional semiconductors. However, it remains a key challenge to epitaxially grow thick, low-doped epilayers repeatably and reliably. Further, many localized energy states can compensate

intentional impurity dopants in the gap for UWBG materials,²³ and characterization and calculation of these deep levels are critical to ensure that the bandgap is “clean”. Finally, wafers can warp, and one needs to be concerned about stress during growth. Dislocations can relieve this, but they are of course undesirable from an electrical standpoint. The consensus of the panel was that larger, more coordinated efforts than what are being undertaken now are needed to solve these problems.

Dielectric and magnetic materials

To realize the tremendous advantages of UWBG semiconductors for the grid and other power applications, it will be necessary to improve the performance of passive components (capacitors and inductors) so they can operate at the higher frequencies involved. Advances in high-frequency magnetic and dielectric materials will require detailed, fundamental investigations that span from advanced synthesis and processing techniques to system- and component-level integration and testing. Mechanisms and methods for tailoring magnetic and dielectric properties through carefully engineered processing methods in magnetic and/or electric fields can also play an important role. Dielectric breakdown mechanisms at high frequencies (kHz-MHz), which may differ from those at DC, must also be carefully explored. For capacitors, both high-temperature operation (>150°C) and low equivalent series resistance (necessary to avoid self-heating due to large ripple currents at high-frequency operation) will be required. An example of a promising research direction is that of novel ceramic dielectrics, such as barium/bismuth-titanate-based compounds, which have recently shown promising results.²⁴⁻²⁶ These materials retain high dielectric constants at temperatures greater than 150°C. However, processing concerns, such as the need to co-fire these dielectrics with base metal electrodes, and lifetime concerns associated with electromigration of point defects, have yet to be addressed. Furthering the understanding of how DC degradation occurs, and how it can be mitigated in these new ceramic dielectrics, is of key interest. Additionally, circuit-level designs that allow for a subset of ceramic DC-link capacitors to periodically relax point defects and repair the degradation are of interest.

Advanced magnetic properties are being realized through novel synthesis techniques that enable micro- and nano-structures not previously attainable by traditional processing methods.²⁷⁻²⁹ In such emergent systems, deeper understanding is needed of the detailed structures, properties, and processing inter-relationships that account for the fundamental thermodynamic and kinetic driving forces of micro- and nano-structure formation. Such increased understanding will enable better control over synthesis, potentially yielding precision engineering of the materials properties. A need also exists for fundamental characterization of advanced magnetic, dielectric, and insulation materials under a wide range of frequencies, temperatures, and excitation conditions that are consistent with the waveforms in emergent UWBG-based power converters. Tailoring of magnetization processes through processing in applied magnetic fields and engineering of magnetic domain structures are examples of promising approaches.³⁰ These can both mitigate against eddy currents and provide desired saturation inductions and tunable permeabilities for a wide range of end-use device applications.

The fundamental mechanisms that are responsible for induced magnetic anisotropies of advanced soft magnetic materials (Figure 4) are notoriously difficult to clarify with certainty, yet are of critical importance for end-use applications. The elucidation is needed of the mechanisms responsible for measured magnetic anisotropy after anisotropic thermal processing strategies. Here, advanced and localized probes of short-range chemical and magnetic ordering, interfacial magnetic and electronic states, and the symmetry of microstructural texture, shape, and even crystal defects are expected to shed light on those mechanisms. Characterization of ferromagnetic exchange strength and distribution in complex micro- and nano-structures and chemistries must also be coupled with fundamental theoretical calculations based upon first principles. Due to the role of eddy currents in the high-frequency losses of magnetic materials, fundamentals of electronic transport and the interplay with electronic structure, magnetic properties, and microstructure must also be understood and

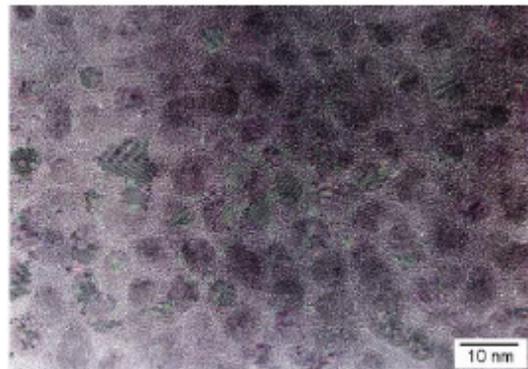


Figure 4. Electron micrograph of soft magnetic material. Courtesy of Paul Ohodnicki, National Energy Technology Laboratory.

described. To fully realize the advanced micro- and nano-structured magnetic materials that satisfy this unique interplay between high resistivity and large saturation induction, advanced processing techniques and synthesis methods, including synthesis under applied magnetic fields for engineered anisotropy, will need to be explored and developed.

Applications

Grid

From a grid applications perspective, several factors need to be considered in facilitating the maturation of power conversion technology. Today's grid is, for the most part, set up with fixed electricity sources and unidirectional power flow. However, that is now starting to change: innovations such as bidirectional flow and distributed generation are being introduced (Figure 5). This results in lower system inertia and faster dynamics of step changes on the grid; thus, faster dynamics for power electronics on the grid are needed. Using an analogy with telecommunications, while the current grid has the same topology as the centralized, unidirectional radio and television broadcast systems of a few decades ago, the new electrical grid will have the topology of the internet – distributed, multidirectional, and dynamic. It will achieve for electrical power what packet switching did for information.

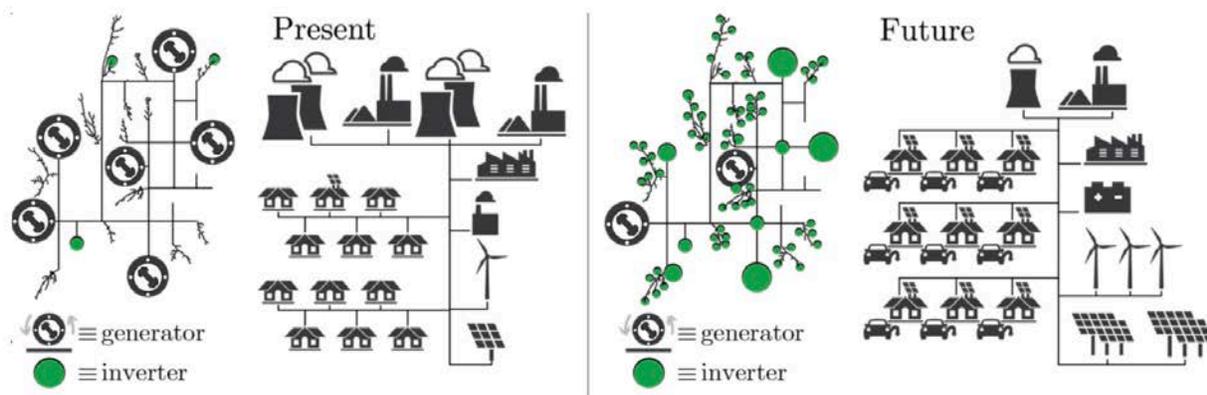
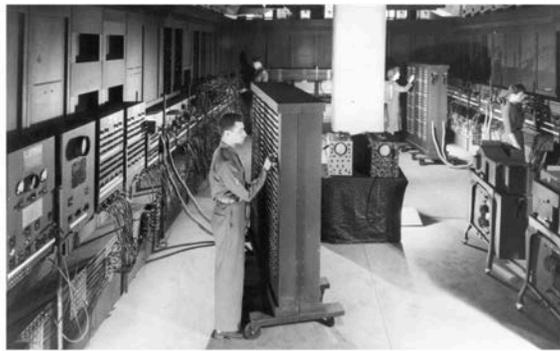


Figure 5. Present-day vs. potential future grid architectures. The former is characterized by fixed sources and unidirectional power flow, while the latter features distributed and renewable sources as well as bidirectional power flow. Courtesy of Brian Johnson, University of Washington.

The grid is also moving towards incorporating more electronic loads, which have impedances that can inject complex harmonics into the grid. The advent of microgrids in rural and urban areas, as well as hybrid AC/DC systems, is also occurring. Power converters can be used to artificially introduce inertia into these systems, increasing their stability, but high-bandwidth power electronics are needed for them to be able to respond on the needed timescales. High-frequency passives are also needed. Such inductors and capacitors will require new materials capable of higher performance, as discussed earlier.

High-voltage blocking devices fabricated from WBG and UWBG semiconductors are also needed. While some progress has been made, reliability is currently a problem at voltages greater than 10 kV, and this needs to be addressed and remedied. High-ratio buck and boost converters are required, but they must also be highly efficient. High-voltage DC transmission will need high-voltage power converters that are at least as small as a refrigerator, and ideally as small as a suitcase (“substation in a suitcase”), requiring daunting power densities. Such a dramatic transformation is challenging but has been achieved in other fields, notably in information processing (Figure 6).



ENIAC, 1946



Today's Microprocessor



Two of five transformers in transmission substation, Melbourne, Australia



Potentially 100 lbs, Solid-State Transformer

Figure 6. "Substation in a suitcase" and analogy with the miniaturization progress for digital computation. Courtesy of Jerry Simmons, Sandia National Laboratories. Images reproduced from <https://en.wikipedia.org/wiki/ENIAC> (upper left), https://www.flickr.com/photos/intel_de/9662276651 (upper right), <https://www2.lbl.gov/mfea/assets/docs/presentations/Steven-Chu-Presentation-MFEA.pdf> (lower right), and https://en.wikipedia.org/wiki/Electrical_substation (lower left).

There is a critical need for high-frequency magnetic materials to fully realize the advantages of power devices based on UWBG semiconductors. Promising candidates, at present, include emerging nanoscale composite magnetic materials as well as new compounds such as iron nitride.³¹ However, it is not clear which of the competing approaches will ultimately end up being best, and fully exploring and evaluating them is a basic research need. Overall, a strong need exists for research on materials for passive inductors and capacitors, which currently lag significantly behind semiconductors, especially in the area of reliability.

Electric and autonomous vehicles

The panel also discussed the application area of electric and autonomous vehicles. One topic was the question of how energy can be delivered to an autonomous vehicle most effectively. Wireless charging is one possibility. Another would involve removing the batteries from the vehicle, and then powering the vehicle directly from the roadway using inductive or capacitive power transfer. For that purpose, 13.56 MHz systems operating at 0.9 kW could be used, with efficiency in the 80-90% range up to a power density of 40 kW/m². The power coupling will depend critically on the misalignment between the vehicle and the chargers embedded in the road. In general, for wireless charging as well as electric powertrains, the power conversion will depend on three main factors: (1) better components, (2) novel circuit topologies, and (3) innovative conversion architectures. The best features of new devices need to be utilized through effective circuit design. A final research need for this application area is systems science. This need can be addressed at different levels; for the electric car, it involves not only the system in the car but the grid as well, and the coupling between the two.

Need for a Comprehensive System-of-Systems Approach

The panel broadly agreed that a high-level system-of-systems modeling approach is needed for all areas of application for power electronics. This includes the electric grid, buildings, transportation, industry, and the military. While these are a diverse set of applications with different concerns and needs, the grid can serve as a prototypical application because most of the other applications touch the grid and will have common performance requirements. Complex power-electronics-based systems have many benefits, such as lower transmission losses and integration of light-emitting diode lighting. Power electronics can link these different pieces, analogous to a router. To answer questions about how the system can be made to work as desired, with optimal performance, a reasonable co-simulation mechanism is required. The different parts of the overall system can be modeled at different levels of fidelity, and then merged. As an example, for the power grid, high-voltage, series-stacked systems need to be designed such that either failure is avoided, or if failure does occur, overall operation can be sustained. Thus, basic research on system-level modeling and optimization, across the range of anticipated applications, is required. There are already urgent challenges now, and they will only become even larger as time progresses.

Regarding the tie between materials/devices and circuits/systems, a need exists to make power converters more compact and efficient at the point of load. UWBG materials and devices will be critical for this purpose. Conduction losses need to be reduced, which can be accomplished by bringing the high voltage closer to the point of use. Existing devices can be improved considerably by systematically reducing parasitic losses.

Another observation is that power management can be simplified by using chip-level packaging and even monolithic integration, which could alleviate problems with the different gate drives required for different device types. A focus needs to be placed on high-temperature, compact, and reliable power electronics systems. Thermo-mechanical reliability is essential; high-temperature operation requires not just the semiconductor material to withstand the environment, but also the other elements in the device, in the package, and in the system. Some areas in need of research include die attachments, bonded interfaces, multi-scale modeling, and metamaterials for heat flow control and dissipation. New ways to integrate disparate materials are needed, as are novel ways of using materials to manipulate heat flow. New materials for package substrates that improve heat flow, including across the interfaces in the system, must be developed. Multi-scale physics-of-failure modeling is also important in the design, development, and fabrication of reliable UWBG devices and packaging.

REFERENCES

1. Quadrennial Energy Review Task Force, *Transforming the Nation's Electricity System: The Second Installment of the Quadrennial Energy Review*, <https://www.energy.gov/sites/prod/files/2017/02/f34/Chapter%20I--Transforming%20the%20Nation%27s%20Electricity%20System.pdf> (2017).
2. Quadrennial Energy Review (QER) Task Force, *Modernizing the Electric Grid*, https://www.energy.gov/sites/prod/files/2015/04/f22/QER%20ch3%20final_0.pdf (2015).
3. U.S. Department of Energy, *Grid Modernization Multi-Year Program Plan*, <https://www.energy.gov/sites/prod/files/2016/01/f28/Grid%20Modernization%20Multi-Year%20Program%20Plan.pdf> (2015).
4. J.G. Kassakian, R. Schmalensee, et al., *The Future of The Electric Grid: An Interdisciplinary MIT Study*, <http://energy.mit.edu/wp-content/uploads/2011/12/MITEI-The-Future-of-the-Electric-Grid.pdf> (2011).
5. J.S. Foster, E. Gjeldel, W.R. Graham, R.J. Hermann, H.M. Kluepfel, R.L. Lawson, G.K. Soper, L.L. Wood, and J.B. Woodward, *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack*, http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf (2008).
6. T. Kimoto and J.A. Cooper, *Fundamentals of Silicon Carbide Technology*, New York: Wiley (2014).
7. M. Meneghini, G. Meneghesso, and E. Zanoni, eds., *Power GaN Devices: Materials, Applications and Reliability*, Switzerland: Springer International (2017).
8. H. Amano et al., The 2018 GaN power electronics roadmap, *J. Phys. D*, 51 (2018) 163001.
9. PowerAmerica, *Strategic Roadmap for Next Generation Wide Bandgap Power Electronics*, https://www.poweramericainstitute.org/wp-content/uploads/2017/01/PowerAmerica_Roadmap_Final-Public-Version-January-2017.pdf (2017).

