



# SEMICONDUCTOR RESEARCH OPPORTUNITIES

## An Industry Vision and Guide

March 2017



## Contributors

Sohrab Aftabjahani, Intel  
Ameen Akel, Micron  
Robert Boland, BAE Systems  
Jeff Burns, IBM\*  
Rosario Cammarota, Qualcomm\*  
Jon Candelaria, SRC  
Gary Carpenter, ARM  
C.-P. Chang, Applied Materials  
An Chen, IBM\*  
Ching-Tzu Chen, IBM\*  
Michael Chen, Mentor Graphics  
Paula Collins, Texas Instruments  
Ken Curewitz, Micron  
Scott DeBoer, Micron  
Robert Doering, Texas Instruments  
Sean Eilert, Micron  
Rich Fackenthal, Micron  
Mike Fitelson, Northrop-Grumman  
Patrick Groeneveld – Synopsys  
James Hannon, IBM\*  
Ken Hansen, SRC  
Daryl Hatano, ON Semiconductor  
C.-M. Hung, MediaTek  
David Isaacs, SIA  
Clas Jacobson, United Technologies Corporation  
Steve Johnston, Intel  
Lisa Jones, Northrop-Grumman  
Marc Joye, NXP  
Ravi Kanjolia, EMD Performance Materials  
Thomas Kazior, Raytheon  
Taffy Kingscott, IBM  
Curt Kolovson, VMWare  
Steve Kramer, Micron\*  
Zoran Krivokapic, GlobalFoundries  
Ming-Ren Lin, GlobalFoundries\*  
Yu-Ming Lin, TSMC  
Scott List, SRC

Sasikanth Manipatruni, Intel  
Venu Menon, Texas Instruments\*  
Celia Merzbacher, SRC  
Susan Moore, AMD  
Richard Murphy, Micron  
Hans Nahata, Verizon  
Mehul Naik, Applied Materials\*  
Om Nalamasu, Applied Materials  
Tod Newman, Raytheon  
Kwok Ng, SRC  
Dimitri Nikonov, Intel  
Modira Pant, Intel  
Chris Ramming, VMWare  
Sandip Ray, NXP  
Juan Rey, Mentor Graphics\*  
Heike Riel, IBM  
Gurtej Sandhu, Micron  
Jay Shenoy, Micron  
David Speed, GlobalFoundries\*  
Hans Stork, ON Semiconductor  
C.-Y. Sung, Lockheed Martin Corporation  
ScottSweetland, BAE Systems  
Dustin Todd, SIA  
Dana Tribula, Applied Materials  
Wilman Tsai, TSMC\*  
Gilroy Vandentop, Intel  
Shi Qing Wang, EMD Perf Materials  
Amy Wolverton, AMD  
Paul Work, Raytheon  
David Yeh, Texas Instruments  
Ian Young, Intel\*  
Todd Younkin, Intel  
Victor Zhirnov, SRC  
Zoran Zvonar, Analog Devices Inc.\*

\* Research Area Section Leads



# Contents

<b>Contributors .....</b>	<b>ii</b>
<b>Executive Summary.....</b>	<b>1</b>
<b>Vision: Introduction and Overview .....</b>	<b>3</b>
<b>Looking Ahead: Research Needs for Future Computing .....</b>	<b>6</b>
<b>Research Areas .....</b>	<b>9</b>
Advanced Materials, Devices, and Packaging .....	9
Interconnect Technology and Architecture .....	14
Intelligent Memory and Storage .....	18
Power Management.....	22
Sensor and Communication Systems .....	26
Distributed Computing and Networking.....	30
Cognitive Computing.....	33
Bio-Influenced Computing and Storage .....	38
Advanced and Nontraditional Architectures and Algorithms.....	41
Security and Privacy .....	45
Design Tools, Methodologies, and Test .....	51
Next-Generation Manufacturing Paradigm .....	56
Environmental Health and Safety: Materials and Processes .....	59
Innovative Metrology and Characterization .....	63
Conclusion: An Industry Vision and Research Guide .....	66



## Executive Summary

Semiconductors are the foundational technology of the digital and information age. This document presents the semiconductor industry's vision for research needed to continue innovation in semiconductor technology, which in turn will open the door for a host of applications and technologies that enable and support many sectors of the economy. While efforts to advance current technology will continue, it is vital to undertake critical research in an array of areas beyond existing technology.

The remarkable pace of innovation in the semiconductor industry has been sustained through high levels of investment in research and development. In 2016, the global semiconductor industry invested 15.5% of revenue, totaling \$56.5 billion, into R&D, a higher percentage than any industry in the world. A critical driver for faster, better, and cheaper computing power and functionality has been the ability over many decades to manufacture chips with twice as many transistors every 18 to 24 months—a principle known as Moore's Law. This conventional silicon based semiconductor technology is maturing. A new roadmap of technology beyond silicon is required. Advances are required in areas of von Neumann computing such as low-power, low-voltage, beyond-CMOS logic and memory devices and associated materials. In non-von Neumann computing, new memory elements and materials have the potential to enable innovation in the semiconductor industry going forward.

For the semiconductor industry to continue achieving performance improvements, the broader research community needs a comprehensive approach that considers all aspects of semiconductor technology, including novel materials, new manufacturing techniques, new structures, systems architecture and applications. Future semiconductor-based systems—whether small sensors, high-performance computers, or systems in between—must maximize performance while minimizing energy use and providing security and assurance.

The semiconductor industry's vision is that the research agenda outlined in this report will enable further ground-breaking advancements in applications such as artificial or augmented intelligence (AI), the Internet of Things (IoT), high-performance computing (HPC) systems and the ever-connected world that society has come to expect and depend upon.

A clear vision of the research needed to bring new applications into reality is fundamental. The goal of this report is to do just that: identify a prioritized set of research investments throughout the semiconductor industry and value chain. A diverse group of industry experts and leaders came together over a nine-month period in 2016-2017 to outline areas in which research is essential to progress. These areas are:

1. Advanced Devices, Materials, and Packaging
2. Interconnect Technology and Architecture
3. Intelligent Memory and Storage
4. Power Management
5. Sensor and Communication Systems
6. Distributed Computing and Networking
7. Cognitive Computing
8. Bio-Influenced Computing and Storage
9. Advanced Architectures and Algorithms
10. Security and Privacy
11. Design Tools, Methodologies, and Test
12. Next-Generation Manufacturing Paradigm
13. Environmental Health and Safety: Materials and Processes
14. Innovative Metrology and Characterization





Nations that recognize the fundamental importance of the semiconductor industry and make the investments to lead in semiconductor research, manufacturing and innovation will reap critical economic and societal benefits. Worldwide competition for innovation and talent underscores the imperative for timely investments in these areas. Industry, academia, and government together must sustain research investments in the areas listed above and detailed throughout report. Investments must be sufficient to drive both innovation and practical applications.



## Vision: Introduction and Overview

We live in a world in which virtually all aspects of daily life are enabled by semiconductor technology. Semiconductors are essential components of everything from the sensors in car tires and medical monitors to massive cloud-based computer systems and reconnaissance satellites. Our semiconductor-enabled world is characterized by prodigious amounts of data and information, hyperconnectivity, and pervasive computing. Emerging now and with increasing velocity are disruptive applications such as driverless vehicles, personalized “electroceutical” medicines, immersive educational tools, and intelligent assistants that can learn and inform the decisions we make every day. Beyond the horizon are myriad, not-yet-imagined applications that will be based on augmented or artificial intelligence (AI) and other semiconductor-based technologies.

Technological advances—from the widespread deployment in the 1960s of mainframe computers that filled a room to today’s powerful mobile devices that fit in our pockets or on our wrists—have been the result of the steady miniaturization, or scaling, of fundamental semiconductor transistor technology, a trend known as Moore’s Law. In addition to increased computing performance, smaller transistors require less energy to operate and operate at higher speeds. As the number of transistors per silicon wafer has grown exponentially, performance has increased significantly and costs per transistor have decreased dramatically.

Advances in semiconductor technology have fueled innovations that led to new products, new businesses and jobs, and entirely new industries. Looking ahead, key factors driving future innovation are the dramatic growth in embedded “intelligence” and connectivity in computing systems, which underpin the expanding Internet of Things (IoT) and the ability to collect and share massive amounts of data, often referred to as Big Data. At the same time, high-performance computing has become more powerful, making it possible to use data analytics and machine learning to extract useful information and knowledge from an avalanche of data.

Until recently, a high proportion of semiconductor research has focused on the many technologies, materials and processes needed to continue scaling, or reducing the size of transistors, to continue “Moore’s Law.” However, the possible end of Moore’s Law in terms of scaling is creating the need for new research directions, with researchers exploring a variety of strategies. For example, new analog technologies offer great advancements in communication technologies. New approaches are emerging for future computing and information processing devices and systems. New devices are needed to augment the silicon-based transistor, and novel computing architectures to replace the traditional von Neumann architecture, leading to entirely new computing paradigms.

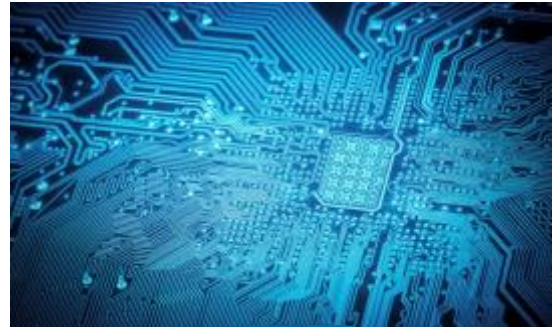
New computing methods and applications require advances in security and privacy. Security is critical, as systems today are rarely stand-alone; society is a system of systems. Even a device that handles what seems like insignificant data can provide a connection to sensitive or private information. Security should

be built-in, not bolted on. Paul Kocher, President and Chief Scientist of Cryptographic Research Inc., has warned, "If we don't fix security issues, the net benefit of new technologies for humanity will *vanish*." <sup>1</sup>

Additional security and privacy challenges are rising as embedded and intelligent systems become ever more pervasive and interconnected. Novel approaches are needed in semiconductor technology and design processes to address the security and privacy challenges posed by the complexity and scale of the IoT ecosystem.

### What is a semiconductor?

*A semiconductor is a material that conducts current, but only partly. The conductivity of a semiconductor is somewhere between that of an insulator, which has almost no conductivity, and a conductor, which has almost full conductivity. Thus its name...semi-conductor. Most semiconductor chips and transistors are created with silicon. Silicon is the heart of an electronic device. Terms like "Silicon Valley" and the "silicon economy," reflect the fundamental role*




*of semiconductors and silicon. Semiconductors are fundamental to society today...imagine life without electronic devices! Semiconductors are all around us. They control the computers we use to conduct business, the phones and mobile devices we use to communicate, cars and planes, the machines that diagnose and treat illnesses, the military systems that protect us, and the electronic gadgets we use every day. Simply put, without semiconductors, the technology that we count on in daily life would not be possible. As a result, semiconductors are highly strategic for societal needs, economic growth and national security.*

The path forward is not as clear as it was during the Moore's Law era. However the enormous potential for economic and societal benefits—some that are envisioned and others yet to be imagined—with continued growth in the ability to collect and process information is driving science and engineering forward. At this pivotal point, progress requires industry, government, and academia to step up. Each has the opportunity to be a key player in the spectrum of activities that will be necessary, from fundamental scientific research to commercial application. The interconnected nature of the IoT and many applications will require even deeper partnerships, across industries and between industry, academia and government in order to succeed.

---

<sup>1</sup> Keynote presentation at 2016 CRYPTO/CHES conference.





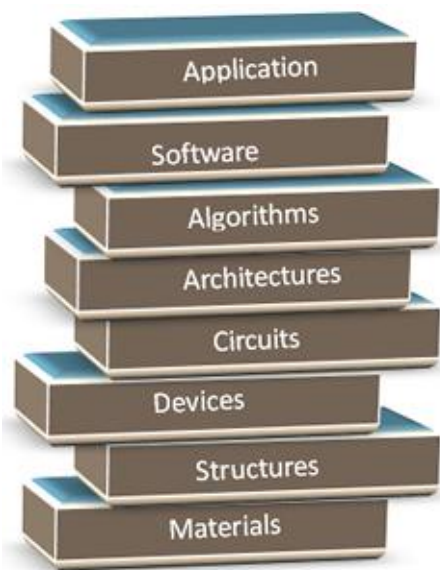
Increasingly, a vertically integrated, multidisciplinary approach will be key to the development of the next generation of technologies. Research on new technologies should be performed with the application in mind and co-designed with software and other system components.

### ***Examples of new computing paradigms being pursued***

- 1. Cognitive systems** are one example of a new approach to using data. Cognitive systems learn, reason, and interact with humans naturally; they support people in decision making. They are not explicitly programmed to perform repetitively a single algorithmic task; rather, they learn and reason from interactions with their environment. They assign probabilities and generate hypotheses, reasoned arguments, and recommendations about complex bodies of data. Cognitive systems understand “unstructured” data, which makes up an increasing portion of the data being generated, and they can keep pace with the volume, complexity, and unpredictability of information and systems. Such systems are well suited, for example, in assisting in medical diagnoses.
- 2. Quantum information systems** promise exponentially more speed and power than traditional computers today. Virtually any area where computing speed and power are critical—from communications to missile detection to encryption to logistics, for example—could benefit from quantum computing. They provide advantages in speed for process-intensive workloads and power to scale out, enabling an entirely new way to tackle problems. *“Quantum computers will open not only a higher processing speed but also applications that we never thought of before.”* Ray Laflamme, Executive Director of the Institute for Quantum Computing, University of Waterloo.<sup>1</sup> Quantum information systems is an area of research that is expected to grow.
- 3. Neuromorphic or brain-inspired computing** is another approach that may lead to novel capabilities and applications. The concept is to design a computer to mimic the way neurons work in the brain and also to achieve power consumption levels that approximate the 20-30 watts used by the human brain. Research is progressing, but the goal of a computer with brain-like functions requires many advances.

## Looking Ahead: Research Needs for Future Computing

Realizing a future in which a combination of distributed, networked sensors, and massive data centers and computing capabilities support innovation and improve our quality of life requires a broad platform of fundamental scientific and technological research. The guiding objective is to enable computing paradigms that radically improve energy efficiency, performance, and functionality while ensuring appropriate security. To achieve this objective, research is urgently needed beyond conventional complementary metal-oxide-semiconductor (CMOS) devices and circuits, beyond von Neumann architectures, and beyond classical approaches to information processing. Additionally, research is needed to develop new materials and scalable processes that lead to a new manufacturing paradigm that can incorporate these new technologies into products.



**Figure 1. The hierarchy of technologies supporting an application.**


The technologies underlying IoT, cloud-based, and high-performance computing and the many applications of semiconductors can be conceived as being organized in a “stack” (Figure 1). Advances in materials, structures, devices and circuits enable architectures that in turn support algorithms and software that have yet to be created, and in turn, these support future applications that will continue to address the economic, social, and security needs of the Nation. Any given application will require specific research and development at many, if not all, levels in the stack.

The diverse group of industry experts who developed this research agenda identified areas that ultimately are interdependent and relate to multiple levels in the technology stack. Progress toward achieving the goals noted above will require multidisciplinary approaches and collaboration among scientists and engineers working at various levels across the entire stack.

The expert industry team identified the following 14 research areas for maintaining U.S. leadership in advanced computing systems:

**Advanced Devices, Materials, and Packaging:** Transistors and other semiconductor devices are the fundamental building blocks of computing, data storage, embedded intelligence, etc. As current technologies approach physical limits and novel architectures are developed, new materials and devices as well as advanced packaging solutions are essential. Advances will enable ultimate CMOS technologies, beyond-CMOS applications, and non-von Neumann computing paradigms.

**Interconnect Technology and Architecture:** Interconnects carry digital information within and between integrated circuits. Limitations of current interconnect technologies are leading to inefficiencies and impacting system performance. Revolutionary advances are needed in interconnect materials, mechanisms, and designs.



**Intelligent Memory and Storage:** The rapidly growing applications based on data analytics and machine learning will benefit from a paradigm shift in how memory is used and accessed. Advances in memory and storage technologies and architectures will improve system performance, enhance security, and enable intelligent systems.

**Power Management:** Critical infrastructure, industrial processes and other systems are powered by electricity. Next generation systems depend on innovations in wide-gap materials, active and passive power devices, designs, and packaging to revolutionize how power is switched, converted, controlled, conditioned, and stored efficiently.

**Sensor and Communication Systems:** A key enabler of the information age and the emerging IoT is the ubiquitous ability to sense and communicate information seamlessly. Future sensor systems will require energy-efficient devices, circuits, algorithms, and architectures that adaptively sense the environment, extract and process information, and autonomously react. Communications systems must be dynamically adaptive and resilient. Efficient spectrum use and interference mitigation will be required to ensure secure service.


**Distributed Computing and Networking:** The growing interconnected web of computing capability, as well as the enormous amounts of data across the IoT create a challenge and an opportunity for distributed computing. Large scale distributed computing systems, supporting very large numbers of participants and diverse applications, require advances in system scalability and efficiency, communications, and system management optimization, resilience, and architecture.

**Cognitive Computing:** Cognitive systems that can mimic the human brain, self-learn at scale, perform reasoning, make decisions, solve problems, and interact with humans will have unprecedented social and economic impact. Creating systems with essential cognitive capabilities requires advances in areas including perception, learning, reasoning, predicting, planning, and decision making; efficient algorithms and architectures for supervised and unsupervised learning; seamless human-machine interfaces; networking cognitive sub-systems; and integrating new cognitive systems with existing von Neumann computing systems.

**Bio-Influenced Computing and Storage:** The convergence of biology and semiconductor technologies has the potential to enable transformational advances in information processing and storage, design and synthesis, and nanofabrication at extreme scale. Examples include DNA-based storage, biosensors, cell-inspired information processing, design automation of biomolecular and hybrid bio-electronic systems, and biology-inspired nanofabrication.

**Advanced and Nontraditional Architectures and Algorithms:** New applications and advanced computing systems require scalable heterogeneous architectures co-designed with algorithms and hardware to achieve high performance, energy efficiency, resilience, and security. Alternatives to the prevalent von Neumann architecture include approximate computing, stochastic computing, and Shannon-inspired information frameworks can provide significant benefits in energy efficiency, delay, and error rates.

**Security and Privacy:** The dependence on interconnected, intelligent systems means that security and privacy need to be intrinsic properties of the components, circuits and systems. Design and manufacture of trustworthy and secure hardware will require design for security, security principles and metrics, security verification tools and techniques, understanding threats and vulnerabilities, and authentication strategies.



**Design Tools, Methodologies, and Test:** Advances in the design and test capabilities are coupled to breakthroughs in materials and architecture, enabling new capabilities to be incorporated in designs and produced at scale. Enormous challenges are posed by growing complexity and the diversity of beyond-CMOS technological options.

**Next-Generation Manufacturing Paradigm:** Advanced manufacturing techniques, including for technologies other than CMOS, as well as tools, and metrologies with high precision and yield, are required to process novel materials, fabricate emerging devices and circuits, and demonstrate functional systems.

**Environmental Health and Safety Related to Materials and Processes:** The semiconductor industry's reputation, freedom to innovate, and profitability depend on a proactive approach to environmental health and safety issues. In addition to developing EHS understanding of new materials and processes early, currently used materials and processes can be improved. Strategies and technologies are sought that minimize waste streams, emissions and occupational risk.

**Innovative Metrology and Characterization:** Semiconductor features are measured in nanometers and the trend is toward 3D stacked structures. Innovative characterization and metrology are critical for fundamental material studies, nanoscale fabrication, device testing, and complex system integration and assessment.

These key topics clearly illustrate the diversity of challenges for future technology and the importance of a coordinated approach by basic researchers and technical experts in universities, government research agencies, and industry.

