Rebooting the IT Revolution:
A Call to Action

August, 2015

This report is based on the Rebooting the IT Revolution Workshop held March 30–31, 2015 in Washington, DC. The workshop was sponsored by Semiconductor Industry Association and Semiconductor Research Corporation and supported by the National Science Foundation.

Acknowledgements

This report was produced by the Semiconductor Industry Association and the Semiconductor Research Corporation in consultation with their member companies and semiconductor technology experts from government and academia, and with support from the National Science Foundation. SIA and SRC would like to acknowledge all those who have contributed to the report.

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/.

Contents

Executive Summary **1**
The Technology Imperative **4**
Intelligent Sensor Nodes 5
Node Characteristics *6*
Materials – Nonsilicon Devices *7*
Manufacturing Needs *7*
Hardware vs. Software *8*
Future Directions *8*
Always Connected: Billions of Connected Nanosystems 8
From Sensor Nodes to the Cloud *8*
Wireless Communications *9*
Infrastructure Technologies for Big Data 10
Examples of Big Data *10*
Data Centers *10*
Big Data vs. High Performance Computing *11*
Technology Drivers and Needs *11*
New Architectures *12*
Brain-Inspired Computing *12*
Elements of an Intelligent Infrastructure 13
Insight Computing Systems *14*
Cyber-physical Systems *14*
Security at the Chip Level *14*
Critical Research Challenges **15**
Energy-Efficient Sensing and Computing 15
Cyber-physical Systems 16
Intelligent Storage 16
Real-Time Communication Ecosystem 17
Multi-level and Scalable Security 17
Next-Generation Manufacturing Paradigm 17
Insight Computing 18
IoT Test Platform 19
What Will It Take? A New Coordinated Research Initiative **20**
Leveraging, Sustaining, and Enhancing Existing Programs 20
A Public–Private Partnership Model 22
Conclusion: A Vision of Future IT Technology and Infrastructure 22
Appendix A: Data Explosion Facts **23**
A1. Storage Needs in 2040 23
A2. Global Silicon Wafer Supply Trend 24
A3. Total Communication Capacity in 2040 25
A4. Total Effective Capacity to Compute Information in 2040 26
A5. The Weight of a Flash Bit 28

Appendix B. Rebooting the IT Revolution Workshop 30
Workshop Steering Committee 30
Workshop Agenda 32
Workshop Participants 33
References 34

Executive Summary

To realize the full benefits of the Internet of Things and Big Data and to build the foundation for future technologies that convert data to "insight", there is an urgent need for a targeted and coordinated government initiative similar to that which sparked the semiconductor revolution fifty years ago—a "National Computing and Insight Technologies Ecosystem" initiative (N-CITE). This new initiative will spur major advances in the science and technology of information systems and unleash broad opportunities for innovation.

Access to information anywhere, anytime, is a hallmark of the world today, but the ability to collect, move, analyze, store, and—most importantly—use the exploding data stream to develop insights and options will require new approaches to computation. In the past, data were generated and communicated primarily among information technology (IT) systems—albeit of diminishing size. In the future, data-producing systems increasingly will involve small, low-power sensors and actuators embedded in the physical world—a network of cyber-physical systems, also referred to as the Internet of Things. At the same time, ever more powerful, high performance computation will analyze and manage the deluge of data, and systems that anticipate needs will deliver useful information when and where it is needed. Beyond information, the combination of distributed, networked sensors with massive data centers and cloud-based computing capabilities will make it possible to obtain *insight* that can inform decision making at all levels.

The combined IT infrastructure—from small, low-cost sensors to massive computing systems—will enable innovation in society writ large, in both government and the private sector. Driverless cars and personalized medicine along with countless other applications of intelligent systems are on the horizon. IT-based solutions for energy generation and delivery, safer air and ground transportation, and national and homeland security are envisioned. Greater availability and access to distributed IT and computing on a global scale, in developed and developing countries and regions, will help to reduce the "digital divide" and provide a platform for locally driven innovation.

However, near term, product-driven investment by the private sector alone is not sufficient to create the significant advances in IT infrastructure and insight technologies needed for these innovations. Private sector research and development must build upon and connect to government-funded programs, including the following:

- The National Strategic Computing Initiative, which lays the framework for major investments in the fundamental technologies that drive computer systems performance.
- Grand Challenges under consideration as part of the National Nanotechnology Initiative point to important areas of inquiry specifically related to semiconductors.
- The BRAIN (Brain Research through Advancing Innovative Technologies) initiative, focusing on better understanding the human brain, could aid the development of brain-inspired computing.

Without substantial United States investment, innovations in future information systems will occur elsewhere, and the economic vitality and national security benefits will flow to other nations.

A workshop on *Rebooting the IT Revolution,* held March 30–31, 2015, in Washington DC, brought together experts from industry, government, and academia to explore leap-ahead IT technologies. Attendees concluded that research is required in the following critical areas:

- **Energy-efficient sensing and computing.** Energy efficiency is vital at all levels—from the smallest sensor to ultra-high performance processors and systems. Sensor nodes also often need to operate without access to the electric grid for power—using batteries or harvesting energy from the environment. Advanced processors, on the other hand, are limited in their performance by energy inefficiencies that result in overheating and thermal management issues. Novel materials, devices, and computational and physical architectures must be developed to reduce the energy used to move data, either on a chip or over longer distance.
- **Cyber-physical systems.** Cyber-physical systems contain networks of collaborating computational elements that control physical entities. They require new engineering models that embrace temporal dynamics and algorithmic computation; indeed, a CPS must have robust control in the face of very dynamic, physical systems that often act randomly. Moreover, safety and security must be designed in to protect lives, personal information, and property.
- **Intelligent storage.** New memory technologies and management systems are needed to store and archive the projected explosion of data. Research is needed on implementation of new memory architectures, including how a new architecture could disrupt the system and what changes are required for optimal memory management by system hardware and software.
- **Real-time communication ecosystem.** Communications bandwidth, especially for wireless communication, must grow along with the amount of data. The ability to provide very high speed communication will be critical for real-time applications. Broadband communication will be greatly aided by advances in technologies for terahertz communication, including novel materials. Research is needed on advanced antenna arrays and the design of reliable and resilient communication networks that are capable of self-organizing and self-reconfiguring.
- **Multi-level and scalable security.** Appropriate levels of security and privacy must be built into hardware and software. A secure system must be trustworthy, assured, and resilient to cyberattack, theft, and counterfeiting. Systems also must be able to adapt over time to meet the dynamic nature of threats.
- **Next-generation manufacturing paradigm.** New processes, such as rapid three-dimensional, additive manufacture with control at the molecular and atomic levels may provide a cost-effective path to novel devices and architectures. In the longer term, the convergence of semiconductor technology and biology offers strong possibilities for disruptive new designs and processes. Manufacturing research is essential to the cost-effective transition of new technologies into practical applications.
- **Insight computing.** Future IT systems and infrastructure will have fundamentally new capabilities based on the creation of insight from data. Insight computing systems will leverage and substantially add to the capabilities of cyber-physical systems and the Internet of Things. Insight computing requires research in machine learning, data analytics, neuromorphic computing, and new approaches for user–machine interfaces.

- **IoT test platform.** A test platform is needed to adequately model the bourgeoning complexity of the IoT. Without such a test platform, solution verification and benchmarking is not possible. Such a platform should be accessible to researchers from academia, industry, and government.

