# A Federal Vision for Future Computing: A Nanotechnology-Inspired Grand Challenge

Collaborating Agencies: Department of Energy (DOE), National Science Foundation (NSF), Department of Defense (DOD), National Institute of Standards and Technology (NIST), Intelligence Community (IC)

## Introduction

This white paper presents a collective vision from the collaborating Federal agencies of the emerging and innovative solutions needed to realize the Nanotechnology-Inspired Grand Challenge for Future Computing. It describes the technical priorities shared by multiple Federal agencies, highlights the challenges and opportunities associated with these priorities, and presents a guiding vision for the research and development needed to achieve key near-, mid-, and long-term technical goals. By coordinating and collaborating across multiple levels of government, industry, academia, and nonprofit organizations, the nanotechnology and computer science communities can look beyond the decades-old approach to computing based on the von Neumann architecture and chart a new path that will continue the rapid pace of innovation beyond the next decade.

## Background

On October 20, 2015, the White House announced "*A Nanotechnology-Inspired Grand Challenge*" to develop transformational computing capabilities by combining innovations in multiple scientific disciplines. The Grand Challenge addresses three Administration priorities—the National Nanotechnology Initiative (NNI),[1] the National Strategic Computing Initiative (NSCI),[2] and the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative[3] to:

> ***Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain.***[4]

While it continues to be a national priority to advance conventional digital computing—which has been the engine of the information technology revolution—current technology falls far short of the human brain in terms of the brain's sensing and problem-solving abilities and its low power consumption. Many experts predict that fundamental physical limitations will prevent transistor technology from ever matching these characteristics.

## Call for a Coordinated Approach

In the announcement, the White House challenged the nanotechnology and computer science communities to look beyond the decades-old approach to computing based on the von Neumann architecture and chart a new path that will continue the rapid pace of innovation in information technology beyond the next decade. There are growing problems facing the Nation that the new computing capabilities envisioned in this challenge might address, from delivering individualized

---

[1] http://www.nano.gov
[2] https://www.whitehouse.gov/blog/2015/07/29/advancing-us-leadership-high-performance-computing
[3] https://www.whitehouse.gov/BRAIN
[4] http://www.nano.gov/futurecomputing

treatments for disease, to building more complex and more reliable systems, to allowing advanced robots to work safely alongside people, to proactively identifying and blocking cyber intrusions. To meet this challenge, major breakthroughs are needed not only in the basic devices, computing architecture, and software that store and process information, and in the amount of energy they require, but in the way a computer analyzes information, images, sounds, and patterns; interprets and learns from data; and identifies and solves problems.

Many of these breakthroughs will require new kinds of nanoscale devices and materials integrated into three-dimensional systems and may take a decade or more to achieve. These nanotechnology innovations will have to be developed in close coordination with new computer architectures, and will likely be informed by our growing understanding of the brain—a remarkable, fault-tolerant system that consumes less power than an incandescent light bulb.

Recent progress in developing novel, low-power methods of sensing and computation—including neuromorphic, magneto-electronic, and analog systems—combined with dramatic advances in neuroscience and cognitive sciences, lead us to believe that this ambitious challenge is now within our reach.

It is likely that future computing may evolve simultaneously in several directions. Traditional compute-intensive and server-based platforms will require continued investment and development. However, another important direction is the development of computing systems based on constrained power, embedded platforms that are aimed primarily at processing sensor data, providing output and system control. These systems will extract complex information from massive sensor data streams. In addition, they will learn and improve their capabilities during operation. A successful result of this Grand Challenge may indeed be the identification of application areas (that could be Grand Challenges themselves) that represent new approaches to computing, and then demonstrating the approach's effectiveness through a physical device technology with scalable manufacturing methods, a compatible computer architecture, and demonstrations of applications performance and capabilities.

Achieving this Grand Challenge would lead to many game-changing capabilities, addressing the following technology priorities shared by multiple Federal agencies:

- Intelligent big data sensors that act autonomously and are programmable via the network for increased flexibility, and that support communication with other networked nodes while maintaining security and avoiding interference with the things being sensed.
- Machine intelligence for scientific discovery enabled by rapid extreme-scale data analysis, capable of understanding and making sense of results and thereby accelerating innovation.
- Online machine learning, including one-shot learning, and new methods and techniques to deal with high-dimensional and unlabeled data sets.
- Cybersecurity systems that can prevent (or minimize) unauthorized access, identify anomalous behavior, ensure data and software code integrity, and provide contextual analysis for adversary intent or situational awareness; i.e., deter, detect, protect, and adapt.
- Technology that enables trusted and secure operation of complex platforms, energy, or weapons systems that require software (or combination of multiple codes) so complicated that it exceeds a human's ability to write and verify the software and its performance.

- Emerging computing architecture platforms, neuromorphic or quantum or others, that significantly accelerate algorithm performance, concurrency, and performance execution while maintaining and/or reducing energy consumption by over six orders of magnitude (from megawatts to watts, as achieved by biology) compared to today's state-of-the-art systems. Indeed, fundamental information theoretic bounds allow such a reduction.
- Autonomous or semi-autonomous platforms supporting the observe-orient-decide-act (OODA) process for both military and civilian purposes, such as transportation, medicine, scientific discovery, exploration, and disaster response.