A successful model for leading the integrated, collaborative research described in this report is the Semiconductor Research Corporation (SRC). SRC is an industry consortium that since 1982 has defined research goals and supported precompetitive university research focused on the semiconductor industry's long-term needs in partnership with government and academia. In addition to increasing the pool of fundamental knowledge, another critical outcome of SRC's work is the education of future industry science and engineering leaders.

Today's semiconductors are approaching the limits of current technology paradigms at the same time as market demands are rapidly evolving and global competition for competitive advantage is intensifying. Significant new technologies, new investments, and new partnerships are required to maintain U.S. leadership. Timely action is needed: industry envisions a 10-year time horizon to realize breakthroughs in the challenges outlined in this report. The time to confront these challenges is now.

## Research Areas

This section describes the 14 key areas in which fundamental research is considered vital to achieving the benefits of next-generation computing, data analytics, and artificial intelligence, and their myriad associated applications. Because the topics are interdependent, there is some overlap in the descriptions. The topics are organized from the lowest level of the technology stack (materials and devices) to the highest (architectures and algorithms), followed by areas that are crosscutting such as security, design tools, manufacturing, and metrology. Each description includes potential topics for research over the next 10 years and recommendations for research strategies. Each research area is essential to achieving the overall vision. Efficient progress will depend on strong coordination of activities across all of the areas.

### Advanced Materials, Devices, and Packaging

#### Introduction/Overview


New information processing systems with dramatically improved energy efficiency and performance will require devices with unique characteristics, possibly based on unconventional mechanisms. In addition to current research needs and challenges in scaled CMOS and conventional architectures, the requirements for novel devices have to factor in advances in and requirements of alternative architectures such as neuromorphic architectures. Such devices may be developed to further improve von Neumann computing (e.g., lower power with steep slope devices) or to enable non-von Neumann computing architectures (e.g., memory-in-logic or neuromorphic computing). Devices leveraging alternative state variables, especially those beyond charge and spin such as exciton and photon, may be used to achieve unprecedented capabilities. The emergence of IoT drives the needs for materials and devices with extremely low power, suitable for flexible substrates, and capable of energy generation or scavenging.



Emerging devices and mechanisms often require materials with different properties, which necessitates extensive research in alternative material systems and the associated interface properties, e.g., III-V, SiGe, carbon-based, low dimensional (2D), multiferroic, ferroelectric, magnetic, phase change, and metal insulator transition materials. To support research and development of new material systems, atomically precise deposition and removal (etch and clean) methodologies are required that are suitable for large areas, low defects, tight geometries (sub-10 nm), and 3D integration, as well as high throughput.

Although security has mainly been implemented in system designs, algorithms, and protocols, materials and devices with intrinsic characteristics suitable for security primitives (e.g., true randomness or unclonability) have the potential to realize robust security features in hardware. Co-optimization of devices and architectures will be critical to fully utilize unique device characteristics and to improve architectural performance.





Advanced 3D integration and packaging techniques enable vertical scaling and functional diversification, offering promising opportunities to enhance system performance and functionalities through heterogeneous integration. In addition to processing innovations, advances in materials and devices may also drive the development of packaging technologies and broaden the applications of 3D integration.

The increasing diversity and complexity of products, from small embedded sensors to heterogeneous “systems on a chip” pose growing technical challenges for packaging. Today’s heterogeneous systems integrate elements that were formerly relegated to board-level integration, such as various passive components (capacitors, inductors, etc.) and active components (antennas and communication devices such as filters), as well as memory and logic. Today’s system-level integration allows more function per unit volume. However, it also accentuates challenges such as within-package power delivery to more functional blocks, thermal density management, and maintenance of signal integrity. This trend also leads to increased assembly complexity and costs, along with associated reliability and testing requirements.

Fast-growing application areas are placing specific demands on packaging technologies. In high-performance computing, I/O bandwidth density bottlenecks and thermal management challenges are limiting overall packaged system performance, driving renewed interest in alternatives to metallic conductors for off-chip communications. Advanced automotive applications are driving the need for new packaging materials able to withstand higher thermal and power densities and more robust material interfaces to improve reliability. Mobile consumer applications are imposing severe form factor constraints and are driving innovation in packaging. Examples of recent advances in the mobile space include advances in chip-scale packaging, fan-out wafer-level packaging schemes, and adoption of 3D/2.5D integration technologies. IoT product concepts require low-cost packaging but at the same time call for innovations, for example, in encapsulation of flexible and stretchable electronics for wearables and other emerging applications.

### **Potential Research Topics**

#### **Low-power, low-voltage, beyond-CMOS logic and memory devices and associated materials for von Neumann computing:**

- Steep-slope devices with large  $I_{on}/I_{off}$  ratios:
  - Tunnel field-effect transistors (TFETs) and other novel tunneling transistors, such as resonant tunneling
  - Transistors with gain in the gate stack, such as negative capacitance FETs
- Devices based on phase transition, lattice distortion, interface mechanisms, and other transduction mechanisms:
  - Mott transition devices, charge density wave (CDW)-based devices, strain-based devices and piezoelectric transistors
- Spin-based logic and memory devices:
  - Sub-nanosecond switching speed with low applied current density
  - Orders of magnitude improvement in tunnel magneto-resistance (TMR) in perpendicular magnetic anisotropy (PMA) junctions
  - High spin polarization through spin filtering barriers

- 10X improvement in charge-spin conversion efficiency, e.g., research on spin orbit coupling, Rashba interface, and topological insulator materials
- Magneto-electric and magneto-strictive switching mechanisms to enable orders of magnitude improvement in energy efficiency
- Spin-based devices, exploiting domain-wall motion and other novel mechanisms
- Magnetic memories in tiered and embedded applications, e.g., ultrafast switching enabled by anti-ferromagnetism, giant spin Hall effects for novel device design, etc.

**Beyond-CMOS devices, memory elements, and materials for non-von Neumann computing:**

- Devices for hardware acceleration of machine learning, suitable for artificial neural networks training and inference, e.g., artificial neurons and synaptic devices, 2-terminal analog resistive devices, memristive devices, spin-based devices, and other emerging analog devices for neuromorphic and bio-inspired information processing, as well as the associated materials development.
- Devices capable of nanofunctions, such as native multiplication, addition, and division, with fidelity.
- Memory elements and 2-terminal selector devices for novel array computing and storage implementations, including 3D adaptations.

**Devices and materials leveraging other state variables in addition to charge and spin** (e.g., electrochemical, electrobiological, photonic, and phase) to achieve significantly improved performance, information density, or energy efficiency.

**IoT-related devices and materials:**

- Ultralow-power devices and designs for sensing, information processing, storage, and communication in sensor node and networks.
- Materials and devices on flexible or other unconventional substrates for large-area in-sensor computing and machine learning.
- Materials and devices for energy generation, scavenging, storage, and management for size- and weight-constrained platforms.
- Beyond-CMOS devices and materials for THz communication and sensors.

**Devices and circuits for security:**

- Devices with unique properties to enable built-in security features, e.g., camouflage, logic encryption, etc.
- Devices to implement security primitives with reduced energy and area overhead, e.g., Physical Unclonable Functions (PUFs), random number generators, etc.

**Materials and devices for power management:**

- Semiconductor materials and devices for improved power performance including low turn-on resistance for a given breakdown voltage and optimal switching figures of merit. These include wide bandgap semiconductors (e.g., GaN and SiC) for high-voltage, high-power devices, and high-mobility semiconductors (e.g., GaAs) for low-voltage power converters.
- Materials and devices for passive components (capacitors, resistors, and inductors) in power conversion systems.
- Materials and packaging techniques for multichip modules.

## Devices enabling high-density, fine-grained, monolithic 3D systems for reducing data movement and communication cost:

- High-performance devices for logic and memory suitable for low-temperature processing used for stacked layer integration; the resulting energy-delay product should be 1000X better than state-of-the-art conventional technology. Candidate materials for next-generation logic devices include 1D nanotubes and nanowires, and 2D graphene, transition metal dichalcogenide (TMD) materials. Suitable memory devices are spin transfer torque random access memory (STTRAM), resistive random access memory (RRAM), and ferroelectric random access memory (FeRAM) or any novel device that can withstand low thermal budget and offer high performance.
- Integration of energy harvesters, connectivity devices, and sensors.
- Heat management of multilayer stack integrated circuits (ICs).
- Novel packaging for small-form-factor, accessible pins in all facets of the package.
- Integration of neuromorphic devices with baseline CMOS.
- Biocompatible packaging options.

### Known Research Activities

#### United States

**Nanoelectronics Research Initiative (NRI):**<sup>2</sup> Industry-government partnership supporting U.S. university research to explore a broad range of beyond-CMOS devices and materials, including steep slope low-power devices, spintronic materials and devices, 2D materials (including graphene and TMD) and devices, and ferroelectric and multiferroic materials and devices. NRI researchers have studied extensively both devices based on non-charge state variables and electrical devices based on novel mechanisms.

**STARnet:**<sup>3</sup> Industry-government partnership supporting U.S. university research in six multiuniversity centers to explore a wide range of topics, including functional materials, spintronic and other devices, architectures, and intelligent networks. The STARnet Function Accelerated nanoMaterial Engineering (FAME) center in particular focuses on novel materials including TMD, functional oxides, and magnetic materials. The Center for Spintronic Materials, Interfaces and Novel Architectures (C-SPIN) focuses on spintronic materials, devices, and architectures. The Low Energy Systems Technology (LEAST) center focuses on low-power devices and architectures.

**Georgia Institute of Technology 3D Systems Packaging Research Center (PRC):**<sup>4</sup> The Georgia Tech PRC takes an integrated, interdisciplinary, and system-level approach with particular focus on: (1) leading-edge electronic and bio-electronic systems research, (2) cross-discipline education of graduate and undergraduate students, and (3) collaborations with companies from the United States, Europe, Japan, Korea, Taiwan, and China. PRC is pioneering transformative systems for a System-On-Package.

---

<sup>2</sup> <https://www.src.org/program/nri/>

<sup>3</sup> <https://www.src.org/program/starnet/>

<sup>4</sup> <http://www.prc.gatech.edu/>

## Europe

**IMEC:**<sup>5</sup> (1) IMEC “Beyond-CMOS” research in the logic program explores a range of steep slope devices (e.g., tunnel FET) and beyond-CMOS devices (e.g., spintronic devices); (2) IMEC Memory program explores both RRAM and STTRAM. Research in these programs is supported by integrated CMOS process and scaling capabilities at IMEC Fab, as well as extensive design expertise at IMEC’s Integrated Solutions for Technology Exploration (INSITE) program.

**CEA-Leti:**<sup>6</sup> A French research-and-technology organization specializing in nanotechnologies and their applications in healthcare, telecommunications, smart devices for medical treatment and safety, transportation, environment, defense, security, and space. NEMS and MEMS are at the core of its research activities.

**European Graphene Flagship program:**<sup>7</sup> A future and emerging technology flagship of the European Commission (EC). With a 1 billion euro budget, the Graphene Flagship explores graphene materials, devices, and applications with the goal of taking graphene from academic laboratories into society in 10 years.

## Japan

**National Institute of Advanced Industrial Science and Technology (AIST):**<sup>8</sup> AIST is one of the largest public research organization in Japan. The Electronics and Manufacturing Department explores nanoelectronics, photonics, advanced manufacturing, spintronics, flexible electronics, and ubiquitous MEMS and microengineering, etc. The Materials and Chemistry Department studies nanomaterials, carbon nanotube applications, computational design of advanced functional materials, etc.

**Center for Innovative Integrated Electronic Systems (CIES)**<sup>9</sup> at Tohoku University: CIES was established for international university-industry research collaboration with the focus on exploring spin-transfer torque magnetic RAM (STT-MRAM) materials, devices, and architectures.

## Research Recommendations

**Conduct fundamental research on devices and materials:** This is critical in the beyond-CMOS era. Design and system innovations based on existing CMOS technology have introduced new applications and market opportunities. However, design and architectural improvements have limits. Sustainable and scalable breakthroughs need to originate from basic device and material research.

**Closely connect device research to innovative design and architecture research:** Co-optimization of novel devices and architectures has potential to improve system performance and efficiency.

**Support device and material research with manufacturing facilities** that are able to test new concepts and innovations in arrays and on test chips.

---

<sup>5</sup> [http://www2.imec.be/be\\_en/home.html](http://www2.imec.be/be_en/home.html)

<sup>6</sup> <http://www-leti.cea.fr/en>

<sup>7</sup> <http://graphene-flagship.eu/>

<sup>8</sup> [http://www.aist.go.jp/index\\_en.html](http://www.aist.go.jp/index_en.html)

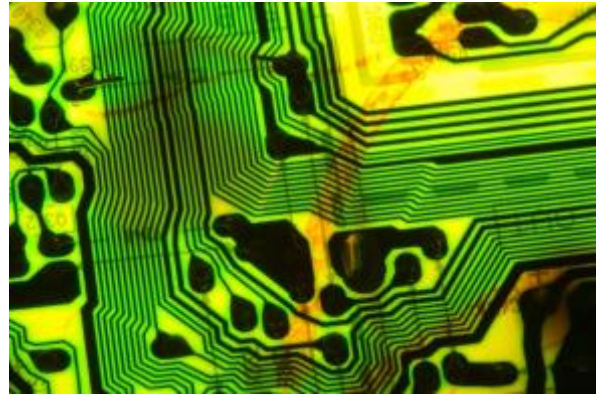
<sup>9</sup> <http://www.cies.tohoku.ac.jp/english/>

**Guide beyond-CMOS device, material, and architecture research by quantitative benchmarks** based on uniform and comprehensive methodologies and criteria.

## Interconnect Technology and Architecture

### Introduction/Overview

Interconnects carry information and power within and between integrated circuits and play a critical role in determining power requirements and performance of semiconductor products. Current interconnect technology is driven by on-chip dimensional scaling, off-chip bandwidth, and energy vs. performance trade-off. These parameters define the scaling of routing design, selection of conductor and dielectric materials, as well as fabrication methods. Beyond conventional scaling, new horizons being pursued include development of new materials, novel interconnect architectures compatible with emerging devices, and alternate compute methods leading to alternate signal conversion and propagation media.



As circuit and system designers seek to continue scaling to smaller dimensions and to increase functionality, interconnect requirements call for lower delay, higher power efficiency, wider bandwidth, and greater reliability. But as interconnects become smaller, RC delay of electrical interconnect increases, dynamic power consumption rises with decreasing dielectric spacing, lower signal integrity and greater cross-talk limit the application of wider bandwidth links, and longer total wire length causes higher rates of failure and lower yield. These trends pose serious challenges to design margins, material property requirements and robustness, and process integration optimization.


Beyond the serious scaling challenges, interconnect architecture must look ahead and adapt to emerging device technologies, new materials, novel computing paradigms, new fabrication schemes, and new application-driven requirements.

Emerging devices, such as new field-effect-transistors (FETs), spintronic devices, and photonic devices, impose new restrictions on interconnects. Integrating interconnects with new devices requires thermal budget matching, RC management, and electrical contact with new materials. For spintronics and photonics, signal conversion and density in interconnect hierarchy need to be considered. Spin-based propagation heavily depends on relaxation length and time, which results in restrictive specifications for signal volatility and integrity. Photonic devices have cut-off length wavelengths that limit density scaling.

A common challenge among all devices is variability. New technologies with promising performance characteristics must be capable of scaling up for manufacture while maintaining acceptable levels of variability. At smaller dimensions, this becomes an even greater issue due to stochastic process variation.

New materials such as 2D materials and 2D hybrid structures have many promising properties, but there are barriers to their wide adoption. Challenges include methods for reproducible large-area deposition,





achieving contact and line resistance competitive against existing materials, packaging techniques, and an inherent tradeoff between carrier density and mobility.

Novel computing paradigms include parallel computation by allowing multivalued logics (e.g., analog and quantum computing) and by leveraging signal weight and convergence (e.g., neural network and neuromorphic computation). These new approaches to computation have implications for interconnect architecture, interconnect materials, signaling mechanisms, and fabrication.

Changes in integration and packaging also lead to interconnect technology disruption. For instance, the transition from 2D to 3D monolithic integrated circuits requires stacked layers of transistors, which poses challenges for interconnect routing and thermal budget control. In addition, heterogeneous integration of multifunctional systems requires co-design of interconnects into active modules. Fan-in/fan-out compatibility needs careful examination at the device-interconnect interface. In the longer term, “intelligent interconnect”, such as dynamically reconfigurable fabrics, are in early stages of research.

Expanding functionality and diversity of applications, such as sensors, biomedical measures, IoT modules, cloud servers, etc., can lead to interconnect customization. In addition to evolving on-chip requirements, off-chip interconnect architecture is confronted with new challenges to connect modules performing a specific function at the board level or in the cloud. Interconnect options to consider include wired versus wireless, electron- versus non-electron-based, and on-chip versus off-chip.

Interconnect in all forms is an important part of next-generation semiconductor products. Advancements in interconnect technology and architecture require careful benchmarking and assessment, including their compatibility with emerging devices.

### **Potential Research Topics**

- **Technologies that enable sub-10 nm electronic interconnect**, including conductors, dielectrics, and their integration methods.
- **Novel metals and composites** to replace current metals, i.e., beyond copper (Cu).
- **Novel interconnection between strata beyond metal vias** (e.g., optical and plasmonic interconnect).
- **Novel self-alignment and self-assembly techniques** to improve integration density.
- **Self-forming barriers and novel 2D barrier materials.**
- **Photonic switching devices and interconnects**, including light sources, detectors, and modulators.
- **Spin-based interconnects**, including novel materials for spin propagation.
- **THz wired and wireless interconnects and line-of-sight/non-line-of-sight (LOS/NLOS) transmission**, optical interconnects, radio frequency (RF) photonics, free space, backhaul transmission, and innovations in air-interface for zero-overhead and scalable transmission, etc.
- **Native interconnects for novel devices that use alternative state variables.**
- **Programmable, high fan-in and fan-out interconnect solutions.**
- **Superlinear, wide-bandwidth electrical and optical links** enabling high modulation formats.

- **Exploration of data center-level interconnect and networking innovations** that dramatically improve scalability and reduce latency and energy consumption.
- **Self-optimizing and resilient networks, reconfigurable interconnect fabrics, and high-speed, secure data links.**
- **Holistic approaches to harness advanced memory/storage devices** and complement CMOS with interconnect and packaging technologies optimized for nontraditional and cognitive computing.

### Known Research Activities

Research activities in interconnect technology include the following:

#### **On-chip core interconnect:**

- **Interconnect technology options for scaling beyond 7 nm or 5 nm technology node:** As an example, a Georgia Tech group (led by Prof. Naeemi) benchmarked interconnect contribution to chip performance, identified scaling limits, and compared between different technology options, including conventional electrical wires, 2D material-based interconnect, spin transport, and domain wall interconnect.
- **Alternate conductor materials as replacement of copper:** As examples, (1) groups at Rensselaer Polytechnic Institute (RPI) (Prof. Gall) and University of Central Florida (Prof. Coffey) surveyed the resistance scaling potential of alternate metals including the majority of transitional metals; their theoretical work defines key properties, including resistivity, Fermi surface, electron mean free path, and conductance isotropy; (2) a group at the College of Nanoscale Science and Engineering State University of New York Polytechnic Institute (Prof. Dunn) studied small dimension (<10 nm) wire resistance and scattering behaviors.
- **Next-generation capacitance solutions, including new ultralow-k (ULK) materials and novel interconnect schemes:** New ULK materials, including periodic mesoporous organosilicates (PMO), metal-organic frameworks (MOF), zeolite, and spin-on glass, recently have been studied in research consortia (e.g., imec) and universities. New curing approaches, such as vacuum ultraviolet (VUV), e-beam, and laser annealing, are proposed to further enhance film mechanical properties. Process protection is also developed to reduce process damages, e.g., low-k repair and pore seal (imec). While results show pristine dielectric constant (k value) below 2.0, process integration impact on these materials needs further study. Alternative integration schemes, e.g., air-gap and low-k replacement, have been adopted in industry. The scaling factors (dimension, material, process impact, and cost effectiveness) face potential limits and require breakthrough technologies.
- **Copper-low-k interconnect reliability:** Numerous groups have been investigating time-dependent dielectric breakdown and electromigration mechanisms within the SRC-NYCAIST (New York Center for Advanced Interconnect Science and Technology). ULK damage and breakdown mechanisms are well understood owing to the work at Columbia University (Prof. Heinz), University of Wisconsin-Madison (Prof. Shohet), RPI (Prof. Plawsky), University of North Texas (UNT) (Profs. Kelber and Du), University of Texas (UT) at Arlington (Prof. Kim), etc. New metrology and test vehicles have been developed at University of Michigan (Prof. Chen), Columbia University (Prof. Venkataraman), and others. For copper electromigration, the UT-Austin group (Prof. Ho), among others, has conducted extensive study in characterization,

failure modes, statistics, and benchmarking of solutions such as metal caps and alternate claddings. Foundries have adopted some of the work, e.g., metal cap, various metal cladding schemes, and low-k repair.