10. J.Y. Tsao, S. Chowdhury, M.A. Hollis, D. Jena, N.M. Johnson, K.A. Jones, R.J. Kaplar, S. Rajan, C.G. Van de Walle, E. Bellotti, C.L. Chua, R. Collazo, M.E. Coltrin, J.A. Cooper, K.R. Evans, S. Graham, T.A. Grotjohn, E.R. Heller, M. Higashiwaki, M.S. Islam, P.W. Juodawlkis, M.A. Khan, A.D. Koehler, J.H. Leach, U.K. Mishra, R.J. Nemanich, R.C.N. Pilawa-Podgurski, J.B. Shealy, Z. Sitar, M.J. Tadjer, A.F. Witulski, M. Wraback, and J.A. Simmons, Ultra-wide-bandgap semiconductors: Research opportunities and challenges, *Advan. Electron. Mater.*, 4 (2018) 1600501.
11. R.J. Kaplar, A.A. Allerman, A.M. Armstrong, M.H. Crawford, J.R. Dickerson, A.J. Fischer, A. G. Baca, and E.A. Douglas, Review – Ultra-wide-bandgap AlGaIn power electronic devices, *ECS J. Solid-State Sci. Tech.*, 6 (2017) Q3061.
12. Y.H. Liang and E. Towe, Progress in efficient doping of high-aluminum-containing group-III nitrides, *Appl. Phys. Rev.*, 5 (2018) 011107.
13. Robert J. Nemanich, John A. Carlisle, Atsushi Hirata, and Ken Haenen, CVD diamond – Research, applications, and challenges, *MRS Bulletin*, 39 (2014) 490.
14. H. Kato, K. Oyama, T. Makino, M. Ogura, D. Takeuchi, and S. Yamasaki, Diamond bipolar junction transistor device with phosphorus-doped diamond base layer, *Diamond Rel. Mater.*, 27 (2012) 19.
15. S.Koizumi, H. Umezawa, J. Pernot, and M. Suzuki, eds., *Power Electronics Device Applications of Diamond Semiconductors*, Elsevier (2018).
16. D. Takeuchi, H. Kawashima, D. Kuwabara, T. Makino, H. Kato, M. Ogura, H. Ohashi, H. Okushi, S. Yamasaki, and S. Koizumi, Proc. of IEEE 27th Inter. Symp. on Power Semiconductor Devices & IC's (IPSPD), May 10–14, 2015, Kowloon Shangri-La, Hong Kong, p. 197 (2015).
17. M. Schreck, J. Asmussen, S. Shikata, J.-C. Arnault, and N. Fujimori, Large-area high-quality single crystal diamond, *MRS Bull.*, 39 (2014) 504.
18. S.J. Pearton, J. Yang, P.H. Cary IV, F. Ren, J. Kim, M.J. Tadjer, and M.A. Mastro, A review of Ga₂O₃ materials, processing, and devices, *Appl. Phys. Rev.*, 5 (2018) 011301.
19. S.J. Pearton, F. Ren, M. Tadjer, and J. Kim, Perspective: Ga₂O₃ for ultra-high-power rectifiers and MOSFETs, *J. Appl. Phys.*, 124 (2018) 220901.
20. F. Kaess, P. Reddy, D. Alden, A. Klump, L.H. Hernandez-Balderrama, A. Franke, R. Kirste, A. Hoffman, R. Collazo, and Z. Sitar, The effect of illumination power density on carbon defect configuration in silicon doped GaN, *J. Appl. Phys.*, 120 (2016) 235705.
21. K. Alberi and M.A. Scarpulla, Effects of excess carriers on charged defect concentrations in wide bandgap semiconductors, *J. Appl. Phys.*, 123 (2018) 185702.
22. J. Simon, V. Protasenko, C. Lian, H. Xing, and D. Jena, Polarization-induced hole doping in wide-band-gap uniaxial semiconductor heterostructures, *Science*, 327 (2010) 60.
23. P. Pampili, P.J. Parbrook, Doping of III-nitride materials, *Mat. Sci. in Semicond. Proc.*, 62 (2017) 180.
24. S. Kumar and K.B.R. Varma, Influence of lanthanum doping on the dielectric, ferroelectric and relaxor behavior of barium bismuth titanate ceramics, *J. Physics D: Appl. Phys.*, 42 (2009) 075405.
25. N. Triamnak, G.L. Brennecke, H.J. Brown-Shaklee, M.A. Rodriguez, and D.P. Cann, Phase formation of BaTiO₃-Bi(Zn_{1/2}Ti_{1/2})O₃ perovskite ceramics, *J. Ceram. Soc.*, 122 (2014) 260.
26. N. Kumar, A. Ionin, T. Ansell, S. Kwon, W. Hackenberger, and D. Cann, Multilayer ceramic capacitors based on relaxor BaTiO₃-Bi(Zn_{1/2}Ti_{1/2})O₃ for temperature stable and high energy density capacitor applications, *Appl. Phys. Lett.*, 106 (2015) 252901.
27. A.M. Leary, P.R. Ohodnicki, and M.E. McHenry, Soft magnetic materials in high-frequency, high power conversion applications, *J. Mater.*, 64 (2012) 772.
28. J.M.Silveyra, E. Ferrara, D.L. Huber, and T.C. Monson, Soft magnetic materials for a sustainable and electrified world, *Science*, 362 (2018) eaao0195.
29. D. Li, H. Yun, B.T. Diroll, V.V.T. Doan-Nguyen, J.M. Kikkawa, and C.B. Murray, Synthesis and size-selective precipitation of monodisperse nonstoichiometric MxFe_{3-x}O₄ (M = Mn, Co) nanocrystals and their DC and AC magnetic properties, *Chem. Mater.*, 28 (2016) 480.

30. S. Flohrer, R. Schäfer, J. McCord, S. Roth, L. Schultz, F. Fiorillo, W. Günther, and G. Herzer, Dynamic magnetization process of nanocrystalline tape wound cores with transverse field-induced anisotropy, *Acta Mater.*, 54 (2006) 4693.
31. S. Bhattacharyya, Iron nitride family at reduced dimensions: A review of their synthesis protocols and magnetic properties, *J. Phys. Chem. C*, 119 (2015) 1601.

This page intentionally left blank.

Panel 4 Crosscutting Themes

INTRODUCTION

In parallel with progress in the fundamental science underlying the architectural requirements, algorithms, and software in microelectronics, the paradigm for future progress will be enabled by advances in the corresponding components, which are co-designed for a tailored application. This section focuses on the foundational materials, emerging device-relevant concepts and phenomena, components, and design methodologies needed for a new era of energy-efficient information processing, spanning the previously discussed research thrusts and priority research directions. These platform-level advances have potential to support transformative advances in exascale computing systems, large-scale “big data” processing, and a more efficient and flexible power grid.

COMPONENT ADVANCES IN CO-DESIGN FRAMEWORK

Current Status and Recent Advances

The end of Dennard scaling marks the end of an era in which a “triple play” of ever faster, cheaper, lower power transistors could be anticipated in each successive microelectronics technology generation. For current technologies at the system level, data transport is not proportional to power consumption, and the energy dissipated per bit of operation is not improving with successive generations. Moreover, current system architectures have minimal prospects for achieving the disruptive advances in energy efficiency which are needed in exascale high-performance computing systems and beyond. In large-scale systems, the power distribution and cooling technologies are suboptimal, with significant energy losses attributable to AC-to-DC energy conversion, air cooling, and heat leakage. Such systems are far from achieving energy efficiencies comparable to fundamental limits.

Scientific Challenges and Opportunities

Current integrated circuits have operating frequencies and power dissipation characteristics that are severely limited by their metallic electrical interconnects. However, the development of conceptually new interconnect technologies could result in the ability to design systems with circuit operating frequencies >100x beyond those of today’s systems, at equivalent power dissipation. This goal will require development of new materials with extreme characteristics in electrical and thermal conductivities, as well as optical absorption and radiative efficiency. Discovery of new electronic and photonic transport phenomena and mechanisms using these materials may cause us to re-envision the physical layer substrate for interconnect technologies. These new phenomena are likely to require novel characterization methods to assess interconnect structure and function, as well as new algorithms and system designs for communication-rich architectures. To a considerable extent, new integrated system architectures will be enabled by unlocking the existing “interconnect bottleneck,” so advances in this area cut across all aspects of system design.

In addition to development of new interconnect technologies, advances in physical substrates and design approaches for large-scale memory could circumvent much of the current need for data movement with systems. Thus, creation of new foundational concepts for memory based on new physical phenomena has considerable potential to enable future big-data processing systems.

Success in discovery of novel switching, transmission, and storage mechanisms in new materials would enable us to overcome the million-fold gap between current computing throughputs and those defined by information theory limits. Thus, component-level advances tailored to system requirements by an accelerated co-design framework, linking together new devices, architectures, and algorithms, have the potential to create conditions where grand-challenge scientific computing problems become tractable.

Progress in both interconnects and memory will motivate researchers to exploit new physical phenomena that enable us to re-imagine device physics characterized by faster and more energy-efficient state transitions, and to develop architectures with intrinsically lower power consumption. Thus, there is considerable impetus to investigate new materials and structures that would facilitate information transmission with low energy dissipation

per bit by using new electrical, optical, and spin phenomena, as well as other degrees of freedom. Approaching energy efficiency limits will require fundamentally oriented exploration of how each of these new concepts couples to the ambient phonon bath in the material or structure employed in a microelectronic system in order to develop radically more efficient cooling and power delivery mechanisms.

Order-of-magnitude improvement in the energy efficiency of electronic systems has the potential to enable revolutionary computing architectures for mobile and power-constrained platforms. Further, because energy dissipation in materials and devices has adverse impacts on reliability, development of systems with improved energy efficiency will enable more reliable high-performance computing systems. Improved understanding and control of energy flows in computing systems will enable development of data centers whose overall energy management is better matched to the electric utility grid.

New components developed in a co-design environment will also catalyze a new era of scientific advances in high energy physics, astrophysics and cosmology, chemistry, climate modeling, and scientific fields that are limited by the frontiers of performance in exascale computing. The seamless integration of large-scale computation with communications and sensing has the potential to unleash the design of, for instance, exascale systems with conceptually new memory and interconnect elements coupled with next-generation experimental detectors in high-energy physics that could enable new science to be done with more comprehensive data capture and analysis.

NEW ELECTRONIC MATERIALS AND PHENOMENA FOR INFORMATION AND ENERGY TRANSFER

Current Status and Recent Advances

In the last ten years, the condensed matter physics community has seen rapid and broad advances in developing new materials and understanding basic phenomena. In the electronics realm, these include two-dimensional and layered material conductors with unrivaled transport properties, such as graphene, Weyl semi-metals, and electronic topological insulators. For two-dimensional and layered materials, the ability to readily adjust the carrier density in the active layers by field effect gating is a powerful approach for exploring fundamental phenomena and also designing novel field-effect electronic devices. Recently, unconventional superconductivity has been discovered in twisted bilayers of graphene.¹

In photonics, the discovery of two-dimensional materials with high radiative efficiency² and large exciton binding energies,³ such as transition metal dichalcogenide semiconductors, has motivated new concepts for information processing based on stable room-temperature exciton states.⁴ The transition metal dichalcogenide semiconductors also exhibit “valleytronic” photonic response that is angular momentum-preserving, sparking interest in use of spin and orbital angular momentum degrees of freedom for information processing.⁵ Design of the photonic density of states in a manner in which a photonic spin element is coupled to the wavevector for an optical mode has enabled the creation of photonic topological insulators, whose topologically protected photonic states and modal band dispersions parallel the physics for the electronic topological insulators.⁶ The vast majority of the new materials and phenomena are just now emerging from the incipient discovery phase, and most have not been assessed carefully for the potential in memory or interconnect for microelectronic system applications (see Figure 1).

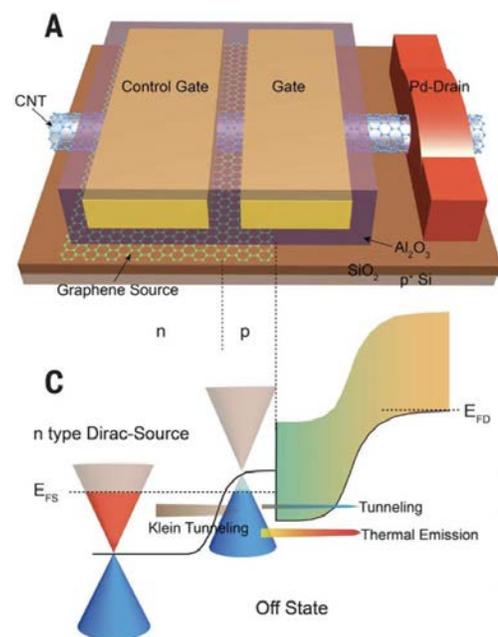


Figure 1. (a) Dirac source FET, a field effect transistor with a Dirac-like electronic density of states in the source where transport is modulated by a gate shown here as a carbon nanotube (CNT). (b) Schematic illustrating the off-state of the Dirac source FET. Reprinted with permission of AAAS from C. Qiu et al., *Science*, 361 (2018) 387–392.

Scientific Challenges and Opportunities

Energy efficient switches by tailoring material density of states

Conventional planar silicon CMOS FETs have a turn-on behavior characterized by a 60 mV/decade sub-threshold conductance slope arising from the electrostatic control of electrons thermionically emitted over the potential barrier separating the source and drain. The electrons have a Boltzmann-like energy distribution that spreads to values exceeding the potential barrier created by the gate.

Unlike the density of states in bulk materials, two-dimensional materials such as graphene have Dirac-like densities of states. Use of a source whose electronic density of states is a decreasing function of energy, characteristic of a Dirac-like material such as doped graphene, allows for conditions where the electron density can decrease super-exponentially with increasing electron energy.⁷ A Dirac source transistor has been shown to exhibit a room-temperature sub-threshold of 40 mV per decade over four decades of current and a current of up to 40 $\mu\text{A}/\text{mm}$ at 60 mV per decade, with an on-state current at 0.5 V comparable to that of a 14-nm node silicon CMOS transistor operating at 0.7 V supply voltage and a sub-threshold slope of 35 mV/decade. This result illustrates that new low-dimensional materials with non-bulk-like electronic densities of states have potential to reduce the power supply voltages and turn-on voltages of FETs and suggests that continued density of states engineering may yield further reductions in sub-threshold slope and operating voltage swing.

Negative capacitance materials and devices

Another approach to overcoming the energy efficiency limit imposed by the 60 mV/decade sub-threshold swing of transistors has been to investigate transistors with gate structures possessing an effectively negative capacitance via use of a ferroelectric element as part of the gate stack.⁸ The switching of the ferroelectric element during transistor turn-on in the sub-threshold regime represents a step-up voltage transformer that amplifies the gate voltage. An attractive feature about this concept is that, other than the polarizable element, it does not require changes to the physics of FET operation and does not alter the on-current state drive. The scientific challenge in design of a negative capacitance transistor lies in the stabilization of the ferroelectric layer in the negative slope region of its polarization curve by using the semiconductor capacitance as a series capacitor. Discovery of other approaches and physical phenomena for generation of negative capacitance could further expand the options for reducing the sub-threshold slope of conventional transistors (see Figure 2).

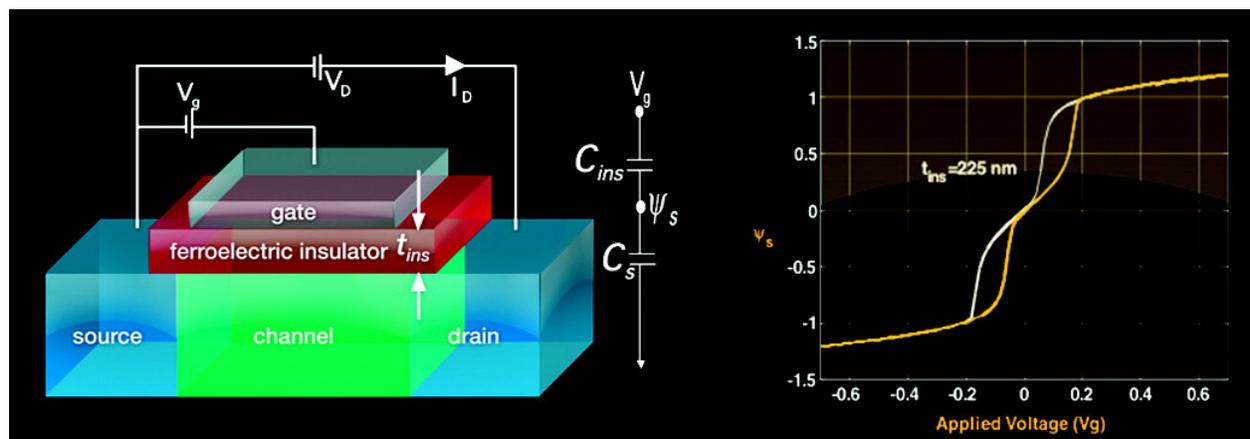


Figure 2. Effect of replacing the standard gate dielectric insulator in a transistor with a ferroelectric insulator. This design has the potential to implement a step-up voltage transformer that amplifies the gate voltage to give sub-threshold voltage swings of lower than 60 mV/decade and enable low voltage/low power operation. The voltage transformer action can be understood intuitively as the result of an effective negative capacitance provided by the ferroelectric capacitor that arises from an internal positive feedback. Reproduced with permission, S.Salahuddin et al., *Nano Lett.* 8 (2008) 2405-410. Copyright (2008) American Chemical Society.

Extreme conductors

Graphene nanoribbons are promising components in future microelectronics, both as active devices and as extreme conductors with high mobility and even ballistic conductance in some length scales and temperature regimes. The extraordinarily high charge carrier mobility in graphene (up to $150,000 \text{ cm}^2/\text{V}\cdot\text{s}$ at room temperature⁹) has also stimulated interest in the ballistic transport regime, where electron transport phenomena follow principles of electron optics, such as novel quasi-photonic device concepts, including Veselago lenses¹⁰ and Klein tunneling transistors.¹¹

The length scale over which charge carriers are transported in graphene is limited by scattering from impurities in the environment on or near the graphene sheet. To achieve extreme conductance in graphene, care must be taken to ensure that charge-scattering impurities are isolated from the graphene conductor, for example, by graphene growth on silicon carbide for which ballistic transport has been observed at lengths (room temperature) exceeding $10 \mu\text{m}$.⁹ Isolation of impurities can also be achieved by cladding the graphene sheet conductor in between charge-free dielectric layers, such as hexagonal boron nitride,¹² as depicted in Figure 3.