Exploratory research indicates the significant potential for new devices, circuit and system architectures, and algorithms to take sensors, communication, actuation, and computing far beyond the limits of today's technology. However, we are at a tipping point. We must lay the foundation for the leap-ahead technologies of the future to realize the vision of a new wave of information and insight technology.

The U.S. semiconductor community—including government, industry and academia—will be able to take these critical steps only through partnership and focused funding. A National Computing and Insight Technology Ecosystem initiative will support development of an aggressive research agenda and a leap forward in new knowledge. Together, the community must exploit the rapidly developing opportunities to reboot, expand, and extend the IT revolution, and thereby ensure the United States of robust, long-term information technology leadership (Figure 1).

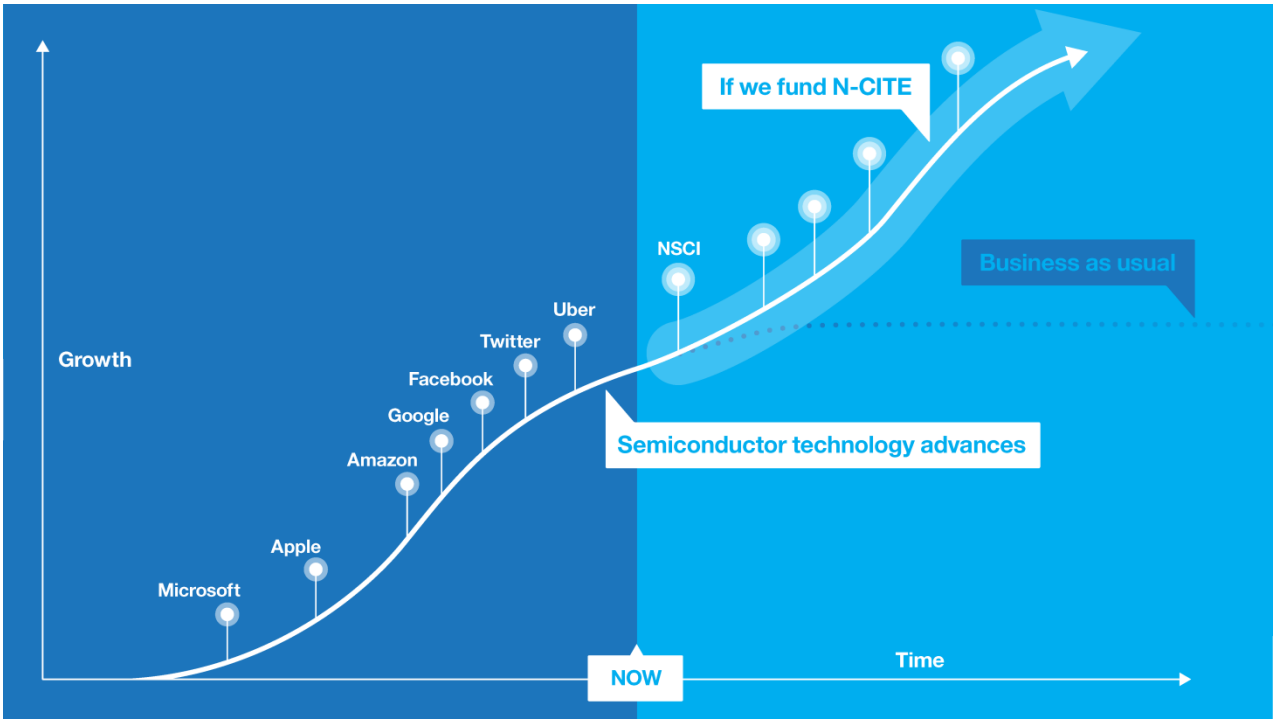

Fig. 1. Growth and advances in semiconductor and information technology have made possible new industries and business models, from personal computing, e-commerce and internet search engines to social networking and the online sharing economy, represented by the companies shown. New research investment in the proposed N-CITE initiative is needed to drive additional growth and support future innovation. N-CITE is aligned with the National Strategic Computing Initiative (NSCI).

The Technology Imperative

Over the last five decades, advances in semiconductor-based electronics have been incorporated into systems that go well beyond computing per se, impacting sectors ranging from aerospace and entertainment to medicine and manufacturing. We rely on semiconductors in almost every aspect of our lives. Over those fifty years, the ability to reduce the size of the individual transistors by half roughly every 18 months has led to increased performance at lower cost and greater functionality in ever smaller form factors, a trend known as Moore's Law.

Semiconductors are the foundation of information technology and have made possible the Internet, online businesses, and social media—connecting people and information across the face of the globe. The information economy is creating vast quantities of data that are growing exponentially; the volume of data created annually is expected to be nearly 30 exabytes by 2017 [1] (Figure 2). The expanding volume of data is paralleled by surging demand for storage (Figure 3). Increases in data also are leading to rising demand for bandwidth; however, the rate of growth in data is outstripping the growth in bandwidth by a factor of two.

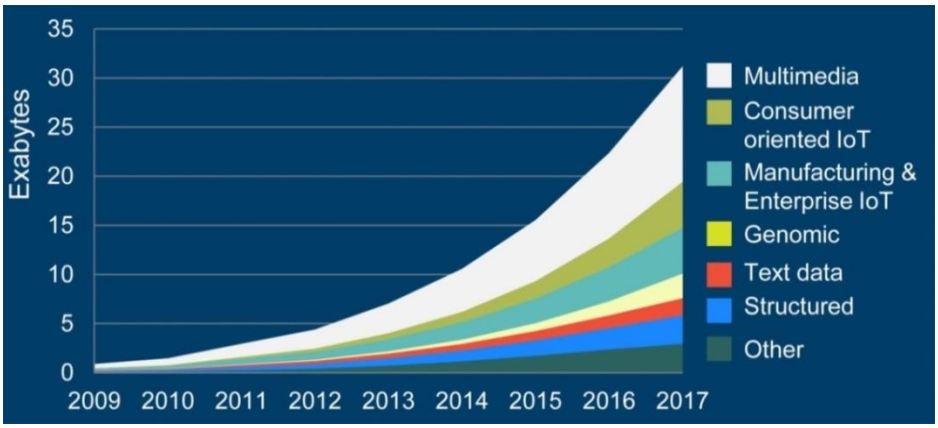

Fig. 2. Estimated data growth in various categories and overall.[1]

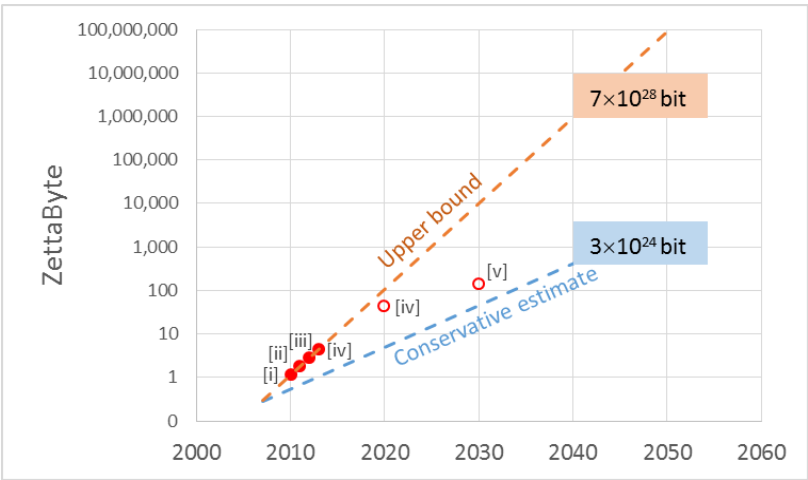

Fig. 3. Estimated and projected global memory demand, including a conservative estimate and an upper bound.[2]

As integrated circuits get smaller and cheaper, they become more ubiquitous, embedded in myriad systems and objects—adding "intelligence" by sensing, communicating, and actuation. They also are increasingly networked, creating the so-called "Internet of Things" (IoT). There are currently an estimated 50 billion sensors connected to objects that are part of the IoT. Some 14 billion of these already are collecting and sending data on the Internet; by 2020, there are expected to be 32 billion connected devices.[3]

At the same time, the capabilities of leading-edge high performance computing systems continue to increase. These powerful machines combine large numbers of the fastest processors to address challenges previously beyond reach. High performance computers are used extensively in applications as varied as nuclear stockpile stewardship, weather prediction, forest fire management, and fundamental materials research. Areas of intense interest and research today are data analytics and machine learning. Advances in these fields transform massive amounts of data into insights used to predict, anticipate, inform, and advise.