### Alternative Interconnect technologies

- **Non-electronic transports (photons, RF signals, spins):** Optical fiber has been used in long-distance broad-bandwidth data transmission. Optical communications have advantages of energy efficiency and high bandwidth by multiplexing. Research frontiers lie in the electrical-optical interface. Research consortia such as those at France's CEA-Leti, Belgium's imec, and Singapore's A\*STAR Institute of Microelectronics are making progress in optical module fabrication and demonstration of data transmission. Photonic components (typically  $\mu\text{m}$  dimension) are not as scalable as electrical wires (on the order of nm). In addition, the fabrication of signal source (laser) faces tremendous challenges, e.g., diode materials, optical alignment, stability, and modulation. RF or high-frequency electromagnetic propagation is in a similar situation. A printed antenna on board serves the transceiver and receiver functions sufficiently; however, it is unclear how on-chip antenna would deliver cost-efficient performance. Spintronic interconnect is critical for spin-based logic. Spin current to electrical current conversion requires spin filters. In addition, spins have relaxation time and length that are dependent on interconnect media, which require careful material engineering and architecture design to preserve signal integrity.
- **2D materials (graphene, carbon nanotubes, TMD, etc.) for interconnect:** Purdue University (Prof. Chen) and Stanford University (Prof. Wong) have made progress in evaluating 2D or 2D-metal hybrid interconnect structures, together with research consortia (imec, CEA-Leti). Advantages include high mobility and thermal conductivity. The roadblocks include lack of large-area deposition techniques suitable for high-volume manufacturing, high out-of-plane contact resistance, and difficulties in controlling multilayer films with suitable electrical properties (mobility, carrier density, resistivity, etc.). Novel approaches have been proposed to address these challenges, including side-wall contact (IBM) and film doping (imec).
- **Interconnect for new computing paradigms**, e.g., multivalued logic, quantum computing: Multivalued logic saves computing power via parallelism, which requires the interconnect to carry and maintain voltage or current levels to avoid losing logic inputs during transport. Research is underway on designing quantum dots and quantum bits to achieve the logic truth table with stability, but compatible interconnect requirements have not been defined. Inevitable IR drop in short wires and vias in the new paradigm may potentially restrict the design of logic front-end.
- **Local interconnect for vertical transistor stacking and monolithic 3D ICs.** CEA-Leti researchers proposed their CoolCube™ program to investigate the possibility of stacking transistors in the vertical dimension. Research includes routing design and thermal budget design across levels to ensure equivalently good transistor behavior at all stack levels. Tungsten was selected as the local interconnect conductor to fill small dimension vias with high aspect ratios. The required routing compromises some of the density benefit from transistor stacking.

## Research Recommendations

**Novel metal barrier and dielectric barrier material / process screening in advanced interconnect**, with scalability requirements (metal barrier < 1 nm, dielectric barrier < 5 nm) and metrologies to understand the fundamental physics of material interaction.

**Alternate conductor materials and alternate inter-layer dielectric (ILD) materials** with considerations of integration impact and correct benchmarking to industrial product records. Scaling benefits from these materials should be leveraged through design rule layout and change of routing / interconnect architecture. Alternate schemes (damascene, subtractive) suitable for new materials should be Identified.

**New integration scheme concepts** to boost yield (e.g., to solve pattern alignment challenges) and performance (e.g., lower RC delay), identification of necessary materials, and demonstration in practical industrial flow.

**Inflection of alternate signal transport (spin, photon, others) techniques on interconnect** (architecture visions, material change, design restrictions, fabrication demands), and definition of hierarchy boundaries versus electrical wires, metrologies, and characterizations.

**Interconnect architecture in advanced packaging solutions**, co-design of local / intermediate / global interconnects and interposers in different fan-out packaging schemes and thermal management solutions.

**Off-chip interconnect**, including wireless, optical links, active interconnects, and resilient data transmission.


**New materials integration** that is inherently robust against variability.

## **Intelligent Memory and Storage**

### Introduction/Overview

Advances in information technology have pushed data generation rates and quantities to a point where memory and storage are the focal point of optimization of computer systems. Transfer energy, latency, and bandwidth are critical to the performance and energy efficiency of these systems. Creating the ideas and tools that are able to sidestep these information throttles has broad implications for future memory and storage. The solutions to many modern computing problems involve many-to-many relationships that can benefit from high cross-sectional bandwidth inherent to a distributed computing platform. As an example, large-scale graph analytics involve high cross-data-set evaluation of numerous neighbor relationships that ultimately demand the highest possible cross-sectional bandwidth of the system.





For radical performance improvements in complex data handling, a holistic, vertically integrated approach to high-performance “intelligent” storage systems is needed, encompassing the operating system, programming models, and memory management technologies, along with a prototype system architecture. Significant advancements may be made by utilizing distributed compute elements and accelerators in close physical and/or electrical proximity to the location of data storage.<sup>10</sup> A more optimal hardware platform may contain a large fabric of interconnected memory and storage devices with large cross-sectional bandwidth and integrated logic and accelerators in conjunction with islands of conventional, high-performance computing elements. The platform might even include sub-elements that allow the computing elements to access all or parts of the shared memory space as cache-coherent memory, and further, features that shuttle and order data from the memory fabric to deliver more information-rich data streams to the high-value processor and accelerator islands. In pursuit of novel platforms, the goal is to achieve orders-of-magnitude gains in power, performance, and density (volumetric as well as processing).

The primary research gap in realizing a paradigm shift in how memory is used and accessed is in establishing an operating system framework allowing run-time optimization of the system based on system configuration preferences, programmer preferences, and the current state of the system. There will be many run-time optimization challenges in such a heterogeneous platform containing near-memory von Neumann and non-von Neumann elements. There are inherently numerous means to achieve the same end on such platforms (e.g., compute slowly locally, transfer data and compute quickly remotely, compute very efficiently but approximately in an accelerator), and proper optimization may differ depending on the current state of the system. For example, execution time and data movement might be automatically balanced in a near-optimal fashion on simple hardware through operating system and hardware hooks, with or without programmer intervention; built-in controls may allow a system administrator to trade off performance, power efficiency, response latency, etc., to operate within the constraints of the specific system installation. Scheduling decisions may involve proper use of local and global metrics of bandwidth, power, latency, temperature, etc.

In order to establish run-time optimization algorithms, suitable metrics for system performance must be established to measure information processing density, e.g., decision rate or correct decision rate per Kg of silicon, cm<sup>3</sup>, or Watt. In a simple example, imagine a single rack unit enclosure containing >100K memory dies interconnected in a high-dimensional fabric, each capable of accessing and operating on data locally, and each consuming up to 1 watt of power. Such a system would require algorithms for power throttling based on temperature and power consumption that would likely involve local and global aspects of the control system. System theoretic bounds on information processing density should be established to demonstrate the research areas of highest potential. To meet performance and resilience targets, replication will be part of such systems and will be co-optimized with retention and endurance

---

<sup>10</sup> Christos Kozyrakis, [DRAF: A Low-Power DRAM-based Reconfigurable Acceleration Fabric](#) (2016); Christos Kozyrakis, [HRL: Efficient and Flexible Reconfigurable Logic for Near-Data Processing](#) (2016); Stratos Idreos, [JAFAR: Near-Data Processing for Databases](#) (2015); Steve Swanson, [SSD In-Storage Computing for List Intersection](#) (2016)



management and authentication strategies (i.e., which users are allowed to run which programs on which data).

Progress will be accelerated through a holistic, vertically integrated approach that focuses on the relevant, emerging memory technologies and their respective, novel system architectures and hierarchies (including subsystems and their caches), as well as the advanced materials and processes required to manufacture them. An eye towards backward compatibility, where possible, is also of importance to facilitate migration of applications to this new framework.

### **Potential Research Topics**

**Compute-in-memory systems:**<sup>11</sup> Compute-in-memory has been demonstrated in many forms, ranging from general-purpose CPUs tightly coupled with, or integrated onto, memory chips, to analog computing using memory arrays. Research is needed to understand algorithms by which systems dynamically balance thermal and power budgets, bandwidth, compute, and memory or storage capacity to best serve metrics of interest such as cost-of-operation, decision latency, and decision bandwidth. This research should address which specific metrics are of interest and how system parameters affect them. Given that there are large similarities between in-memory compute systems and Compute+Memory+Sensor nodes (simple, low-cost, low-energy) capable of making basic observations and/or decisions that report to a larger system, it is envisioned that there may be compelling research in which these are regarded as the same concept.

**New architecture and programming paradigms for memory and storage in a system; self-optimizing (semi-autonomous) systems allowing for appropriate programmer control:** Imagine a large multidimensional network fabric of thousands or millions of memory nodes in a single appliance within a system that is capable of managing power, thermal, bandwidth, compute, and storage constraints autonomously. For this system to be possible, a programming paradigm is required that allows programmers to articulate algorithms in such a way that the underlying system can dynamically optimize the execution of these algorithms. Features of this paradigm might include ways in which the programmer can, or even be required to, specify critical code paths and parallelization opportunities. It is conceivable that movement of programs to data may be even more common than moving data to programs. As such, both programs and the data they operate on must be enclosed in well-defined containers that can be handled in a modular fashion. Feature extraction local to sensor nodes is conceptually similar to in-memory computation or data filtering; thus, it is expected that research in this area could significantly benefit both Big Data analytics and intelligent memory, perhaps affecting the same workload in multiple ways.

**Authentication, resilience, and coherence:** Inherent in the operating system must be controls such that only “allowed” programs are permitted to run and such that they are only permitted to run on “allowed” data. The integrity of both the programs and data must be ensured to a degree agreed upon by the programmer and the administrator. Additionally, it is expected that due to the scale of the system, it will

---

<sup>11</sup> Onur Mutlu, [LazyPIM: An Efficient Cache Coherence Mechanism for Processing-in-Memory](#) (2016); Onur Mutlu, [PIM-Enabled Instructions: A Low-Overhead, Locality-Aware Processing-in-Memory Architecture](#) (2015); Onur Mutlu, [A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing](#) (2015).

be advantageous to use program replication to resolve resilience and performance challenges. The added feature of replication makes the issue of coherence, both among replica copies and in the context of processor and accelerator caches, a challenging area of research and innovation. This will also necessitate algorithms that determine when and how to spawn and collect replica copies.

**IO, networking:** Again, imagine a system consisting of a large multidimensional network fabric of thousands or millions of memory nodes in a single appliance within a system that is capable of managing power, thermal, bandwidth, compute, and storage constraints autonomously. For this system to be possible, a very efficient, globally asynchronous network will be required that can gracefully optimize around failing links and over-provisioned bandwidth while dynamically throttling within local and global thermal and power constraints.

**New technologies, materials, and processes:** It is expected that as solutions to the problems associated with intelligent memory and storage materialize, there will be new requirements on device technologies, materials, and processes. Solutions may come from existing and emerging technologies that become useful in these new environments, while others may need to be newly created.

### **Known Research Activities**

**Hybrid Memory Cube Consortium**, a working group of industry leaders  
(<http://www.hybridmemorycube.org/>)

**Center for Automata Processing**, University of Virginia, Kevin Skadron, Director (<http://cap.virginia.edu/>)

**Multiscale Architecture & Systems Team (MAST)**, Stanford University, Christos Kozyrakis, Director  
(<http://web.stanford.edu/group/mast/cgi-bin/drupal/>)

**SAFARI Research group**, Carnegie Mellon University, Onur Mutlu, Director  
(<http://www.ece.cmu.edu/~safari/>)

**Parallel Data Lab (PDL)**, Carnegie Mellon University, Greg Ganger, Director (<http://www.pdl.cmu.edu/>)

**Data Systems Lab (DASlab)**, Harvard University, Stratos Idreos, Director (<http://daslab.seas.harvard.edu/>)

**Non-Volatile Systems Lab (NVSL)**, University of California, San Diego, Steven Swanson, Director  
(<http://nvsl.ucsd.edu/index.php?path=home>)

**Scalable Energy-efficient Architecture Lab (SEAL)**, University of California, Santa Barbara, Yuan Xie, Director (<https://seal.ece.ucsb.edu/>)

**Utah Arch** (Computer Architecture Research), University of Utah, co-led by Rajeev Balasubramonian, Mahdi Nazm Bojnordi, and Erik Brunvand (<http://arch.cs.utah.edu/>)

**High Performance Computing Lab**, Pennsylvania State University, Chita Das, Director  
(<http://www.cse.psu.edu/hpcl/index.html>)

**System Energy Efficiency Lab (SEE)**, University of California, San Diego, Tajana Šimunić Rosing, Director  
([http://seelab.ucsd.edu/mem\\_arch/overview.shtml](http://seelab.ucsd.edu/mem_arch/overview.shtml))

**EcoCloud**, École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland, Babak Falsafi, Director (<http://www.ecocloud.ch/about/>)

**Multimedia VLSI Laboratory (M VLSI)**, KAIST, Republic of Korea, Lee-Sup Kim, Director  
(<http://mvlsi.kaist.ac.kr/research>)

**Fast-Forward**, U.S. DOE Extreme Scale Technology Acceleration award to IBM Research, Jaime Moreno  
([http://researcher.watson.ibm.com/researcher/view\\_person\\_pubs.php?person=us-jhmoreno&t=1](http://researcher.watson.ibm.com/researcher/view_person_pubs.php?person=us-jhmoreno&t=1))

### Research Recommendations

**Realizing a comprehensive overhaul of the memory access and use architecture:** This key technological challenge necessitates a new operating framework that provides for run-time optimization of the system based on system configuration preferences, programmer preferences, and the current state of the system's run-time optimization.

**Actualizing truly intelligent memory and storage systems:** This will require a coordinated program involving a team of scientists with expertise in operating systems, programming, architecture, circuits, devices, and materials.

## Power Management

### Introduction/Overview

Essentially all systems are dependent on electrical power for their operation. New and innovative systems and applications will both drive and be enabled by revolutionary advances in how that power is switched, converted (AC to DC, DC to DC, etc.), controlled, conditioned, and stored.

Efficiency of these operations is one of the most important characteristics. For example, we need power switches that more closely approach the ideal of zero resistance when ON and infinite

blocking voltage when OFF and switching between these two states in almost zero time with zero losses. Over the next decade, we need improved power conversion efficiency from low power levels (e.g., battery-operated personal devices) to very high power levels (e.g., solar conversion, high power transfer, and grid-renewable energy connectivity). Devices will be needed to cover voltage ranges from a few volts up to thousands of volts at power levels from a few watts to multiple kilowatts. Even though the power levels will be different, almost all of these applications will be driven by the requirement for higher power density at lower costs than are achievable today.



The key challenges that need to be addressed over the next decade are:

1. Significant improvement in power conversion indices like power density, conversion efficiency, failure rate, conversion cost, and weight while reducing environmental impact (e.g., reducing the amount of aluminum, copper and iron per kW).
2. Continued dynamic performance improvement of the overall system that makes it more fault tolerant and robust, increases switching speeds by more than 10–50 times, and reduces system size and cost.

3. High-performance advanced packaging that drives extremely low inductance and lower thermal resistance with excellent thermal cycling reliability.
4. Analysis and design of more complex and more connected power systems.

Breakthroughs in conversion efficiency, linearity, etc., need to happen, and these would need significant improvements in materials, devices (e.g., Si/GaN/SiC/GaAs transistors, passives, and magnetics), and circuit topologies (e.g., for power device drivers). Advances are also needed in power conditioning and power storage, such as in more-efficient charging of batteries. Novel methods of heat reduction and removal are also important. Finally, on-chip/in-package integration of new technologies will be required for significant improvements in power density.

The focused research areas that are needed are:

1. Improvements in semiconductor, as well as passives (capacitor, magnetic), material, and device architectures to get extremely low  $R_{DS(on)}$ , high voltage-tolerant, high switching speed devices that can operate reliably and efficiently at very high operating temperatures.
2. Innovation in converter topologies like multi-cell architectures, resonant topologies, and CAD tools for their analysis and optimization. Also, new active control topologies to improve system fault tolerance.
3. Thermally efficient and highly reliable packaging with low inductance and compatibility with different semiconductor materials
4. System-oriented analysis and virtual prototyping of complex power systems: more comprehensive electromagnetic interference (EMI) and reliability modeling of the entire system, including the semiconductor, the package, and the printed circuit board.

### **Potential Research Topics**

**Converter topologies:** Typical converter topologies, such as buck, boost, and buck-boost, have well-established performance metrics and have been used widely for decades. With the development of improved power transistors over the last 10 years, the opportunity to explore more complicated converter topologies now presents itself. Improved converter topologies allow for significant improvements in power density and EMI. This in turn will support improved power delivery efficiencies, enable higher-power systems, and improve overall system ease of use. Take, for example, a mobile battery charger where solution volume and thermals are of utmost importance: With battery sizes growing and battery charge current rates increasing, the ability to support higher charge currents through power conversion techniques is the limiting factor in phone charging time. New topologies offer the opportunity to improve charging time without violating stringent size, volume, and thermal constraints. Within the topology categories for low voltage systems, there are two primary suggested segments:

- **Switched capacitor converters:** Development of high-density, switched capacitor converters has demonstrated the ability to dramatically improve power density and conversion efficiency. The remaining challenges with switched capacitor conversion are interconnect and conversion ratio.
  - With interconnect, the ability to connect power transistors to multiple discrete capacitors limits the number, type, and complexity of the switched capacitor converters employed. By seeking improved switched capacitor topologies combined with on-chip or monolithic

integration of high density capacitors, topologies previously considered prohibitively complicated could become more realistic.

- With conversion ratio, it is well known that switched capacitor converters offer very high efficiency at fixed conversion ratios. In practical applications, the ability to operate over wide input and output voltage ranges is a necessity. One increasingly interesting area for research is incorporating resonant techniques to enable wide-conversion-range switched capacitor converters. To overcome limitations, these techniques specifically employ very small inductors, often less than 10 nH, to supplement traditional switched capacitor converters.
- **Resonant converters:** Resonant converters have well-established topologies, which add value by eliminating losses occurring during each switching period, often referred to as switching losses. These losses limit the operating frequency of the converter and, as a result, also limit the achievable power density. The main challenges with resonant converters lie in achieving excellent performance over a wide range of operating conditions. Research is needed on resonant converters for low-voltage systems that can maintain high efficiency over wide  $V_{in}$ ,  $V_{out}$  and load current ranges. These topologies should not add substantial complexity to the system so that the required power density improvements can be maintained.

**Passive component integration:** Passive components such as capacitors, resistors, and inductors are a fundamental and critical part of any power conversion system. Typically, the two most critical passive elements are inductors and capacitors. In many cases, these passives actually limit the achievable performance of the system due to their size, losses, and cost. Significant research is needed to improve passive components, as well as overall system performance.