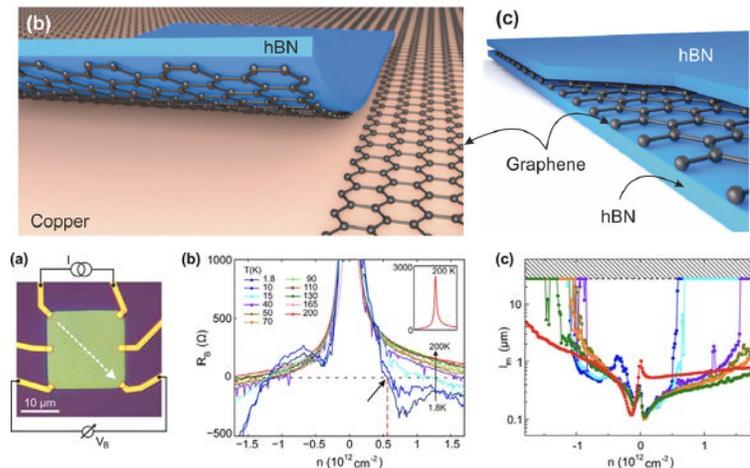


Figure 3. (Top) Illustrations of graphene delamination from copper foil by van der Waals pick-up by hexagonal boron nitride (hBN). The graphene sheet is stamped out along the edges of the hBN flake. (a) Optical image of a square-shaped hBN/chemically vapor deposited graphene/hBN device with Cr/Au side contacts. Ballistic transport is probed along the diagonal (dashed line) over $28 \mu\text{m}$. (b) Bend resistance as a function of charge carrier concentration for temperatures ranging from 1.8 to 200 K. Inset: Complete RB trace at 200 K. Charge transport is fully diffusive at this temperature, as seen by the positive RB. (c) Elastic mean free path as a function of charge carrier concentration at 1.8 to 200 K (same color coding as image b). Reproduced with permission, L. Banszerus et al., *Nano Lett.* 16 (2016) 1387–1391. Copyright (2016) American Chemical Society.

Chemical vapor deposition of graphene followed by transfer onto a desired substrate represents a scalable method for formation of large-area graphene sheets. Recently, ballistic transport in graphene synthesized by chemical vapor deposition was reported up to 200 K over a length of $1 \mu\text{m}$, and ballistic transport over distances exceeding $28 \mu\text{m}$ was achieved at 1.8 K. For densely arrayed interconnects in nanoelectronics, a further requirement is to consider edge scattering from nanoscale linewidth graphene ribbons. Hence, the edge roughness or effective edge roughness induced by scattering, along with the intrinsic mean free path, will affect the diffusive scattering probability and, thus, the total mean free path in graphene lines as a function of graphene width.

Unconventional superconductivity and correlated insulators in layered materials

Unconventional superconductivity has been studied for 30 years beginning with the discovery of superconductivity in cuprate oxide in the 1980s. Since that time, the mechanisms for unconventional superconductivity have been extensively investigated but are still not well understood. Recently, a new window for exploring unconventional superconductivity has been opened by the discovery of superconductivity in twisted-bilayer graphene structures with specific “magic” twist angles, which cannot be explained by a Bardeen-Cooper-Schrieffer-type mechanism for electron phonon coupling.¹ Superconductivity in twisted-bilayer graphene bears similarities to that observed in cuprate oxides, such as the

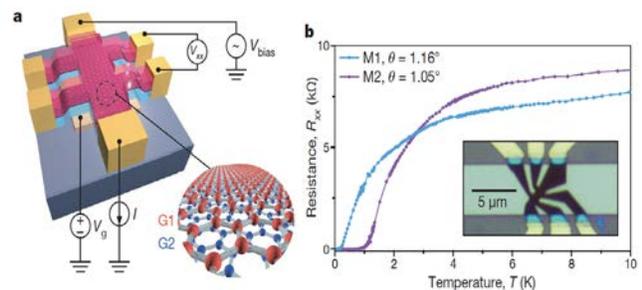


Figure 4. (a) Unconventional superconductivity in magic-angle graphene superlattices, where the electronic band structure of this “twisted bilayer graphene” exhibits flat bands near zero Fermi energy, resulting in correlated insulating states at half-filling with electrons. Electrostatic doping moves electron occupancy of the material away from these correlated insulating states, giving rise to a superconducting state (b) with a critical temperature of up to 1.7 K. From Y. Cao et al., *Nature*, 556 (2018) 43.

dependence of carrier density on temperature and the existence of small Fermi surfaces near the correlated insulating states. Notably, unlike the case for cuprate oxides, the ability to create gate-tunable twisted-bilayer graphene structures, such as the one depicted in Figure 4, enables the doping of the material to be varied within a single sample, allowing the phase behavior for superconductivity and insulating states to be studied in a single sample unhampered by sample-to-sample variation in crystal quality and defect density. The remarkable observation of a relatively high superconducting critical temperature at a low charge carrier density ($n = 10^{11} \text{ cm}^{-2}$) suggests that twisted bilayer graphene has among the largest pairing strengths between electrons among known superconductors.

Similar twisted-bilayer graphene heterostructures also exhibit correlated insulator behavior,¹³ as shown in Figure 5. When the twisted-bilayer graphene twist angle is close to the theoretically predicted magic angle, the electronic coupling and hybridization between the graphene layers induces nearly flat bands at low energy relative to the Dirac point, and this quenching of the quantum kinetic energy for half-filled bands leads to a correlated insulating phase corresponding to a Mott insulator arising in the localized flat bands. While such correlated superconducting and insulating phases have been observed for graphene, the Moiré structure in other bilayer two-dimensional lattices can be expected to lead to other hybridized electronic states with interesting properties. The ability to precisely control the twist angle in bilayer two-dimensional structures enables control of materials on a very interesting length scale, greater than the two-dimensional monolayer unit cell and smaller than the mean free path for scattering in many materials.

Topological insulators

Topological insulators are band insulator materials that support Dirac-like surface and edge electronic conduction channels in which electrons with a given wave vector have only one spin degree of freedom. These spin-momentum locked surface states are protected against backscattering by time-reversal symmetry, and therefore, charge transport is expected to be less sensitive to defect-related scattering processes than conventional or topologically trivial materials. Topological insulators have been identified as potential candidates for spintronic devices and quantum computation with dissipationless transistors using the quantum spin-Hall effect and quantum anomalous Hall effect.¹⁴ An example of such a device is depicted in Figure 6. Topological insulator states have been extensively studied with angle-resolved photoemission and transport measurements, especially in the narrow-bandgap semiconductors Bi_2Se_3 and Bi_2Te_3 . For these and other materials, the ability to observe and exploit the

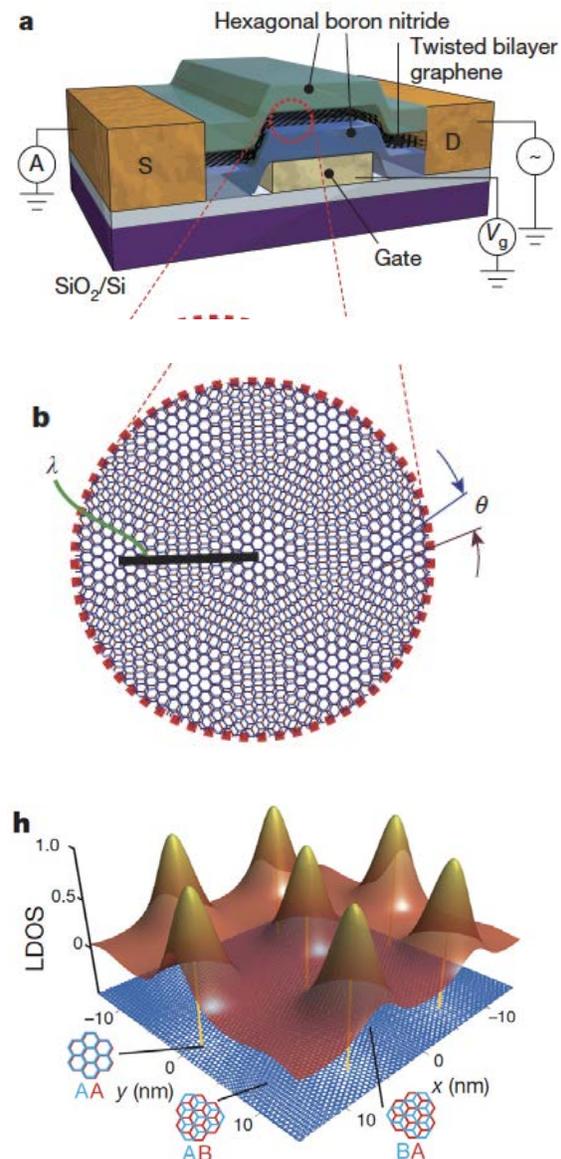


Figure 5. (a) Design of twisted-bilayer graphene heterostructure. Correlated insulator states created in this structure for magic twist angles (b) give rise to spatially correlated peaks in electron density corresponding to regions of A-A unit cell stacking in the graphene twisted bilayer. LDOS = local density of states. From Y. Cao et al., *Nature* 556 (2018) 80.

properties of the topological surface state is also dependent on the control of residual doping of the underlying bulk material. However, ternary and quaternary alloys such as $\text{Bi}_2\text{Te}_2\text{Se}$ and $\text{Bi}_{2-x}\text{SbxTe}_{3-y}\text{Sey}$ have been shown to exhibit high bulk resistivity, enabling observation of quantum transport features such as Shubnikov–de Haas oscillations arising from the surface states.

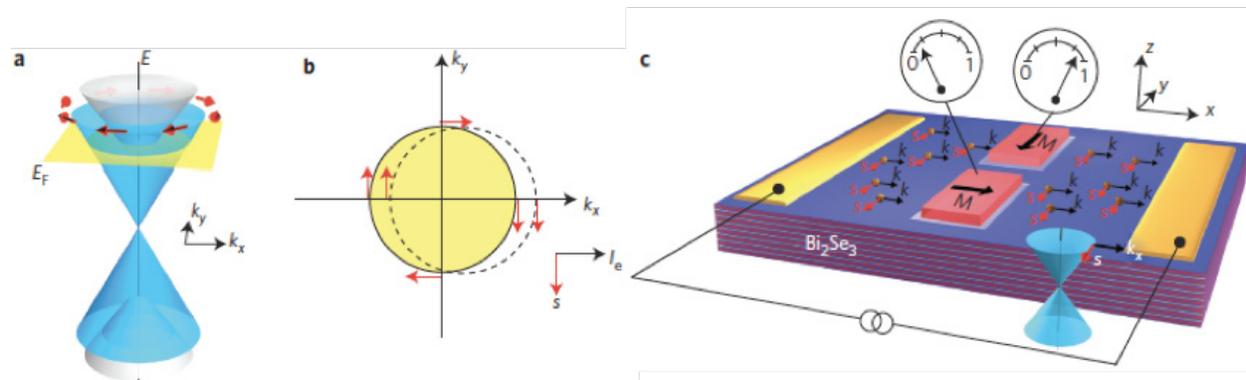


Figure 6. Schematic of TI surface bands and an experimental device for electrical detection of charge-current-induced spin polarization due to spin-momentum locking. (a) Dirac cone of the TI surface states (blue), with the spin at right angles to the momentum at each point. The bulk conduction and valence bands are shown in grey. (b) Top view of the TI surface states. An applied voltage produces a net momentum along k_x , and spin momentum locking gives rise to a net spin polarization oriented in-plane and at right angles to the current. (c) Concept drawing of the transport experiment. The voltage measured at the ferromagnetic detector is proportional to the projection of the current-induced TI spin polarization onto the contact magnetization. From C.H. Li et al., *Nature Nanotechnology*, 9 (2014) 218.

NEW PHOTONIC AND OPTOELECTRONIC MATERIALS AND PHENOMENA FOR INFORMATION AND ENERGY TRANSFER

Polaritons in Two-Dimensional Materials

Current status and recent advances

Polaritons are collective states of light and matter, and the recent growth of interest and activity in exploration of two-dimensional materials has precipitated intensive investigation of a wide variety of polaritonic states in these materials, including exciton polaritons, surface plasmons, magnons polaritons, and phonon polaritons.¹⁵ Polaritons in two-dimensional and layered materials can exhibit extreme-deep-subwavelength light confinement and tailorable, nonlinear, and anisotropic optical dispersion relations. Because of the ability to confine light at deep subwavelength dimensions, polaritonic media represent potential building blocks for deep-subwavelength waveguide interconnects (see Figure 7).

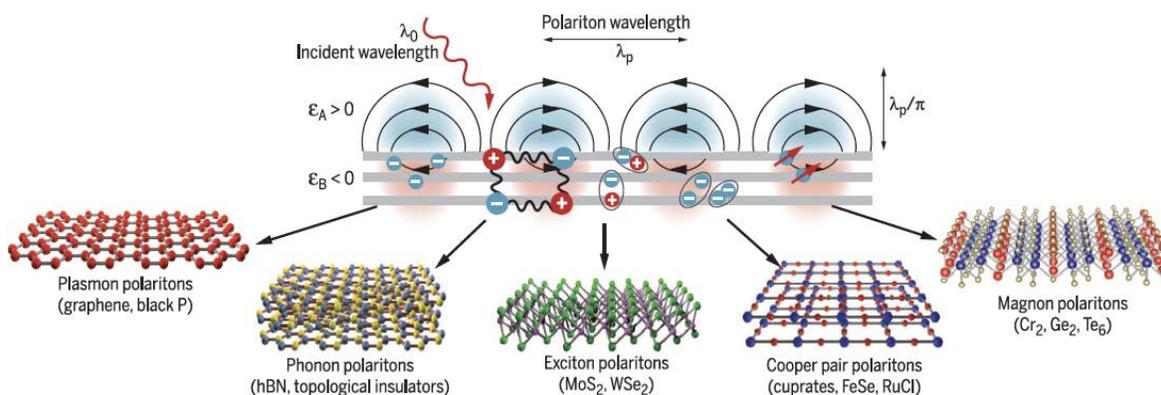


Figure 7. Structures and optics of polariton media. Polaritons are light-matter hybrid states that can exhibit strong optical confinement relative to the free space wavelength. Polaritons arise from coupling of light to electron in a conductor (surface plasmon polaritons), to optical phonons (phonon polaritons), to excitons (exciton polaritons) in semiconductors, to Cooper pairs or to magnetic resonances (magnon polaritons). Two-dimensional and layered materials can exhibit each of these hybrid states, which propagate as a surface wave at the interface between the materials and their surrounding environment. From D.N. Basov et al., *Science*, 354 (2016) 195.

Scientific challenges and opportunities

Polaritons in two-dimensional and layered materials seem well-suited for exploration as components for future photonic interconnect architectures, active switches, and memory devices. Polaritonic phenomena are observed across the electromagnetic spectrum from the visible, near-infrared, mid-infrared, and terahertz regime. Nanoscale structures such as tips, antennas, and cavities can be used to couple light from free space into polaritonic waveguide structures. Because van der Waals layered material heterostructures are not limited in their design by epitaxial lattice matching constraints, they can be composed of complex assemblies of dissimilar conducting, semiconductor, magnetic, and insulating materials such as graphene, black phosphorus, and transition metal dichalcogenides. This ability to form complex heterostructures enables flexible tailoring of the modal dispersion properties of polaritons.

Because of the inherent capability for modifying the optical properties by gate tuning, polaritons in van der Waals two-dimensional and layered materials have considerable potential as components of novel optoelectronic devices. The gate tuning can be used to electrostatically and independently control the quantum-confined sub-band density of states and the carrier density, giving the ability to tune the optical dispersion over negative and positive permittivity values both in-plane and out of the layer plane. This is exemplified by the gate-tunable dichroism exhibited by black phosphorus,¹⁶ which also has the potential for tuning dispersion from the elliptical to in-plane hyperbolic dispersion regime, where polaritonic modes could exhibit efficient waveguide propagation that is not limited by the diffractive effects seen in conventional isotropic media. Gate tuning and materials design also facilitate the tailoring of nonlinear optical properties, where out-of-plane electric fields can be used to break the symmetry in electronic structure. Moreover, synthesis of Janus-like layered two-dimensional structures,¹⁷ such as transition metal dichalcogenides with dissimilar anion on the top and bottom of the layer sheet, can give rise to large built-in dipoles and, thus, potential large nonlinear optical coefficients (see Figure 8).