## Research and Development Focus Areas

The research and development needed to achieve the Grand Challenge can be categorized into the following seven focus areas:

1. Materials
2. Devices and Interconnects
3. Computing Architectures
4. Brain-Inspired Approaches
5. Fabrication/Manufacturing
6. Software, Modeling, and Simulation
7. Applications

These focus areas are discussed in detail below, including near-, mid-, and longer-term goals for each area that will be significant advances in their own right.

### *1. Materials*

Discovery, understanding, and optimization of novel functional materials, as well as innovative materials integration, are needed for incorporation into advanced devices and architectures. Specific needs include materials for ultra-low-power digital switches, for interconnects below 10 nm scales, for intra-chip optical communication, and for new architectures such as neuromorphic computing, quantum computing, etc. New two-dimensional (2D) materials are attracting considerable research interest due to their potential for nanoelectronic applications. Graphene and other 2D materials (e.g., BN, $MoS_2$, $WS_2$, and fluorographene) are currently being considered for nanoelectronic logic applications. Individual 2D materials can be arranged into heterostructures with atomic precision, thus creating stacks with novel electronic properties. Superstructures built from combinations of different 2D materials offer dramatically richer opportunities in terms of physics and transport properties than each of the individual materials. Since the band structure of 2D materials depends on the number of layers, simply by changing the thickness of one of the components, one can fine tune the resulting electronic and optical properties.

An important task for sub-10 nm nanoelectronics is recognition of the fundamental role of "defects" that should not be treated as imperfections but instead as thermodynamically controllable entities. In fact, thermodynamics dictates the theoretical impossibility of "defect-free" materials of finite size. Due to nonstoichiometric defects, materials often behave as doped semiconductors, and can be described using the classical semiconductor model. A full understanding of nonstoichiometric and doping effects in metal oxides is required in order to control and optimize these and other properties for practical devices. The scaling limits of electron-based devices such as transistors are known to be on the order of

5 nm due to quantum-mechanical tunneling. Smaller devices can be made if information-bearing particles with mass greater than the mass of an electron are used. Therefore, new principles for logic and memory devices, scalable to ~1 nm, could be based on "moving atoms" instead of "moving electrons;" for example, by using nanoionic structures. Examples of solid-state nanoionic devices include memory (ReRAM) and logic (atomic/ionic switches). A critical issue for any new device material is the ability to achieve control of the properties of "semiconductor" materials that is comparable to that realized in conventional semiconductors.

Solid-state systems are, in general, not best suited for moving atoms, and concepts of fluid nanoelectronics/nanoionics from liquid media may offer a new path to replace the foundation of today's computing technologies. Ions in liquid electrolytes play an important role in biological information processors such as the brain or living cells. Based on this analogy, a binary state could be realized by a single ion that can be moved to one of two defined positions, separated by a membrane (the barrier) with voltage-controlled conductance. Although at an early stage, fluid nanoelectronics could allow for powerful and energy-efficient computing. Fluid nanoelectronics systems could be reconfigurable, with individual elements strung together to create wires and circuits that could be reprogrammed. Such flexibility would be in distinct contrast to conventional electronic circuits, which are hardwired by a fixed network of interconnects. Examples include nanoionic devices based on electrolyte-filled nanochannels, and protonic transistors based on ionomers (polymers with ionic properties that may offer a merger of solid-state and fluid nanoionics). In principle, such structures might be used to make devices scalable to ~1 nm or below. Such methods and principles also offer the potential for self-healing of nanodevices and more efficient heat removal.

Large-scale computational efforts are needed for understanding and predicting material properties for future information technologies because: (1) the structures themselves are neither atom-like nor bulk-like; (2) interfaces modulate device properties; and (3) systems operate under non-equilibrium conditions. A full understanding of the thermodynamics and kinetics of point defects in metal oxides would open the way to precisely engineered semiconductor properties. Development of computational models for tunneling in metal oxides is a critical task, both for the ultimate scaling of transistors and flash memory devices and for new emerging devices. Another example of a mission-critical task is to develop a better physical understanding to enable predictive models for the heat transport properties of semiconductors at the nanoscale. Computational explorations of emerging two-dimensional (e.g., graphene, BN, $MoS_2$, $WS_2$, fluorographene), one-dimensional (carbon nanotubes, etc.), and zero-dimensional (quantum dot) materials can potentially lead to new insights and discoveries of new thermal, electronic, and optical phenomena. Accurate first-principle models that address realistic structure sizes and that operate at multiple scales are needed to support further developments in nanoscale information technologies, such as those being addressed by the Materials Genome Initiative.[5]