- **Capacitor integration:** Integration of capacitors can provide substantial improvements in power management system performance. Locating the capacitor as close to the switching device as possible helps with improved ease of use, higher efficiency, and higher power density. High-density capacitors integrated into silicon can be used for bootstrap capacitors, input capacitors, output capacitors, and decoupling capacitors. The performance advantages of each of these possible applications differ in specifics, but in general, they result in an overall more efficient and higher-density system. Research is needed not only to look at improving the capacitance density (new materials, for example) but also to look at manufacturability, including cost of manufacture.
- **Inductor integration:** Power inductors are necessary in many switching power supply designs. The inductors provide filtering of large signal converter waveforms and improve efficiency and controllability. Similar to capacitors, despite significant advances in process technologies, magnetics can often be the limiting factor for system performance. There has been significant effort towards integrating magnetics into silicon, with limited commercial success. The barriers to entry in these applications are performance and cost.

Performance-related challenges typically center on materials. In order to build a high quality inductor, thick materials (copper, magnetic material, etc.) are necessary. The thicker the materials that can be deposited, the lower the resistance and the higher the saturation currents. Research is needed on new magnetic materials, inductor architectures, and deposition/etch techniques to make integrated inductors practical from performance (efficiency, current levels) and cost perspectives.



**Multi-chip module (MCM) technologies:** The need to co-package dissimilar chips is becoming important. The ability to do so provides the freedom to optimize individual die for its primary purpose while enabling improved power density, system solution cost, and ease-of-use. This need becomes even more necessary with the use of GaN and SiC. Research is needed into module materials, modeling of multi-chip package stress, thermal effects and performance, and simulation tools to emulate multi-chip modules.

**Power device technologies:** Power converter performance is fundamentally dictated by the performance of the switching transistors. These transistors typically have a few key figures of merit (FOMs) that help designers quickly assess the quality of the technology. The first is  $R_{sp}$  for a given breakdown voltage ( $BV_{dss}$ ).  $R_{sp}$  is a metric for determining the resistance caused by a unit area of the FET. The lower the  $R_{sp}$ , the better the device (smaller die area for the same resistance). The second is a set of switching FOMs that determine the amount of energy that needs to be provided to charge/discharge the capacitive elements of the power device. The lower these charge-based FOMs, the better the overall performance.

Although silicon continues to be the dominant semiconductor material used for power devices, there has been an increasing use of wide bandgap semiconductors—such as GaN and SiC for high-voltage, high-power devices—and of high mobility semiconductors like GaAs for low-voltage power converters. These devices offer substantially improved  $R_{sp}$  and switching FOMs, which in turn enable higher switching frequency and reduced power losses.

To increase the acceptance of these new materials in the power management area, significant research is needed over the next decade in device/process modeling, understanding the fundamental device operation and reliability aspects, improved growth techniques, reducing defectivity, and improving and scaling the manufacturing to 200 mm and even 300 mm to help improve the cost-performance metric.

It is critical to innovate on new device architectures using these materials to continue improving the key device FOMs. In addition to GaN, SiC, and GaAs, there are other materials with even wider bandgaps, such as AlN and diamond, which could help achieve higher levels of performance. Device and manufacturing technology as well as reliability understanding for such new materials should also be subjects of research over the next decade.

Additionally, as power devices are improved, there will be a requirement to develop new packaging techniques that are compatible with both Si and non-Si materials, to take advantage of their improved performances.

### **Known Research Activities**

**PowerAmerica Manufacturing USA Institute:** a public-private partnership sponsored by the U.S. Department of Energy and led by North Carolina State University to advance wide bandgap semiconductor technologies, including those based on SiC and GaN.

<https://www.poweramericainstitute.org/>

**FREEDM Systems Center:** At the Future Renewable Electric Energy Delivery and Management (FREEDM) Systems Engineering Research Center, U.S. universities have joined forces with industry partners to develop a more secure, sustainable environmentally friendly electric grid. Research priorities include power electronics packaging, controls theory, solid state transformers, fault isolation devices, and power systems simulation and demonstration.

<https://www.freedm.ncsu.edu/>

**CPES:** The Center for Power Electronics Systems is dedicated to improving electrical power processing and distribution that impact systems of all sizes, from battery-operated electronics to vehicles to regional and national electrical distribution systems. <https://www.cpes.vt.edu/>

### **Research Recommendations**

**Improved power conversion and management systems** with innovations in semiconductor materials, converter topology, passive component integration, and packaging.

**Wide bandgap semiconductors and devices with power FOM significantly improved beyond those of Si.**

**Advanced MCM and packaging techniques that are thermally efficient and highly reliable** in order to integrate different power units, devices, and materials.

## **Sensor and Communication Systems**


### **Introduction/Overview**

A key enabler of the information age is the ubiquitous ability to seamlessly sense and communicate information. The explosion of information and the emergence of the Internet of Things are straining the capability of current technologies. As a result, the electromagnetic spectrum is becoming increasingly congested and new sensing and communication solutions are required to satisfy growing demand for information. Two synergistic application areas are sensors and communications operating at microwave, millimeter wave, or THz frequencies. Advances are critical for continued innovation in wireless communications, radar, computer vision, and reconnaissance and imaging systems used in consumer, military, industrial, scientific, and medical applications.



As one example, future sensor systems operating in RF to THz frequency bands will require novel, energy-efficient devices, circuits, algorithms, and architectures for adaptively sensing the environment, extracting and processing information, and autonomously reacting or responding. As another example, cognitive communication systems that operate in complicated radio environments with interference, jamming, and rapidly changing network topology will be expected to sense information about their environment and available resources and dynamically adjust their operation. In addition, efficient spectrum use, interference mitigation, and spectrum prioritization will be required in order to ensure secure services to end users.

Creating these new system capabilities requires new approaches for devices and circuits, for example, breakthroughs in mixed signal and analog circuit performance, with very low noise, high sensitivity, and low power dissipation characteristics while maintaining practical operating characteristics. In addition, many sensors operate in the analog domain at relatively high voltages, so new approaches to analog signal processing, analog-to-digital conversion, and efficient power management (detailed in another section) are essential.



On the system side, the emphasis is to create the end-to-end systems that will provide a complete solution for a given application. In wireless communications the requirement is also to influence future standardization, thereby enabling global proliferation of the technology.

### **Potential Research Topics**

To address these applications, breakthrough ‘integrated’ research is required, addressing all aspects of sensor and communications systems including materials, devices, components, circuits, integration and packaging, architectures, and algorithms. In addition, the key of the research theme is a holistic approach to development and design by applying optimization across all the elements of the protocol stack and taking advantage of true cross-layer design.

Research is needed to achieve and/or improve the following:

#### **New classes of sensors that will emphasize agility, compact design, computation at the sensor node, and privacy, among other characteristics:**

- Agility: reconfigurable, adaptive, multifunction, multimode, self-calibrating sensors with increased degrees of freedom for efficient use of the electromagnetic (EM) spectrum, including spectrum agility, instantaneous bandwidth/waveform agility, (very) wide bandwidth, and high dynamic range.
- Compact and inexpensive with ability to sense and measure multiple variables.
- Coordinated use of multiple sensor modalities.
- Resilient, with trust and privacy.
- Incorporate internal sensor memory and compute.
- Enable collaboration of sensors and compute through flexible and resilient architectures and interface protocols.

**Superlinear communication links** that enable high modulation formats and integrated communications components for IoT, **and distributed sensor systems** that enable ultralow-power, high-data-rate, long-range sensor communications with highly linear up/down conversion:

- Large format (e.g., 3 m x 3 m) flexible smart sensors and sensor arrays.
- Highly linear, high-efficiency (low power dissipation), (ultra)wide-bandwidth, ultralow-noise analog/RF devices and circuits (e.g., PAs, LNAs, switches, mixers, and IF amplifiers) with practical operating characteristics such as reliable operation at room temperature or higher.
- Breakthroughs in mixed signal and analog circuits (e.g., high dynamic range, low power dissipation ADCs, DACs, etc.), including new approaches (circuits and architectures) for analog signal/information processing.

#### **Architectures and circuit solutions for sensors and communications:**

- Agile (adaptive and dynamically reconfigurable) RF devices, circuits, and architectures including self-optimizing and resilient networks and dynamically reconfigurable interconnect fabrics.
- Scalable, adaptive behavioral learning and cognitive processing (information extraction) architectures for real-time and reliable sensing, operation, and communication (e.g., feature-

driven "intelligent" sensing that achieves the ultimate energy efficiency by only detecting salient features, patterns, and/or thresholds).

- Novel circuits, architectures, and/or algorithms for precision timing for sensor and communication system operation in PNT (positioning, navigation, and timing)-denied environments.
- Novel, energy-efficient transceiver modules, subsystems, arrays, and architectures including high-power (high power density) mw to THz transmitters, receivers, and communication links such as massive yet scalable multiple-input, multiple-output (MIMO) and beam formers.

**Co-design simulation tools and methodologies:** Development of rapid, accurate, multiphysics, multidimensional, multidomain (co-)design (RF/analog/EM, thermal, and mechanical) and simulation tools and methodologies for complex (e.g., 3D and heterogeneous) multifunction RF systems.

**New system level management of spectrum utilization** to improve spectrum occupancy and allow "smart" frequency allocation (including noncontiguous band aggregation and use of new waveforms) and/or sharing to overcome interference (including self-induced) and jamming problems (e.g., scan and detect transmission across very wide range of frequencies to avoid interference as well as to make efficient use of spectrum resources).

**Communication and reconnaissance approaches** to mitigate threats and enable rapid transmit and receive of critical data and constant streaming, autonomous operation, and decision making including embedded real-time learning, ability to recognize threat scenarios, and ability to do local processing before transmitting the data or information. Autonomous operation also may require integrated power generation and harvesting, distribution, control and regulation, and modulation.

#### **Materials and devices enabling RF to THz operation**

- Novel materials (semiconductors, packages, substrates, magnetics, metamaterials, etc.); devices (diodes, transistors, switches, varactors, etc.); components (High Q inductors, filters, low loss transmission lines, couplers, duplexers, antennas/radiating elements, etc.); and associated fabrication processes and manufacturing methodologies for realization of RF to THz circuits and systems.
- New technologies, materials (metals, dielectrics), components (sources, detectors, propagation media, etc.), fabrication processes, and architectures to realize (ultra)low-loss, wide-bandwidth interconnects and communication links (e.g., THz wired and wireless interconnects and LOS/NLOS transmission, optical interconnects, RF photonics, free space, backhaul transmission, innovations in air-interface for zero-overhead and scalable transmission etc.) that dramatically reduce energy consumption in communication or distribution of data/information.
- Materials, devices, components, and packaging for extreme operating conditions (e.g., high temperature) and harsh environments.
- Materials, devices, circuits, subsystems, and approaches for package-level or integrated (e.g., on-chip) energy generation (harvesting), storage, distribution, management, and regulation.
- Material, devices, circuits, packaging, assembly, and systems for spectroscopy and mWave and (sub)THz imaging for military and commercial applications (e.g., for vital signs, security, non-invasive bio-signal characterization, etc.).
- Novel packaging and integration approaches, including 3D and heterogeneous integration of diverse materials (e.g., Si, III-Vs, II-Vis, 2D material, magnetics, ferromagnetic, ferroelectric,

multiferroics, etc.); and devices and functions to improve performance, reduce cost, create new capabilities, and/or increase functional density (e.g., fully integrated logic, RF, and mixed signal functions).

### **Known Research Activities**

The area of sensors and communication systems has attracted much interest in the research community. Some relevant examples include:

**Platforms for Advanced Wireless Research (PAWR):** National Science Foundation (NSF) program that aims to support advanced wireless research platforms conceived by the U.S. academic and industrial wireless research community. PAWR will enable experimental exploration of robust new wireless devices, communication techniques, networks, systems, and services that will revolutionize the nation's wireless ecosystem, thereby enhancing broadband connectivity, leveraging the emerging Internet of Things, and sustaining U.S. leadership and economic competitiveness for decades to come.

**Spectrum Collaboration Challenge (SC2):** DARPA's first-of-its-kind collaborative machine-learning competition to overcome scarcity in the radio frequency spectrum.

**STARNet program:** TerraSwarm center co-funded by Semiconductor Research Corporation and DARPA explores the potentials and risks of pervasive integration of smart, networked sensors and actuators in the connected world.

**Horizon 2020 ICT-Leadership in Enabling and Industrial Technologies (LEIT) Work Programme:** European Community (EC) funded programs within Horizon 2020 ("H2020") provide a balanced response to the main challenges faced by Europe in the field: firstly, the need to maintain a strong expertise in key technology value chains; secondly, the necessity to move quicker from research excellence to the market.

### **Research Recommendations**

Research investment should focus on:

**Breakthrough and disruptive technologies underpinning system solutions that will enable end-to-end services and support.** For worldwide deployment, ultimately, approaches are required that have potential to become standard platforms.

**Advanced sensors with greater agility, compact design, computation at the sensor node, and privacy.**

**Mixed signal and analog circuits with significantly improved performance,** including very low noise, high sensitivity, and low power dissipation, as well as practical operating characteristics. Also needed are new approaches to analog signal processing and analog-to-digital conversion.

**Innovations at all levels of the technology stack** must occur, from materials and devices through architectures and system-level management; co-design across these levels will be essential.



## Distributed Computing and Networking

### Introduction/Overview


The use of computing to support enterprises and communities in social interaction, commerce, defense, and governance requires large-scale distributed computing systems. These systems support very large numbers of participants and very large numbers of diverse applications.

Data center examples of distributed systems include warehouse-scale public cloud computing infrastructures and high-performance computing clusters used for scientific research and for commercial applications such as oil and gas exploration. Emerging big-data workloads execute on distributed clusters. At the edge of the network, IoT processing is based on distributed systems in which the edge and endpoint devices are much smaller in performance, capability, and energy consumption than the compute nodes in a data center cluster. Between the data center tier and the edge tier, one or more aggregation tiers may exist in a distributed system, providing intermediate processing and storage to balance the high capabilities in the data center or cloud, with the relatively modest capabilities of endpoint and edge devices. While centralized computing systems such as symmetric multiprocessors play a critical role in workloads such as transaction processing, the memory consistency requirements of symmetric multiprocessors limit their scalability to a few tens of processors. Distributed systems do not require a similar consistency model and, as a consequence, can be scaled to much larger numbers of processors. As a result, the ongoing growth in computing capacity worldwide is based on distributed architectures.

In distributed architectures, resources including compute, memory, storage, sensors, and actuators are spread throughout the system and are interconnected using a variety of networking architectures and technologies. Because an application must access these networked (distributed) resources in the course of its execution, the cost and overhead of communication becomes a significant factor in the performance and efficiency of the system. New end-user applications require significant improvements in communications, namely lower cost and latency, and increased bandwidth. Physics-based implementation constraints on latency and energy exacerbate the challenges in improving communications. Hence, innovations in networking are an imperative at all levels, from on-chip to between data centers, using both wired and wireless technologies and protocols.

There are significant imbalances in the costs and in the constraints on performance growth and energy consumption of compute, storage, and networking that will require novel distributed-system advances well beyond today's hardware and distributed architectures. In order to support the continued growth in demand for computing, for existing workloads, and especially for emerging applications that leverage big data and deliver cognitive computing capabilities, dramatic advances must be realized in distributed





architectures and systems. The purpose of this research is to explore the challenges of extremely large-scale distributed architectures with the objective of unlimited scalability. This scalability must be practical and realistic in terms of energy consumption and cost in order for these new distributed systems to be economically viable and deployable. Due to the maturation of Moore's Law, improvements in performance and energy consumption will increasingly rely on accelerators—of which today's graphics processing unit (GPUs) are an example—and other heterogeneous architecture techniques. As a result, future distributed systems must increasingly enable heterogeneity without adding a large software development or programming burden. Novel, multitier, wired and wirelessly connected systems are required; tiers may comprise sensors and/or actuators, aggregation, cloud or data center, or combinations thereof. All tiers are expected to be highly scalable, and heterogeneity is expected both within and across the tiers.

The endpoint interfaces to the physical world are often analog in nature; however, this section focuses primarily on digital computing, addressing all processing once the endpoint signals are converted to the digital domain where the vast majority of the processing occurs in most systems. Sensors and actuators that interface to analog signals are addressed in another section. Workload- and system-management aspects of distributed systems pose many challenges. Security and privacy also are key topics that must be addressed, including safety and security in multi-tenant, multiworkload scenarios. Research is needed to address the key new challenges in resource and applications management that must effectively be overcome in order for these novel, very-large-scale systems to support a wide range of existing and emerging workloads.


### **Potential Research Topics**

**System performance, efficiency, and applicability:** To enable and deliver substantial improvements in system value and capability, research is required to advance distributed-system performance growth and to achieve dramatic improvements in energy efficiency at all levels, from individual components up to the system level. Relevant goals are to:

- Demonstrate breakthrough levels of scalability and efficiency for large-scale systems of multiple, heterogeneous tiers (e.g., sensor, aggregation, data center and/or cloud).
- Achieve major improvements in system performance, energy efficiency, and robustness across compute, networking, and storage.
- Include both nontraditional and traditional architectures in tasks in order to push the boundaries for distributed-system software research and application.

**Networking and communication:** Distributed systems require highly effective communications mechanisms and technologies at all levels of the system hierarchy, from on-chip communication to the wide-area networks between data centers. Research in networking and communications is central to this research. Key research areas include:

- Architectures, protocols, algorithms, and systems to support >10X improvements in power efficiency and latency for wired and wireless communication.
- Exploration of data center-level interconnect and networking innovations that dramatically improve scalability and reduce latency and energy consumption.
- Self-optimizing and resilient networks, reconfigurable interconnect fabrics, and high-speed, secure data links.



**System management:** Distributed systems management has multiple facets. The systems themselves must be monitored for proper operation, and they must be tuned and optimized to achieve the necessary levels of performance and quality of service. Security and privacy of data must be monitored, managed, and guaranteed. Workloads must be scheduled, monitored, and tuned. Research is required to advance the state-of-the-art in all of these areas. Suggested areas of emphasis are to:

- Deliver and demonstrate innovations for management of privacy and authentication protocols across the system, with provable properties and guarantees.
- Deliver and demonstrate innovations for configuration, workload, and data management for large-scale, multi-tenant systems.
- Develop system instrumentation and analytics for automatic prognosis, diagnosis, reconfiguration, optimization, and repair of large-scale distributed systems.

**Distributed systems architectures and fundamentals:** To enable the continued improvement, scalability, and broader deployment of heterogeneous distributed systems, research is encouraged in the following core areas of computer science and computer engineering for distributed systems:

- Resilient distributed computing fundamentals.
- Programming paradigms and languages for distributed and networked systems.
- Software-defined infrastructure and resource virtualization.
- Distributed decision and optimization support.
- Novel computing architectures to reduce the energy and time used to process and transport data, locally and remotely, for hyperspectral sensing, data fusion, decision making, and safe effector actuation in a distributed computing environment.
- Cooperative and coordinated distributed-system concepts that are scalable and function properly in communications-challenged and isolated environments where both wired and wireless environments are not guaranteed to be available, reliable, or safe, and that can intelligently synchronize when communications are restored or partially restored.

### **Research Recommendations**

**Develop energy-efficient heterogeneous distributed systems** with breakthrough levels of scalability, performance, and robustness. Develop highly efficient and secure communication mechanisms and technologies. Research on system management, monitoring, and optimization.

**Undertake fundamental research** on distributed computing, programming paradigms, infrastructure, resource virtualization, novel architectures, resilience, etc.