The convergence of these trends poses both opportunities and challenges at the infrastructure level. "Insight technologies" include everything from intelligent sensor nodes to scalable data centers. Delivering both efficiency and performance in these technologies will require a hierarchical, tiered architecture. This new IT infrastructure also will require advances in wireless communication hardware. And at every level, there must be appropriate security, privacy, and assurance. The sections that follow describe essential components of the next IT revolution, that is, the insight technology revolution.

Intelligent Sensor Nodes

Sensor technologies are experiencing exponential growth, with future volumes driven by applications not yet envisioned, as depicted in Figure 4. Further development of extremely small-scaled intelligent nodes is critical to realizing concepts such as ubiquitous computing and the Internet of Things.

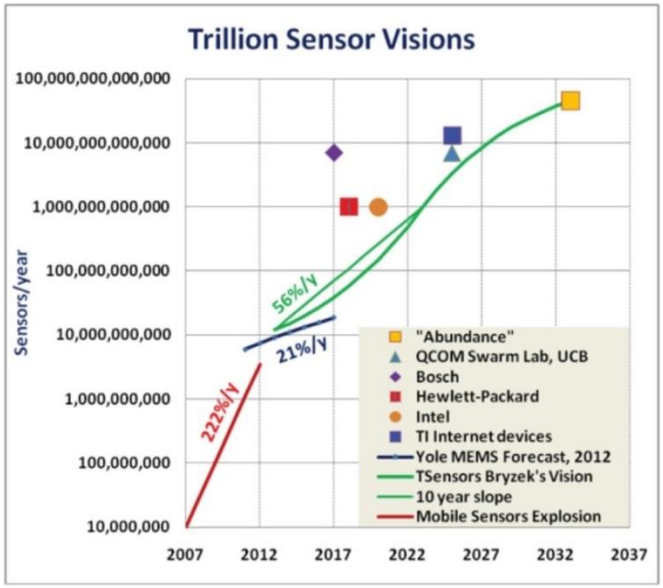

Fig. 4. Forecasts by various organizations for increases in the number of sensors deployed per year.[4]

Size reduction in sensors falls within the trend of general compaction of information, communications, and telecommunications (ICT) system scaling known as Bell's Law, illustrated in Figure 5.[5, 6, 7] For example, the volume of a general-purpose computing unit has scaled down from ~6.53×10^8 cm^3 in 1958 (IBM 709) to ~80 cm^3 in 2007 (iPhone1) to ~2 cm^3 in 2014 (Intel's Edison chip). A related trend is dramatic unit cost reduction derived in part from an exponential increase in production volume.

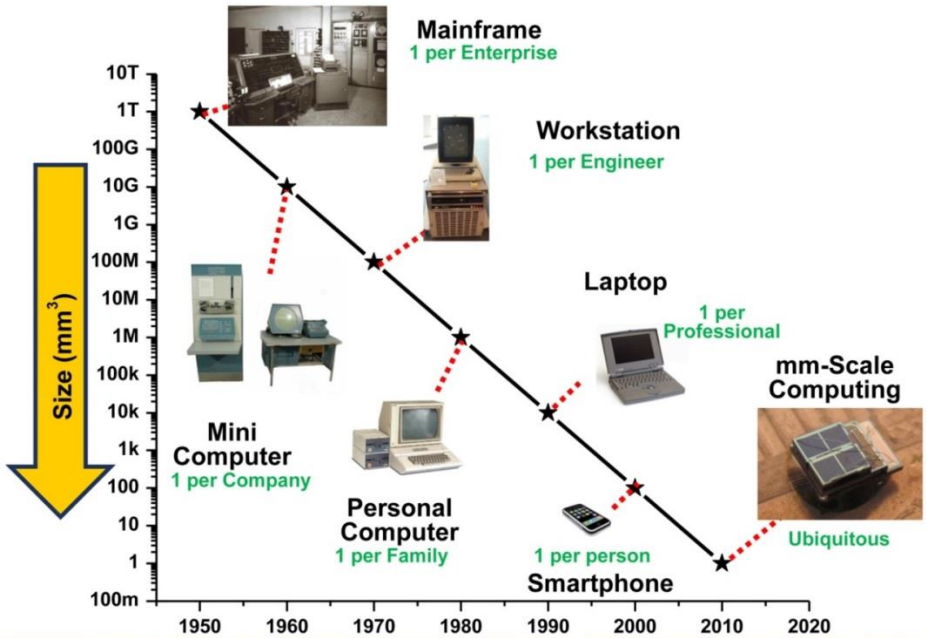

Fig. 5. Illustration of trends over the last 60 years in ICT systems: A new class of computing systems emerging every decade (Bell's law),[6] devices getting 100 times smaller every decade,[8] and the number of units produced and used rising as size and price fall and functionality increases.[7] (Figure courtesy of D. Sylvester)

Node Characteristics

An individual sensor node should include components for transduction, computation, memory, communication, and power supply. Nodes may need to act autonomously, be programmable via the network to increase flexibility, and support communication with other network nodes—all while maintaining security and privacy. In many applications, nodes will be embedded in the environment and should not interfere with other functionality. Moreover, many autonomous nodes will be constrained by size and energy; therefore smart energy management will be needed, for example, to support periodic bursts of operation. In the case of small-scale "nano-nodes" that are not connected to external power, the available volume for on-board energy storage will be extremely limited.

The smallest node that has been reported is ~ 10 mm^3,[7] and the challenging target of demonstrating a fully operational node 1 mm^3 in size has yet to be achieved. A nearer-term application driver could be intelligent image sensors with full scale communication capabilities.[9] It should be noted that while the scaling limits of individual electronic devices have been estimated from physics-based considerations, the question of how small a system can be and still offer useful functionality remains open. Research is needed to understand the possible physics-based estimates for the smallest-size intelligent sensor nodes.

Typically, nodes need to operate without connection to outside power supply, relying instead on batteries or harvesting energy from the local environment. Requirements for battery storage capacity and rate of power delivery can place severe constraints on system design. The theoretical upper bound on power densities for electrochemical batteries is ~10^4 J/cm^3, and modern practical batteries provide up to 550 Wh/L ≈ 2×10^3 J/cm^3.[9] Therefore, room for further improvement is limited. Energy harvesting concepts, even in the best case scenario, yield orders of magnitude less power than electrochemical sources.[7, 10] As a result, intelligent nodes can be realized only if significant energy efficiency improvements in sensing, computation, and communication are achieved. Recent progress in low power circuits is impressive; use of analog-to-digital converters has increased by an order of magnitude in the last six years. Radio and microcontroller units also show significant gains.[7] Examples of reduced energy use in modern ICT are shown in Figure 6.