An important attribute of all biological computing systems is the use of inherently three-dimensional (3D) materials and structures. Therefore, methods for 3D nanofabrication are critical for future brain-inspired computing technologies. Possible directions include 3D lithographic patterning, 3D self-assembly (including programmable, DNA-controlled self-assembly), and inkjet printing. Also, biological 3D nanofabrication may serve as an inspiration for future manufacturing technologies: the living cell is capable of fabricating amazingly complicated structures with high yield and low energy utilization. How

[5] https://www.whitehouse.gov/mgi

can an understanding of such "cellular factories" be used to guide substantial improvements in the processes now used in semiconductor manufacturing? It is known that silicon-based memory may become prohibitively expensive for zettascale "big data" deployments in a decade or two. However, DNA could be a candidate for scalable, random-access, and error-free information storage. DNA research has demonstrated an information storage density that is several orders of magnitude higher than any other known storage technology. Potentially, a few tens of kilograms of DNA could meet all of the world's storage needs; moreover, DNA can store information stably at room temperature with zero power requirements, making it a suitable candidate for large-scale archival storage.

A new materials base may be needed for future electronic hardware. While most of today's electronics use silicon, this approach is unsustainable if billions of disposable and short-lived sensor nodes are needed for the coming Internet-of-Things (IoT). To what extent can the materials base for the implementation of future information technology (IT) components and systems support sustainability through recycling and bio-degradability? More sustainable materials, such as compostable or biodegradable systems (polymers, paper, etc.) that can be recycled or reused, may play an important role. The potential role for such alternative materials in the fabrication of integrated systems needs to be explored as well.

- 5-year goal: Identify promising emerging materials systems suitable and with high potential for device fabrication and CMOS integration. Concurrently, begin development of the measurement science and technology required to determine materials properties and scaling effects.
- 10-year goal: Enable physical modeling and simulation at scales that will allow for the characterization, simulation, and prediction of potential device behavior and performance for future circuit designs and analysis. Conduct parallel efforts to address multiple aspects of the materials problem (discovery, characterization, manufacturability), where integration of such efforts will inform the direction of each individual effort.
- 15-year goal: Achieve a fundamental understanding of materials properties, scaling, and prediction for the properties of new materials systems and their performance and characterization; of their suitability for the design, fabrication, and scalability of new devices; and of their integration with CMOS.

### *2. Devices and Interconnects*

In an ongoing Nanotechnology Signature Initiative, *Nanoelectronics for 2020 and Beyond,*[6] the efficacy of using non-charge-based devices as a replacement for conventional charge-based transistors has been explored. There the goal is to determine if other state variables—electron spin, magnetization, strain, phase, molecular conformation, or yet other physical quantities—could be used as a variable for switching, and thus replace the role of electric quantities (charge, voltage, current) in transistors, and to ultimately determine if significant benefits could be derived from such novel devices. The outcome of this effort has been a gamut of potential non-charge-based devices, including essentially mechanical devices that exhibit switching phenomena, but as yet no single candidate has emerged as an ideal alternative for today's silicon CMOS transistors. The reason could be that while an alternative switch must function at low power levels, for it to be successfully integrated into a computing architecture it

[6] National Nanotechnology Initiative Signature Initiative: *Nanoelectronics for 2020 and Beyond* (National Science and Technology Council, Committee on Technology, Subcommittee on Nanoscale Science, Engineering, and Technology, July 2010: http://www.nano.gov/NSINanoelectronics).

must also satisfy a large number of other criteria, such as scalability, reliability, crosstalk, manufacturability, and feasibility of integration into a common platform.

Current transistors operate at about one volt; with the emergence of millivolt switches the task of organizing and interconnecting them into higher-level functional blocks calls for fundamentally new computer architectures. Besides an overall change in architectural framework involving devices, memory, interconnects, and modes of data transfers between them, new methods of error correction are also needed to reach optimum performance levels. In this respect, fundamental innovations in computing technologies will require that future device research be guided by the feasibility of these new devices being accommodated in newer architectures of the future.

The performance of today's advanced digital circuits is highly constrained by the need to limit switching energy dissipation, and many new, more energy-efficient device concepts have been proposed that would greatly reduce this constraint. The fundamental theoretical limit of power dissipation for switching is known as the Landauer limit, measured in bit flips, and it is still approximately five orders of magnitude lower than current technologies. In modern processors, the metal *interconnects* between the switching devices are known to be a greater source of power dissipation than the switching devices themselves. The recent advent of multicore/many-core architectures and network-on-chip technologies have made it necessary to have efficient intra-chip and inter-chip communication. However, the requirements for power, bandwidth, latency, throughput, and scalability of this new development in turn require innovations from materials to circuits to microarchitectures and even systems, encompassing design tools, smart network topologies, and new parallel algorithms and software.

A heterogeneous mix of traditional and novel technologies (and materials) for interconnects needs to be explored. This mix includes conventional technologies such as electronics, 3D stacking, radio frequencies, photonics, and silicon nanophotonics, as well as emerging technologies based on novel materials and devices such as carbon nanotube-based interconnects, terahertz solutions exploiting surface plasmonics, and metamaterials. Current device research is primarily focused on single device demonstration. However, equally important is the task of integrating billions of nanometer-scale devices and interconnects into a computing architecture while assuring the availability of suitable programming models, software, and manufacturability—downstream needs that have not historically been addressed in parallel development of both software and hardware.