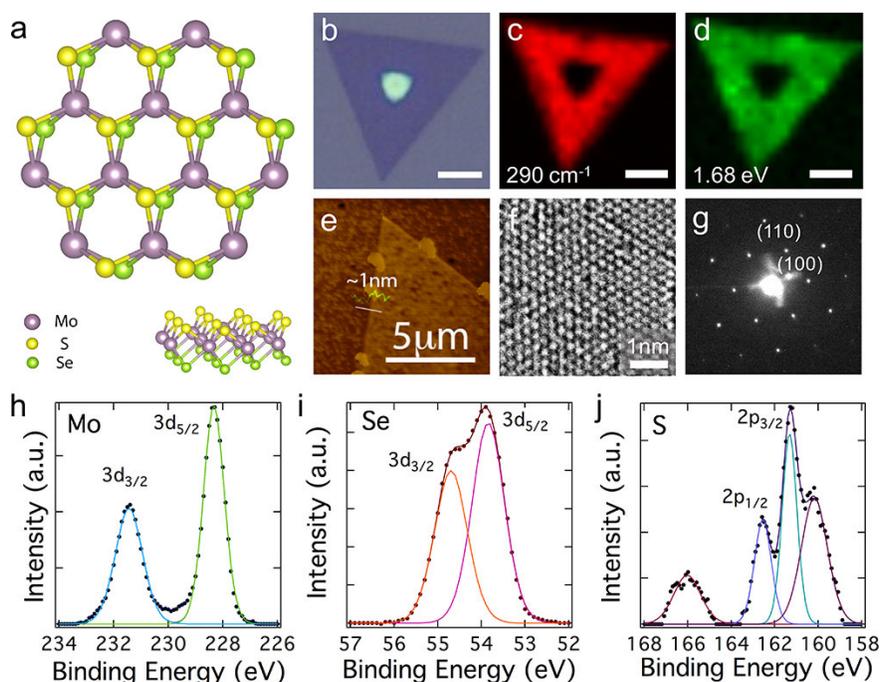


Figure 8. Monolayer Janus SMOSe characterizations. (a) Off-angle top view and side view of an eight-unit-cell Janus SMOSe monolayer. The purple, yellow, and green spheres represent molybdenum, sulfur, and selenium atoms, respectively. (b) Optical image of a Janus SMOSe triangle. The purple and the central island with high contrast is the monolayer and bulk crystal region, respectively. (c, d) Raman and photoluminescent peak intensity mappings of the Janus SMOSe triangle in (b). The mapping shows uniform distribution of the identical Raman peak at 287 cm⁻¹ and photoluminescent peak at 1.68 eV. (e) Atomic force microscopy topography image of the Janus SMOSe triangle. The profile shows that the thickness of the flake is <math>< 1\text{ nm}</math>. (f) Transmission electron microscopy image of the Janus SMOSe lattice. The atom arrangement indicates the 2H structure of the monolayer. (g) Corresponding selected area electron diffraction pattern of the monolayer. (h–j) X-ray photoelectron spectra of the Mo 3d, Se 3d, and S 2p core level peaks for the Janus SMOSe monolayer. From J. Zhang et al., *ACS Nano* 11 (2017) 8192-8198.

Excitonic Switching and Transistors

Current status and recent advances

Monolayer-thickness transition metal dichalcogenides such as MoS_2 , MoSe_2 , WS_2 , and WSe_2 are direct bandgap semiconductors, which can exhibit high radiative efficiency in high-quality samples. These materials also exhibit exciton binding energies of >250 meV, so that excitons are stably bound even at room temperature, and the absorption and luminescence features are determined by an excitonic absorption and emission feature well separated from the band edges. Thus, transition metal dichalcogenides have potential for realization of excitonic devices, such as optical interconnects where signal transduction occurs via excitonic absorption and emission. Interposing a gate to modulate diffusive exciton transport between a source where exciton absorption occurs and a drain where exciton emission occurs can enable a transistor-like modulation or switching of excitons.⁴

Scientific challenges and opportunities

Previously, exciton-based transistor behavior was demonstrated in coupled quantum wells in III-V compound semiconductor heterostructures, but the weak excitonic binding energies limited device operation to low temperatures. However, materials with large exciton binding energies, such as the transition metal dichalcogenides, have potential for creation of excitonic devices and circuits that can operate at room temperature. Long-lived strongly bound excitons can be created, for example, at type-II heterojunction interfaces such as that between MoS_2 and WSe_2 , since the heterojunction band offsets naturally lead to spatial separation of the electron and hole. This separation dramatically increases the exciton recombination lifetime and facilitates exciton diffusion over distances of many microns. Using these spatially indirect exciton states consisting of MoS_2 – WSe_2 van der Waals heterostructures encapsulated in hexagonal boron nitride, researchers have recently reported room-temperature electrically controlled transistor action and manipulation of the exciton dynamics by using gate-tuned confining and repulsive potentials for the exciton flux (see Figure 9).

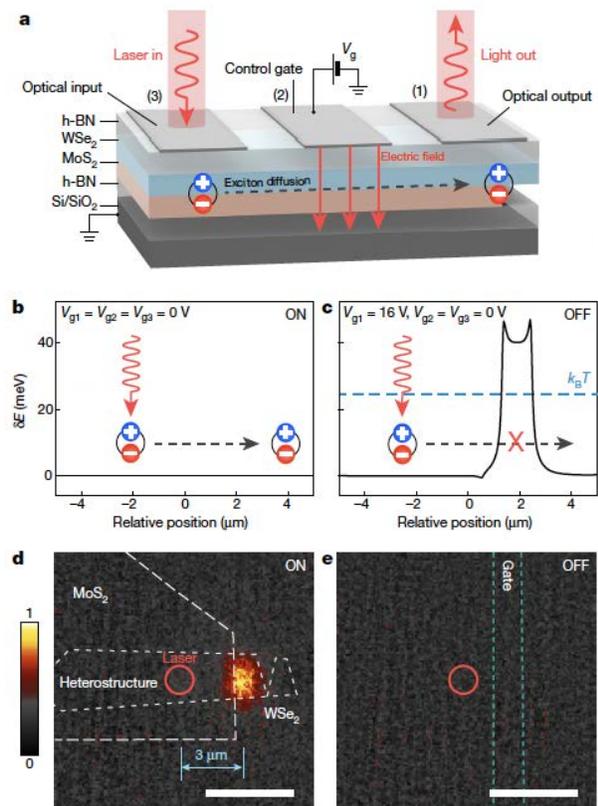


Figure 9. Excitonic transistor operation at room temperature. (a) Schematic illustrating that the application of gate voltages (V_{g1} , V_{g2} , V_{g3}) to transparent graphene electrodes (gates 1–3) can engineer a potential landscape for the diffusion of excitons, controlling their flux through the device. (b,c) Calculated energy variation δE for the excitons in the ON state (free diffusion b) and OFF state (potential barrier c). Red arrows represent laser excitation; the bound charges and black dashed arrows denote the excitons and their diffusion, respectively. (d,e) Corresponding images of exciton emission. Dashed lines indicate the positions of the different layers that form the heterostructure and the top graphene gate (gate 1). The laser spot is represented by the red circle. Color scale indicates the normalized photoluminescence intensity. Scale bar = 5 μm . From D.Unuchek et al., *Nature*, 560 (2018) 340-344.

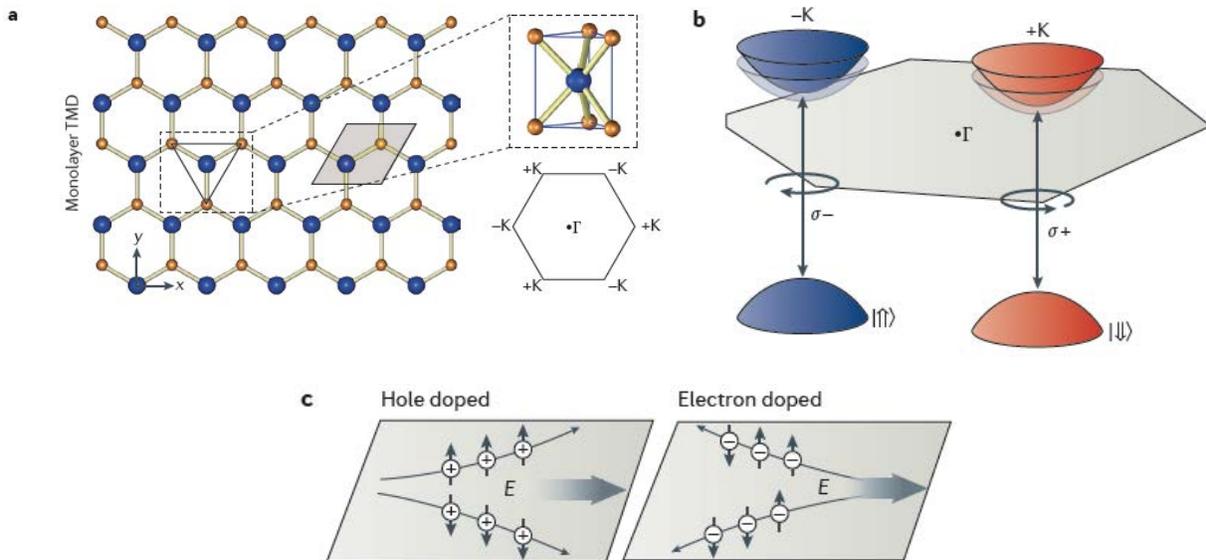


Figure 10. Valley-dependent carrier transport. (a) The 2D hexagonal crystal structure of a monolayer transition metal dichalcogenide (TMD) composed of transition metal atoms (blue) and chalcogen atoms (orange) resembles that of graphene but with broken inversion symmetry. A side view shows the 3D structure. The hexagonal Brillouin zone is shown by labelling the Γ point and the two inequivalent $+K$ and $-K$ points. (b) Valley-dependent optical selection rules for interband transitions in monolayer TMDs. The σ^+ polarized light couples to the $+K$ (red) valley, and the σ^- polarized light couples to the $-K$ (blue) valley. (c) Illustration of the valley unpolarized electron and hole Hall effect, originating from the Berry-curvature effects, when an in-plane electric field, E , is applied. The arrows depict the real spins for the electrons and holes that are coupled to the valley degree of freedom and accumulate on sample edges. From J.R. Schaibley et al., *Nature Reviews: Materials*, 1 (2016) 16055.

Valleytronics

Current status and recent advances

While “valley degree of freedom” has been understood for some time, the ability to exploit valley polarization in optoelectronics has been limited in most materials. With the advent of two-dimensional and layered materials with hexagonal lattices, such as graphene and monolayer transition metal dichalcogenides, the electronic band structure is marked by $+K$ and $-K$ points situated at two inequivalent valleys of the reciprocal lattice.⁵

The $+K$ and $-K$ point valleys carry a binary valley-specific pseudospin that behaves like a spin-1/2 system; the electrons in the $+K$ valley can be regarded as pseudospin up, and the electrons in the $-K$ valley can be regarded as pseudospin down. When electrons populate the two valleys, the electron distribution polarized in a $+K$ or $-K$ valley can store binary information.

In materials such as the monolayer transition metal dichalcogenides, the crystal lacks inversion symmetry, so electrons and holes in $+K$ and $-K$ valleys experience opposing effective magnetic fields in the momentum space known as Berry curvature, giving rise to electron motion along the directions governed by semiclassical equations of motion. The electrons also experience a valley-specific magnetic moment corresponding to the orbital angular momentum arising from the electron motion.

The valley Hall effect occurs when opposite Hall currents are carried by electrons and holes located in opposite valleys, as shown in Figure 10, analogous to the spin Hall effect, which enable electrical generation of spin polarized electrons. The valley Hall effect can be measured by using optoelectronic devices in which valley-polarized electrons and holes are injected optically according to valley optical selection rules. Field effect transistors in transition metal dichalcogenides can be used to sense the valley Hall effect as a Hall current that changes sign with the polarization of the excitation laser when a population imbalance occurs between the valleys. Circularly polarized optical excitation gives rise to a transverse Hall voltage, which changes sign when the helicity of the optical excitation changes from left- to right-hand circularly polarized. The spin and valley polarization lifetime can be enhanced by the strong spin–valley coupling in transition metal dichalcogenides, which also suggests the potential to manipulate spin via use of valley properties. Devices consisting of a hetero-bilayer with Type II band alignment, such as the $\text{MoS}_2/\text{WSe}_2$ system, can support long-lived spatially indirect excitons that are bound to the hetero-bilayer interface. These interlayer excitons are “dark” excitons because of the electron-hole momentum mismatch. Interlayer excitons flowing in the direction of an applied

field experience an anomalous transverse velocity from the valley Hall effect, which spatially splits the interlayer exciton populations based on the valley polarization.

Scientific challenges and opportunities

Valleytronics has the potential for new device concepts of interest for optical interconnect applications. For example, a reversible valley-polarizer pair of optical absorbers and emitters can serve as a valley optical interconnect that interconverts valley polarization with the circular polarization state of the absorbed or emitted photons. To date, optoelectronic devices such as valley light-emitting diodes with gate-controlled polarization of emitted photons,¹⁸ photodetectors,¹⁹ field effect phototransistors,²⁰ and photo-pumped lasers have been developed, but electrically pumped transition metal dichalcogenide lasers remain an outstanding challenge. Dual-gated, electrostatically induced p–n junctions in WSe_2 have shown tunable emission of excitons, and trions and bi-excitons indicated the potential for controllably reconfiguring the valley emission circularly polarized state to linearly polarized coherent states.¹⁸

Photonic Topological Insulators

Current status and recent advances

Photonic topological insulators are metamaterials that support propagating channels for electromagnetic radiation in surface and edge modes in the frequency range where the bulk material has a photonic bandgap.⁶ By analogy with electronic topological insulators, topologically nontrivial protected surface and edge modes can be designed at the interfaces between bulk metamaterials that have dissimilar associated topological charges. These helical interface and edge modes can exhibit spin-polarized unidirectional propagation of photons in a manner that is robust with respect to interface orientation and is also robust against interfacial disorder, as shown in Figure 11.

Scientific challenges and opportunities

Photonic topological insulation represents a method for achieving unidirectional photon transport without breaking time-reversal symmetry by use of a nonlinear medium, a time-varying input, or by application of an external magnetic field. This spin-polarized photon propagation has many interesting properties, including the potential to combine to “cloak” multiple spin-protected photon sources on a chip in a manner such that they do not interact with one another. Photonic crystal topological insulator structures using all-dielectric materials have been fabricated at optical and infrared frequencies and have been coupled to quantum emitter sources of near-field radiation operating at infrared frequencies in the telecommunication band. These emitter-coupled structures have demonstrated unidirectional propagation of radiation along the interfaces between dissimilar photonic crystals.

Recently reconfigurable²¹ and three-dimensional photonic topological insulators²² have been realized at microwave frequencies, which have potential as structures to guide radio-frequency signals on and off chip while avoiding interchannel interference and crosstalk.

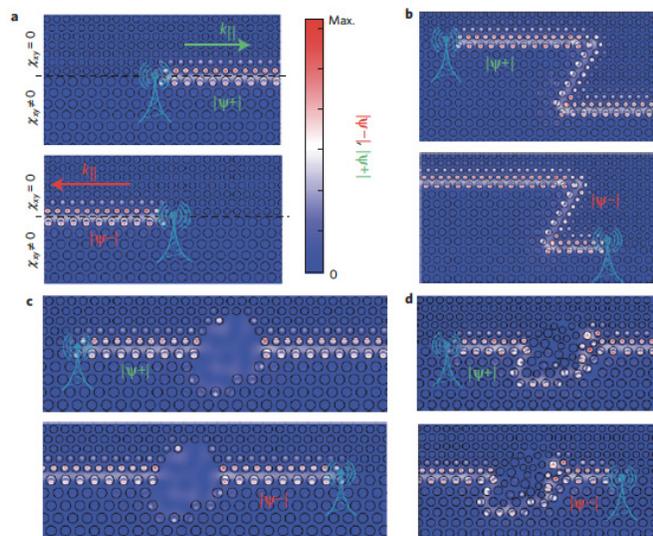


Figure 11. When surface waves are excited by a point dipole source at the interface between topologically trivial and non-trivial photonic insulators, the emitted radiation is guided selectively and, in a manner, robust with respect to defects: (a) selective excitation of spin-up and spin-down photonic one-way edge states along a straight interface; (b) demonstration of the robustness of the edge modes against different types of defects, sharp bends at the interface; and (c,d) a cavity obstacle and a strongly disordered domains in the adjacent media. From A.B. Khanikaev et al., *Nature Materials*, 12 (2013) 233–239.