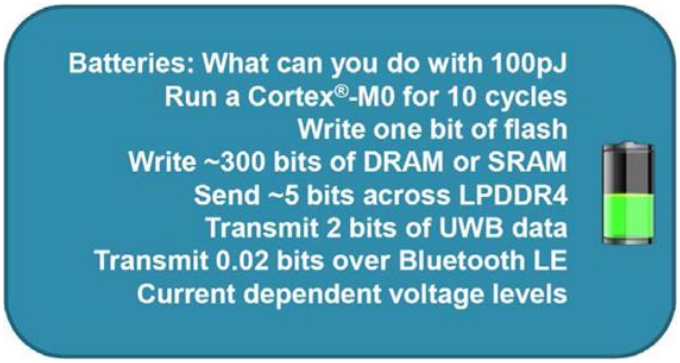

Fig. 6. Examples of energy uses in modern ICT.[10] (© & courtesy, ARM, The Architecture for the Digital World)

Materials – Nonsilicon Devices

Another pertinent challenge in the area of sensors is materials. While most of today's sensors have been implemented using silicon, this may not be a sustainable approach if billions of sensor nodes with short lifetimes and disposability are needed. Research is needed regarding more sustainable materials, such as carbon-based systems that are biodegradable or can be reused or recycled (as in nature), or potential roles for alternative materials such as polymers, paper, or cellulose.[11, 12] New materials also have the potential for novel approaches to sensing—from bio-based materials to other physical or chemical sensing materials.

Manufacturing Needs

A major challenge is fabrication of ultimately compact, heterogeneous hardware that includes sensors. In the nearer term, dramatic assembly and packaging innovations are required, including automation, parallelization, thin wafer handling, and 3D system modeling.[13] In the longer term, a new paradigm for semiconductor manufacturing must be created, because the existing mainstream planar technology is not well suited to the manufacture of heterogeneous systems 1 mm^3 or smaller in size. As space becomes limited, 3D rather than planar 2D architectures allow for more compact designs. In addition, due to space limitations, discrete components are not practical.[7] The widespread use of new additive manufacturing technologies is anticipated, inspired by the recent successes in 3D printing. Ink-jet printed radio frequency (RF) antennas and other passive components with typical feature sizes of a few

microns have been demonstrated; further improvement of this technique is expected to yield feature sizes as small as 100 nm and reconfigurable structures that respond to the environment.[11] Three-dimensional fabrication below 100 nm will require radical advances. Examples include atomic-level 3D printing methods, biological approaches such as programmable "cellular factories" using engineered microorganisms, and programmable DNA-directed self-assembly.[14]

Hardware vs. Software

Sensor node intelligence and functionality stem from both hardware and software components. Hardware offers efficiency and reliability but relative inflexibility and higher cost. Software, on the other hand, is flexible, provides more functionality, and is relatively low-cost; however, it is slower, more error-prone and more susceptible to attack. There are many hardware innovations that could abet intelligence in sensor nodes, for example, low-power switches, low-power high-density nonvolatile memory, small-form-factor solid-state oscillators, efficient sensors, and hardware-based security. These innovations could enable new sensor node platforms with functions that now are implemented in system software. This would represent a shift of the hardware/software functional boundary to include more responsibility in the hardware systems.

Future Directions

The application space today is quite different from just a decade ago. For example, sensor networks may have a higher-efficiency, lower-performance target than mobile hardware platforms. Existing devices could be optimized differently, e.g., near-threshold operation, for better performance. Emerging devices with novel capabilities (e.g., nonvolatility in logic gates, and steep subthreshold slope) can provide even greater advantages, especially when optimized from a system perspective.

There is remarkable growth in the range of functionalities of sensor nodes today, mostly based on existing device and manufacturing technologies. With more advanced hardware components, even more capable and sophisticated sensor systems could be developed. Advances are needed in high efficiency sensors/actuators, ultralow-power logic and memory, sustainable energy supply, and scalable communication units, to name a few. Research is required to understand the fundamental science and engineering underlying these opportunities.

Always Connected: Billions of Connected Nanosystems

From Sensor Nodes to the Cloud

The emerging IoT will consist of billions of sensor nodes, enabling deeply interconnected systems that anticipate, adapt, and control, while being autonomous and dependable.[15] Massively connected systems of autonomous sensor nodes could one day form a ubiquitous, embedded computing platform. Such sensor networks will require unprecedented compute and communication capabilities[16] and sophisticated management of complex flows of information. A nearer-term application of such systems is next-generation industrial electronic control systems that have a real-time set of operational constraints.

Cloud-based platforms show considerable promise for the operation of certain intelligent sensor networks.[17, 18] Cloud platforms will support the assessment and monitoring of sensory data at runtime,

and also are affordable and a proven, reliable method of information processing.[18] Opportunities will arise for sensor-assisted cloud processing in which feedback generated from the sensors can be used to tune and focus front-end signal processing and filtering.[17] However, cloud-based platforms may not be suitable for applications requiring extremely fast response times or the utmost reliability.

Wireless Communications

An essential element of massively connected systems is communications. Unfortunately, neither existing technologies nor current deployment models will be able to support the skyrocketing demand for communication, especially in the wireless sector.[19, 20] Demand for wireless connectivity is expected to grow rapidly (Figure 7), while the world's capacity to broadcast and telecommunicate information is projected to rise to ~ 10^{23} bit/year in 2040 (see Appendix A, Section A3).

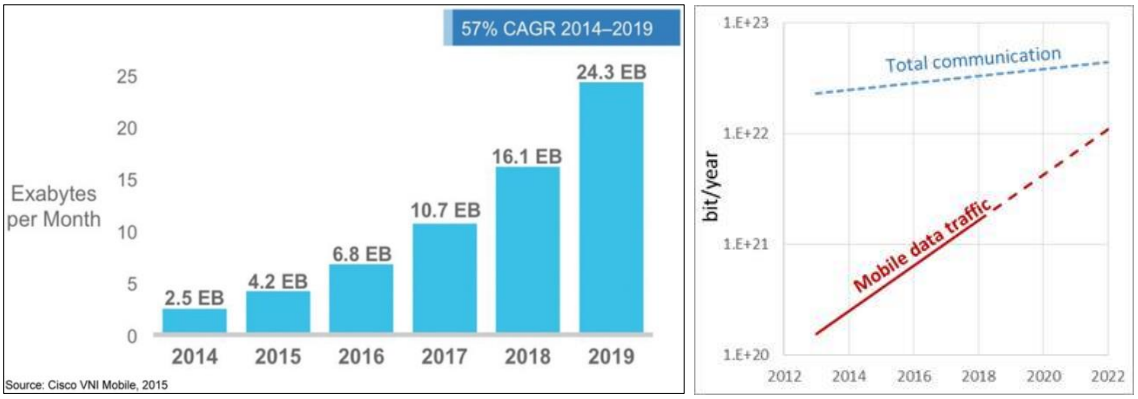

Fig. 7. (*Left*) Annual growth of mobile data traffic;[19] (*right*) mobile data traffic vs. total communication.[2]

Trends illustrated in Figure 7 may in fact underestimate the anticipated "data deluge" resulting, in good measure, from massive deployment of various sensors. Today, the number of sensors deployed exceeds 50 billion, and it is expected to grow exponentially, as depicted in Figure 4. Depending on the nature of the sensors and systems, the demand for communication and connectivity could potentially increase considerably.