- 5-year goal: Fabricate and characterize emerging devices, circuits, and interconnects with promising scalability properties and potential integration with CMOS. Develop open-sourced device models and simulation techniques, and integrate with open and industry standard circuit design and simulation tools and environments. It will be critical to understand and incorporate reliability fundamentals from the start, considering lifetimes and degradation as soon as promising new materials, devices, interconnects, and architectures are identified.
- 10-year goal: Develop standard libraries incorporating nonlinear phenomena and fabrication variations. Develop design and simulation environments suitable for large-scale circuit architectures in both analog and digital domains.
- 15-year goal: Enable device and circuit design, modeling, and simulation environments with the capability of predicting device structure, behavior, and performance based on future computing system requirements. The ultimate goal is to minimize expert knowledge requirements in

materials or device physics in order to create and design devices based on new materials systems and circuits driven by desired user properties, behaviors, and applications.

### *3. Computing Architectures*

The basic architecture of computers today is essentially the same as those built in the 1940s—the von Neumann architecture—with separate compute, high-speed memory, and high-density storage components that are electronically interconnected. However, it is well known that continued performance increases using this architecture are not feasible in the long term, with power density constraints being one of the fundamental roadblocks.[7] Further advances in the current approach using multiple cores, chip multiprocessors, and associated architectures are plagued by challenges in software and programming models. Thus, research and development is required in radically new and different computing architectures involving processors, memory, input-output devices, and how they behave and are interconnected.

Application-specific integrated circuits (ASICs) can provide significant performance improvement for specific tasks compared with programmed general-purpose computers. However, ASICs' relatively high non-recurring engineering and design costs only make them commercially economical for large-volume applications (e.g., mobile computing). The development of domain-specific architectural design principles is one approach currently being used to lower this cost, including accelerator-rich architectures. Such design approaches are not only useful for ASICs, but also for broader applications areas including biomedical imaging, financial modelling, and DNA sequencing. Other techniques using 3D integrated circuit technologies, crossbar architectures, microfluidic cooling technologies, and software-hardware codesign principles can be leveraged to optimize future computing architecture performance as well. Computing architecture design innovations are needed to address the power dissipation challenge in the near future in order to ensure the continued economic growth of the IT industry.

The rise of solid-state, nonvolatile memory devices provides the opportunity to collapse the traditional computer data hierarchy and to store with immediate availability all the data required for computation directly adjacent to the processors. This is the processing in memory (PIM) approach to overcoming the von Neumann bottleneck. Other paths, such as approximate, probabilistic, and stochastic computing methods, use a variety of approaches to trade off precision in the result for reducing the time to provide a result, or to relax computational determinism for energy efficiency, or to minimize the data required for computation. Algorithms, architectures, and technologies that are developed for these approaches can also have significant benefit for brain-inspired approaches by minimizing training data requirements, improving performance, and reducing energy.

Architectural design innovations are needed to sustain the growth of computing performance and tackle power dissipation challenges in the near future. An important research goal will be to architect machines that will leverage ultralow power consumption devices built from new material systems that offer alternatives to CMOS technologies based only on silicon. The task of integrating billions of emerging nanometer-scale devices and interconnects into new computing architectures that can be

[7] *The Future of Computing Performance: Game Over or Next Level?* (National Research Council, Computer Science and Telecommunications Board, Committee on Sustaining Growth in Computing Performance, 2011: http://www.nap.edu/catalog/12980/the-future-of-computing-performance-game-over-or-next-level).

readily manufactured, while simultaneously developing the required programming models, software, etc., is an essential research priority.

- 5-year goal: Enable large-scale design, modeling, characterization, and verification of future computing architectures in both digital and analog domains. Leverage advances in high-performance computing platforms to enable parallel, high-concurrence, and large-scale simulations beyond exascale performance. This will enable the hybridization and interfacing of current digital computing with quantum- or biology-inspired computing approaches that require analog and other novel interfaces.
- 10-year goal: Be able to predict the performance of new architectures incorporating new material systems and physical nonlinear phenomena.
- 15-year goal: Be able to predict the design and characterization of computing architectures based on user applications needs. These results should enable ready-to-fabricate designs and specifications.