THERMAL ENERGY MANAGEMENT: MATERIALS, STRUCTURES, AND ARCHITECTURES

Current Status and Recent Advances

The thermal management requirements for electronics and high performance computing have surpassed the needs of other technologies and challenge conventional approaches, such as natural or forced convection air cooling with fans.

These needs have motivated the development of specialized high-performance methods for thermal management in microelectronics, including microchannel heat sinks and micropumps, jet impingement, flat heat pipes, and phase-change solid and fluid media for energy storage and contact conductance. Other novel concepts have included integrated liquid microchannel cooling systems, microscale ion-driven air flow, and piezoelectric coolers. Microchannel heat sinks and micropumps have been widely investigated and provide very high heat transfer coefficients. Microchannel heat sinks are very compact in size, which enhances their suitability to electronics cooling²³ (see Figure 12).

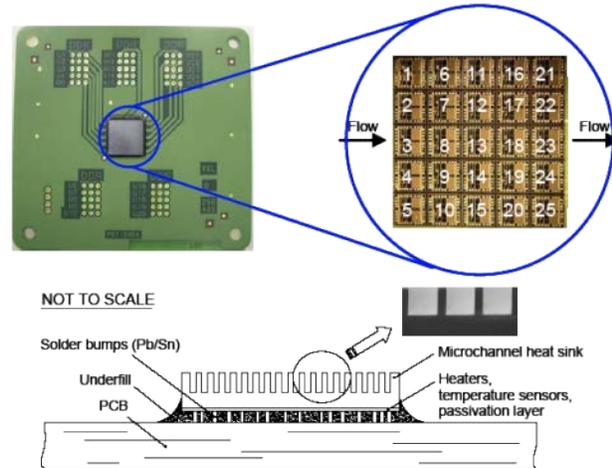


Figure 12. Test chip for evaluation of two-phase flow in microchannels. Chip features in-situ measurement of local wall temperature and flux, high-speed visualizations, flow regime determination, and regime-based modeling for dielectric fluids. From T. Harirchian and S. V. Garimella, *J. Electronic Packaging*, 133 (2011) 011001.

Scientific Challenges and Opportunities

Materials and interfaces

Silicon microelectronic devices are now in the deep nanoscale regime, and experiments have demonstrated that the nanoscale proximity of interfaces and the extremely small volume available for heat dissipation strongly modify thermal transport, thereby aggravating problems of thermal management.²⁴

Advanced chips and thermal conduction packages for microelectronics feature many interfaces, and thus fundamental understanding and characterization of interfacial heat transport are of significant importance. Thermal transfer at interfaces is affected by the character of interfacial bonding and thermal conductance at the atomic level. Experimental methods utilizing a combination of ultrafast pump-probe time-domain thermoreflectance, picosecond photoacoustic measurements, and laser spallation measurements on thin metallic films transfer-printed to a self-polymer assembled monolayer have elucidated fundamental processes in interfacial heat transport, which has enabled a correlation of changes in interface bond strength and heat flow at the interface. These approaches have also established a correlation between the covalent bond density in the bonding layer and both interfacial stiffness and interfacial thermal conductance.

While individual transistor devices are nanoscopic, microelectronic packages are large compared to the atomic scale and are not yet amenable to first-principles atomic-level simulation, so calculations of thermal transport are currently largely performed with the Boltzmann transport equation, or via classical molecular dynamics simulations. An area of opportunity is in prediction and calculation of microscopic phonon scattering rates needed as inputs for assessment of thermal conductivity, and these are poorly known for most materials. Fundamental issues also remain in establishing definitions of temperature in nonequilibrium nanoscale systems.

As noted above, layered and two-dimensional materials such as graphene and carbon nanotubes can exhibit ballistic electron transport, and therefore, these materials also feature extremely high thermal conductivities. Notably, the van der Waals bonding characteristics of layered and two-dimensional materials give rise to extremely anisotropic thermal transport.

Architecture

Acceleration of performance in scientific supercomputing and data centers almost inevitably comes at the cost of increasing power consumption and leads to computing systems that are ultimately thermally limited. The Top500 list ranks worldwide supercomputers based on their peak performance as determined by floating point operations per second (flops)—when running a LINPACK benchmark (<https://www.top500.org>). Today, the world’s most powerful supercomputer is Summit, which consumes 9.8 MW while delivering 143.5 petaflops, giving a power efficiency of 14.6 gigaflops/watt. This architecture is accelerated via GPUs. Alternative architectures such as many core processors have a measured power efficiency of about 2.6 gigaflops/watt, demonstrating the efficiency gains of a high-throughput processor design that can trade the costs of instruction decode for more processing elements and can minimize data motion. Modern processors have dynamic thermal and power management and can dynamically scale voltage and clock time to attempt to maintain optimum operating parameters over highly variable workloads.

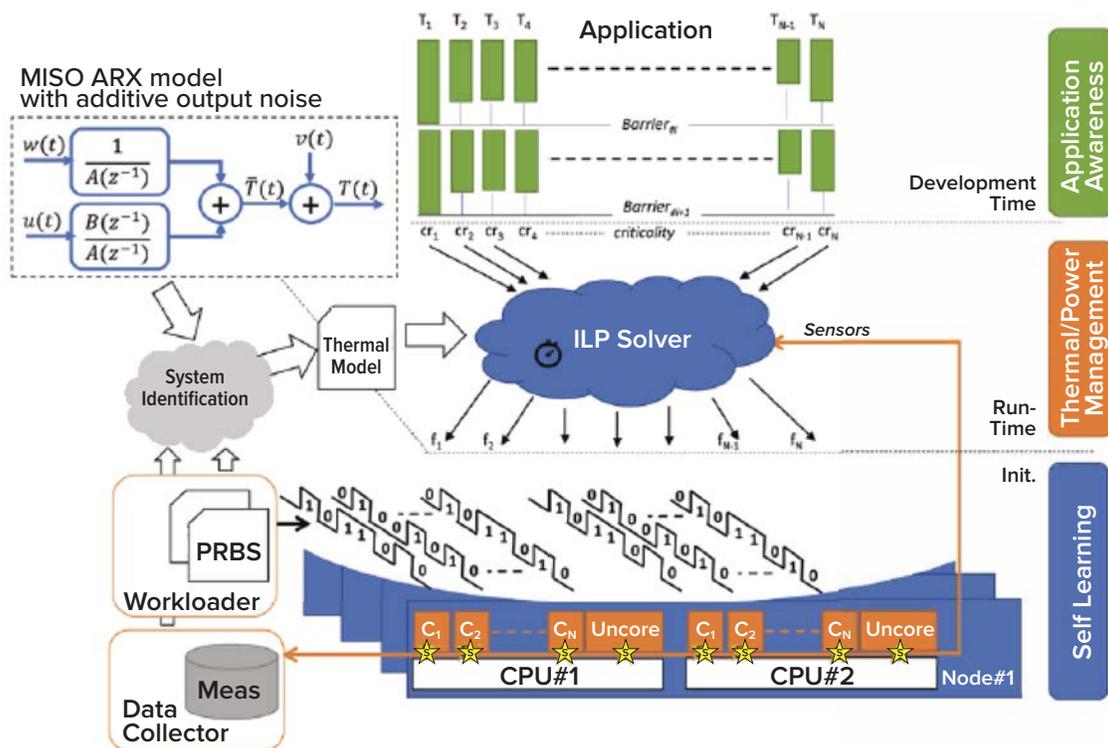


Figure 13. Block diagram of a self- and application-aware power and thermal management framework for a high-performance computing system. MISO = multiple input single output; ARX = autoregressive exogenous input; ILP = integer linear programming; PRBS = pseudorandom binary sequence. Reproduced with permission, A. Bartolini et al., *IEEE Design & Test*, 17 (2017) 2168-2356. Copyright (2017) IEEE.

Thermal management in multicore high-performance computing systems is a very complex endeavor due to lack of communication with the utility power grid, variable application program loads, inhomogeneous thermal dissipation in systems, and long thermal transients. Multicore processors in high-performance computing systems are equipped with thermal sensors for power management to regulate the power consumption of the processor and peripherals. While computing systems are equipped with sophisticated sensors, the thermal sensor data are “noisy,” and thus there is an opportunity for development of machine learning algorithms that can be used to build algorithms for a dynamic optimization strategy²⁵ (see Figure 13), and that is responsive to and interactive with the power grid (see sidebar).

A PROTOTYPE HOLISTIC DATA CENTER DESIGNED TO STUDY ENERGY MANAGEMENT



Current Status

The power demands of large-scale high-performance computing data centers can "swing" the power grid by multiple megawatts, and an energy waste of 1 metawatt-year comes at a cost of ~1M\$. Currently, the electric power infrastructure is conservatively designed and utilizes AC rather than DC power, with attendant AC/DC power conversions in data centers. Today, the demand created by applications running on high performance computing does not actively interact with the electric power grid.

Opportunity

There is an opportunity for design of a data center as a scientific laboratory for studying energy management, beginning with design of data center-optimized cooling and power delivery, monitoring and analysis systems to break the silos between electric power facilities, and computing system operations and application job programming. Such a project would accelerate development of data center-level energy optimization systems, responsive to the power grid, weather, scheduling, and cooling systems.

Impacts

Such a prototype holistic data center would greatly improve energy efficiency, and enable understanding of the relationship between energy usage and application throughput. Data centers of the future would serve as core ecosystem player and foster new responsive cooling and power delivery technologies.

ACCELERATED CO-DESIGN OF NOVEL MATERIALS, DEVICE CONCEPTS, AND SYSTEM ARCHITECTURES FOR CHANNEL OPERATION NEAR QUANTUM NOISE/DISSIPATION LIMITS

While significant progress has been made in exploring new materials and new device concepts, a major challenge is to understand these materials and devices in the context of a computing system and to relate the performance of materials and devices to the performance of the end-use application. In the field of embedded systems design, where the challenge is often to balance cost, performance, space (memory), and power, the design methodology called “co-design” iterates on the overall product design space by exploring alternative partitioning of the problem from software and general-purpose compute elements to collections of specialized compute elements that may reduce the software footprint (and flexibility) in exchange for improved power and performance. A key question is, can this co-design concept be extended to include analysis of the effect of novel switching device and material alternatives on end-use application performance, power, cost, etc.? An important note is that new materials and device concepts will also drive new computing abstractions and new applications, so while it is tempting to look at this co-design problem as understanding the impact of alternatives on a relatively static systems architecture and applications set, there is the possibility for discovering new computing models, architectures, and algorithms built on novel devices and materials. The later opportunity might be significant and should be included in our thinking.

Scientific Challenges and Opportunities

The capability is needed to rapidly iterate from materials choices and device parameters up through a series of machine abstractions and compute models to systems models and applications. This could be accomplished by large-scale modeling and simulation of the full system stack from the lowest level quantum behavior of the materials through to applications. This grand challenge in multi-scale simulation, practically, could be done as a loosely coupled problem driven by key parameters (switching times, density, power, etc.) across scale boundaries. This formulation of the problem would also admit to including a variety of search strategies for materials and device parameters that optimize a given compute design as well as to optimizing an abstract compute model for a given set of materials. Creating new building block abstractions (perhaps systematically and automatically) and composing them to create new computing models and architectures would be one route to evaluating the implications of new physical devices.

In addition to a pure simulation-based approach, it would be fruitful to develop means to experimentally validate new materials and devices in some kind of a test and evaluation harness, aimed at rapidly characterizing the key device parameters that drive co-design. Nearly every new device is evaluated in some form of a test harness. The advantages to creating more standardization and scalability of these test harnesses and standard ways to measure and characterize the devices and materials (so that data can be compared and “plugged” into a common co-design process) need to be explored.

Current Status and Recent Advances

Existing co-design practice is generally limited to working through tradeoffs at the architecture, software, and applications interfaces.²⁶⁻²⁹ This approach is widely used in the embedded computing industry from devising a range of special-purpose architecture extensions in processors in cell phones to assessing the tradeoffs in sensors among computing, sensing, and communication.^{30,31} In current co-design practice, the abstract computing model is rarely changed or challenged, and the assumption generally is that the materials and device building blocks are fixed (CMOS, fin FETs, DRAM, and NAND memory, etc.).³² An example is the co-design of tightly coupled (single die) with algorithms targeting small/modest local footprints focusing on weak relative interconnects and exploring the opportunity of a new nonvolatile memory technology that could be integrated on the die to change memory capacity, or investigating the implications of system-on-a-chip designs on applications performance. The discussion is essentially between the application and architecture, where the software component serves as the means of implementing choices between what to implement in hardware and what to keep in software. To extend this paradigm to include materials and devices — and by implication new functional abstractions and models of computation — we need to expand the tools to support design space tradeoff analysis.³³ As we introduce more degrees of freedom in the design space, the complexity of effectively searching this space for good designs increases, and the need for automation of this analysis becomes clear.

The desired workflow would be along the following lines:

1. Define the software and simulation framework for co-design, enabling end-use application consequences of physical-layer design.
2. Automatically generate mathematical operators that enable abstraction of device/circuit level function.
3. Automatically measure and generate models of noise and error, data transforms, performance, bit resolutions, etc., of the new materials/device.
4. Automatically re-factor known algorithms to use these new structures and to automatically generate software stacks to permit experimentation.

A significant gap exists in our ability to synthesize (automatically or manually) robust abstractions that bridge between new device concepts and applications. If the new device/material implements classical switching abstractions, then the analysis is straightforward so that a wide range of existing computer-aided design tools can be applied; however, once we move away from existing switching and circuit concepts, the tool chain becomes more challenging. Intermediate situations, especially in the near term, are also likely to be important where we want to rapidly test new structures and new materials while reusing some existing abstractions, tools, and software stacks. Current approaches to generating and testing software abstractions are highly labor intensive, though machine learning is beginning to have an impact on software design and optimization, particularly in artificial intelligence applications. These approaches could be harnessed to improve productivity in co-design. The result would be a path for immediate utilization of new physics breakthroughs in an end-to-end scenario. This approach could enable potentially much more energy-efficient and cost-effective systems for a class of problems. The benefits are that the approach supports accelerated co-design by providing rapid, iterative feedback to/from materials, devices, abstractions, and algorithms. In turn, this approach helps drive constraints on devices and structures from top-down algorithm and application requirements, and it could dramatically accelerate the time to market for technologies that demonstrate improvement in application solutions.

REFERENCES

1. Y.Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero, Unconventional superconductivity in magic-angle graphene superlattices, *Nature*, 556 (2018) 43.
2. M. Amani, D.-H. Lien, D. Kiriya, J. Xiao, A. Azcatl, J. Noh, S.R. Madhupathy, R. Addou, Santosh KC, M. Dubey, K. Cho, R. M. Wallace, S.-C. Lee, J.-H. He, J.W. Ager, X. Zhang, E. Yablonovitch, and A. Javey, Near-unity photoluminescence quantum yield in MoS₂, *Science*, 350 (2015) 1065–1068 (2015).
3. M.M. Ugeda, A.J. Bradley, S.-F. Shi, F.H. da Jornada, Y. Zhang, D.Y. Qiu, W. Ruan, S.-K. Mo, Z. Hussain, Z.-X. Shen, F. Wang, S.G. Louie, and M.F. Crommie, Giant bandgap renormalization and excitonic effects in a monolayer transition metal dichalcogenide semiconductor, *Nature Materials*, 13 (2014) 1091–1095.
4. D. Unuchek, A. Ciarrocchi, A. Avsar, K. Watanabe, T. Taniguchi, and A. Kis, Room-temperature electrical control of exciton flux in a van der Waals heterostructure, *Nature*, 560 (2018) 340–344.
5. J.R. Schaibley, H. Yu, G. Clark, P. Rivera, J.S. Ross, K.L. Seyler, W. Yao, and X. Xu, Valleytronics in 2D materials, *Nature Rev. Mater.*, 1 (2016) 16055.
6. A.B. Khanikaev, S.H. Mousavi, W.-K. Tse, M. Kargarian, A.H. MacDonald, and G. Shvets, Photonic topological insulators, *Nature Materials*, 12 (2013) 233–239.
7. C. Qiu, F. Liu, L. Xu, B. Deng, M. Xiao, J. Si, L. Lin, Z. Zhang, J. Wang, H. Guo, H. Peng, and L.-M. Peng, Dirac-source field-effect transistors as energy-efficient, high-performance electronic switches, *Science* 361 (2018) 387–392.
8. S. Salahuddin and S. Datta, Use of negative capacitance to provide voltage amplification for low power nanoscale devices, *Nano Lett.* 8 (2008) 405-410.
9. J. Baringhaus, M. Ruan, F. Edler, A. Tejada, M. Sicot, A. Taleb-Ibrahimi, A.-P. Li, Z. Jiang, E.H. Conrad, C. Berger, C. Tegenkamp, and W.A. de Heer, Exceptional ballistic transport in epitaxial graphene nanoribbons, *Nature*, 506 (2014) 349.