Sustained growth in wireless communication capacity demands both new communication technologies and high quality, affordable network access. New wireless technology will be needed, including new devices, materials, and process integration. [21] One approach that could help meet these demands is self-organizing, fully wireless networks that grow and improve organically with the addition of extra hardware at low incremental cost.[22]

It is envisioned that the mm- and sub-mm spectrum (0.1–10 THz) will be utilized for communication in the future, [21, 22] providing an unprecedentedly large bandwidth able to support ultra-broadband links, such as terahertz virtual wires (TVWs).[23] Advances in (sub)terahertz technologies likely will require new nanomaterials, such as graphene, that exhibit novel physical phenomena, including in their electronic, photonic, and plasmonic characteristics.[23]

Advanced antenna technologies are among the most critical enabling components for future wireless systems and networks.[15, 22] Effective and innovative use of antenna arrays could dramatically improve

the capacity of wireless channels: the concept of antenna swarms or seas of antennas would allow highly flexible directional wireless links, thereby forming a hierarchical mesh network with dynamic management of traffic using all available parameters: frequency, space, and time.[22] Swarms consisting of trillions of deployed antennas offer new opportunities for collecting intelligence from networks of interacting intelligent nodes; realization of such concepts will require development of low-cost manufacturing and deployment solutions such as printed wallpaper antennas.[22]

Infrastructure Technologies for Big Data

Future information and communication technologies (ICT) will generate enormous amounts of data, far surpassing today's data flows. Much of the data will be collected, analyzed, and stored in data centers, driving exponential growth in demand for computing, communication, and storage capabilities (see Appendix A, Section A1). In the past, expansion of storage and computation capabilities was based on aggressive feature scaling, manifested in Moore's Law. However, scaling will not be able to address the upcoming needs in ICT performance and utilization of energy resources. Therefore, radically new technologies for energy-efficient analysis and storage of massive volumes of data are needed in the face of a growing flood of data.[24] New classes of data processors and underlying technologies are needed to provide actionable insight and inferences from the data, which cannot be handled by traditional methods of storage and computation.

Examples of Big Data

An example of technology that will lead to Big Data is the Square Kilometer Array (SKA),[25] a radio telescope to be built in 2018–2030 for studies of galaxy evolution, dark energy, gravitational waves, etc. SKA will contain thousands to millions of receiving antennas spread over thousands of kilometers and will generate up to 10^{20} bit of data each day.[26]

Defense applications are another significant driver of Big Data technologies, including data related to intelligence, surveillance, and reconnaissance (ISR).[27] For example, today's Global Hawk unmanned ISR aircraft system generates 500 Mbps, five times the bandwidth used by the entire U.S. military in the Gulf War. Another example, the ARGUS (autonomous real-time ground ubiquitous surveillance) aerial imaging ISR system, outputs 1,854 Gbps and uses a common data link of 274 Gbps, a ratio of 10^3 between need and capability, even with 10:1 compression.[27] Finally, the "poster child" of Big Data is the Internet of Things, in which billions of online sensors will generate and transmit unprecedented amounts of data from which intelligence must be derived, often in real time.

Data Centers

To manage data flow and storage, data centers continue to grow. Each data center containing 50,000 or more servers costs from $50 million–$1 billion to build and consumes on the order of 20–50 MW. Cloud data centers are typically realized in standard, custom, or commodity configurations.[26] Input data rates are in the range of 100 Tb/s over proprietary fiber, and other networks and facilities are provided to cache and manage the latency of this data. A cloud data center also must provide resilience at scale via failure detection and center level management processes.

Big Data vs. High Performance Computing

Big Data and high performance computing (HPC) have many technical commonalities and are likely to evolve to become more similar, yet there also are differences. HPC, which has been developing for decades, focuses on computationally intensive tasks. Big Data deals with large amounts of data of various types and is a topic that is emerging in many fields and application areas, such as genetics and astronomy. Big Data uses massive storage networks, whereas HPC uses local storage; also, they each use different software development tools and are guided by different cultures and expectations. HPC also encompasses science and engineering computing that is primarily model-based and offline (although cyber-physical computing is a notable exception), whereas Big Data cloud servers are transaction-based and operate in nearly real time. In the world of software, abstractions have been the key to success, yet metrics like operations/joule drive software specialization, not generalization. The conflicting pressures between the model-based, specialized software for HPC versus real-time computing and generalized software typical of Big Data need to be addressed through further research.

Technology Drivers and Needs

A new conceptual abstraction is needed for emerging information processors, but the most commonly examined model of computation, the Turing model,[28] is deeply embedded in computer science and engineering culture. Effecting a radical departure from current processing hardware is difficult. Possible alternate computing models include analog "approximate computing" and quantum computing. A juxtaposition or hybridization of computing models may drive the development of brain-like computation, which could be a significant component of future hybrid systems.

In the nearer term one can expect the emergence and widespread use of micro-servers and cloud appliances. Micro-servers use very small chips to produce smaller, lower-cost, lower-power servers[29] that may be capable of operating with renewable energy sources. Cloud appliances cache data as it is moved to and from the cloud to enhance performance.[30] It should be noted that use of these systems would entail data flows crossing jurisdictional boundaries and thus provoke many legal and/or privacy issues.

In many cases, data acquisition is relatively easy, but maintenance of "cold data" is more difficult. Increasingly, data are stored in mediums such as magnetic disks, which are cost-effective but slow to access. Future Big Data systems will benefit from research and development in areas such as nanophotonics for communication and new memory/storage devices (ReRAM, PCM, FeRAM, etc.) that address data movement energy costs and access latency.

A further conceptual breakthrough is required for data storage solutions, given the collision of ecosystems driven by competing technical and business factors, e.g., general-purpose computing versus accelerators; special-purpose computing versus radical alternatives; and systems aimed at volume sales versus high margins and value-add. The greatest opportunities lie at the extremes: on one end are cloud data centers and exascale HPC, while on the other are mobile and IoT prospects. Advances may arise from quantum computing or superconducting devices in data centers and neuromorphic/analog mobile devices.[26]

New Architectures

Radically new processor–memory architectures must be developed to address the current mismatch between processor speed and memory access time. The significant energy costs in time and memory to move data will be one of the most difficult barriers for future HPC. As a goal, novel memory management technologies could enhance programmer productivity by using programming methods promoting data locality, by auto-tuning to eliminate the need for low-level optimization, and by managing software layers to ensure reliability.[31]

The bandwidth-power dilemma in today's computing is due to the mismatch in transfer times between the various memory layers, e.g., in transfers from CPU to RAM, from RAM to disk, and from disk to tape.[32] One approach to managing the disparity among transfer times is to add solid-state devices between the RAM and disk memory systems, thus increasing memory hierarchy depth. It may be that in the future, storage-class memory such as STTRAM, PCRAM, and RRAM will also be inserted into the memory hierarchy between the RAM and disk memory systems. Finally, innovations in memory structure such as the hybrid memory cube could augment the RAM system.

Processing-in-memory (PIM) might offer the best solution to the memory bandwidth issue.[32] The idea is to greatly reduce or eliminate data movement in order to unleash the raw bandwidth of memory at a power envelope not heretofore achievable. The central concept is to locate a bit element on every bit-line to afford massively parallel and scalable computing technology. It is estimated that the resulting very low data movement would lead to operating energies in the femtojoule range. PIM architecture, which is related to automata theory,[33] would be extendable to other emerging memory technologies. A key to its success is in firmware and software.