### *4. Brain-Inspired Approaches*

Neuroscience research suggests that the brain is a complex, high-performance computing system with *low energy consumption* and incredible *parallelism*. A highly plastic and flexible organ, the human brain is able to grow new neurons, synapses, and connections to cope with an ever-changing environment. Energy efficiency, growth, and flexibility occur at all scales, from molecular to cellular, and allow the brain, from early to late stage, to never stop learning and to act with proactive intelligence in both familiar and novel situations. Understanding how these mechanisms work and cooperate within and across scales has the potential to offer tremendous technical insights and novel engineering frameworks for materials, devices, and systems seeking to perform efficient and autonomous computing.
This research focus area is the most synergistic with the national BRAIN Initiative. However, unlike the BRAIN Initiative, where the goal is to map the network connectivity of the brain, the objective here is to understand the nature, methods, and mechanisms for computation, and how the brain performs some of its tasks. Even within this broad paradigm, one can loosely distinguish between neuromorphic computing and artificial neural network (ANN) approaches. The goal of neuromorphic computing is oriented towards a hardware approach to reverse engineering the computational architecture of the brain. On the other hand, ANNs include algorithmic approaches arising from machine learning, which in turn could leverage advancements and understanding in neuroscience as well as novel cognitive, mathematical, and statistical techniques. Indeed, the ultimate intelligent systems may as well be the result of merging existing ANN (e.g., deep learning) and bio-inspired techniques.

High-performance computing (HPC) has traditionally been associated with floating point computations and primarily originated from needs in scientific computing, business, and national security. On the other hand, brain-inspired approaches, while at least as old as modern computing, have traditionally aimed at what might be called pattern recognition applications (e.g., recognition/understanding of speech, images, text, human languages, etc., for which the alternative term, *knowledge extraction*, is preferred in some circles) and have exploited a different set of tools and techniques. Recently, convergence of these two computing paths has been mandated by the *National Strategic Computing Initiative Strategic Plan,*[8] which places due emphasis on brain-inspired computing and pattern

[8] *National Strategic Computing Initiative Strategic Plan* (National Strategic Computing Initiative Executive Council, July 2016: https://www.whitehouse.gov/sites/whitehouse.gov/files/images/NSCI%20Strategic%20Plan.pdf).

recognition or knowledge extraction type applications for enabling inference, prediction, and decision support for big data applications. DOE and NSF have demonstrated significant scientific advancements by investing and supporting HPC resources for open scientific applications. However, it is becoming apparent that brain-like computing capabilities may be necessary to enable scientific advancement, economic growth, and national security applications.

- 5-year goal: Translate knowledge from biology, neuroscience, materials science, physics, and engineering into useable information for computing system designers.
- 10-year goal: Identify and reverse engineer biological or neuro-inspired computing architectures, and translate results into models and systems that can be prototyped.
- 15-year goal: Enable large-scale design, development, and simulation tools and environments able to run at exascale computing performance levels or beyond. The results should enable development, testing, and verification of applications, and be able to output designs that can be prototyped in hardware.

### *5. Fabrication/Manufacturing*

The National Strategic Computing Initiative recognizes the importance of support for all aspects of computing, including fabrication. From the NSCI fact sheet of July 29, 2015: *Sustaining this capability requires supporting a complete ecosystem of users, vendor companies, software developers, and researchers. The Nation must preserve its leadership role in creating HPC technology and using it across a wide range of applications.*[9] Access to advanced nanofabrication capabilities for the research community is key for ensuring ongoing improvements in technology, along with collaborations with industry to transition new fabrication methods to commercial scale.

A comprehensive industrial ecosystem built around a nanofabrication paradigm will bridge the gap between one-off experimental fabrication and high-volume manufacturing production, and will shorten the product development cycles. Academic and corporate users, including those from small and medium-sized enterprises, will use the increased accessibility to high-performance nanofabrication capabilities to support the missions of the NNI agencies and to help realize the potential of nanotechnology to benefit society. To maximize the societal impact and commercial potential of nanotechnology, we need a new paradigm for nanofabrication that better matches the diverse needs of emerging nanotechnologies.

The natural fault tolerance and stochastic nature of neuromorphic computing or bio-inspired circuitry may allow the implementation of a whole new family of “approximate” manufacturing techniques that cannot be leveraged by existing digital computing structures.

- 5-year goal: Develop tools and fabrication capabilities able to integrate new materials systems, potentially similar to additive manufacturing, at scales relevant to this challenge.
- 10-year goal: Achieve the ability to prototype new computing architectures incorporating new materials systems and nonlinear phenomena with relatively fast turnaround times (from years to months) compatible with state-of-the-art microelectronics practices.

[9] https://www.whitehouse.gov/sites/default/files/microsites/ostp/nsci_fact_sheet.pdf

- 15-year goal: Develop a cost-effective foundry process and design methodology widely accessible to a broad range of research and development groups suitable for both low- and high-volume device fabrication/manufacturing.

### *6. Software, Modeling, and Simulation*

It is important to realize that progress in future computing will continue to rely and depend upon further improvements in digital computing systems. It is critical to be able to simulate and emulate at scale any new future computing system, and doing so will require current petascale and near-future exascale systems. Therefore, important research and development should continue and potentially increase on the scalability, portability, usability, verification, and validation of extreme-scale computing architectures that will enable future computing to become a reality. In current HPC systems, parallelism and concurrency have become critically important. Breakthroughs will require a collaborative effort among researchers representing all areas—from services and applications down to the nano-architecture and materials level—to research, discover, and build on new concepts, theories, and foundational principles. Approaches to achieving beyond-exascale performance and usability will require new abstract models and algorithms; new programming environments and models; and new hardware architectures, compilers, programing languages, operating systems, and runtime systems; and each must exploit domain- and application-specific knowledge.