10. V.V. Cheianov, V.Fal'ko, and B.L. Altshuler, The focusing of electron flow and a Veselago lens in graphene pn junctions, *Science*, 315 (2007) 1252.
11. Q. Wilmart, S. Berrada, D. Torrin, V. Hung Nguyen, G. Feve, J.-M. Berroir, P. Dollfus, and B.A. Placais, Klein tunneling transistor with ballistic graphene, *2D Mater.*, 1 (2014) 011006.
12. L. Banszerus, M. Schmitz, S. Engels, M. Goldsche, K. Watanabe, T. Taniguchi, B. Beschoten, and C. Stampfer, Ballistic transport Exceeding 28 μm in CVD grown graphene, *Nano Lett.*, 16 (2016) 1387–1391.
13. Y. Cao, V. Fatemi, A. Demir, S. Fang, S.L. Tomarken, J.Y. Luo, J.D. Sanchez-Yamagishi, K. Watanabe, T. Taniguchi, E. Kaxiras, R.C. Ashoori, and P. Jarillo-Herrero, Correlated insulator behaviour at half-filling in magic-angle graphene superlattices, *Nature*, 556 (2018) 80.
14. C. H. Li, O. M. J. van 't Erve, J. T. Robinson, Y. Liu, L. Li, and B. T. Jonker, Electrical detection of charge-current-induced spin polarization due to spin-momentum locking in Bi_2Se_3 , *Nature Nanotechnol.*, 9 (2014) 218.
15. D.N. Basov, M.M. Fogler, and F.J. Garcia de Abajo, Polaritons in van der Waals materials, *Science*, 354 (2016) 195.
16. M.C. Sherrott, W.S. Whitney, D. Jariwala, S. Biswas, C.M. Went, J. Wong, G.R. Rossman, and H.A. Atwater, Anisotropic quantum well electro-optics in few-layer black phosphorus, *Nano Lett.* 19 (2019) 269–276.
17. J. Zhang, S. Jia, I. Kholmanov, L. Dong, D. Er, W. Chen, H. Guo, Z. Jin, V. B. Shenoy, L. Shi, and J. Lou, Janus monolayer transition-metal dichalcogenides, *ACS Nano*, 11 (2017) 8192-8198.
18. A. Pospischil, M.M. Furchi, and T. Mueller, Solar-energy conversion and light emission in an atomic monolayer p–n diode, *Nature Nanotechnol.*, 9 (2014) 257–261; see also: B.W.H. Baugher, H.O.H. Churchill, Y. Yang, and P. Jarillo-Herrero, Optoelectronic devices based on electrically tunable p–n diodes in a monolayer dichalcogenide, *Nature Nanotechnol.*, 9 (2014) 262–267.
19. C.H. Lee, G.H. Lee, A.M. van der Zande, W.C. Chen, Y.L. Li, M.Y. Han, X. Cui, G. Arefe, C. Nuckolls, T.F. Heinz, J. Guo, J. Hone, and P. Kim, Atomically thin p–n junctions with van der Waals heterointerfaces, *Nature Nanotechnol.*, 9 (2014) 676–681.
20. Z.Y. Yin, H. Li, H. Li, L. Jiang, Y.M. Shi, Y.H. Sun, G. Lu, Q. Zhang, X.D. Chen, and H. Zhang, Single-layer MoS_2 phototransistors, *ACS Nano*, 6 (2012) 74–80.
21. X. Cheng, C. Jouvaud, X. Ni, S. Hossein Mousavi, A.Z. Genack, and A.B. Khanikaev, Robust reconfigurable electromagnetic pathways within a photonic topological insulator, *Nature Materials*, 15 (2016) 542.
22. L. Lu, C. Fang, L. Fu, S.G. Johnson, J.D. Joannopoulos, and M. Soljačić, Symmetry-protected topological photonic crystal in three dimensions, *Nature Physics*, 12 (2016) 337–340.
23. S. Garimella, J.D. Killian, and J.W. Coleman, An experimentally validated model for two-phase pressure drop in the intermittent flow regime for circular microchannels, *J. Fluids Eng.*, 124 (2001) 205-214.
24. D.G. Cahill, W.K. Ford, K.E. Goodson, G.D. Mahan, A. Majumdar, H.J. Maris, R. Merlin, and S.R. Phillpot, Nanoscale thermal transport, *J. Appl. Phys.*, 93 (2003) 793.
25. A. Bartolini, R. Diversi, D. Cesarini, and F. Beneventi, Self-aware thermal management for high-performance computing processors, *IEEE Design & Test*, 17 (2017) 2168-2356.
26. G. De Micheli, R. Ernst, W. Wolf, and M. Wolf, Readings in hardware/software co-design, San Francisco, CA, Morgan Kaufmann (2002).
27. G. De Micheli and R.K. Gupta, Hardware/software co-design, *Proceedings of IEEE*, 85(3) (1997) 349–365, <http://doi.org/10.1109/5.558708>.
28. G. DeMicheli and M.G. Sami, eds., *Hardware/Software Co-design*, Dordrecht, Netherlands, Kluwer Academic Publishing (1996).
29. W.H. Wolf, Hardware-software co-design of embedded systems, *Proceedings of IEEE*, 82(7) (1994) 967–989, <http://doi.org/10.1109/5.293155>.
30. I. Bolsens, H.J. De Man, B. Lin, K. Van Rompaey, S. Vercauteren, and D. Verkest, Hardware/software co-design of digital telecommunication systems, *Proceedings of IEEE*, 85(3) (1997) 391–418, <http://doi.org/10.1109/5.558713>.
31. M. Chiodo, P. Giusto, A. Jurecska, H.C. Hsieh, A. Sangiovanni-Vincentelli, and L. Lavagno, Hardware-software codesign of embedded systems, *IEEE Micro*, 14(4) (1994) 26–36, <http://doi.org/10.1109/40.296155>.
32. W. Wolf, A decade of hardware/software codesign, *Computer*, 36(4) (2003) 38-43.
33. J. Teich, Hardware/software codesign: The past, the present, and predicting the future, *Proceedings of IEEE*, 100(Special Centennial Issue) (2012) 1411–1430, <http://doi.org/10.1109/JPROC.2011.2182009>.

Appendix A Preparatory Material for DOE Office of Science Basic Research Needs for Microelectronics Workshop

DOE Office of Science Basic Research Needs for Microelectronics Workshop

Compiled by
Gil Herrera and Jerry Simmons
Sandia National Laboratories

INTRODUCTION

As long as it has existed, the Department of Energy Office of Science (DOE-SC) has been at the leading edge of microelectronics, both as a consumer and as an engine of scientific understanding that has enabled many of the technological breakthroughs adopted by industry. Since the invention of the integrated circuit in 1960, advances in microelectronics have followed Moore's Law and other scaling laws, leading to circuit density and device performance improvements of 10^9 over this time period. In turn, strong commercial demand fueled the pace of scaling, and assured that the needs of Office of Science facilities were met. That time is coming to an end, however, as scaling is approaching its physical and economic limits.

As we approach the end of that era, we can now no longer rely on scaling to support our ever-increasing need for better microelectronics devices and the systems they enable. As a result, the challenges we face in meeting our needs are growing. As one example, advances in instruments used for scientific inquiry have created a 'big data problem' spanning from data movement to storage to analysis – a problem that has the potential to impede scientific progress in high-energy physics, analytical chemistry, and many other fields. A further example is that continued progress beyond Exascale Computing will require novel architectures and devices that radically reduce the power required for communication and processing, and yet still achieve requisite increases in performance absent a concomitant feature size reduction. A final example is that the nation – led by the DOE – must evolve to an electrical grid that is energy efficient, resilient to both naturally-occurring phenomena and intentional attack, and agile in adapting to fluctuations in demand and power generation by all types of sources. Indeed, the challenges are large and diverse.

This workshop is not the first to attempt to address these problems. Many organizations have identified potential technologies, devices, and architectures to continue advancing microelectronics performance after scaling reaches its limit. Many of the participants in this have participated in developing these roadmaps, and thus bring extensive background knowledge to the workshop. The aim of this workshop is to focus on scientific issues associated with advanced microelectronics technologies for applications relevant to the DOE mission, with an emphasis on pioneering work and bold, game-changing research, and attention to the underlying science that needs to be understood and determined. The DOE Office of Science is a logical sponsor of this work as we have some of the most challenging problems and a critical mass of world-class facilities and researchers capable of overcoming these challenges. The workshop chairs have also focused on bringing together an integrated team of participants with diverse backgrounds in order to have a cohesive understanding of the research needs that address issues of interdependency between the materials, devices, architecture, algorithms and work tasks.

This document includes preparatory material intended to help participants at the workshop. It covers (a) the workshop charge and structure, (b) a summary of the present state, (c) details on the scope of the four breakout panels, (d) the anticipated outcomes of the workshop, and (e) a list of supplemental reference materials that are available on a SharePoint site established by DOE for the workshop (access information provided separately to workshop participants).

WORKSHOP CHARGE AND STRUCTURE

The three chairs were charged with convening a panel of expert scientists to hold a workshop and write a report that provides a thorough assessment of the scientific issues associated with advanced microelectronics technologies for applications relevant to the DOE mission. The workshop should focus on identifying critical scientific challenges, fundamental research opportunities, and priority research directions that require further study as a foundation for advances in microelectronics over the next decade and beyond. Particular emphasis should be placed on energy-relevant applications, and those areas that are aligned with the missions and needs of the DOE Offices of Advanced Scientific Computing Research (ASCR), Basic Energy Sciences (BES), and High Energy Physics (HEP) including data management and processing, power electronics, and high-performance computing. The workshop should examine research that is relevant to both the extension of CMOS and beyond CMOS technologies; however topics of direct relevance to Quantum Information Science and Quantum Computing are outside the scope of this workshop. Finally, participants should focus on a co-design innovation ecosystem in which materials, chemistries, devices, systems, architectures, and algorithms are researched and developed in a closely integrated fashion.

Breakout Panel I: Big Data Collection, Analytics, and Processing for future SC Facilities

Topics consist of priority research directions to enable electronics (including logic, memory, data communications, and architectures) for collecting, storing, interrogating, and processing large volumes and rates of data, particularly those generated by DOE facilities such as BES light and neutron sources or HEP experiments. Edge computing is critical to enable early data reduction and to meet real-time constraints. Devices and architectures for non von-Neumannian approaches are relevant to data intensive computing. Also of interest are capabilities for early stream data processing and low voltage, low power electronics.

Breakout Panel II: Co-Design for High Performance Computing beyond Exascale

Topics consist of priority research directions to enable logic, memory, architectures, and novel algorithms for future high-performance computing technologies, both traditional HPC beyond exascale and new generations of machines aimed at artificial intelligence, machine learning, and analytics. This should focus on approaches not solely consisting of on variants of current architectures implemented in new technologies, but rather on new ideas for architectures. Also consider the relationship to future technologies – connections to quantum, neuromorphic, etc. How will future HPC machines interface with these technologies? The discussions should take into account the interdependencies between needs from the systems level (algorithms, architecture and micro-architecture) and what the hardware (circuits, devices, materials, physics & chemistry) can provide in that context.

Breakout Panel III: Power control, conversion, and detection

Topics consist of priority research directions to enable electronics for the control and conversion of high power, high voltage and high current for applications on the electrical grid, in transportation and machinery, and in DOE facilities such as accelerators. Panel discussions should include, but not be limited to, energy-efficiency and portability enabled by new generation of wide bandgap and ultra-wide bandgap semiconductor materials (SiC, GaN, diamond, c-BN, Ga₂O₃, ...), and novel device designs that can fully exploit their superior properties, for power electronics, sensing, optical devices, and other DOE-relevant applications. Ways to harvest, at a system level, unanticipated opportunities arising from these developments (e.g. high temperature electronics in harsh environments) are of interest.

Breakout Panel IV: Crosscutting Research

The crosscut panel will focus on the bi-directional, closed-loop cycle that includes materials/chemistry, devices, circuits, architectures, and algorithms. The opportunity is to promote a co-design ecosystem in which research and innovation in these spaces is carried out in parallel but in a closely connected/coordinated fashion.

SUMMARY OF PRESENT STATE

Evidence strongly suggests we are near the end of Moore's Law. Some aspects of scaling stopped over a decade ago when the difficulty of operating circuits below a volt ended voltage scaling, resulting in the end of frequency scaling. There is strong evidence that dimensional scaling is reaching its limit. For example:

- Intel has been at the 14nm node for 5 years, longer than any node in their 50-year history
- GlobalFoundries announced in 2018 that they will no longer pursue dimensional scaling.
- Problems with scaling of lithography are becoming dire; industry is running out of optical tricks to extend 193nm immersion lithography, and Extreme Ultraviolet (EUV) Lithography still faces major challenges, with tools costing ~\$200M/each and not planned for volume production until the 7nm node
- Samsung suggested in a 2016 SEMICON-West presentation that cost scaling stopped after the 28nm node

Industry and government agencies have been preparing for this inevitability for some time. Industry and government have long partnered on developing technology roadmaps to guide research. The first semiconductor roadmap workshop (MICRO TECH 2000) was held in 1991 and co-sponsored by the White House Office of Science and Technology Policy and the National Advisory Committee on Semiconductors. This led to the creation by SEMATECH of the National Technology Roadmap for Semiconductors, followed by the International Technology Roadmap for Semiconductors (ITRS), and now the IEEE-sponsored International Device Roadmap for Semiconductors (IDRS). Sections from the last ITRS and most recent IDRS can be found in the supplemental reference material on the SharePoint site.

Considerable effort has gone into advancing several candidate technologies to replace CMOS transistors in microelectronics. Through entities like the Semiconductor Research Corporation (SRC), a sponsor of academic research, and by conducting internal research, industry has been sponsoring research in post-CMOS device technology for about two decades. Likewise, government agencies have initiated multiple programs to sponsor and conduct research in post-CMOS devices and non-Von Neumann computer architectures. Additionally, novel approaches have been explored by industry and government to continue increases in performance through architectural advances, new materials and CMOS device designs, new non-volatile memories, 3-D integration, and improved chip-chip, board-board, and chassis-chassis communications. Detailed information regarding these efforts can be found in the supplemental reference material on the SharePoint site.

Significant time and investment is required to mature a technology into production. It took 10 years and billions of dollars for industry to bring FinFETs into production. The technology maturation process started after the FinFET was invented and demonstrated at Berkeley under DARPA and SRC sponsorship. Similarly, it took close to 20 years to bring high dielectric constant insulator based MOSFETs from the stage of discovery science to that of a product. When the EUV Lithography Consortium was established in 1997, they projected that EUV would be in production by 2003 to support the 100 nm node. At best, it will be in volume production 16 years and 7 nodes later at 7 nm. The high cost of technology maturation, process development, and production facilities mandate that markets must exist to cover the high investment and to guarantee a reasonable return on investment.

Electricity generation currently accounts for 40% of end-use energy consumption in the US, and is expected to grow by 50% globally by 2045. Power electronics play the role of controlling and converting electrical power into forms most suitable for the separate processes of transmission, distribution, and end-use consumption. Estimates are that as much as 80% of U.S. electricity will pass through power electronic converters by 2030. Fast-switching power devices are the key enablers of high efficiency, compact electronic power conversion. While the performance of Si-based power electronic devices has steadily increased over the past couple of decades, there are signs that Si power technology is now reaching its physical limits. Alternative wide bandgap (WBG) semiconductor materials, such as SiC and GaN, are enabling a new generation of power devices with orders of magnitude improvements in performance. Farther out on the horizon are the ultra-wide band gap (UWBG) semiconductors such as diamond, AlN, c-BN and Ga₂O₃, which promise additional factors of ten in improvement. However, these materials are not well developed and will need significant fundamental materials research to reach fruition.

Now is the time to identify and focus resources on the discovery science for new materials, devices architectures and algorithms to deliver the microelectronics technologies and systems needed by future Office of Science facilities and programs. We must conduct research to mature emerging technologies to assure that the Office of Science facilities meet all known challenges, and to conduct research to explore beyond the known challenges and roadmaps of today.

Discussion Scope for Breakout Panel I: Big Data Collection, Analytics, and Processing for Future SC Facilities

Authors: Kerstin Kleese Van Dam (BNL) and Sayeef Salahuddin (UC Berkeley)

Synopsis - This panel will not discuss the many challenges presented by big data collections and analytics at the extreme scales, but will focus instead on the hardware innovations necessary to enable future big data science. Specifically we will discuss the research needed to drive progress in the field of microelectronics for new sensors, compute and storage devices, appraising research needs ranging from novel materials to devices and systems.