## Cognitive Computing

### Introduction/Overview

Cognitive computing refers to intelligent information processing systems that mimic the human brain. Such systems can self-learn at scale, interpret data proactively, perform reasoning and decision making with purpose, solve unfamiliar problems using acquired knowledge, and interact with humans in real time naturally.<sup>12,13</sup> They can generate not only answers to well-formulated numerical problems, but also uncover subtle patterns in largely unstructured data to provide “hypotheses, reasoned arguments, and recommendations”<sup>12</sup> for human decision makers.



The target is to create these autonomous intelligent machines that can operate in the real world by forming and extending models of the social environment they perceive, interacting with local human operators and with global intelligent networks to perform highly complex tasks.

Cognitive systems bear the potential for boosting economic competitiveness and contributing to social good. For instance, they could allow us to automate manufacturing, administrative, and financial processes at scales hitherto impossible. Advancing precision medicine, developing personalized educational programs, deploying government resources and social services much more efficiently, promptly detecting anomalies in context, effectively executing disaster prevention and environmental protection, and conducting defense and military operations more humanely and ethically are only some of many huge leaps that could be enabled by cognitive systems.<sup>13,14</sup> These applications rely on the capability of cognitive systems to absorb relevant information swiftly and extract meaningful insight from a deluge of loosely structured or unstructured data.

Enabling the ambitious goal of creating cognitive systems requires that the following key challenges be addressed: (1) developing systems with essential cognitive capabilities, including perceiving, learning, generating knowledge, reasoning, predicting, planning, and decision making; (2) developing efficient algorithms and architectures for both supervised and unsupervised learning; (3) developing seamless human-machine interfaces; (4) developing networks of cognitive sub-systems; and (5) integrating the new cognitive systems with the existing von Neumann computing systems. This entails a full-system approach, including information processing, programming paradigms, algorithms, architectures, circuits, device technologies, and materials development. On top of the technical challenges, security issues must be addressed in order to provide safe and trusted cognitive computing systems and networks.

---

<sup>12</sup> [https://www.research.ibm.com/software/IBMRResearch/multimedia/Computing\\_Cognition\\_WhitePaper.pdf](https://www.research.ibm.com/software/IBMRResearch/multimedia/Computing_Cognition_WhitePaper.pdf), John Kelly, “Computing, cognitive, and the future of knowing.”

<sup>13</sup> <https://www.whitehouse.gov/sites/whitehouse.gov/files/images/NSCI%20Strategic%20Plan.pdf>, “National Strategic Computing Initiative Strategic Plan” (2016).

<sup>14</sup> <http://www.research.ibm.com/cognitive-computing/ostp/rfi-response.shtml>, IBM, “Response to – Request for Information Preparing for the Future of Artificial Intelligence” (2016).



## Potential Research Topics

**Perceiving and learning:** For the cognitive systems to be versatile and capable of broadly serving our society, they must be (1) perceptive to the environment and the persons/objects that they interact with, (2) distill useful knowledge from the sensor data, and (3) learn from the human-machine interactions, from past experiences, and from other environmental input. To this goal, specific research needs may include:

- Pursuing breakthrough advances in multimodal sensing technologies such as developing perception algorithms to enable understanding of the environment from raw sensor data.
- Developing algorithms that:
  - allow learning from unstructured, unlabeled data, such as via human-machine interaction or other environmental feedback
  - enable context-aware learning
  - enable transfer of learned knowledge or concepts to new domains
  - allow online learning and real-time inferencing
- Exploring the fundamental limits of existing learning and deep learning algorithms,<sup>15,16,17</sup> developing theoretical foundations for deep neural networks, and deriving insights from neuroscience for unsupervised learning.
- Developing new architectures and algorithms that allow hitting a required test accuracy with a significantly reduced training set.
- Conducting research beyond the state-of-art deep learning models to (a) overcome the dependence on offline learning, (b) reduce the need for labeled training data, and (c) improve capability to capture higher-order structure in the data.
- Developing energy-efficient, low-cost analog techniques for cognitive workloads, neural networks, and other brain-inspired computing applications.
- Developing foundational theory for neuromorphic and bio-inspired computing based on novel devices including novel implementation of neurons/neural circuits, and programming paradigms. Particular interest lies in beyond existing state of art in deep learning networks, recognition beyond Deep Neural Network (DNN)/Convolutional Neural Network (CNN), sparse coded data methods.
- Developing reconfigurable networks suitable for Neural Network applications, e.g., Artificial Neural Network (ANN), Recurrent Neural Network (RNN), Cellular Neural Network (CNN), etc.
- Developing neural associative models, sparse coding, sparse distributed coding.
- Exploring computer vision: A detailed understanding of environments is required to navigate through them (e.g., in autonomous vehicles). Information about the environment could be collected by image sensors, processed and analyzed by a computer vision system, and output to actuators and robotic functions. Therefore, computer vision is sometimes considered to be part of artificial intelligence.

---

<sup>15</sup> Deep learning, Y. LeCun, Y. Bengio, G. Hinton, Nature **521**, 436-444 (2015).

<sup>16</sup> BlackOut: Speeding up Recurrent Neural Network Language Models With Very Large Vocabularies, Shihao Ji, Swaminathan Vishwanathan, Nadathur Satish, Michael Anderson, Pradeep Dubey - Intl. Conference on Learning Representation (2016).

<sup>17</sup> Dynamic Network Surgery for Efficient DNNs Yiwon Guo, Anbang Yao, Yurong Chen, NIPS (2016).



**Hardware acceleration of learning:** To deploy cognitive computing systems at scale requires fundamentally new computing hardware for the above-mentioned learning algorithms that are orders of magnitude more efficient in speed, size of training data, energy consumption, footprint, and cost than current technologies, without compromising performance. Research topics may include:

- Developing beyond-CMOS devices, materials, memory elements, circuits and architectures suitable for training and inference using artificial neural networks<sup>18</sup> and other brain-inspired algorithms to mimic the actions of a neuron by means of:
  - New devices, memory elements, and circuits that perform native operations frequently used in learning with fidelity, e.g., (matrix) multiplication, addition, and division
  - New topologies like spiking neural-nets and inference augmented neural-nets that significantly advance efficiency, accuracy, scalability, and latency
  - New topologies that allow variations in numbers of neurons depending on learning feedback and variable synaptic connections
  - New memory and data representations in beyond-CMOS hardware for much higher learning efficiency than direct digital representation
  - Reconfigurable networks for various classes of artificial neural networks
  - Implementation of neuromorphic primitives (synapse, neuron, spiking and continuous variable oscillators and their networks) in beyond-CMOS hardware, utilizing the unique stochastic and interconnect fabric enabled by beyond-CMOS technologies
- Developing energy-efficient, low-cost techniques for neural networks and other brain-inspired computing, such as by leveraging the intrinsic stochasticity of beyond-CMOS nanodevices and by efficient mapping of deep learning models to beyond-CMOS hardware.
- Developing hardware fabrics for machine learning.
- Connecting established neural and cortical mechanisms to functional beyond-CMOS hardware.

**Decision making:** When tackling highly challenging, complicated tasks, we rely on cognitive computing systems to analyze options and tradeoffs, evaluate risks, detect anomalies in context, and provide recommendations for decision making under uncertainties.<sup>14</sup> Research should develop algorithms to address these needs. In addition, there exist needs to:

- Develop architectures to accelerate tactical decision making—delegating decisions to machines in situations that require faster-than-human reaction and nontraditional chain-of-command responses.
- Compute on encrypted data for supporting safety of scalable decision making systems.

**Enabling trust:** To establish trust in the cognitive systems, researchers must develop reliable techniques to ensure that training and testing data remains unbiased, complete, and uncompromised. Also in crucial need are methods to verify that learning, reasoning, and decision making algorithms are objective, robust, deterministic, resilient, and accurate.

**Other important topics:**

---

<sup>18</sup> [http://science.energy.gov/~media/bes/pdf/reports/2016/NCFMtSA\\_rpt.pdf](http://science.energy.gov/~media/bes/pdf/reports/2016/NCFMtSA_rpt.pdf). “Neuromorphic computing: from materials to systems architecture – Report of a roundtable convened to consider neuromorphic computing basic research needs.” DOE, Office of Science (2016).

- Human-machine interface:
  - Develop seamless human-machine interface for autonomous systems, including high-accuracy sensor-feedback IoT systems
  - Develop innovative collaborations between manned and unmanned platforms
- Network of cognitive sub-systems:
  - Develop architectures for resilient self-optimizing and self-healing networks, memories, and compute elements to connect billions of devices in intelligent systems
  - Develop society-scale applications and data collection systems that can interact with local cognitive systems to optimize decision support
- Security and safety:
  - Investigate the nature of malicious and/or destructive cognitive systems
  - Understand the sources and risks of unintended harmful cognitive systems

### **Known Research Activities**

In the United States, several initiatives directly support cognitive computing related research:

**National Strategic Computing Initiative (NSCI)**, announced in 2015, lists the following task as one of its five strategic objectives: “Computing beyond Moore’s Law,”<sup>19</sup> with a time horizon of 10–20 years to “explore and accelerate new paths for future computing architectures and technologies, including digital computing and alternative computing paradigms” [such as neuromorphic computing]. It includes two parallel efforts: (1) beyond-CMOS for digital computing and (2) beyond-von Neumann for large-scale computing.

- DOE, IARPA, NIST, NSF, and DOD will “support the development of non-CMOS technologies and non-classical CMOS technologies, with early efforts by NSF, NIST, and IARPA.”<sup>20</sup>
- DOD, DOE, NSF, IARPA, and NIST support the development of “alternative computing paradigms or the underlying technologies.” IARPA will increase investment in alternative computing paradigms, including neuromorphic.
- NSF, in partnership with SRC, initiated the “Energy-Efficient Computing: from Devices to Architectures (E2CDA)” program as a first step to develop future computing systems, from devices to architectures.<sup>21</sup> Program kicked off in October, 2016.

**The Networking and Information Technology Research and Development (NITRD) Program:**<sup>22</sup> Among the 10 program component areas (PCAs) for 2017, Robotic and Intelligent Systems, Human Computer Interaction and Information Management, Large-Scale Data Management and Analysis, High-Capability Computing Systems Infrastructure and Applications, and Large Scale Networking appear most relevant.<sup>23</sup>

**IARPA’s Machine Intelligence from Cortical Networks (MICrONS) program:** “Achieve a quantum leap in machine learning by creating novel machine learning algorithms that use brain-inspired architectures and

<sup>19</sup> Intel maintains that Moore's law, which is about innovation and improving cost, will continue well beyond CMOS.

<sup>20</sup> [https://www.whitehouse.gov/sites/whitehouse.gov/files/images/NSCI%20Strategic%20Plan\\_20160721.pdf.pdf](https://www.whitehouse.gov/sites/whitehouse.gov/files/images/NSCI%20Strategic%20Plan_20160721.pdf.pdf)

<sup>21</sup> <https://www.src.org/program/nri/e2cda-nri/>.

<sup>22</sup> <https://www.nitrd.gov/>.

<sup>23</sup> <https://www.nitrd.gov/subcommittee/pca-definitions.aspx#LSDMA>.

mathematical abstractions of the representations, transformations, and learning rules employed by the brain.”<sup>24</sup>

**National Nanotechnology Initiative (NNI)** participates in supporting cognitive computing research through its Nanotechnology-Inspired Grand Challenge for Future Computing<sup>25</sup> initiated in late 2015.

- NSF, Expeditions in Computing program: “[P]ursue ambitious, fundamental research agendas that promise to define the future of computing and information.”<sup>26</sup>
- NSF, Robust Intelligence program: “The RI program encompasses all aspects of computational understanding and modeling of intelligence in complex, realistic contexts, advancing and integrating across the research traditions of artificial intelligence, computer vision, human language research, robotics, machine learning, computational neuroscience, cognitive science, several areas of computer graphics, and related areas.”<sup>27</sup>
- AFOSR, Computational Cognition and Machine Learning program: “This program supports innovative basic research on the fundamental principles and methodologies needed to enable intelligent machine behavior in support of autonomous and mixed-initiative (i.e., human-machine teaming) systems. The overall vision of this program is that future computational systems will achieve high levels of performance, adaptation, flexibility, self-repair, and other forms of intelligent behavior in the complex, uncertain, adversarial, and highly dynamic environments faced by the U.S. Air Force.”<sup>28</sup>
- ONR, Nanoscale Computing Devices and Systems program: “[S]tudy the electronic, optical and magnetic properties of, and potential device functionalities in, nanometer-scale materials and structures with a view toward building novel computing devices, circuits and architectures.”<sup>29</sup>
- DARPA Unconventional Processing of signals for Intelligent Data Exploitation program.<sup>30</sup>

In Europe, several funding programs target hardware implementations of non-von Neumann computing concepts:

**Human Brain Project**<sup>31</sup> (since 2013) is under the EC Future and Emerging Technologies Flagship; it tackles several computing aspects for brain research, cognitive neuroscience, and brain-inspired computing—among them the design, implementation, and operation of the project’s **Neuromorphic Computing Platform**.<sup>32</sup>

**Spiking Neural Network Architecture (SpiNNaker)**<sup>33</sup> is a many-core computer architecture to massively scale up hardware implementations of spiking neural networks as “a platform for high-performance

---

<sup>24</sup> <https://www.iarpa.gov/index.php/research-programs/microns/microns-baa>.

<sup>25</sup> <https://www.whitehouse.gov/blog/2015/10/15/nanotechnology-inspired-grand-challenge-future-computing>.

<sup>26</sup> [https://www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=503169](https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503169).

<sup>27</sup> [https://www.nsf.gov/cise/iis/ri\\_pgm12.jsp](https://www.nsf.gov/cise/iis/ri_pgm12.jsp).

<sup>28</sup> <https://community.apan.org/wg/afosr/w/researchareas/7679.computational-cognition-and-machine-intelligence/>.

<sup>29</sup> <http://www.onr.navy.mil/Science-Technology/Departments/Code-31/All-Programs/312-Electronics-Sensors/Nanoscale-Electronics.aspx>.

<sup>30</sup> <http://www.darpa.mil/program/unconventional-processing-of-signals-for-intelligent-data-exploitation>.

<sup>31</sup> <https://www.humanbrainproject.eu/>.

<sup>32</sup> <https://www.humanbrainproject.eu/ncp>.

<sup>33</sup> <http://apt.cs.manchester.ac.uk/projects/SpiNNaker/>

massively parallel processing appropriate for the simulation of large-scale neural networks in real-time” and “an aid in the investigation of new computer architectures.”

**Projects focused on adding memristors as synaptic elements to extend the implementations of neural networks with standard CMOS circuitry:** FP7 project "DIASPORA" (2013–2017); ERC project "NEURAMORPH" (2015–2020); French ARN project "MEMOS" (2014-2019); EU-H2020 projects "ULPEC" (2017–2019); and "NeuRAM3" (2016–2018). These projects fund research on novel materials to realize artificial synapses. Some projects target the co-integration of memristors with CMOS-based neurons to realize trainable neural networks.

**Projects focused on reservoir computing as an example of implementing neural networks in the hardware:** FP7 project "ORGANIC" (2009–2012) was focused on establishing neuro-dynamical architectures as an alternative to statistical methods for speech and handwriting recognition; FP7 project "PHOCUS" (2010–2012) and ERC project "NaResCo" (2010–2014) were focused on mapping algorithms and mathematical models into optical systems, resulting in the first demonstrations of photonic reservoir systems; the follow-up H2020 project "PHRESCO"<sup>34</sup> aims to develop scalable and CMOS-compatible implementation of reservoir systems.

### Research Recommendations

To extend cognitive computing capabilities beyond what today’s technologies offer, the following tasks are crucial:

**Mapping proven learning algorithms into proper, efficient new hardware.**

**Inventing learning/deep learning algorithms suitable for existing hardware, and demonstrating that new algorithms perform better than the existing ones.**

**Maintaining resiliency in long-term prediction in the face of noisy data and environments.**

**Pursuing deep understanding of the processes of reasoning and prediction by machine intelligent systems.**

## Bio-Influenced Computing and Storage

### Introduction/Overview

At the convergence of biology and semiconductor science and engineering is a new multidisciplinary field that has the potential to provide transformational advances in the design and manufacture of information processing systems. Advances will build upon breakthroughs in DNA synthesis and characterization, bidesign automation, nanoscale manufacturing, and understanding of biological processes for energy-efficient information processing. It is an opportune time to capitalize on this emerging field, sometimes



<sup>34</sup> [http://cordis.europa.eu/project/rcn/198823\\_en.html](http://cordis.europa.eu/project/rcn/198823_en.html)

referred to as semiconductor synthetic biology, to enable the next generation of information processing and storage.

Future ultralow-energy computing systems may be built on principles derived from organic systems that are at the intersection of chemistry, biology, and engineering. New information technologies are envisioned based on biological principles and using biomaterial for fabrication of devices and components that could yield storage capacity a thousand times greater than today's storage technology can provide. Such advances would enable compact high-performance computers that operate at a million times less power than today's computers.

### **Potential Research Topics**

**DNA-based massive information storage:** Device scaling and energy consumption during computation and storage has become a matter of strategic importance for modern information and communication technologies. For example, nucleic acid molecules have an information storage density that is several orders of magnitude higher than any other known storage technology. In theory, one kilogram of DNA has storage capacity of  $\sim 2 \times 10^{18}$  Mbit, which is equivalent to the total projected world's storage requirement from 2035 to 2040.<sup>35</sup> Recent progress in DNA synthesis and sequencing has made it possible to experimentally explore DNA storage beyond biological applications. Major breakthroughs occurred in the period 2012–2016 when several groups demonstrated DNA-based information storage compatible with mainstream digital formats.<sup>36,37,38,39</sup> If an integrated DNA memory technology can be developed, this class of memory systems may find widespread use, particularly for archival applications.

**Energy-efficient, small-scale, cell-inspired information and systems:** Understanding principles of cellular information processing could enable new generations of computing systems. Among the most promising characteristics of biological computing is the extremely low energy of operation, close to thermodynamic limits. What lessons can be transferred from biological information processing to future highly functional, space-limited, digital and analog semiconductor systems that combine high information density with extremely low energy consumption? Nature appears to have successfully addressed the submicroscopic design challenge and may suggest new solutions for future microsystems for information processing. Advances in the science of synthetic biology are beginning to suggest possible pathways for future semiconductor technologies. For example, there are indications that biochemical reactions perform information processing at energy efficiencies that are a few orders of magnitude beyond those projected for advanced semiconductor nanotechnologies of the future.

---

<sup>35</sup> V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church, W. L. Hughes, "Nucleic Acid Memory", *Nature Materials* 15 (2016)


<sup>36</sup> G. M. Church, Y. Gao, K. Yuan, S. Kosuri, "Next-generation digital information storage in DNA", *Science* 337 (2012) 1628

<sup>37</sup> N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA", *Nature* 494 (2013) 77

<sup>38</sup> E. Strickland, "Tech Companies Mull Storing Data in DNA", *IEEE Spectrum*, 20 Jun 2016  
<http://spectrum.ieee.org/biomedical/devices/tech-companies-mull-storing-data-in-dna>

<sup>39</sup> A. Exantane, "How DNA could store all the world's data", *Nature*, 31 Aug 2016  
<http://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496>





**Cell-semiconductor interfaces and hybrid semiconductor-biological systems:** Hybrid biological-semiconductor platforms can leverage both natural and synthetic biological processes and semiconductor technologies. In such hybrid platforms, living cells and tissues can function as a “biological front-end” layer with the cellular biochemical processes serving as an organic interface to the external environment and performing biological sensing, actuation, signal processing, synthesis, and energy harvesting. In parallel, the underlying semiconductor platforms can form a “semiconductor back-end” layer for information computation, control, communication, storage, and energy supply. Self-powered, on-chip intelligent sensor systems that integrate biological sensing functions and energy generation with inorganic information and computation capabilities enable diverse new applications. For example, advances could stimulate developments of self-powered intelligent sensor systems that integrate biological sensing functions and energy generation with inorganic computation capabilities, enabling diverse new applications. Example applications include fast, high-throughput chemical screening for drug discovery, diagnosis and therapy planning for personalized medicine, detecting chemical and biological agents for defense and environmental needs, and novel microscopic biological actuators or robots.