The development and deployment of new materials, models, algorithms, and hardware architectures is expected to decrease latency and energy consumption by several orders of magnitude in overall system computing efficiency. Also, with the present trend for performance improvement, a 50 exaFLOP computer that runs within a 20 MW power envelope could be at the head of the TOP500 list by 2025. However, to go above and beyond that performance will require a serious research effort that begins now. New software stacks, compilers, data management, analytics, visualization, programming models, languages and environments, extreme-scale emulation, and user interfaces that leverage the full capabilities of the new computing paradigm are required as well. In particular, several important aspects need to be researched and understood; for example, computational theory, including formal methods, modeling, verification and simulation, and metrics for evaluating “brain-like” systems.

- 5-year goal: Create programing and development languages and environments, libraries, solvers, and compilers that do not require deep knowledge and expertise to use. Resulting software and solutions must support state-of-the-art and beyond-exascale high-performance computing platforms.
- 10-year goal: Incorporate nonlinear physical and materials phenomena within modeling and simulation systems capable of design, simulation, and verification of future computing architectures, including accurate prediction of performance. Systems should be capable of application exploration and demonstration at large scales.
- 15-year goal: Develop software methods and techniques capable of automated discovery and exploration of large, complex parameter spaces from a mathematical, materials, physical, biological, fabrication, or computing architecture point of view.

### *7. Applications*

The general vision for computing systems enabled by the Nanotechnology-Inspired Grand Challenge for Future Computing includes the ability to process, analyze, and eventually understand multi-modal sensor data streams and complex workflows. Such systems will be able to learn online and in real time

using unsupervised training and will be able to capture and understand complex data structures. These systems will then also be able to plan and generate complex actions in response to the input data. Some examples of such applications enabling new capabilities include the following:

- Scientific discovery and analysis of very large and complex datasets, including automatic discovery from published literature and experimental research facilities.
- Information and data integrity assurance enabled by tamper-resistant, self-protecting software codes.
- Autonomous robotics and vehicles and intelligent prosthetics integrating motion, vision, sound, planning, and understanding.

Cybersecurity will be an important application area for future computing systems. Security can be considered as a design attribute as well, much like performance or power dissipation. From this perspective, we need systems that adapt quickly to threats, faster than human system administrators can respond to prevent (or minimize) unauthorized access, identify anomalous behavior, and provide contextual analysis for adversarial intent or situational awareness.

Future cybersecurity systems will need to provide analytics for modeling and predicting incidents that will enable us to deter adversaries from attack. Also needed are advanced capabilities for detecting adversarial activity with automated response for protection across a wide range of platforms, including network infrastructure, critical national infrastructure, and emerging Internet of Things and supervisory control and data acquisition (SCADA) platforms. Real-time fusion of disparate data will enable reasoning about the state of the system and selection of optimal defense actions or system adaptations for anomalous conditions. Dynamic response mechanisms intended to influence adversarial actions or confuse and deceive sophisticated attackers can provide an asymmetric advantage for cyber defense.

- 5-year goal: Achieve autonomous capabilities for routine attack scenarios, utilizing enterprise-level computing resources, and human-machine augmented capabilities for sophisticated attack scenarios.
- 10-year goal: Achieve autonomous capabilities for sophisticated attack scenarios, utilizing enterprise-level computing resources, with routine attack scenarios resolved with compact and energy-efficient computing resources.
- 15-year goal: Achieve fully autonomous capabilities for sophisticated attack scenarios with compact and energy-efficient computing resources.

## Conclusion

This future computing Grand Challenge offers a great opportunity to advance computing to a new historical level, enabling a new computing paradigm able to deliver human brain-like performance in terms of sensing, problem-solving abilities, and low energy consumption. This exciting opportunity requires new approaches to understand new materials, devices, algorithms, software, and their integration within new computing architectures. Research and development in radically new and different computing architectures involving processors, memory, devices, materials, and the way they are interconnected are required. In summary, a significant and well-coordinated national effort across multiple levels of government, industry, academia, and nonprofit organizations is necessary to achieve success in this very important Grand Challenge.

## Agency Interests

The following table outlines the interests of the participating agencies in the seven focus areas outlined above.

| | Materials | Devices and Interconnects | Computing Architectures | Brain-inspired Approaches | Fabrication/ Manufacturing | Software, Modelling, and Simulation | Applications |
|---|---|---|---|---|---|---|---|
| **NSF** | x | x | x | x | x | x | |
| **DOE** | x | x | x | x | x | x | x |
| **DOD** | x | x | x | x | x | x | x |
| **NIST** | x | x | x | x | x | x | x |
| **IC** | x | x | x | x | x | x | x |

## Related Agency Activities

### *1. Materials*

The Division of Materials Research (DMR)[10] and the Division of Civil, Mechanical, and Manufacturing Innovation (CMMI)[11] at NSF have core programs supporting materials science and materials engineering research. Additional efforts at NSF can be found in the Division of Chemistry (CHE)[12] and in the Division of Electrical, Communications and Cyber Systems (ECCS).[13] The NSF Science and Technology Center (STC) on Integrated Quantum Materials at Harvard, Howard, and MIT[14] is currently working on graphene, topological insulators, and nitrogen vacancy centers in diamond, as well as their integration. Developing materials, techniques, and simulation methods for controlled evolution of quantum mechanical states of multiple to many qubit systems is one of the emphasized topics for the upcoming competition of the Materials Research Science and Engineering Centers in DMR.[15] The competitions in FY 2014 and FY 2015 on Two-Dimensional Atomic-layer Research and Engineering (2-DARE)[16] represented an example of close collaboration between NSF and AFOSR.