Finally, at the system level, hierarchical, tiered networks for storage, management, and analysis of data could greatly improve the overall energy efficiency, as well as speed and performance. A more distributed architecture also could provide greater security by decentralizing data.

Brain-Inspired Computing

The rise of Big Data is creating huge new opportunities, but the challenges in dealing with the volume, velocity, variety, and veracity of data are straining current systems and architectures. Radically new computer architectures are needed, perhaps that operate in a manner similar to how the brain functions, in order to process large amounts of data, particularly unstructured data. The neocortex, nature's cognitive processor, is an engineering marvel, comprising a folded 2D planar sheet with a six-layer structure. If unfolded, the brain has an area of about 0.7 m^2 and is 3-4 mm thick. It has approximately 10^{10} neurons, each of which has about 10^4 connections. It requires about 25 watts of continuous power and "computes" at an equivalent of 25 PetaOp/sec.[27] Figure 8 offers a comparison of the brain and a conventional inorganic serial architecture computer. What is needed is a model for brain-like computation for which appropriate technologies could be developed.

New applications call for new approaches. In many cases, the value is not in high-precision right or wrong answers based on an entire data set, but rather, in finding interesting patterns and connections leading to new solutions and insights. The differentiators will be how fast one can find a "better" idea, not an exact conclusion.

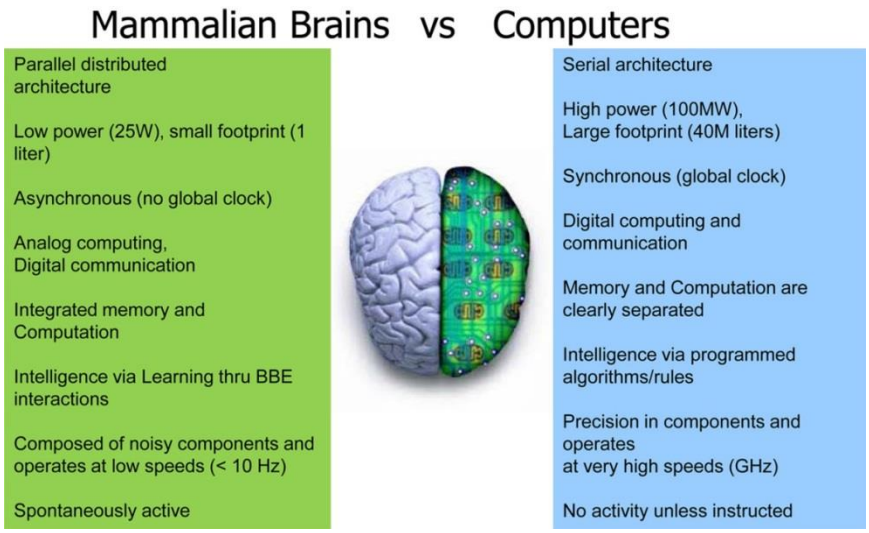

Fig. 8. Comparison: Why can't today's computers process data like the brain?[27]

The brain provides useful guidance: it trades off raw speed, computational capability, and precision for much lower power, rapid decisions, and (when we are lucky) astonishing insights. While mimicking the brain structure completely is probably unnecessary (even nature takes 9 months to build the human brain and some 18 years to program it sufficiently to make it useful), experimenting with computation architectures that are slower, lower in power, densely interconnected, and more tightly integrated with memory is an approach that already is showing promise.

As the demand for more energy-efficient, yet more powerful computing grows, new approaches such as brain-inspired computing have the potential to transform the way systems are designed and manufactured. A device that integrates memory and logic, or enables higher-level functionality (potentially sacrificing accuracy) at lower power, may be a fruitful choice. Device technology research needs to be conducted in concert with circuit research to insure the right trade-offs are made to achieve the best system solution. In the end, we should not seek to replace the precision and computational performance of our existing systems, but rather to complement those systems with the ability to find new insights and solutions in an increasingly power-constrained, mobile world awash in data.

Elements of an Intelligent Infrastructure

The ultimate intelligent infrastructure will comprise pervasive and intelligent sensors connected to each other, to local networks, and to cloud-based systems. This complex, hierarchical interconnected system of systems should be capable of analysis to provide insight and actionable advice or control. Although progress is being made in some areas, there are enormous challenges that must be overcome to realize the full potential of such a system. Many of the technologies that exist today simply cannot be adequately scaled. For example, transmitting every piece of data to the cloud for analysis and storage is neither feasible nor practical.

Insight Computing Systems

A new era of computing is emerging, moving from the era of programmable systems to the era of "insight computing" systems. Attributes of insight computing include context and learning, visual analytics and interaction, software-defined environments, data-centric systems, and nano- and atomic-scale components. Sensors are critical in such systems, collecting the data from which insights are generated. For certain real-time applications, the value of sensory data may be brief, perhaps only a few milliseconds, and the data must be utilized within that time frame.

Industry is making strides toward advanced dialog, reasoning, image recognition, and breakthroughs in knowledge representation and machine learning. With sufficient research, these advances could lead to neurosynaptic computing, quantum computing, and design of brain-inspired systems within a decade.

Cyber-physical Systems

Cyber-physical systems will be an important component of the upcoming IT revolution. According to the National Science Foundation, "cyber-physical systems (CPS) are engineered systems that are built from, and depend upon, the seamless integration of computational algorithms and physical components…CPS technology will transform the way people interact with engineered systems."[34] A CPS contains networks of collaborating computational elements that control physical entities. They are fundamentally different from computational systems and from physical systems, and they require new engineering models that embrace temporal dynamics and algorithmic computation; indeed, a CPS must have robust control in the face of very dynamic, physical systems that often act randomly. Additionally, they must satisfy security and safety constraints.[35]

Security at the Chip Level

As reliance on information systems continues to expand, security is a growing concern. At the same time, a number of trends are increasing the vulnerability of systems to an attack—whether economically or politically motivated. The design and manufacture of integrated circuits and systems involves a global supply chain of many businesses and hundreds of engineers, creating opportunities for malicious tampering at a number of points along the path. Complexity of high end processors continues to grow, and designers are using more and more third party designs, or IP, that may be unvetted when it comes to security. Counterfeit parts that may not be what they purport to be are entering the supply chain. Reverse engineering of existing products exploits the intellectual property of the semiconductor producers and results in fraudulent products that may not perform as intended. Perhaps the greatest risk is from designs that provide unintended access to data. The current verification process is primarily aimed at insuring that a design does what was intended. It does not ask, "Does it do anything else?"

Security is projected to become an even bigger challenge in the future as the number of interconnected devices increases. Intelligent chips for anticipatory sensing will be embedded and integrated in everything from space-based platforms to the human body, making data and information security and assurance critical. In fact, the Internet of Things can be viewed as "the largest and most poorly defended cyberattack surface conceived by mankind."[36] In addition to the standard requirements of low cost, simplicity, and low power for the IoT, a multitude of security requirements must be addressed. A short list of requirements includes tamper resistance and secure communications and storage.

Critical Research Challenges

As outlined above, future IT technologies and systems are facing significant practical challenges that could limit their growth if left unaddressed.[37] Radically new information and communication technologies will be needed to analyze, store, and enable decision making based on increasingly massive data streams. Research is needed in a host of foundational technologies.

Energy-Efficient Sensing and Computing

Greater energy efficiency is mandatory in order to fully realize future IT capabilities. Without progress, power demands at the transistor, chip, and system levels will become prohibitive. Energy efficiency is vital at all levels—from the smallest sensor to ultrahigh performance processors and systems.