**Design automation for hybrid electronic-biological systems:** New methodologies and design principles will be needed for the complex electronic-biological systems. While ad hoc synthetic biology has demonstrated many impressive proof-of-concept circuits, full-scale computer-aided design tools will be needed for reliable design of larger and more complex systems such as whole-cell models. Leveraging advanced electronics design automation (EDA) tools and concepts for complex design could enable a radical increase in the complexity of biological design automation (BDA) capabilities. Currently demonstrated are  $\sim 10^4$  BDA equivalent “bits” (e.g., DNA base-pairs) versus  $\sim 10^9$  EDA “bits” (e.g., binary switches on a chip). The scope of this topic includes theoretical foundations, design methodology, and standards aiming at development of new engines for transformation and integration of synthesis artifacts, effective methods for programmer interaction and feedback that embrace multiscale processes, and automated program synthesis tools to create software that meets specifications for complex biological-electronic systems.

**Biological pathways for electronic nanofabrication and materials:** Biomolecules, such as DNA, RNA, or proteins, can provide a programmable mechanism for development of a wide variety of structures and shapes. In principle, cells fabricate amazingly complicated new structures with high yield and low energy utilization. Biological assembly occurs at an assembly rate of  $\sim 10^{18}$  molecules per second (at biological growth rates a 1 Gb chip could be built in about 5 s), and energy of  $\sim 10^{-17}$  J/molecule, which is a hundred times less than that in conventional subtractive manufacturing. Based on demonstrated DNA-controlled self-assembly of increasingly complex structures, such approaches have the potential for making complex sub-10 nm semiconductor structures. Also, engineered microorganisms can be used to produce a range of important chemicals and materials for semiconductor processes that have desired chemical composition and morphology. Methods need to be developed in which bacteria, viruses, etc., are used to self-assemble, pattern, organize, or repair organic polymers, inorganic materials, biopolymer materials, functional circuits, and/or electrical components.

## Known Research Activities

**ONR Metabolic Engineering program:** Target the fundamental understanding of metabolic processes in microbes or plants for the production of chemicals.

**ONR Synthetic Biology program:** Extend the natural capabilities of living organisms such as viruses, microbes, algae, and plants using synthetic biology.

**DARPA Living Foundries:** Develop the tools, technologies, methodologies, and infrastructure to increase the speed of the biological design-build-test-learn cycle while significantly decreasing the cost and expanding the complexity of systems that can be engineered.

**NSF Evolvable Living Computing project:** Develop the ability to “program biology” and democratize the process, similar to electronic computing.

**SRC Semiconductor Synthetic Biology (SemiSynBio) thrust:** Stimulate nontraditional thinking, concentrating on synergies between synthetic biology and semiconductor technology that could lead to novel, breakthrough solutions for the semiconductor and other industries.

## Research Recommendations

**Address the still many unknowns regarding DNA operations** in-cell and the potential of DNA technology for massive storage applications.

**Advance understanding of the principles of cellular information processing** in order to enable new generations of computing systems, whether they are based on semiconductor or biological materials, or a combination of both.

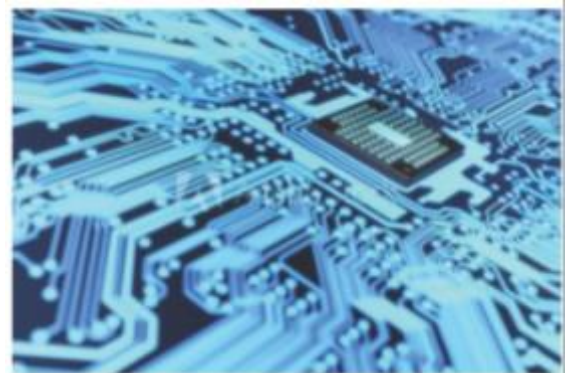
**Focus on the difficult technical challenge of electrical connections between the cell and the external world.** In principle, cells can be “programmed” for electrical connectivity using a synthetic biology approach.

**Tackle programming DNA to control the assembly of complex sub-10 nm heterogeneous structures—**another difficult technical challenge.

## **Advanced and Nontraditional Architectures and Algorithms**


### Introduction/Overview

Research is needed to lay the foundations for new paradigms in scalable, heterogeneous architectures, co-designed with algorithmic implications and vice versa. For example, graph processors utilize heterogeneously integrated accelerator-based architecture, a sparse matrix-based graph instruction set, and randomized communication to handle graph computation more efficiently than von Neumann architecture.<sup>40</sup> A major goal is to develop approaches and



---

<sup>40</sup> W.S. Song, et al, Novel Graph Processor Architecture, *Lincoln Lab Journal* **20**, 92 (2013).



frameworks to design and integrate a broad variety of application-specific computing architectures in concert with algorithmic and system software innovations. Energy-efficiency, resiliency, and security of the new architectures are paramount. Architectures and algorithms that are significantly lower in power consumption and capable of scaling to  $10^{12}$  devices/CPU and/or  $10^{12}$  nodes/network using CMOS and beyond-CMOS technologies are targets for research. Depending upon the application, a reexamination of the computation itself might lead to new ways to approach the design of the signal-to-information path.

Heterogeneous architectures and elements such as accelerators will increasingly be required to enable performance, power, and cost scaling. A major goal and requirement for new architectures is to address the design and integration of a broad variety of accelerators, both on-chip and off-chip, along with algorithmic and system software innovations. Broadly applicable architectures for compute in and near data are needed (movement of data is costly in terms of energy and latency), as are abilities to auto-configure and auto-tune system parameters.

Modern computing has operated on a nearly error-free device layer since the advent of Von Neumann computing. Computational architectures that rely on approximate computing (AC) and on stochastic computing (SC) can allow a significant scalability in error rates. Application of AC, SC, and Shannon-inspired information frameworks can provide significant benefits in energy, delay, and error rates or enable scalable architectures on erroneous hardware layers.

Currently, circuit and system architectures are developed independently from algorithms, leading to less than optimal performance and energy use. Research is needed on approaches for novel architecture and algorithm co-design to bridge this gap and thus improve the power-performance metrics of applications such as optimization, combinatorics, computational geometry, distributed systems, learning theory, online algorithms, and cryptography. Specific accelerators and heterogeneous integration of components dedicated to algorithms can play an important role. Innovative computing architecture techniques will need to address a shift from CMOS towards beyond-CMOS devices, which may operate at a hundred times lower energy and nearly 10 times lower speed. Theoretical computer science and modeling is expected to play an important role as well. The innovations that result from this foundational theme are expected to impact a broad variety of workloads.

### **Potential Research Topics**

In addition to the broad themes described above, possible research tasks of interest include but are not limited to:

#### **Co-designed hardware (HW) and algorithms for nontraditional computing:**

- Computational theory of computing using Beyond CMOS devices.
- Computational theory for neuromorphic computing.
- Computational complexity (time and memory) of non-von Neumann computing.
- Computational Complexity (time and memory) of neuromorphic computing.
- Computational Complexity (time and memory) of stochastic and random computing.
- Complexity of classifiers, complexity of associate processors.
- Scalability of Neuromorphic and Stochastic computation.

**Programming paradigms and languages for emerging technologies and architectures.**

**Fundamental understanding of the new architectures for scaling**

**Approximate/stochastic/Shannon-Inspired compute** suitable for traditional applications in extreme environments or near end of roadmap Si/CMOS limits.

**Computational systems with nontraditional thermal management, power dissipation and delivery, energy recovery, data collection or storage, and communications features offering overall greater efficiency and or performance:**

- Alternative models of data representation, storage, transportation, and endpoint interpretation.
- New memory and data representations in beyond-CMOS hardware that capture higher order structure and information content that can be more efficiently processed than a direct digital representation of data.
- Novel architectures with reduced data movement and communication needs, and novel algorithms to optimize data movement strategies adaptively as system loads and usage patterns vary.
- Novel computing and storage paradigms in which the hardware security is intrinsic to the architecture.
- Computing with devices near thermodynamic limits.
- Evaluation of system performance under extreme but managed nontraditional operating conditions.

**Hardware-software (HW-SW) co-design:**

- Design and integration of accelerators, from circuit level up to system and software levels for heterogeneous systems, based on both existing and emergent devices.
- Development of nonconventional technologies for accelerators.
- Development of software mechanisms to enable transparent usage of accelerators.

**Co-designed hardware and algorithms:**

- Energy-efficient circuits, architecture, algorithms, and software: Hardware-software-algorithm co-design methodologies and tools, including underlying physics and physics of failure; these are required for energy-efficient sensor systems, information extraction, and autonomous systems.
- Architectures that explicitly scale to  $10^{12}$  devices per processor (not including memory).
- Computational theory for encryption and security.
- Verification and validation.

**Heterogeneous systems:**

- Systems on chip with novel memory hierarchy.
- Nonvolatile computing (HW and SW).
- Advanced power management methods.
- In-memory computing.
- Reconfigurable computing.
- Computing on nontraditional fabrics, e.g., cross points.





### Modeling of the new architectures:

- Simulate the performance of the new heterogeneous architectures, and benchmark them against von Neumann machines.
- Model performance: Demonstrate needs for a broad enough class of workloads.
- Validate modeling results against hardware performance: Prove the relevance of models to real-world applications. (The deliverable of the theme is system hardware, not just a simulator or modeling results.)

### Integration and verification of nontraditional computing with Von Neumann systems to support realistic, complex workflows:

- The theoretical basis and support for nontraditional computing methods.
- The breadth of applicability of the proposed nontraditional computing methods.
- Their integration into relevant workloads and systems.
- The resiliency and reliability of the proposed nontraditional computing methods.

### Research Recommendations

#### **New paradigms of scalable, heterogeneous architectures co-designed with algorithms and system software.**

**Co-design of hardware** (including emerging technologies and architectures), **algorithms, and software** for nontraditional computing paradigms.

## Security and Privacy

### Introduction/Overview

Technology advances create business opportunities for the semiconductor industry where billions of connected *entities* act and interact autonomously<sup>41,42</sup> as a whole complex ecosystem. *Entities* may be computing systems, embedded systems, systems-of-systems,<sup>43</sup> cloud-based systems, human beings, and surrounding environments or a combination thereof, and they may be connected using a diverse set of communication technologies. Development of such an ecosystem, including the



---

<sup>41</sup> Marwedel. "Embedded System Design." 2nd ed., Springer 2011

<sup>42</sup> [Toward the definition of the Internet of Things](#), IEEE Internet Initiative, May 2015

<sup>43</sup> Maier. Research Challenges for Systems-of-Systems. SMC 2005: 3149-3154

“Internet of Things” and more, is expected to impact previously independent market segments, such as cloud computing, healthcare, and automotive markets, with an estimated economic impact of between \$3.9 and \$11 trillion dollars by 2025.<sup>44</sup>

Technology advances also create new vulnerabilities and opportunities for attackers. Adding complexity in the enabling technologies provides attackers with new tools, methods, and achievable results. As the IoT expands, an attacker can leverage more attack vectors and pursue new goals at a lower risk. The complex ecosystem of connected entities was labeled as “the largest and most poorly defended cyberattack surface conceived by mankind.”<sup>45</sup> Recent news highlights vulnerabilities in automobiles, power stations, implanted medical devices, and transportation systems.<sup>46,47,48,49</sup> Reported cybersecurity incidents indicate that existing practices in building systems are insufficient to encompass the inherent complexity of the ecosystem, and that such practices will be insufficient in the foreseeable future to properly address emerging security and privacy challenges. As more connected entities play an active role in the ecosystem, attackers will be able to target an even larger and more vulnerable attack surface—leading to loss of data, compromised privacy, hostile exploitation, and fraud at a massive scale,<sup>50,51</sup> with unprecedented economic and social impact. Even personal safety could be at risk.

To embrace the complexity of the envisioned ecosystem and to address its security and privacy challenges, government and industry stakeholders must facilitate the development of effective security and privacy mechanisms from their theoretical foundations to their effective implementations and lifecycle management. This calls for novel approaches to (a) threat modeling to embrace the complexity and the scale of the ecosystem; (b) security-aware execution models to counter both known and unexpected threats; (c) security-aware hardware and software development design/co-design tools to build security in our systems at their foundations; and (d) security and privacy foundations, i.e., cryptography and cryptographic implementations, to name a few areas of concern.

### **Potential Research Topics**

A big concern of system security and privacy remains the possibility that software execution can be subverted via either software or hardware attacks, such that arbitrary software execution can be triggered and private data (user location, biometrics, etc.), metadata (software versions, type of query, etc.), and key materials are accidentally or forcefully leaked to unintended recipients or abused by attackers. This concern is magnified because of the potential social and economic impacts of the diffusion

---

<sup>44</sup> Unlocking the potential of the Internet of Things, Report, McKinsey Global Institute, June 2015

<sup>45</sup> [Rebooting the IT Revolution: A Call to Action](#), August 2015

<sup>46</sup> [The hosting provider OVH continues to face massive DDoS attacks launched by a botnet composed at least of 150000 IoT devices](#). 2016


<sup>47</sup> ['Smart' home devices used as weapons in website attack](#). 2016

<sup>48</sup> [Safer Sky \(SPECIAL Topic: AVIATION 100\)](#). 2015

<sup>49</sup> [Why transportation networks are especially vulnerable to ransomware](#). CNBC November 2016

<sup>50</sup> [Mirai Botnet Linked to Dyn DNS DDoS Attacks](#), October 2016

<sup>51</sup> [An After-Action Analysis of the Mirai Botnet Attacks on Dyn](#), October 2016



of system misbehaviors (whether attacker-induced or unintentional) through the ecosystem—the impact of denial of Internet and other public services, frauds at a massive scale, etc.

The following are priority research areas to address security and privacy challenges in this coming world of “hyper-”connected, tightly coupled entities:

**Threat modeling:** The steady drumbeat of news stories make it clear that the coexistence of current security practices is insufficient to mitigate cyberattacks. One reason that current practices fall short is that the threat models on which the security of systems rely fail to account for the big picture. Moreover, many systems were not designed to operate securely in the current, highly networked operating contexts.

A critical requirement for security is that it cannot be isolated from other constraints that arise when an entity operates in the shared ecosystem, with aggressive connectivity, intelligence, and energy requirements. Significant trade-offs become immediately apparent. For example, system requirements for an intelligent system that collects, communicates, and analyzes a significant amount of information have conflicting implications for security and privacy needs. Unfortunately, current research and practices in security design do not address these interoperability requirements early enough in the design cycle. This often leads to joint requirements being comprehended and validated later in the design process, with a resulting patchwork in architecture and implementation, as well as in the potential for larger attack surfaces and increased system vulnerabilities.


New research must develop methods to approach threat modeling at a higher level—the ecosystem level. New research must develop methods to approach threat modeling at a higher level—the ecosystem level. There is the need to develop threat modeling techniques and tools—frameworks—capable of embracing the complexity, scales, and trade-offs across different application domains. This is especially valuable if patching or updating mechanisms are unavailable or insufficient to address a specific vulnerability. Furthermore, such frameworks must be extensible and open to updates, as new groups of threats emerge. In addition, a standardization effort must be undertaken in the appropriate standardization bodies.

**Security-aware execution models:** Research in microarchitectural side-channel and fault analysis has revealed weaknesses in modern computer system design (e.g., the design of microprocessor and memory hierarchy) and execution models. Current systems exhibit a larger attack surface than the semiconductor industry could have ever imagined. For example, it has been shown that physical faults can be induced remotely with consequences that conventional software mitigations and containerization cannot mitigate;<sup>52</sup> that software-only mitigations against induced physical faults are insufficient when the software implementation executes on a modern pipelined microprocessor;<sup>53</sup> and that shared resources on modern computer systems, e.g., memory hierarchy organization and mechanisms, expose behaviors that can be abused by attackers to leak valuable information even across virtual machines or containers, hence diminishing the value of the virtual compartments (e.g., sandboxes and virtual machines, and some

---

<sup>52</sup> Xiao et. al. One Bit Flips, One Cloud Flops: Cross-VM Row Hammer Attacks and Privilege Escalation. USENIX Security Symposium 2016.

<sup>53</sup> Yuce et. al. Software fault resistance is futile: effective single-glitch attacks. FDTC 2016.



properties of secure execution environments).<sup>54,55</sup> These results indicate that current practices in microarchitecture design, software development, and execution models must evolve to mitigate the emerging threats.

Moving-target mitigations aim to make each instance of a system unique under the attacker's eye, even though the overall system component functionality does not change. Moving-target mitigations in software can provide resilience to code subversion on conventional hardware, whereas moving-target mitigations in hardware could provide resilience against information leakage to nearly unmodified software. An appropriate combination of such mitigations can lead to the design, deployment, and operation of new systems capable of raising the security bar higher to greatly reduce the current attack surface, the efficacy of class-breaking attacks, and the spread of cyber-epidemic diseases. However, theoretical foundations in this research are missing and current implementations lack consensus and adoption. A companion to moving-target research is the development of distributed mechanisms to reveal implications of system failures to security and privacy.

**New hardware and software design frameworks and processes with security in mind:** Validation is a key component of security design and will remain so in the foreseeable future. Unfortunately, validation technologies and tools today are hardly security-specific. In industrial practice, current security verification is usually performed locally, with a lot of assumptions derived from the system-level security requirements. Even with these rudimentary security verifications completed at the local level, it is difficult to have assurance of security at the system level or of the communication of the system with the overall connectivity infrastructure. In part, this happens because architecture, design, and implementation, whether at the device or application level, do not comprehend broader validation needs.

New research needs to focus on developing techniques and tools to ensure that validation is performed from the ground up, and that the architecture and the hardware and software design is performed to facilitate validation. Specifically, there is a need for research to create new development environments that incorporate security specifications to generate, or guide generation of, alternative architectures for components of security subsystems and their communications as well as the corresponding communication interconnects and protocols.

**Research in cryptography and cryptographic implementations:** New mathematics and algorithms in the security and privacy space need to be developed to face future challenges to the pillars of secure systems. The impact that a large quantum computer would have on the current cryptographic landscape cannot be ignored. While it appears it would be sufficient to double the key length for symmetric cryptography, all public-key cryptography used in practice would be broken.<sup>56</sup> For our current infrastructure this means no more secure HTTPS, no encrypted emails, and no digital signatures. The good news is that there are alternatives to the public-key cryptographic schemes in use today, including

---

<sup>54</sup> Inci et. al. Cache Attacks Enable Bulk Key Recovery on the Cloud. CHES 2016.

<sup>55</sup> Xu et. al. Controlled-Channel Attacks: Deterministic Side Channels for Untrusted Operating Systems. IEEE Symposium on Security and Privacy 2015.

<sup>56</sup> [Report on Post-Quantum Cryptography](#), NISTIR 8105, February 2016

research on post-quantum cryptography.<sup>57</sup> However, their efficiency is suboptimal and their standardization and support (e.g., libraries) are largely lacking.

Research in side-channel analysis of cryptographic implementations—both attacks and mitigations—must evolve to match the stringent requirements of resource-constrained devices such as smart appliances. Side-channel attacks rely on the use of noninvasive approaches to observe physical phenomenon of the system and infer secret key material by using statistical analysis and some assumption on the implementation. Side-channel attacks are particularly problematic because once a side-channel leakage and its characteristics are correctly identified, they often are relatively easy to exploit and apply widely. For example, if the communication bus in an automobile is cryptographically secured but its signal levels can be sensed by a small recording device, the security of the safety features of the car is undermined. The same holds for a smart door lock that may leak its secrets and hence hand access to an adversary. Resilience against side-channel attacks must be taken into account during both hardware and software design. However, effective mitigations are generally too demanding of resources. Practical, less expensive, and better side-channel resilience is a major research topic.