Basic Energy Sciences (BES) at DOE[17] has core programs supporting extensive materials research related to this Grand Challenge. These programs include, but are not limited to, quantum materials broadly categorized in the areas of superconductivity, magnetism, topological materials, quantum coherence, and low dimensions (0, 1, and 2D). BES supports synthesis, characterization, and theory related to these materials at universities and DOE-supported national laboratories, and has recently sponsored reports

[10] http://www.nsf.gov/funding/programs.jsp?org=DMR
[11] http://www.nsf.gov/funding/programs.jsp?org=CMMI
[12] http://www.nsf.gov/funding/programs.jsp?org=CHE
[13] http://www.nsf.gov/funding/programs.jsp?org=ECCS
[14] http://ciqm.harvard.edu/
[15] http://www.nsf.gov/pubs/2016/nsf16545/nsf16545.htm
[16] http://www.nsf.gov/pubs/2015/nsf15502/nsf15502.htm
[17] http://science.energy.gov/bes/

directly impacting materials underpinning the needs for future computing. These reports include the 2015 BES Advisory Committee report, *Transformative Opportunities for Discovery Science*[18] and *Neuromorphic Computing – From Materials Research to Systems Architecture Roundtable*[19] (co-sponsored with the DOE-Advanced Scientific Computing Research program office). In addition, three BES-sponsored Basic Research Needs workshops on (1) quantum materials; (2) synthesis science; and (3) instrumentation science have been completed, with reports forthcoming. BES also collaborates with NSF to support National Academy of Sciences studies on materials, many of which are related to future computing needs.

NIST has core programs that support the development of the materials science and engineering foundation for future electronics with measurement science, data, and standards covering a broad range of nanoscale and low-dimensional materials, including carbon nanotubes, graphene and related 2D layers, magnetic materials, thin film oxides, interconnects, dielectrics, superconductors, and organic/molecular semiconductors. These efforts span materials structure fundamentals, fabrication process measurements and control, device reliability, and electrical characterization of devices and circuits. Further, through the Materials Genome Initiative (MGI),[20] NIST is building the materials innovation infrastructure that closely integrates advanced computation, data management, and informatics to enable the discovery and deployment of advanced materials such as those needed for future electronics.

### *2. Devices and Interconnects*

Several NSF programs within multiple directorates are currently supporting research in this area. Notable among them are the core programs in the Computing and Communication Foundations (CCF)[21] and ECCS divisions. The NSF STC at the University of California, Berkeley on "Energy Efficient Electronic Systems (E3S)"[22] has supported work on novel materials, quantum tunneling field-effect transistors, nanoelectromechanical switches, and interconnect research for the last 6 years. Together with the Semiconductor Research Corporation (SRC), NSF has recently announced a joint program on "Energy Efficient Computing: from Devices to Architectures (E2CDA),"[23] which, among other aspects of low-power computer design, is largely focused on devices and interconnect research as well. ONR is presently supporting graphene research and plans to exploit its superior functionalities to develop electronic, optoelectronic, magnetic, and mechanical devices.

NIST has core programs that support this area, including measurements for the development of "superfill" mechanistic models for the metallization of high-aspect-ratio trenches in interconnects and through silicon vias and the characterization of nanoporous low-k dielectric thin films. NIST has performed innovative work on the fundamentals of fabrication processes for reliable interconnects, utilizing the concept of "building in reliability," wherein design of metal deposition processes is informed by measurement of the resulting interconnect microscopic structure and subsequently by performance in service. This approach of establishing a feedback method of improving reliability is based on cyclic ties

[18] http://science.energy.gov/~/media/bes/besac/pdf/Reports/Challenges_at_the_Frontiers_of_Matter_and_Energy_rpt.pdf
[19] http://science.energy.gov/~/media/bes/pdf/reports/2016/NCFMtSA_rpt.pdf
[20] https://mgi.nist.gov/
[21] http://www.nsf.gov/funding/programs.jsp?org=CCF
[22] https://www.e3s-center.org/
[23] http://www.nsf.gov/pubs/2016/nsf16526/nsf16526.htm

among material processing/resulting material structure/resulting material properties. The approach is extendable to any materials system that may emerge as promising for future interconnect systems.