Industry's ability to follow Moore's Law has led to smaller transistors but greater power density and associated thermal management issues. More transistors per chip mean more interconnects— leading-edge microprocessors can have several kilometers of total interconnect length. But as interconnects shrink they become more inefficient. Today more than half of the energy used by a processor is to move data via interconnects, e.g., between memory and logic. Conventional approaches are running into physical limits. Reducing the "energy cost" of managing data on-chip requires coordinated research in new materials, devices, and architectures. Whereas in the past these topics have been addressed separately, meaningful progress will require a multidisciplinary and coordinated approach.

Energy efficient sensors are essential to the IoT. Distributed sensors with significant on-site "intelligence" and communications ability that allow decisions and actuation to be handled locally will reduce the energy needed to manage the anticipated volume of data. Remotely located sensors will need to be securely accessible for calibration and validation checks, and if they are stand-alone for long durations, they must have mechanisms for efficient energy collection and utilization. Moreover, intelligent sensor devices could be networked to perform collective tasks, creating distributed information computing and storage capacity on demand and further reducing energy consumption.

Fundamentally, there must be a new conceptual model and new supporting architecture to enable processing the vast Internet data flows and to enable extracting and delivering timely insights. This new technology and architecture needs to be several orders of magnitude more energy efficient than best current estimates for mainstream digital semiconductor technology if energy consumption is to be prevented from following an explosive growth curve.

Additionally, there is a critical need for miniature energy sources (both batteries and harvesting) with intelligent power management. In order to minimize the volume of transmitted data, sufficiently high local nonvolatile memory density will be needed at the sensor node.[15]

The core technology that is in high performance computers and large data centers also must be made more energy efficient if more capable systems are to be built, e.g., beyond exascale computing. The most cost-effective solution would be a single technology that could be implemented in high performance, mobile, and embedded sensor systems. However, optimal solutions in such diverse

applications likely will take different approaches. Therefore, multiple technologies should be further explored.

Focus areas for research should include extremely low power devices for sensors and advanced processors, novel energy-harvesting devices, novel circuit architectures that reduce data movement, explorations to overcome physical limits due to noise fluctuations without significant power trade off, and nanophotonics and integration of photonics on-chip. Next-generation electronic design automation (EDA) tools are needed, including improved models for sensors, processors, and systems, and validation techniques that measure the effectiveness of energy reduction efforts. At the system level, research is needed to build intelligence and energy efficiency into the IoT at all levels.

Cyber-physical Systems

Cyber-physical systems are a special and critical case within the future IT infrastructure. To realize successful CPS, research must address a number of important design considerations:

- Design of robust, efficient, effective cyber algorithms that accommodate unanticipated events in the physical system
- Trade-offs between cost-effectively exploiting existing networked resources, such as cloud computing, vs. conceiving a stand-alone system independent of network resources
- Achieving high performance with minimal energy consumption
- Ability to operate asynchronously and also to respond collaboratively to effect system control
- Maintenance of repeatability of processes—for example, in a manufacturing environment—in the face of unexpected system dynamical behavior
- Assuring safety and security along with connectivity
- Trade-offs between open CPS architectures that exploit the benefits of many contributors vs. protection of data and intellectual property

Intelligent Storage

New memory technologies and management systems are needed to store and archive the explosion of data that is projected. Advances in data storage require research from the device to the system levels. These new memory systems will need to provide access to the data for processing and be capable of archival operations, and yet they must be managed in such a way that intelligent data purges can be implemented to reduce total storage volume.

New memory architectures, such as processor-in-memory, will be highly disruptive and will require hardware and software system-level research in order to optimize memory management.

Future intelligent chips offer unprecedented opportunities for combining the information computing and storage capacity of billions of distributed, networked systems. Fundamental understanding of scalability issues, as well as the design and operation of such massive, distributed systems is needed. Addressing challenges in this area may lead to a new field of Very Large System Science.

Real-Time Communication Ecosystem

Research and innovation is needed in communication network architectures and technologies so that it is possible to transmit and receive critical data quickly for analysis while preventing network traffic jams due to massive data transmissions. In addition to being fast, communication networks must minimize energy utilization and transmission errors at all levels, from network fringes to centralized processing systems.

The value of data will be critically dependent on the delivery time, with some data losing its value in milliseconds. Thus, research is needed in the areas of real-time data streaming and data analytics. No matter how fast communication becomes, the anticipated amount of data will be disruptive and overwhelm the entire IT industry unless new storage, communication, and processing technologies are invented.

Multi-level and Scalable Security

Security and privacy are two of the biggest challenges for future systems. The Achilles heel of current information technologies is their vulnerability to attack, which can lead to loss of data, compromised privacy, and potentially hostile exploitation. Research is needed at a fundamental level, far beyond cryptographic protocols for links and storage.[17]

The intense connectivity enabled by the IoT poses additional challenges. In order to be acceptable to users and to protect property and even lives, the IoT will require a very high level of immunity from attack. Otherwise, worldwide distributed systems of communicating intelligent chips could become the greatest and easiest target for malicious actions. Appropriate levels of security must be embedded in each piece of hardware, as well as in the network; security must be designed in, not bolted on as an afterthought. The limits of security need to be understood based on fundamental physics and information science.

IoT researchers and designers should explore answers to the following questions regarding IoT security.[36]

- How can security be defined and measured?
- Can hardware provide greater security at the software level?
- Must connectivity always imply vulnerability?
- To what extent should IoT data retention mechanisms incorporate data erasure?
- Are there analogs for social trust that apply to IoT design?
- Should security be regarded as a graded service and priced to end-users accordingly?
- To achieve a trustworthy IoT, are more standards needed?

Next-Generation Manufacturing Paradigm

Existing mainstream semiconductor manufacturing technologies do not support fabrication of the heterogeneous 3D systems needed for future intelligent chips. Fundamental research is needed in materials, fabrication, assembly, and packaging to enable new IT hardware—from the smallest sensor node to ultrahigh performance processors. Examples that have demonstrated promise include

semiconductor inkjet printing that could ultimately achieve 100 nm features, atomic layer deposition reactors reaching molecular scales, flexible and reconfigurable 3D structures, and biologically inspired technologies.

Current semiconductor manufacture requires large volume production to be economically feasible. The principle of hardware modularity, i.e., a distributed network of manufacturing cells in which new designs can be downloaded and produced, could enable much more rapid application-specific development of new hardware components. Manufacture of hardware modules that can be combined in novel or custom ways would support a diversity of innovation, including small volume product development, and could make possible hardware security "patches" for current systems. Modular manufacturing, combined with novel design and test tools, would allow businesses of all sizes to develop initially niche or small scale applications and to iterate designs quickly.

In the longer term, a new paradigm is needed for semiconductor manufacturing, because the existing mainstream planar technology is not well-suited for heterogeneous systems 1 mm^3 and smaller. At these sizes, space is the critical asset and the most compact designs are three-dimensional. In addition, discrete components are not possible under such severe space limitations.[7]

Manufacture of sensor/actuator systems that generate measured data and/or respond to actuation commands pose particular challenges. Today, the fabrication processes and the packages for sensors are typically specific-sensor-dependent, which is inefficient and costly. Research is needed into new approaches to manufacturing and packaging for sensor/actuator systems.