Privacy-preserving computing becomes very important in any context where data are created, communicated, stored, and processed at different system hops owned by different entities. Examples include relays and consumers of healthcare data, storage and computation off-loaded to the cloud (e.g., machine learning), and object tracking, to name a few. While theoretical foundations in this space exist, e.g., homomorphic encryption<sup>58</sup> and Yao's garbled circuits,<sup>59,60</sup> the market largely lacks practical implementations and consensus on existing implementations in specific use cases. It is up to the semiconductor industry to innovate and help to set the proper standards as privacy regulations emerge and are mandated.<sup>61</sup>

### **Known Research Activities**

**NSF-SRC Secure, Trustworthy, Assured and Resilient Semiconductors and Systems (STARSS):** This is a collaborative government-industry effort that supports fundamental university research to improve hardware security. Information about STARSS projects can be found in the NSF online awards database. STARSS research includes microarchitectural side-channel attacks and mitigations; security-aware development, design and verification tools; supply chain assurance; security primitives and privacy-preserving computing; and anti-counterfeiting, cloning and aging.

**Security-aware execution models:** Existing moving target mitigations in the literature feature implementations at the algorithmic level (e.g., forms of data randomization),<sup>62</sup> or in software (at run-time and/or compilation time),<sup>63,64</sup> or in hardware (microarchitecture and circuit design mechanisms).<sup>65,66</sup>

---

<sup>57</sup> [Post-Quantum Cryptography](#).

<sup>58</sup> Gentry. [A Fully Homomorphic Encryption Scheme](#). PhD Thesis. 2009

<sup>59</sup> Yao. How to generate and exchange secrets. In FOCS, pages 162–167. IEEE, 1986

<sup>60</sup> Songhori et. al. TinyGarble: Highly Compressed and Scalable Sequential Garbled Circuits. IEEE Symposium on Security and Privacy 2015

<sup>61</sup> [Reform of EU Data Protection Rules](#).

<sup>62</sup> Gross et. al. Compact Masked Hardware Implementations with Arbitrary Protection Order. IACR Cryptology ePrint Archive 2016: 486 (2016)

<sup>63</sup> Lee. Rethinking Computers for Cybersecurity. IEEE Computer 48(4): 16-25 (2015)



- Existing research in microarchitecture physical security shows that current practices in microarchitecture and software design must move toward more security-aware and adaptive designs and execution models.<sup>63,67</sup>

**Research in cryptography and cryptographic implementations:** Current cryptography research and cryptographic implementation aim to diversify the set of public key cryptographic algorithms in protocol implementations. These include PQCRYPTO<sup>68</sup> and SAFECrypto,<sup>69</sup> which aim for the development of cryptography secure after the dawn of quantum computers.

- The project HEAT<sup>70</sup> is concerned with homomorphic encryption. The cryptographic building blocks developed in HEAT are, for example, relevant for privacy-preserving smart metering.
- The European Union funds a large-scale research theme on the topic areas of side-channel and fault analysis of cryptographic implementations.<sup>71</sup>
- Other interesting projects in the area of security and cryptography include the projects ECRYPT and ECRYPT II,<sup>72</sup> which provided independent recommendations for the lengths of cryptographic key material in different scenarios. ECRYPT-CSA<sup>73</sup> and ECRYPT-NET<sup>74</sup> are the successors of ECRYPT II and currently active.

### **Research Recommendations**

Addressing security is never simple, and addressing privacy raises additional challenges. The race between attackers and defenders across the ecosystem of "hyper"-connected, tightly coupled entities highlights that building-in security and privacy during technology development requires more than applying a series of isolated mitigations. The growing complexity and scale of the ecosystem and the myriad stakeholders calls for a broad approach that involves government, industry, and academia. In particular, government-sponsored programs and regulatory actions can make a huge difference, and the semiconductor industries can play a key enabling role by working together to develop the tools and technologies to ensure systems are trustworthy and secure.

Making integrated circuit and systems secure by design is more effective than finding and addressing vulnerabilities, whether inadvertent or intentional, after deployment. The goal is to develop strategies, techniques, and tools that avoid and mitigate vulnerabilities and lead to semiconductors and systems that are resistant and resilient to attack or tampering. Priority research areas include:

---

<sup>64</sup> Fuchs et. al. Disruptive prefetching: impact on side-channel attacks and cache designs. SYSTOR 2015

<sup>65</sup> Szefer and Lee, Hardware-Enhanced Security for Cloud Computing. Secure Cloud Computing 2014

<sup>66</sup> Leieron et. al. Gate-Level Masking Under a Path-Based Leakage Metric. CHES 2014

<sup>67</sup> Yuze et. al. FAME: Fault-attack Aware Microprocessor Extensions for Hardware Fault Detection and Software Fault Response. HASP@ISCA 2016: 8:1-8:8

<sup>68</sup> <http://pqcrypto.org/>

<sup>69</sup> SAFECrypto.

<sup>70</sup> HEAT.

<sup>71</sup> Horizon 2020.

<sup>72</sup> ECRYPTII.

<sup>73</sup> ECRYPT-CSA.

<sup>74</sup> ECRYPT-NET.

**Architecture and design:** Approaches to architecture and design that are studied at the system level; the impact of security at the level of circuits and processors must be understood in terms of system-wide functionality, performance, and power goals.

**Principles, properties, and metrics:** Hardware security design principles, semiconductor-specific properties, and security metrics for evaluating or comparing designs that are extensible and potentially useful for privacy composition or for providing trust evidence at the system level.

**Verification:** Tools, techniques, and methodologies for verifying hardware-specific security properties and enforcing security design principles; innovative approaches to establish safety properties without knowing all aspects of the design, and thereby providing strong provable assurance; and approaches to increase automation of security verification and analysis.

**Embedded software and firmware:** Strategies and techniques to reduce vulnerabilities in embedded software and firmware, and for providing updates to address known vulnerabilities discovered after deployment in the field.

**Authentication and attestation:** Models for the insertion of artifacts or design elements that are verifiable during design and throughout the life cycle; and supporting issues such as generation, protection, and establishment of trust models for hardware-implemented keys.


## Design Tools, Methodologies, and Test

### Introduction/Overview

Current design tools and methodologies have evolved over the last fifty years to enable creation of systems based on Von Neumann architectures. Delivering solutions for major challenges, such as those derived from the collection of nonstructured data, which is growing at a much faster pace than structured data, requires different types of architectures for extracting meaning and effectively processing data. Alternative computing systems, such as highly distributed and heterogeneous systems, systems based on non-Von Neumann architectures (for example, analog computing, stochastic processing, approximate computing, and bio/brain-inspired models such as neuromorphic computing), will require major innovation in the design flows and tools that will enable the creation of such systems.



The design community takes advantage of the most advanced computing technology to deliver timely solutions. As an example, when processor frequency stopped increasing and architectures moved from single processor to many-core architectures, electronic design automation (EDA) algorithms were redesigned to scale properly on the new systems. Similarly, design flows and methodologies will need to change to take advantage of future computing hardware and systems.



Because many of the alternatives that must be explored will include the discovery of new materials and processes, there are needs not only for the development of computational capabilities—ranging from *ab initio* calculations to physics-based effective models—that provide predictive treatment of material discovery and manufacturing process simulation capabilities, but also for EDA tools that can deliver complete systems for the novel computing paradigms they support, including logical and physical design, simulation, and verification.

Due to the high complexity and high manufacturing cost of all leading-edge semiconductor processes, the entire industry relies on fundamental models and design methodologies that accurately describe and prescribe each step, from the system-level design of complete electronics systems all the way to the materials and processes used during semiconductor manufacturing. These models and methodologies are essential enabling components of semiconductor production. For example, they are used to validate designs before manufacturing, to verify specification compliance during each stage of the design and manufacturing flow, to improve the speed of failure detection, and to continuously improve the robustness of complete systems.

Modeling is key in benchmarking, characterization, fabrication, data analysis, parameter extrapolation, and process simulation and control. To obtain accurate material and device models, both clear understanding of physical mechanisms and verification based on experimental feedback are essential.

In addition, to extend Moore's Law to its ultimate limit, radically different materials and processes are being explored, which in turn have implications for design tools and methodologies. Therefore, fundamental research is needed on design tools and methodologies at all levels to enable ultimately scaled CMOS. In order to manufacture complex integrated circuits and integrated systems, efficient test and validation techniques also will be needed to reach end markets.

### **Potential Research Topics**

To deliver today's tremendously complex advanced integrated circuits and systems, the semiconductor design process relies on well-established paradigms that are fundamental to the design, verification, and test activities. Concepts such as separation of design flows from verification flows, hierarchical design (i.e., well-defined devices, standard cells, IP blocks, chip assembly, board-level, system-level), simulation (at multiple abstraction levels), and design for test, are pervasive throughout design tools and methodologies, and have enabled relatively small engineering teams to create components with billions of transistors. The growth in system complexity is expected to continue at an exponential rate, allowing future systems to meet the increasing demands for functionality and to enable entirely new applications. Furthermore, circuits and systems are becoming more adaptable, which will require significant adaptations of verification, validation, and test methodologies. These questions arise:

- Which of these paradigms will remain unchanged?
- Which ones will remain but will need to adapt? And how?
- What new paradigms will need to be created?
- Which ones will need to be abandoned?



These fundamental questions may have different answers for the multiple alternatives of the post-von Neumann era.

To illustrate how the new computing systems will impact the current state of the art in design tools and methodologies, it is beneficial to explore brain-inspired architectures in general, and specifically, one of its promising alternatives, neuromorphic computing.

## Neuromorphic computing

Neuromorphic systems emulate how neurons operate at the individual unit up to the complete system level, how they connect to each other, and how their responses change over time. A list of some of the most salient characteristics includes:

- Large number of computational units that can compute, store, and communicate.
- Large number of inputs (tens of thousands) to a single output.
- Single output connected to a large number of computational units.
- Reconfigurable interconnections.
- Asynchronous operation.

Each of the characteristics listed above has a profound effect on design automation; taken as whole, they indicate the need for a complete rethinking of current approaches.

For example, the vast majority of current computing systems are based on synchronic timing operation. The existence of a clock to synchronize all signals in a circuit is a key technology that has been entrenched in the design flow since the earliest stages of design creation activity. This concept is deeply embedded in multiple design flow concepts (e.g., timing verification and closure) and the design tools used in those flows (e.g., static timing analysis). How the timing verification aspects will need to change for these new circuits that may contain billions of components is an open question at this point.

Looking at the human brain is a long-term objective for novel computing, but there are multiple intermediate steps that could be taken to deliver cognitive computing functions. It would be beneficial to establish intermediate cognitive functionalities of increasing complexity that inspire the transition from current technology, design tools, and methodologies to those required for the long-term goal.

At a high level, it is possible to identify some major categories that will require coordination of multiple research activities:

### Device-level to system-level design:

- Comprehensive design, simulation, and planning techniques, as well as overall approaches for realizing new microsystem functions and capabilities.
- Design tools and methodologies as well as rapid, accurate simulation approaches for multiple heterogeneous systems that consider performance, power, thermal, and reliability; validation and verification of such systems.
- High-level synthesis-based design methodology to improve performance and productivity of the overall system, particularly for cognitive and autonomous computing.
- Design approaches for specialized hardware components of future autonomous systems, including machine intelligence.
- Deeper understanding of cognitive computing techniques, such as deep neural networks, and their relationships to hardware specification.
- Novel methodologies for design testability to accelerate diagnostics and debugging.
- Novel cognitive computing approaches and design-relevant hardware.
- Definition of key metrics to benchmark and drive efforts in the design tools and methodologies space against established technologies and existing design practices.



## Predictive models:

- Comprehensive modeling, design, and simulation methods for novel devices to increase performance, decrease power usage, and improve yield and manufacturability.
- Comprehensive (physics-based) modeling approaches that account for the complexity of nanoelectronic and atomic-level systems.
- First-principles computational techniques, such as density functional theory (DFT), to screen potential realizations of novel materials and combine them with microscopic physics-based models.
- Validation of the accuracy of electron transport methods for treatment of transition metal oxides and dichalcogenides in finite nanoscale and atomic-scale devices.
- Atomistic modeling of materials to probe ferroelectric, magnetic, and strain order parameters; spin, charge, and ionic transport via atomistic interfaces; atomistic design and modeling of multilayered, heterostructure materials; and microscopic modeling of interfacial spin torques and magnetic anisotropy.
- Multiscale modeling of novel material properties for novel device designs, including charge, spin, ionic transport, magnetic susceptibility, damping, exchange, dielectric constant, carrier mobility, band gaps, crystalline or amorphous structure, etc.
- Modeling of magnetic materials, including domain nucleation, pinning, domain wall patterns, and dynamics; modeling of ferroelectric and multiferroic materials, including domain nucleation, pinning, reversal, and domain wall motion; and models to span multidomain, single-domain, and bulk vs. surface area-dominated regimes.
- Microscopic simulation of the switching kinetics of multiferroic materials, including distribution of order parameters and thermal fluctuations.
- Multiscale modeling of multiphase boundaries and other discontinuous fronts and filaments; and modeling of resistive random-access memory (RRAM) materials, including conductive-bridging RAM (CBRAM) (e.g., oxides, chalcogenides, etc.), including vacancy and ionic migration, formation of chains, and kinetically driven structures.
- Compact models of novel devices that comprehend magnetization, carrier spin, charge, strain, and polarization dynamics.
- First-principles modeling of defect and trap states in novel materials.
- Reliability models of failures, e.g., imprint, fatigue, and breakdown in ferroelectric devices.
- Experimental calibration and validation of models to allow benchmarking of scaled devices.
- Models that account for roughness, shape, and size effect, along with surface or interface, phonon, and grain boundary scattering in realistically long interconnects composed of novel materials and composites.
- *Ab initio* estimates of surface-interface chemistry and kinetics during fabrication processes.
- Simulations of atomic layer selective deposition and atomic-level etching for novel materials and multilayered material stacks in gas/plasma/solution-based chemistries.
- Based on Monte Carlo methods, feature profile evolution of high-aspect-ratio structures that captures structural, mechanical, chemical, electrostatic, fluid dynamics, and physical factors during wet and dry processing.
- Design tools and methodologies on new compute platforms.
- Algorithms that take advantage of massive compute clouds.

- Algorithms that take advantage of heterogeneous architectures such as a combination of large numbers of Graphic Processing Units (GPUs), Field Programmable Gated Arrays (FPGAs) or other Application Specific Integrated Circuit (ASIC) accelerators, as key elements to drive performance.
- Algorithms that run effectively on cognitive and “brain-inspired” computers.
- Algorithms that take advantage of new memory-processor architectures.
- Application of machine learning and other nontraditional techniques to electronic design automation tools and methodologies.

### **Research Recommendations**

**Develop computational capabilities and design tools** for novel materials, processes, and computing paradigms beyond von Neumann.

**Develop predictive models** for novel devices, physical mechanisms, and material properties that are validated with experimental results.

## **Next-Generation Manufacturing Paradigm**


### **Introduction/Overview**

Advanced manufacturing and processing is needed to enable the fabrication of emerging devices and systems. Novel materials and devices often require novel fabrication technologies, e.g., atomic-scale precision patterning, deposition, and etching. Additive processes, such as directed self-assembly (DSA) and atomic-scale placement, may enable high-precision material and device engineering. Advanced integration technologies (e.g., 3D) may enable functional diversification.



With the number of processing steps approaching one thousand and the number of critical lithographic layers over fifty, manufacturing control and costs are a growing challenge for the semiconductor industry. As feature sizes approach molecular dimensions, optical resolution limits are no longer the only roadblock. Patterning pitfalls can be mitigated by new patterning material paradigms, precise material placement, and planarization and etching techniques (e.g., atomic level selective deposition and atomic resolution etching). Other avenues for improvement include all mainstream materials, intelligent functionalizing or templating materials and bio-inspired manufacturing research. Manufacturing of high-voltage, power-management chips using wide-bandgap materials such as GaN and SiC present unique technology and cost challenges.

Control of processing, interfaces, and defects is also crucial and will require advanced metrology tools capable of nanometer resolution. Stringent form factors for novel cyber-physical systems will require novel packaging techniques. Manufacturing research in new materials, processes, and tools with control at the



molecular and atomic levels is critical to provide a cost-effective path to producing reliable future device and interconnect architectures. Use of novel material and processes may require novel testing techniques.

Additionally, innovative metrology and characterization techniques are required to measure unique properties of novel materials and devices. Uniform test platforms, standards, and methodologies are essential for the benchmarking of materials, devices, and systems, providing important research guidelines. *In situ* metrology is important for process monitoring and high-precision nanofabrication.

### **Potential Research Topics**

#### **Nanometer (nm)-size feature patterning:**

- Lithographic innovations in photoresist materials to reduce molecular-scale side-wall roughness with the minimum exposure dose possible, novel materials for thinner hard mask, use of DSA materials with atomic-level precision, and 3D scaffolding for selective layer self-assembly.
- Innovation in etching methods (e.g., atomic layer etching) to improve uniformity, profile control, selectivity, and manufacturability.
- Techniques with atomic level precision and high throughput, and innovation in material and process selectivity to improve post-chemical-mechanical polishing (CMP) variation.
- Low defectivity growth of heterostructures on silicon; integration of the most suitable materials for monolithic 3D integration; and processes for 3D integration of logic, memory, analog, and RF devices and components to reduce cost, improve performance, and increase functional density.

#### **Deposition and interfaces:**

- Materials and techniques for film deposition, including strain mediated growth, with atomic level precision, uniformity, high selectivity, using high throughput.
- Single crystalline material deposition, especially for correlated electron oxides.
- Novel self-aligned materials and processes (e.g., cuts, vias), including selective dielectric and metal deposition, deep-via etch, and electroless/CVD metal-fill.
- Self-assembly techniques to improve density: novel materials, chemistry, processes for the directed self-assembly of nanoscale objects from solution with atomic level precision, including biological templating, chemical recognition, and co-polymer systems.
- Lower thermal budget processes, and expanded material availability for work function selection.
- Improve fundamental understanding of surface reconstruction and interface control (including cleaning and passivation) for gate stack, contacts, BEOL vias, and low  $D_{it}$  channel materials.
- Precision placement of dopants and their activation:
  - Sub-nanometer scale electrical, magnetic, and optical metrology with 3D capability and high throughput with real-time feedback and feed-forward on critical processes
  - Manufacturing processing for large-scale production of soft materials for flexible electronics
  - Manufacturing processing for displays
  - Manufacturing processes for integrated energy harvesting



- Integration of on-chip optoelectronic devices
- Novel design for early detection of chip anomaly caused by issues related to reliability, defectivity, electrical functionality, and local layout effects

## Research Recommendations

**Conduct research on atomic-scale precision material deposition, placement, patterning, and etching techniques with high throughput, high yield, and low defectivity.**

**Develop 3D integration and packaging technologies, advanced metrology for manufacturing, and modeling support.**

## **Environmental Health and Safety: Materials and Processes**

### Introduction/Overview

The semiconductor industry has a sustained record of proactive environmental, health, and safety (EHS) accomplishments. These accomplishments include a coordinated worldwide reduction of fab greenhouse gas emissions, elimination of chemicals of concern such as perfluorooctanol sulfonate (PFOS), and proactive development of industry EHS standards for development of manufacturing tools to protect employees and the environment. As the industry continues to adopt new manufacturing processes and materials to meet ever-increasing performance demands, the industry's reputation, freedom to innovate, and profitability are dependent on a proactive EHS R&D program to address current and future challenges.