The exponential growth in computing power, sensor and instrument technologies over the last five decades has led to the availability of abundant data. The concomitant emergence of big data analytics has brought about unparalleled changes in our way of life. Intelligent use of data, and data intensive computing are transforming education, healthcare, business, sustainable energy, and national security. It is also radically altering the way scientific research is done. At the same time, the incessant forward march of computing hardware faces fundamental roadblocks. The semiconductor electronics community is searching for a path forward for scaling beyond 2025. On the one hand this uncertainty, coupled with assured and significant increase in the demand for computing power, provides a unique opportunity to bring fundamental changes to how electronics is implemented today. On the other hand, as fierce competition in future computing hardware research continues to level the playing field, it raises the possibility of a potential inflection point in the unquestionable leadership position that the US has traditionally held in computing. In parallel, big data analytics is driving fundamental changes in data storage technologies. While communities have looked to long term preservation of data in the past, they are now demanding fast random access and processing of petabyte to exabyte scale data in minutes to hours. This change can no longer be supported by traditional memory, network and storage technologies, and new systems and devices need to be designed. For example, new technologies and materials such as DNA based storage require novel approaches to the embedded and connecting microelectronics.

Within this context, the objective of this panel is to discuss the basic scientific research needs in terms of device physics, material discovery and design, circuits, systems and architecture innovations to accelerate the advance in hardware for next generation computers, networks and storage. This next generation will undoubtedly be focused on data-abundant applications such as data collection (e.g., sensors and sensor networks), data analytics, data communications and processing (microprocessors). Many of these systems will have to work within significant energy constraints, which will drive power efficiency and low voltage operation requirements in devices. This will require new approaches and fresh thinking in transport physics, magnetism and photonics.

Moreover, to ensure intelligent collection and analysis of data, local and distributed processing will be necessary, in many cases, by enabling life-long learning. Such learning machines will eventually inspire new computer architectures compared to what we use today. For this to materialize, the memory bottleneck of today's computers will need to be overcome, requiring completely new, integrated, non-volatile memory devices, new approaches in the physics of high data rate information transfer, and novel computing-in-memory concepts. As we collect more and more data for immediate as well as for long term consumption in data analytics pipelines, we also need to look at long term, high volume (well beyond exabyte) storage media, that at the same time can support fast – and often random – access to small data or knowledge samples. Today tape technology is at a critical cross road as almost all vendors have left the market, and disks have not been able yet to replace tape in terms of longevity, reliability and price. The access speed to both is still too limited to support true high throughput data analysis at an affordable price point.

Panelists by categories

The panelists represent a number of general categories of expertise:

- Device Physics
- Material Synthesis
- Lithography
- Thermal management
- Electronics
- Neuromorphic computation
- Memory and Storage systems
- Microscopy
- System Architecture and Design

Discussion Scope for Breakout Panel II: Co-design for High Performance Computing beyond Exascale

Authors: Jim Ang (PNNL) and Tom Conte (Georgia Tech)

Given the multi-disciplinary nature of Co-design, our panelists will identify, discuss, and prioritize recommendations for DOE-SC's Microelectronics research strategy in the areas of:

- Materials for microelectronics
- Microelectronics circuits and devices
- Computer and System Architectures – including architectural analysis and simulators, hardware/device design and synthesis tools
- System Software and Tools – including compilers, runtime and operating systems
- Applied Math and Algorithms
- HPC application development to participate in collaborative co-design
- A co-design innovation ecosystem that crosscuts all of these DOE-SC disciplines

For over 25 years, DOE's HPC community has leveraged the commodity computing ecosystem. Moore's Law technology advances and the clock speed multiplier of Dennard Scaling gave rise to the *Killer Micros* that led to the corresponding fall of Cray's custom vector supercomputers. In this timeframe, DOE supercomputers were built from the integration of commodity computing components. DOE practiced multi-disciplinary co-design among different software development activities: system software, tools, algorithm and application development. A focus for these multi-disciplinary collaborations was improving the scalability performance of DOE's HPC application portfolio with capabilities such as message passing communication libraries and computational domain partitioning tools. The predictability of performance improvements for general purpose, CPU processor technologies provided DOE with a solid foundation for the evolution and co-design of the HPC applications with the software stack. The DOE HPC community did not need to close a loop with hardware architecture designers. Recall that during the Cray vector supercomputer era, this loop *was* closed – it allowed Cray's computer architects to prioritize among options as they developed their *custom* hardware designs.

The end of Dennard Scaling was marked by the appearance of the first dual-core processors in 2005. While the core counts in CPUs have steadily increased, as long as compute nodes were based on the integration of general-purpose, multi-core CPUs and commodity DRAM, DOE's HPC needs were met by continuing the strategy of leveraging commodity computing technology. DOE's Computer Science investments continued to focus on realizing the compute performance of multi-core processors – to address changes driven by reduced memory bandwidth and capacity, and reduced interconnect bandwidth, on a per core basis. Overall, the application software and system software stacks were remarkably stable over this timeframe. A major disruption in computing models was made when the first large-scale, heterogeneous compute node architecture was introduced by the NNSA/ASC Roadrunner system at Los Alamos National Laboratory in 2008. Roadrunner included two dual-core AMD Opteron CPU processors and four IBM Cell eDP GPU video game processors per compute node. With nearly 3.5k compute nodes, the Roadrunner system was the first Petaflop HPC system. While IBM ended the Cell processor line, the concept of heterogeneous compute nodes continued when Nvidia general-purpose GPUs were deployed on Oak Ridge National Laboratory's Titan supercomputer in 2012. GP-GPUs provided a dramatic increase in energy efficient processor performance; but the heterogeneous mix of CPUs and GP-GPUs in these "advanced architecture" compute node designs were very disruptive to the application development community. The complexity of these node architectures has driven much of the subsequent ASCR and ASC investment in applied math algorithms, and in computer science programming models and tools.

DOE's HPC community is still leveraging commodity computing. The recent DOE ASCR workshop on extreme heterogeneity provides additional guidance in this space. A key research topic is emerging: do we need to remain dependent on commodity commercial ecosystem? If we cannot afford this path, can the DOE and US Government at least influence the designs of future commodity computing ecosystem components? Can DOE once again close the loop in co-design and either directly or collaboratively develop custom architectures that can reduce the software development burden, while continuing to meet our energy efficiency performance goals? The DARPA Electronics Resurgence Initiative (ERI) is also making strategic investments that pave the way towards an era with lower barriers to the development of new architecture designs.

Key Concepts for a Co-design Innovation Ecosystem: Integrate Co-design and Lead User Concepts. ASCR and DOE Exascale Computing Project have been adopting co-design principles.

- *Democratizing Innovation* – Lead User vs Manufacturer
 - Lead User is a key concept from *Democratizing Innovation* by Eric Von Hippel
 - Lead Users *have present strong needs that will become general in a marketplace months or years in the future*
 - Lead Users are *a source of novel product concepts*
 - This is in contrast to manufacturer-designed novel product concepts
- Open Innovation
 - Open Source Software – Catalyst for collaboration and shared development
 - Open Source Hardware – Open System on Chip ecosystems: ARM and RISC-V
 - Open software and hardware can drive a culture change – Acceptance of integrating technology and designs from others
- Purpose designed hardware, supported with an integrated system software stack
 - The microelectronics community understands the separation of manufacturing capability from design innovations – *fabless design*
 - We can envision that successive generations of system-on-a-chip (SoC) processors will support a growing portfolio of co-designed IP Blocks/Subsystems to accelerate specific computational kernels
 - Within a SoC eco-system, we can also envision a growing portfolio of special purpose IP blocks, not special purpose accelerators. Moore's Law like performance increases due to increase in IP block reuse, where IP blocks are co-designed with the software stack.

Topics for Focused DOE-SC research on key enabling technologies:

- New transistor materials, materials by design, guided by machine learning
- New packaging technologies to increase transistor design by moving from 2D to 3D
- Optical networking: socket, node, chassis, rack, and system scales
- Advanced memory technologies: materials, protocols and new memory interfaces
- New, high data-rate and high density interconnect approaches particularly for connecting memory to processor
- System software support for accelerator frameworks
- Broad range of processor and accelerator concepts:
 - Non-von Neumann architectures
 - Neuromorphic computing
 - FPGA accelerators
 - Data flow architectures
 - Adiabatic / Reversible Processors
 - Cryogenic Processors
 - Processing in/near memory

Discussion Scope for Breakout Panel III: Power Control, Conversion, and Detection

Authors: Robert Kaplar (SNL) and Debdeep Jena (Cornell)

Energy-efficient computation with electrons has traditionally focused on the transistor, but the control of energy – from its generation to final dissipation in computation – is slated for a major upheaval. This panel will address basic research needs in the energy supply chain – power conversion and control – from the power station to individual transistors. A strong focus will be on the electric grid, due to the recognition that today’s grid has not undergone any major technology evolution from the grid of a century ago – certainly not to the same degree as computational electronics, in which for example FinFETs have evolved from the transistors of 50 years ago. The technology for controlling energy is thus in need of a major overhaul to deal with emergent changes such as the widespread integration of renewable energy sources and energy storage, and increasing vulnerability to natural disasters and/or malicious attack. Important architecture-level issues couple to the other panels - for example, widespread monitoring of the health of the grid will likely involve big data, and topology optimization of the grid for maximally efficient performance may require advanced algorithms and post-exascale computing power. A complementary effort will involve research and implementation of new power conversion technologies, including new semiconductor materials and devices, as well as other materials and components in power conversion circuits. New circuit designs that make maximally efficient use of the new components are also required. These same components and designs should be such that they can be used in other power conversion application areas as well, such as vehicle electrification in the transportation sector.

One approach to increasing the resiliency and efficiency of the electric grid is to replace large power transformers (LPTs) with solid-state power converters. LPTs can weigh hundreds of tons, are difficult to transport, and are typically custom-ordered with fabrication lead times of one year or more. The loss of multiple LPTs on the grid could produce catastrophic consequences, with region-wide power outages of months. Indeed, a National Research Council study* included as its number one recommendation that the Department of Homeland Security should work with DOE and industry to “develop and stockpile a family of easily transported high-voltage recovery transformers and other key equipment.”

This vulnerability can be mitigated by the replacement of LPTs with solid-state transformers (SSTs). In addition to being an order of magnitude smaller in size and weight and modular in design, such SSTs would greatly accelerate the development of smart grid architectures. This is because they can have full programmability, active voltage regulation, unity power factor, programmable AC frequencies and DC capability, rapid disconnect during a fault, and do not require phase coordination between input and output. The extraordinary controllability of SSTs will be a key ingredient to building the smart, agile, topologically reconfigurable, and highly responsive modern grid architecture of the future. Just like the initial computer that filled an entire room has been replaced by a chip, SSTs are slated to shrink the large power transformer to much smaller form factors, while simultaneously dramatically increasing performance and efficiency.

Unfortunately, the current generation of Si power devices (MOSFETs, insulated-gate bipolar transistors, and diodes) enable construction of solid-state power converters at only modest power levels, and are woefully inadequate for grid-scale SSTs. A new generation of power devices using wide band gap (WBG) and ultra-wide band gap (UWBG) materials are required to replace LPTs, since they enable higher power densities and faster switching speeds. The WBG and UWBG materials are rather immature (especially the latter), and will require significant additional R&D if their promise is to be realized for grid applications.

The WBG semiconductors Silicon Carbide (SiC) and Gallium Nitride (GaN) have been developed over the past few decades, and are now being deployed for a wide variety of power conversion applications, leading to greater conversion efficiency and improved size, weight, and power. Moreover, even as the WBG semiconductors continue to mature, research into the next generation of semiconductors, the UWBG materials, has commenced. This family of materials includes Aluminum Nitride / Aluminum Gallium Nitride (AlN/AlGaN), hexagonal and cubic Boron Nitride (c-BN), Diamond, Gallium Oxide (Ga_2O_3), and others. These materials have bandgaps exceeding that of SiC and GaN (3.4 eV), with a bandgap as high as 6.4 eV for c-BN. The critical electric field (defined as the electric field required to induce avalanche breakdown) of these materials is expected to scale as a power-law with bandgap, $E_c \sim E_G^n$, with n in the range 2.0-2.5. Thus, the breakdown voltage of devices fabricated from these materials may be dramatically enhanced compared to SiC and GaN. Moreover, Figures of Merit (FOMs) that assess the trade-offs between breakdown voltage and conduction and switching losses will be similarly enhanced. For example, the unipolar FOM, which measures the trade-off between breakdown voltage and on-resistance or conduction loss, scales as the 3rd power of the critical electric field, and thus at least as the 6th power of bandgap. Hence, the unipolar FOM is expected to be dramatically enhanced for the UWBG materials relative to GaN and SiC, with commensurate benefits to the electric grid and other power conversion applications.

* National Research Council, et al. “Terrorism and the Electric Power Delivery System.” Committee on Enhancing the Robustness and Resilience of Future Electrical Transmission and Distribution in the United States to Terrorist Attack; Board on Energy and Environmental Systems; Division on Engineering and Physical Sciences. ISBN: 978-0-309-11404-2. <http://www.nap.edu/catalog/12050/terrorism-and-the-electric-power-delivery-system>. Accessed Oct. 7, 2018.

However, many challenges concerning materials growth, doping to create carriers of both polarities, device processing, etc. exist at present for the UWBG semiconductors, and much fundamental research is needed to further mature these materials, as highlighted in a 2016 workshop[†] sponsored by the Air Force Office of Scientific Research and the National Science Foundation. As ultrawide bandgap semiconductors are probed at electric field strengths that have not been possible in earlier semiconductors, new and unexpected physical phenomena can emerge. It is expected that several aspects of traditional power device design developed primarily for Silicon will have to be re-examined and re-done for these new materials. How can we as a community encourage and educate researchers to be alert to such opportunities? Since these devices, based on the bold new generation of ultra-wide bandgap semiconductors, will convert and manipulate large amounts of power in very small volumes, handling of the heat dissipated is anticipated to become a challenge, just as a major bottleneck in ultradense CMOS is its heat dissipation. Concurrently, research into novel device architectures, optimized circuit designs, and clever means of thermal dissipation are required in order to take maximal advantage of these emerging semiconductor materials.

Panelists by categories

The panelists represent a number of general categories of expertise:

- Grid architecture
- Power converter design
- Semiconductor materials and devices for power electronics
- Materials for passive components (dielectrics, magnetics)
- Thermal management

Discussion Scope for Breakout Panel IV: Crosscutting Research

Authors: Rick Stevens (ANL) and Harry Atwater (Caltech)

The crosscut panel will explore the common research and scientific threads that run through the other three panels. The focus will be on the bi-directional, closed-loop cycle that includes materials/chemistry, devices, circuits, architectures, and algorithms. The opportunity is to promote a co-design ecosystem in which research and innovation in these spaces is carried out in parallel but in a closely connected/coordinated fashion.

ANTICIPATED OUTCOME OF THE WORKSHOP

During the 3-day workshop and subsequent interactions, the panels will identify and prioritize the key research focus areas for their topic, aligned with Office of Science facility and program needs. Crosscutting issues will also be identified and addressed. As the key research focus areas and crosscutting issues are identified, participants will emphasize pioneering work and bold, game-changing research. The participants will try converge on a limited set of Priority Research Directions (PRDs) that capture the majority of critical research needs identified during the workshop. The panel leads and a subset of panelists will stay during the afternoon of the last workshop day to begin writing the workshop report. A final report, including concise descriptions of the Priority Research Directions as well as longer Panel Summaries, will be delivered to DOE Office of Science early in 2019.

[†] J.Y. Tsao et al., "Ultra-Wide Bandgap Semiconductors: Research Opportunities and Challenges," *Adv. Electron. Mater.* **2018**, 4, 1600501.