IARPA's Cryogenic Computing Complexity (C3) program[24] is developing materials, device designs, and processes for superconducting computing that could offer an attractive low-power alternative to CMOS with many potential advantages. Josephson junctions, the superconducting switching devices, switch quickly (~1 ps), dissipate little energy per switch (< $10^{-19}$ J), and communicate information via small current pulses that propagate over superconducting transmission lines nearly without loss.

### *3. Computing Architectures*

The E2CDA program jointly announced by NSF and SRC is intended to support innovative device and architecture research. The NSF core programs in the Computer & Information Science & Engineering (CISE) Directorate[25] have ongoing research in design of circuits, systems, and architectures relevant to future computing needs. Several Defense Advanced Research Projects Agency (DARPA) programs[26] have also focused research on reliability and PIM-type architectures.

### *4. Brain-Inspired Approaches*

The DARPA SyNAPSE program[27] is an early example of recent efforts showing the path forward, with the goal to meet certain architecture constraints. The software and hardware partition for brain-inspired architectures could be very different than what exists today. For example, an entire algorithm could be built into the hardware (e.g., in DARPA's UPSIDE program[28]), where in a sense, the hardware itself becomes the algorithm, and where sometimes the physics does the computation. Albeit somewhat specialized, such systems could have orders of magnitude energy and/or speed performance improvement over current systems, and yet have some flexibility or "programmability" for domain-specific tasks by virtue of network parameters that are adapted and pre-loaded during training. Industry has recently taken steps in this direction; for example, the IBM TrueNorth is an advanced implementation of a similar idea in silicon CMOS technology. Further experimentation with platforms of this type is needed by researchers to fully exploit their potential.

The DOE national laboratories have created spiking neural network models that are running on large-scale HPC systems with the potential to significantly enhance scientific discovery on high-dimension and highly connected data sets. They have also developed specialty hardware to accelerate adaptive neural algorithms and improve memory and logic interaction on microprocessors, and have developed the *Xyce* open-sourced software[29]—a SPICE-compatible, high-performance analog circuit simulator capable of solving extremely large circuit problems by supporting large-scale parallel computing platforms. DOE has demonstrated an evolutionary "deep learning" optimization that runs on thousands of graphic processing unit (GPU) processors to find the best configuration of the hyper parameters for a network,

[24] https://www.iarpa.gov/index.php/research-programs/c3
[25] http://nsf.gov/funding/programs.jsp?org=CISE
[26] http://www.darpa.mil/our-research
[27] http://www.darpa.mil/program/systems-of-neuromorphic-adaptive-plastic-scalable-electronics
[28] http://www.darpa.mil/program/unconventional-processing-of-signals-for-intelligent-data-exploitation
[29] https://xyce.sandia.gov/

with the goal of enhancing scientific discovery. DOE researchers are also working with academia to explore new methods on neuromorphic and quantum D-Wave processors.

Discovering and creating brain-like algorithms is an important area to be addressed by this Grand Challenge. The development of such algorithms is not a priority of the BRAIN Initiative (or the European Human Brain project), and thus this Grand Challenge can fill an important gap. To this end, the DARPA UPSIDE Program Cortical Processor Study[30] and the IARPA MICrONS program[31] are two current efforts beginning to address the issue. The latter program is directed towards discovering mesoscale, brain-like machine learning algorithms by establishing a dialogue between data science and neuroscience. The Cortical Processor Study is aimed at taking existing neural-inspired techniques and merging them with traditional machine learning to create a new set of algorithms that address a number of perceived limitations in existing machine learning, and then to apply these hybrid algorithms to real-world applications.

The Collaborative Research on Computational Neuroscience (CRCNS) program[32] at NSF, a decade-long program that is contributing to the BRAIN Initiative, is fostering transformative research on methods for quantifying and predicting neural and behavioral data in biological systems from cellular to human-level brain function. The Neural and Cognitive Systems (NCS) program[33] at NSF also considers technological innovations in neuroengineering and brain-inspired concepts and designs to inform the development of neuromorphic or neural-inspired chipsets and computing devices. Finally, the recently initiated E2CDA program at NSF, while broadly focused on low-power computing, also entertains brain-like algorithms and their architectural implementation at the nanoscale.

### *5. Fabrication/Manufacturing*

Several efforts within NIST address nanoscale 3D self-organization, defect detection, and critical dimension confirmation for new fabrication and computing paradigms. For example, computational methods are coupled with advanced experimental tools to build predictive design modules for the directed self-assembly of block copolymer thin films to enable the fabrication of ultrasmall patterns needed in future computing nodes. The NIST Center for Nanoscale Science and Technology (CNST) is a user facility that supports the development of nanotechnology by providing industry, academia, and other government agencies access to nanoscale measurement and fabrication methods and tools.

[30] http://rebootingcomputing.ieee.org/images/files/pdf/RCS4HammerstromThu515.pdf
[31] https://www.iarpa.gov/index.php/research-programs/microns
[32] https://www.nsf.gov/pubs/2015/nsf15595/nsf15595.htm
[33] http://www.nsf.gov/pubs/2016/nsf16508/nsf16508.htm