Advances in semiconductor technology increasingly are dependent on innovations in novel materials and processes, which can only be used commercially with appropriate EHS information and controls. The goal is always to select the lowest-risk material that is effective for an application, optimize processes for minimum waste, and implement occupational exposure and environmental controls that are protective of human health and the environment. However, the materials and processes employed by the semiconductor industry are often novel or used in a unique manner that requires research and development of fundamental EHS information and control technology. Moreover, because of the rapid pace of innovation, the semiconductor industry sometimes is ahead of the “regulatory envelope”, and it becomes incumbent on the industry not only to develop its own information, but also to provide the information to government agencies on which future regulations will be based. Much of the fundamental EHS R&D needed by the semiconductor industry has been conducted by university researchers supported by the SRC in collaboration with NSF.

### Potential Research Topics

**Assessing what EHS information and technology is needed:** Determining what information and technology is needed to enable a new process or material use is the initial step in an EHS evaluation. Toward that goal, the check list below presents a simple set of declaratives that can be used to evaluate



the state of EHS knowledge regarding any new material or process. To the extent that each of these can be answered with affirmative answers, there is a positive outlook for using the material or process in a manner that is protective of human health and the environment. To the extent that these questions cannot be answered affirmatively, additional information, and possibly R&D, is needed.

**Check list for EHS evaluations of new materials and processes:**

- Chemical handling, usage, and safety information:
  - Knowledge of hazardous properties and risks
  - Well-defined safe handling practices
  - Well-understood and optimized manufacturing processes
  - Use of the least hazardous, most benign material that is effective for the application
- Occupational exposures and controls:
  - Knowledge of exposure routes
  - Defined measurement and detection methods for relevant exposure routes
  - Protective exposure threshold values
  - Appropriate personal protective equipment (PPE) for relevant exposure routes
  - Appropriate engineering controls for relevant exposure routes
- Environmental discharges and controls:
  - Characterized air, water, and waste effluent streams
  - Metrology for the substances in relevant waste and emission streams
  - Protective discharge and emissions threshold concentration levels
  - Cost-effective effluent and discharge controls
  - Known fate and behavior in the environment


**Availability of information:** The availability of reliable information regarding the toxicity and behavior of a chemical is often a limiting factor in an EHS evaluation. Knowledge of the respiratory toxicity of a chemical, for instance, is a prerequisite for determining whether an appropriate level of occupational exposure controls has been implemented. Likewise, knowledge of the aquatic toxicity of a chemical is a prerequisite to determining the appropriate level of wastewater treatment. However, there is a substantial gap between the number of chemicals in use and those for which there is a full complement of EHS information available. Of the approximate 90,000 chemicals in the European Union's REACH database, a recent study reported that a full suite of measured EHS data existed for less than 0.1 % of the chemicals, and aquatic toxicity information was available for less than 3 % of the chemicals.<sup>75</sup>

The current paradigm for determining the toxicity of a chemical to humans is primarily based on the results of animal testing. However, animal testing requires on the order of 2 to 3 years to complete and may cost several million dollars. Given the tens of thousands of chemicals for which information is needed and the rapid pace of materials innovation, it is both logistically and financially impractical to determine the toxicity of all chemicals solely using conventional animal testing methods. Further, public pressure for reduced use of animal testing is an important driver for the development of alternatives.

A number of efforts, including the EPA's ToxCast program, are working toward the development of rapid, automated toxicity assays using simpler organisms, which can be used as an initial screen to identify

---

<sup>75</sup> Stempel, S., M. Scheringer, C.A. Ng, and K. Hungerbühler. "Screening for PBT Chemicals among the 'Existing' and 'New' Chemicals of the EU." *Environmental Science & Technology* 46 (2012) 5680-5687.



chemicals that may need to undergo more rigorous animal testing. Likewise there are many ongoing efforts to develop computational algorithms to predict toxicological effects *in silico*, based on a molecule's structure and/or chemical properties. The progress of these efforts has been slow and is gated by the limited availability of substantive empirical data sets for the types of chemicals for which predictive capability is needed.


Similarly, various software methods (*in silico*) are available as means of predictively estimating the bioaccumulative and other properties of chemicals. However, current software tools are limited with regard to the range of chemical types for which they can reliably produce estimates, as well with regard to the particular properties that they can estimate. For instance, many such tools use the propensity of a chemical to partition to fats as a paradigm for whether they will be bioaccumulative. However, many perfluorinated compounds, such as PFOS and PFOA (perfluorooctanoic acid), which are currently an escalating concern, appear to bioaccumulate in protein-rich areas of the body, not in the fat-rich areas as assumed under the conventional paradigm.

The difficulty in obtaining relevant EHS information for novel materials is exemplified by engineered nanoparticles. After more than 10 years of concerted effort, the international research community and regulatory agencies worldwide continue to struggle with fundamental questions regarding how to measure workplace exposures, evaluate toxicity, and determine what respiratory exposure threshold limit values to apply to this complex class of materials.

In summary, the rate of new chemical development outpaces the rate at which chemical behavior and toxicity information can be determined by empirical methods. Advances are being made in the use of high-throughput screening techniques and computational tools, but the progress of these developments has been slow relative to this growing need. Whereas the prioritization of publicly funded efforts tends to be towards high-production-volume chemicals, the semiconductor industry often uses specialized classes of chemicals, many of which do not sit high on priority lists due to their limited use, often in highly specialized processes. There is a need to steer the development of new chemical behavior prediction tools and toxicity assay methods toward novel chemicals with strong potential application and use in semiconductor manufacture.

**Process optimization and selection of benign materials:** The use of optimized manufacturing processes and more benign materials are important goals for our industry. However, the ability to optimize the operations of semiconductor manufacturing tools to minimize chemical and resource use is challenging when the capability of conventional manufacturing tool sets has been stretched in the effort to produce devices at increasingly more advanced technology nodes, with feature dimensions nearing the atomic level. In addition, there is an ever increasing demand for higher purity materials in contact with the wafer. In some CVD deposition processes, for instance, less than 0.5 % of the feed material is effectively delivered to the wafer surface, while more than 99.5 % is carried out of the process chamber and deposited in downstream appurtenances and/or removed in exhaust abatement systems. Similarly, in some wet clean processes, less than 10% of the hydrogen peroxide is utilized at the wafer surface, while the remainder, which is often mixed with high concentrations of sulfuric acid or ammonium hydroxide, is directed to waste.

There is a need for R&D that leads to improved process understanding, enabling the design of new manufacturing tool sets that are engineered from inception to minimize resource use and waste generation, and to maximize the ability to reclaim, recycle, and reuse valuable resources. Likewise, R&D



needs to be directed toward the development of more advanced sensors, controls, metering, and analytics that enable wafer surface level information regarding the status of the process, and which enable minimization of chemicals and resource use.

The design of more benign and lower risk processing materials is also an important goal that can be facilitated by improved fundamental process understanding, in combination with improved metrology and predictive simulation tools that can guide the design of molecules to achieve a desired function and to have minimal toxicity and environmental impacts.

**Water and waste treatment and recycling technology:** Wastewater treatment and recycling is difficult to achieve when the feed streams contain chemicals that are difficult to measure, exert toxicity and/or inhibitory effects at very low concentration levels, or are not readily removed by existing wastewater treatment technology. Research is needed to develop advanced oxidation process (AOP), biotreatment, and other technologies that can render organic chemicals to harmless forms like carbon dioxide and water. Currently available AOP technologies are challenged by a tendency to form complex byproducts, which may be difficult to measure and may be as toxic as the initial target chemical.

The challenge presented by refractory organic chemicals can also be a limiting factor in water recycling efforts, where a typical reverse osmosis reject stream contains elevated concentrations of chemicals that are difficult to remove or destroy.

Similarly, removal of metals from wastewater has limitations. Most current wastewater treatment technologies create concentrated metal oxides or other bulky sludges that must be disposed. A better solution would be wastewater treatment methods that cost-effectively recover the metals into pure forms suitable for reuse.

**Nanoparticle occupational exposure and control:** Advanced semiconductors are designed and manufactured at the nanoscale, and therefore manufacturing processes increasingly use engineered nanomaterials. Presently, the only engineered nanoparticles that are used in substantial quantities by the semiconductor industry are the alumina, ceria, and amorphous silica particles used in chemo-mechanical polishing (CMP). However, there are a great number of potential nanoparticle applications including for extreme ultraviolet (EUV) resists, in self-assembling materials, and in carbon nanotube (CNT)-based transistors.

Knowledge regarding the EHS characteristics of nanoparticles has been slow to develop. At present, occupational exposure limits have only been published for two nanomaterials, carbon nanotubes and titanium dioxide (TiO<sub>2</sub>) nanoparticles. Uncertainty remains regarding how best to quantify toxicity and develop occupational exposure limits for the wide variety of nanoscale materials. Techniques have been developed to characterize nanoparticles in workplace environments, but they require the use of an overlapping suite of complementary sampling and analytical methods that are not easily adaptable to personal exposure monitoring. Until the toxicity of nanoparticles can be determined, occupational exposure limits established, and occupational exposures measured, it is not possible to confirm that the appropriate exposure controls are in place and effective. A significant effort is being made towards addressing these challenges by NIOSH and its partners. The semiconductor industry and its suppliers must continue to take steps to support these efforts so that the nanoparticles used by the semiconductor industry are used properly.

## Known Research Activities

**Engineering Research Center for Environmentally Benign Semiconductor Manufacture:**<sup>76</sup> Led by the University of Arizona, this center’s objectives are to (1) improve performance, reduce cost, and lower EHS impact; (2) incorporate EHS principles in engineering education; and (3) promote “design for environment and sustainability” as a driver, not a burden.

**Center for Environmental Implications of Nanotechnology:**<sup>77</sup> led by the University of California at Los Angeles, this center works to ensure the responsible use and safe implementation of nanotechnology in the environment.

## Research Recommendations

**High-throughput screening methods and computational models** for determining and predicting toxicity and behavior—in the body and in the environment—of new materials under consideration for use in semiconductors and their manufacture.

**Research on less hazardous or nonhazardous substitutes for chemicals and materials of concern.**

**Optimized plasma etch and deposition chamber clean processes and feed-gas substitutes** that use lower global warming potential (GWP) gases and/or reduce emissions.

**Manufacturing process and tool research** to reduce material and energy use and waste.

**Novel waste stream processing** to reduce and eventually eliminate contaminants and to convert them into useful by-products.

## **Innovative Metrology and Characterization**

### Introduction/Overview

The metrology and characterization needs of the semiconductor industry span a wide range of topics, from fundamental materials properties, to patterning, mask metrology, and process fidelity. Future needs are likely to be driven by two dominant themes in semiconductor manufacturing: complex (nonplanar) device architectures that enable continued scaling, and new device concepts and materials for “more-than-Moore” post-CMOS electronics and memory. Generally speaking, the push is for better 3D spatial resolution, more comprehensive characterization of physical properties (crystal structure, composition, magnetization, and polarization), and higher throughput. Finally, as wafer sizes



<sup>76</sup> <https://www.erc.arizona.edu/>

<sup>77</sup> <http://www.cein.ucla.edu/new/>

increase, the cost of destructive testing grows and nondestructive metrology methods are even more desirable.

### **Potential Research Topics**

**Advanced materials property characterization:** In conventional semiconductor manufacturing, the introduction of new materials has enabled continued device scaling.<sup>78</sup> This trend is likely to continue, especially for memory elements. Novel memory concepts based on phase change, magnetic tunneling, and bi-stable resistive (memristive) elements are rapidly reaching maturity. Some of these new approaches are not charge-based and instead exploit physical phenomena such as crystalline state, magnetization, and oxygen vacancy structure (conducting filaments). The same is true for emerging post-CMOS logic devices, which make use of unconventional materials including carbon nanotubes, semiconductor nanowires, 2D materials, and magnetic materials.<sup>79</sup> The incorporation of these exotic new materials in manufacturing demands better chemical and structural characterization.<sup>80</sup> Some specific needs include:

- Sub-nanometer film thickness measurement, including that of buried layers, in layered magnetic devices.
- High-resolution, high-throughput scanning capacitance and spreading resistance measurements for characterizing dopants.
- Measurement of magnetic properties, e.g., magnetic exchange stiffness, in ultrathin (~ 1 nm thick) magnetic layers.
- Larger-area characterization of surface chemistry for self-assembly and patterning.

**Nanoscale structure and composition:** As conventional devices evolve towards more complex geometries (e.g., III-V FinFETs and TriGates), quantitative structural and compositional characterization becomes even more important.<sup>81</sup> In addition, novel memories and post-CMOS logic incorporate new materials and exploit different physical properties, including magnetization, polarization, and crystallinity.<sup>82</sup> Many of the new memory concepts involve structural phase transformations, or the evolution of polarization or magnetization domains. These dynamic processes also need to be characterized. Specific examples include:

- Nanoscale characterization of composition and structure.
- Nanoscale characterization of electrical properties, including polarization, charge density, conductivity, and carrier mobility.

---

<sup>78</sup> "Metrology and Characterization Challenges for Emerging Research Materials and Devices," C.M. Garner, D.J.C. Herr, Y. Obeng, 2011 International Conference on Frontiers of Characterization and Metrology for Nanoelectronics, MINATEC Campus, Grenoble, France, May 24, 2011.

<sup>79</sup> Ibid.

<sup>80</sup> "Metrology, Analysis and Characterization in Micro- and Nanotechnologies - A European Challenge," L. Pfitznera *et al.*, ECS Transactions, 10 (2007) 35.

<sup>81</sup> "Microscopy needs for next generation devices characterization in the semiconductor industry," L. Clement *et al.*, Journal of Physics: Conference Series 326 (2011) 012008.

<sup>82</sup> "Metrology and Characterization Challenges for Emerging Research Materials and Devices," C.M. Garner, D.J.C. Herr, Y. Obeng, 2011 International Conference on Frontiers of Characterization and Metrology for Nanoelectronics, MINATEC Campus, Grenoble, France, May 24, 2011.



- Dynamics of ions and vacancies in nanoscale memristive elements, including conductive filament formation and dissolution.
- Domain wall dynamics in magnetic and ferroelectric devices.

**Critical dimension and overlay metrology:** The measurement of critical dimensions (CD) and overlay (OVL) are well-established tasks in semiconductor manufacturing.<sup>83</sup> New technologies such as directed self-assembly have enabled patterning at length scales approaching 10 nm half-pitch, making CD measurement both more critical and more challenging. In addition, the emergence of extreme ultraviolet lithography (EUVL) has increased the need for CD metrology, especially for masks. In fact, mask inspection and cleaning is a key challenge facing EUVL.<sup>84</sup> Specific needs in this area include:

- Higher-spatial-resolution 3D scatterometry.
- Nondestructive 3D imaging of EUV masks and advanced devices/interconnects using scanning electron microscopy (SEM)- or helium ion microscopy (HIM)-based 3D reconstruction.

**Rapid wafer-scale inspection and metrology:** In a manufacturing environment wafer-scale metrology is often needed for process control and defect identification (e.g., voids in interconnects). Automation of metrology and processing is also desired. Some specific needs in this area include:

- Void detection and grain orientation in interconnects.
- Mitigation of charging effects and drift in electron backscatter diffraction (EBSD).
- Automatic process control (APC) linking metrology and processing.
- Automated sample preparation for destructive 3D imaging, e.g., focused ion beam (FIB) and SEM.
- In-tool integration of metrology capabilities, e.g., scatterometry.
- Rapid wafer-scale surface roughness measurements at speeds far surpassing conventional atomic force microscopy (AFM).
- Design of test structures for specific metrology needs.

### **Known Research Activities**

The European Union has funded several industry-centric programs aimed at metrology and nanotechnology, including:

#### **The European Nanotechnology Landscape Report:**

<http://www.nanowerk.com/nanotechnology-report.php?reportid=145>

#### **The Analytical Network for Nanotech (ANNA):**

<https://data-minalab.fbk.eu/anna/context1cba.html>

#### **The SEAL Project:** SEA-NET: 01/2006 - 06/2009, SEAL 06/2010 - 05/2013

<http://www.seal-project.eu/project.html>

---

<sup>83</sup> "Development of Metrology at NIST for the Semiconductor Industry," Stephen Knight, in *Characterization and Metrology for VLSI Technology: 2003 International Conference*, Eds: D. G. Seiler, A. C. Diebold, T. J. Shaffner, R. McDonald, S. Zollner, R. P. Khosla, and E. M. Secula.

<sup>84</sup> "EUV masks under exposure: practical considerations," E. Gallagher *et al.*, *Extreme Ultraviolet (EUV) Lithography II*, edited by B.M. La Fontaine, P.P. Naulleau, Proc. of SPIE 7969 (2011)



## **Research Recommendations**

The refinement of advanced characterization tools and techniques is an active area of research in materials science and physics. The semiconductor industry continues to benefit from these efforts, especially in the early stages of basic research. However, the impact of many of these advances in characterization in semiconductor manufacturing is limited. The task of bringing new characterization tools into a manufacturing environment is both costly and complex. University researchers often do not have the required knowledge or experience. Given the cost of creating commercial tooling, vendors are necessarily conservative, focusing on technology for which customers already exist. There may be an opportunity for government agencies such as NIST to help identify promising new characterization methods and to aid tool vendors in bringing these technology to market.

## **Conclusion: An Industry Vision and Research Guide**

As we said in our Rebooting the IT Revolution report of September, 2015, "the computing systems of today have delivered tremendous economic and societal benefits by automating tabulation and harnessing computational processing and programming to deliver enterprise and personal productivity. The new insight computing systems of tomorrow will forever change the way people interact with computing systems to help them extend their knowledge and make complex decisions involving extraordinary volumes of fast moving data. These future systems offer a multitude of opportunities to improve our society and our daily lives. However, much research is required to meet the runaway challenges of an increasingly data-rich world."

Information systems based on semiconductor technology are at the heart of electronic infrastructure that enables our society. The needs today are even more acute than they were eighteen months ago when we wrote those words. Tomorrow is right around the corner. We need to "adopt and fund an innovation agenda—built upon fundamental research— that creates a new engine to drive the next generations of human experience, economic and societal progress, security and sustainability. "



## About SIA

The Semiconductor Industry Association (SIA) is the voice of the U.S. semiconductor industry, one of America's top export industries and a key driver of America's economic strength, national security and global competitiveness. Semiconductors – microchips that control all modern electronics – enable the systems and products we use to work, communicate, travel, entertain, harness energy, treat illness, and make new scientific discoveries. The semiconductor industry directly employs nearly a quarter of a million people in the U.S. In 2014, U.S. semiconductor company sales totaled \$173 billion, and semiconductors make the global trillion dollar electronics industry possible. Founded in 1977 by five microelectronics pioneers, SIA unites companies that account for 80 percent of America's semiconductor production. Through this coalition, SIA seeks to strengthen U.S. leadership of semiconductor research, design and manufacturing by working with Congress, the Administration and other key industry stakeholders to encourage policies and regulations that fuel innovation, propel business and drive international competition. More information about SIA is available at <http://www.semiconductors.org>.

---

## About SRC

The Semiconductor Research Corporation (SRC) is a nonprofit consortium of companies having a common interest in accelerating the progress of research in semiconductor science and engineering. SRC defines industry needs, invests in and manages the research that gives its members a competitive advantage in the dynamic global marketplace. SRC partners with Federal agencies that also fund basic research and have an interest in semiconductor-related science and engineering. The SRC Nanoelectronics Research Initiative (NRI) partners with the National Institute of Standards and Technology (NIST) to fund three multi-university centers, and with NSF to fund about a dozen individual projects. STARnet, a program jointly supported by MARCO (a subsidiary of SRC) and the Defense Advanced Research Projects Agency (DARPA), supports six multi-university centers across the country. SRC's core program, Global Research Collaboration, has partnered with NSF on a number of joint programs. SRC expands the industry knowledge base and attracts premier students to help innovate and transfer semiconductor technology to the commercial industry. SRC was awarded the National Medal of Technology, America's highest recognition for contributions to technology. More information about SRC is available at <https://www.src.org/>.