APPENDIX: SUPPLEMENTAL REFERENCE MATERIAL

The reference documents listed below are available to all participants on the workshop SharePoint Site (<https://login.ornl.gov/landing>); access information has been provided separately. These can be located by clicking on “Factual Document” on the top rail and then looking in the following folder: “/Reference Documents cited in Factual Document/”

- 2017 IRDS Moore and Beyond CMOS reports
- 2015 ITRS 2.0 More Moore and Beyond CMOS reports
- SRC Vision and Guide
- SRC JUMP Needs Document
- DARPA ERI Program Overview
- Nature Electronics Perspectives
- Modernizing The Electric Grid, Quadrennial Energy Review Chapter 3 (2015)
- Grid Modernization Multi-Year Program Plan (2015)
- International Technology Roadmap on Wide Bandgap Semiconductors
- Ultrawide-Bandgap Semiconductors: Research Opportunities and Challenges
- A Roadmap for HEP Software and Computing R&D for the 2020s (CERN)
- Future High Performance Computing Capabilities (ASCAC)
- BES Exascale Requirements Review
- HEP Exascale Requirements Review
- DOE Workshop on the Management, Analysis, and Visualization of Experimental Data:
The Convergence of Data and Computing

Appendix B Workshop Participants

CHAIR:

Cherry Murray, Harvard University

CO-CHAIRS:

Supratik Guha, Argonne National Laboratory and University of Chicago

Dan Reed, University of Utah

TECHNOLOGY LIAISON:

Gil Herrera, Sandia National Laboratories

PANEL LEADS:

Panel 1: Microelectronics for Big Data at Future Facilities: Memory and Storage

Kerstin Kleese van Dam, Brookhaven National Laboratory

Sayeef Salahuddin, University of California at Berkeley

Panel 2: Co-design for High Performance Computing beyond Exascale

James Ang, Pacific Northwest National Laboratory

Thomas Conte, Georgia Institute of Technology

Panel 3: Power Control, Conversion and Detection

Debdeep Jena, Cornell University

Robert Kaplar, Sandia National Laboratories

Panel 4: Crosscutting Themes

Harry Atwater, California Institute of Technology

Rick Stevens, Argonne National Laboratory

PLENARY SPEAKERS:

Dushan Boroyevich, Virginia Tech University

William Chappell, DARPA

Tsu-Jae King Liu, University of California Berkeley

Justin Rattner, Intel (retired)

Michael Witherell, Lawrence Berkeley National Laboratory

INVITED PARTICIPANTS

Simon Ang, University of Arkansas

Ron Brightwell, Sandia National Labs

William Camp, Sandia National Labs (retired)

Gabriella Carini, Brookhaven National Lab

Barbara Chapman, Stony Brook/Brookhaven National Lab

Zhihong Chen, Purdue University

Alok Choudhary, Northwestern University

Srabanti Chowdhury, University of California Davis

Giri Chukkapalli, Marvel/Cavium

Ramon Collazo, North Carolina State University

Jim Cooper, Purdue University

Suman Datta, University of Notre Dame

Nathan DeBardleben, Los Alamos National Lab

Peter Denes, Lawrence Berkeley National Lab

Dave Donofrio, Lawrence Berkeley National Lab

Nic Dube, Hewlett Packard

Keith Evans, Kyma

Farah Fahim, Fermi National Accelerator Lab

Jack Flicker, Sandia National Lab

Josh Fryman, Intel

Al Gara, Intel
Reenu Garg, Infineon
Maya Gokhale, Lawrence Livermore National Lab
Sam Graham, Georgia Tech
Robert Hoekstra, Sandia National Labs
Axel Hoffmann, Argonne National Lab
Adolfy Hoisie, Brookhaven National Lab
Mark Hollis, MIT
Norbert Holtkamp, SLAC National Accelerator Lab
Zhenyu Huang, Pacific Northwest National Lab
Conrad James, Sandia National Lab
Noble Johnson, PARC
Ken Jones, Army Research Lab
Peter Kogge, University of Notre Dame
Jing Kong, MIT
Sriram Krishnamoorthy, Pacific Northwest National Lab
John Leidel, Tactical Computing Labs
Heiner Litz, University of California Santa Cruz
Ted Liu, Fermi National Accelerator Lab
Daniel Lopez, Argonne National Lab
Andrew Lumsdaine, University of Washington-Seattle/PNNL

INVITED OBSERVERS

Vivek Agarwal, Idaho National Lab
Ken Alvin, Sandia National Labs
Jim Amundson, Fermi National Accelerator Lab
Dan Armburst, Lawrence Berkeley National Lab
David Asner, Brookhaven National Lab
Kevin Barker, Pacific Northwest National Lab
Tiziana Bond, Lawrence Livermore National Lab
Babu Chalamala, Sandia National Labs
Hans Christen, Oak Ridge National Lab
Yong Chu, Brookhaven National Lab
Antonio Ferreira, National Energy Technology Lab
John Field, Lawrence Livermore National Lab
Peter Fischer, Lawrence Berkeley National Lab
William Gu, Jefferson Lab
Salman Habib, Argonne National Lab

Mark Lundstrom, Purdue University
Matthew Marinella, Sandia National Labs
Paul McIntyre, Stanford University
Baxter Moody, Hexatech
Sreekant Narumanchi, National Renewable Energy Lab
Bob Nemanich, Arizona State University
Sanghamitra Neogi, University of Colorado Boulder
Nag Patibandla, Applied Materials
Nigel Paver, ARM Research
Steve Pawlowski, Micron Technology
Ramamoorthy Ramesh, University of California Berkeley
Michael Schuette, AFRL
Shadi Shahedipour-Sandvik, SUNY Polytechnic
Davood Shahjerdi, New York University
John Shalf, Lawrence Berkeley National Lab
Tom Theis, IBM
Jeffrey Vetter, Oak Ridge National Lab
Michael Vildibill, Hewlett Packard
Grace Xing, Cornell University

Chi-Chang Kao, SLAC National Accelerator Lab
Mark Kemp, SLAC National Accelerator Lab
Thomas Koschny, Ames National Lab
Katrina Krulla, National Energy Technology Lab
David Lawrence, Jefferson Lab
Ho Nyung Lee, Oak Ridge National Lab
Pat Looney, Brookhaven National Lab
Christian Mailhot, Sandia National Labs
Petra Merkel, Fermi National Accelerator Lab
John Mitchell, Argonne National Lab
Kelly Perry, SIA
Michael Rabin, Los Alamos National Lab
Robert Rallo, Pacific Northwest National Lab
Curt Richter, NIST
Andreas Roelofs, Los Alamos National Lab

Lindsay Roy, Savannah River National Lab
Liz Sexton-Kennedy, Fermi National Accelerator Lab
David Sickinger, National Renewable Energy Lab
Nikolai Sinitsyn, Los Alamos National Lab

Valerie Taylor, Argonne National Lab
Jana Thayer, SLAC National Accelerator Lab
Jack Wells, Oak Ridge National Lab
Eric Whiting, Idaho National Lab

DOE ATTENDEES

Drew Baden, Office of High Energy Physics
Steve Binkley, Office of Science
Laura Biven, Office of Advanced Scientific Computing
Tof Carim, Office of High Energy Physics
Christine Chalk, Office of Advanced Scientific Computing
Alan Cohen, Office of Fossil Energy
Eric Colby, Office of High Energy Physics
Claire Cramer, Office of Advanced Scientific Computing
Glen Crawford, Office of High Energy Physics
Thomas Cabbage, Office of Science
Jim Davenport, Office of Basic Energy Sciences
Greg Fiechtner, Office of Basic Energy Sciences
Bruce Garrett, Office of Basic Energy Sciences
Kurt Heckman, Office of Science
Barb Helland, Office of Advanced Scientific Computing
Linda Horton, Office of Basic Energy Sciences
Jim Horwitz, Office of Basic Energy Sciences
Rob Ivester, Office of Energy Efficiency and Renewable Energy

Harriet Kung, Office of Basic Energy Sciences
Ted Lavine, Office of High Energy Physics
LK Len, Office of High Energy Physics
George Maracas, Office of Basic Energy Sciences
Helmut Marsiske, Office of High Energy Physics
Raul Miranda, Office of Basic Energy Sciences
Jim Murphy, Office of Basic Energy Sciences
Robinson Pino, Office of Advanced Scientific Computing
Nirmol Podder, Office of Fusion Energy Sciences
Tom Russell, Office of Basic Energy Sciences
Andy Schwartz, Office of Basic Energy Sciences
Michelle Shinn, Office of Nuclear Physics
Jim Siegrist, Office of High Energy Physics
Ceren Suset, Office of Advanced Scientific Computing
Paul Syers, Office of Energy Efficiency and Renewable Energy
Bill Vanderlinde, Office of Advanced Scientific Computing
Bruce Walker, Office of Electricity

This page intentionally left blank.

Appendix C Agenda

Basic Research Needs for Microelectronics – Workshop Agenda

Bethesda North Marriott Hotel & Conference Center – October 23-25, 2018

Tuesday, October 23, 2018

- 7:00 – 8:00 a.m. Registration/Breakfast
Opening Plenary Session – Salon D
- 8:00 – 8:15 a.m. DOE Welcome
Steve Binkley, Deputy Director, DOE Office of Science
- 8:15 – 8:35 a.m. Chair Welcome
Cherry Murray, Harvard University (Workshop chair)
- Session Moderator: Supratik Guha, Argonne National Lab (Workshop Co-Chair)**
- 8:35 – 9:10 a.m. Plenary #1
Justin Rattner, Intel (retired)
- 9:10 – 9:45 a.m. Plenary #2
Michael Witherell, Lawrence Berkeley National Lab
- 9:45 – 10:15 a.m. Break
- Session Moderator: Dan Reed, University of Utah (Workshop Co-Chair)**
- 10:15 – 10:50 a.m. Plenary #3
William Chappell, DARPA
- 10:50 – 11:25 a.m. Plenary #4
Tsu-Jae King Liu, University of California Berkeley
- 11:25 – 12:00 p.m. Plenary #5
Dushan Boroyevich, Virginia Tech University
- 12:00 – 12:15 p.m. Workshop Goals, Expectations and Process
Cherry Murray, Harvard University (Workshop chair)
- 12:15 – 1:15 p.m. Working Lunch – Informal panel discussions in breakout rooms (Lower Level)

- 1:15 – 5:30 p.m. Parallel Panel Sessions
- Panel 1: Big Data Collection, Analytics & Processing for Future SC Facilities – White Oak A*
Kerstin Kleese van Dam, Brookhaven National Laboratory
Sayeef Salahuddin, University of California-Berkeley
- Panel 2: Co-Design for High Performance Computing beyond Exascale – White Oak B*
James Ang, Pacific Northwest National Laboratory
Thomas Conte, Georgia Tech
- Panel 3: Power Control, Conversion and Detection – Brookside A*
Robert Kaplar, Sandia National Labs
Debdeep Jena, Cornell University
- Panel 4: Crosscutting Research – Brookside B*
Harry Atwater, Caltech
Rick Stevens, Argonne National Laboratory
- 3:00 – 4:00 p.m. Refreshments available (Lower Level)
- 5:30 – 7:00 p.m. Closed Executive Session Working Dinner
(Chairs, Panel Leads, Tech. Liaison, DOE) – Oakley
- 5:30 – 7:00 p.m. Break for Dinner (on own)
- 7:00 – 10:00 p.m. Parallel Panel Discussions continued (as needed, at the discretion of the panel leads)

Wednesday, October 24, 2018

- 7:00 – 8:00 a.m. Breakfast – Lower Level
- 8:00 – 10:30 a.m. Parallel Panel Sessions for discussion/preparation of preliminary reports
- 10:30 – 10:45 a.m. Break
- Interim Plenary Session – Salon D**
- 10:45 – 11:15 a.m. Report from Panel 1
Big Data Collection, Analytics & Processing for Future SC Facilities
- 11:15 – 11:45 a.m. Report from Panel 2
Co-Design for High Performance Computing beyond Exascale
- 11:45 – 12:15 p.m. Report from Panel 3
Power Control, Conversion and Detection
- 12:15 – 1:45 p.m. Working Lunch (Lower Level)
- 1:45 – 2:00 p.m. Report from Panel 4
Crosscutting Research
- 2:00 – 2:15 p.m. Proposed consolidated Priority Research Directions (PRDs) and PRD leads
Workshop Chairs

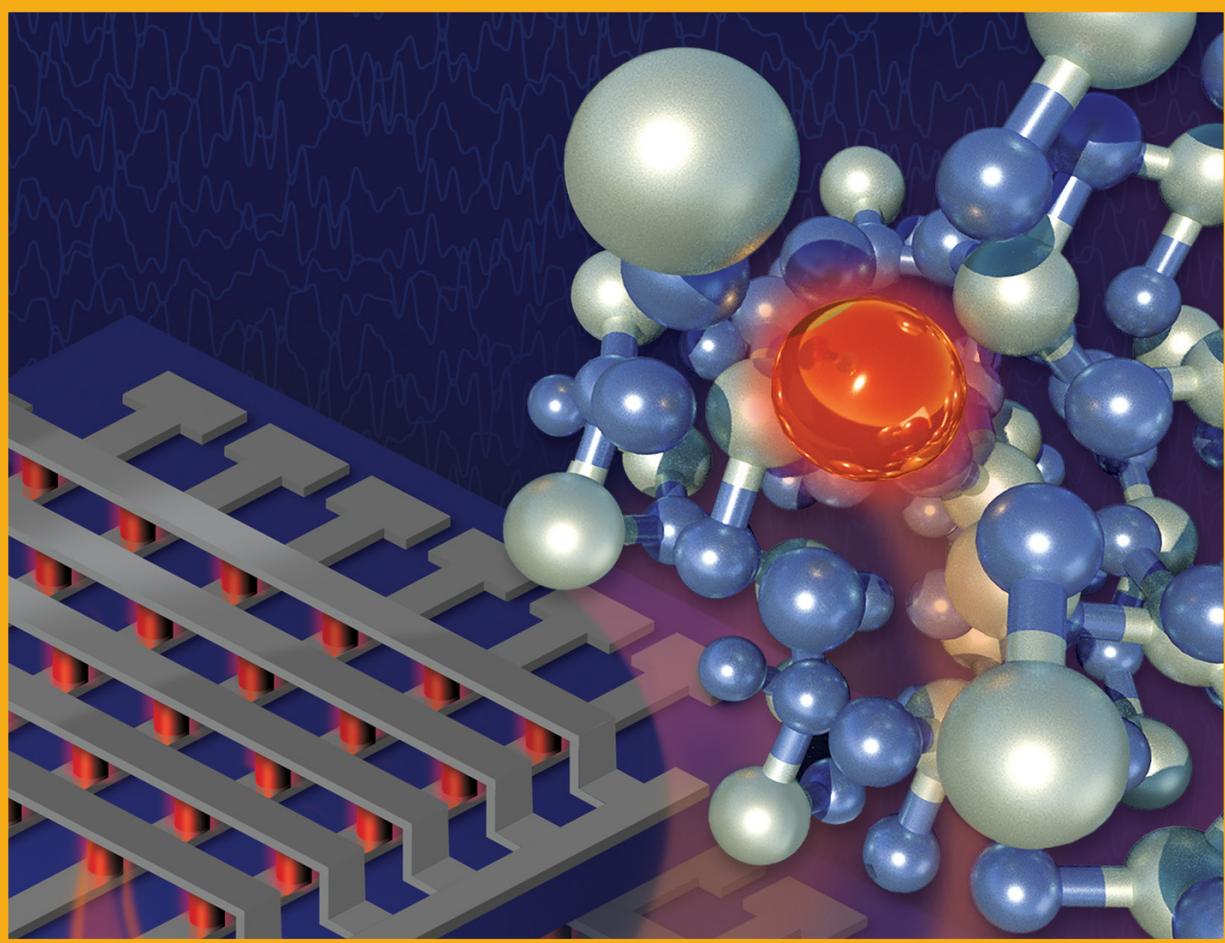
- 2:15 – 5:30 p.m. Breakout Panel/PRD Discussions (continued) – Breakout Rooms
- 3:00 – 4:00 p.m. Refreshments available (Lower Level)
- 5:30 – 6:00 p.m. Closed Executive Session (Chairs, Panel & PRD Leads, Tech. Liaison, DOE) – Oakley
- 5:30 – 7:00 p.m. Break for Dinner (on own)
- 7:00 – 10:00p.m. Parallel Discussions & Preparation for final PRD report-out
(as needed, at the discretion of the panel/PRD leads)

Thursday, October 25, 2018

Closing Plenary Session – Salon E

- 7:00 – 8:00 a.m. Breakfast
- 8:00 – 9:30 a.m. PRD Report-out Session 1 (Schedule/Format TBD)
- 9:30 – 9:45 a.m. Break
- 9:45 – 10:45 a.m. PRD Report-out Session 2 (Schedule/Format TBD)
- 10:45 – 11:45 a.m. Discussion
- 11:45 – 12:00 p.m. Closing Remarks (Workshop chairs)
- 12:00 noon Workshop Adjourned
- Noon – 5:00 p.m. Working Lunch & Writing (chairs, panel leads, and designated writers only)

This page intentionally left blank.



DISCLAIMER: This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government.



U.S. DEPARTMENT OF
ENERGY

Office of
Science