Another significant manufacturing challenge regarding sensors is the robust and economic fabrication of multi-die assemblies. Different essential components of the intelligent sensor node such as CMOS circuitry, MEMS transducers, Flash memory, and energy storage/harvesters are fabricated today using separate processes. No single ubiquitous process exists analogous to CMOS for digital systems.[17] In addition, new packaging technologies will be needed to address small form-factor devices and sensors.

Finally, the materials base for ultimately compact hardware such as intelligent sensor nodes presents a major challenge. Today's uses of silicon may not be sustainable if up to a trillion sensor nodes are deployed, especially if many are designed to have short lives and to be disposable. We need more sustainable materials, for example, carbon-based systems that can be recycled or reused, as in nature. The potential role of alternative materials like polymers and paper in fabrication of integrated systems needs to be explored.[11]

Insight Computing

Future IT systems and infrastructure will have fundamentally new capabilities based on the creation of insight from data. Insight computing requires a paradigm shift from bit generation to insight generation and from raw data to actionable intelligence. Insight computing systems will leverage and substantially add to the capabilities of cyber-physical systems and the Internet of Things.

Insight computing requires research in machine learning, data analytics, and neuromorphic computing. Novel algorithms for addressing complex problems and new approaches for user–machine interfaces are needed. Ultimately, insight computing requires advances in all of the areas outlined in this report, from

energy efficient sensing to communications and novel manufacturing. In addition, there is need for coordination among researchers, for example, in materials, devices, and architectures, or in software and hardware. Advances in each of these areas will depend critically on coordination among the other areas, which have not been integrated traditionally.

IoT Test Platform

Due to the bourgeoning complexity of the IoT, there is a significant requirement for infrastructure to test new technologies. Today there is no representative test platform or system demonstrator to stand in for the anticipated IoT of the future. Without such a test platform, solution verification and benchmarking is not possible. Definition and support of such a system demonstrator will be required for rapid and meaningful progress. Such a platform should be accessible to researchers from academia, industry, and government.

What Will It Take? A New Coordinated Research Initiative

To reboot the IT revolution and to ensure the Nation of unassailable long-term leadership in information technology and information management areas critical to economic vitality and national security, a new targeted research initiative is required—a National Computing and Insight Technology Ecosystem (N-CITE) initiative.

The goal of the initiative should be to coordinate research in the areas outlined in this report:

- Energy efficient computing and sensing
- Cyber-physical systems
- Intelligent storage
- Real-time communication ecosystem
- Multi-level and scalable security
- Next-generation manufacturing paradigm
- Insight computing

The breadth of technical challenges that must be overcome to fully achieve the benefits of the Internet of Things and the massive amounts of data that are anticipated require a significant effort with support by both the public and the private sectors. To expedite critical science and engineering advancements and efficiently transition results to practical use, the initiative should include a government-industry-academic partnership.

The initiative must bridge multiple disciplines and areas of research that typically have not collaborated. Progress demands iterative and coordinated research in materials, circuits, and systems. Hardware and software will need to be co-designed. Security will need to be considered at multiple levels, from the chip to the application layer. Design should be done with manufacture in mind.

A next step is for government, academia, and industry to develop a more detailed set of research priorities based on the recommendations outlined in this report. Such a plan, or roadmap, can be used to assess gaps not adequately addressed by current research programs and can serve to guide future programs. The N-CITE initiative will provide a framework for interagency, as well as public-private, collaboration and coordination.

Leveraging, Sustaining, and Enhancing Existing Programs

The N-CITE initiative should build on existing programs supported by industry and government that focus on long-term research for information technology, semiconductors, and nanotechnology.

The National Strategic Computing Initiative (NSCI) established in July 2015 by Executive Order [38] aims to maximize benefits of high performance computing (HPC). The NSCI lays the framework for research and development of exascale computing systems that deliver approximately one hundred times the performance of current systems. In addition, looking even further out, NSCI is to create a path toward future HPC systems with capabilities that go beyond the limits of current semiconductor technology. These objectives correspond with those of the proposed N-CITE initiative.

The Federal government's National Nanotechnology Initiative (NNI) includes significant investments in nanoelectronics and nanomanufacturing research. In 2010, the NNI established a number of Signature Initiatives to focus research in "technology areas of national importance that may be more rapidly advanced through enhanced interagency coordination and collaboration," including Nanoelectronics for 2020 and Beyond and Sustainable Nanomanufacturing. The NNI is developing Grand Challenges to guide greater public–private engagement toward "audacious, yet achievable" nanotechnology-inspired goals.[39] Grand Challenges under consideration include:

- Create devices no bigger than a grain of rice that can sense, compute, and communicate without wires or maintenance for 10 years, enabling an "internet of things" revolution.
- Create computer chips that are 100 times faster, yet consume less power.

These Grand Challenges would directly support the proposed N-CITE initiative as well.

Other national efforts are targeting the convergence of biology and technology. The Administration's BRAIN (Brain Research through Advancing Innovative Technologies) initiative, with its focus on better understanding the human brain, may contribute to the development of brain-inspired computing. In addition, the semiconductor industry is leading the development of a roadmap of technologies at the intersection of electronics and advanced biology.[40] The five themes of the roadmap are: (1) DNA-based Massive Information Storage, (2) Energy Efficient, Cell-inspired and Cell-based Physical and Computational Systems, (3) Intelligent Sensor Systems, (4) Biological System Design Automation, and (5) Biological Pathways for Semiconductor Fabrication and Integration.

The proposed N-CITE initiative complements these interagency and industry-led activities. Further, directly stimulating development of next-generation insight computing systems will provide capabilities that will support many U.S. government agencies to accomplish their respective missions, as indicated in Figure 10.

Fig. 10. Potential N-CITE applications corresponding to missions of selected agencies.

A Public–Private Partnership Model

Achieving true insight computing systems that will be capable of addressing government needs as well as commercial applications will be greatly accelerated through government-industry collaboration. The semiconductor industry has a long history of working with government to identify and support critical platform technology research that has helped keep the U.S. semiconductor industry competitive and the world leader. The Semiconductor Research Corporation (SRC) was created in 1982 to manage this type of research on behalf of and with support of the industry. Today, SRC comprises a broad portfolio of research programs, many of which are supported through partnerships with Federal research agencies. The Semiconductor Industry Association, formed in 1977, provides public policy issue development, education, and advocacy for initiatives that advance semiconductors through research and technology.

For example, DARPA and MARCO (a subsidiary of SRC) jointly fund the Semiconductor Technology Advanced Research Network (STARnet) program, including the TerraSwarm Research Center focused on pervasive integration of smart, networked sensors and actuators. The Nanoelectronics Research Initiative (NRI) is a consortium of SRC member companies that collaborates with the National Institute of Standards and Technology (NIST) and the National Science Foundation (NSF) to focus on discoveries that will propel computing beyond today's technology. In addition, SRC and NSF have jointly funded numerous targeted research efforts in areas such as Secure, Trustworthy, Assured and Resilient Semiconductors and Systems, Failure Resistant Systems, and Multi-core Chip Design and Architecture. Finally, SRC is leading the development of a consortium and roadmap at the convergence of semiconductor technology and advanced biology with support from the NIST Advanced Manufacturing Technology Consortium (AMTech) program.

The N-CITE initiative will require strong coordination among government and industry partners in order to address the research challenges outlined herein. The semiconductor industry model not only provides funding for research that benefits the entire industry sector, it enhances university research and student education, informs Federal research programs, builds public support for research, and removes barriers to technology transfer from "lab to fab".

In addition to building on existing research programs, the N-CITE initiative also will leverage related activities such as the IEEE's corporate-wide initiative on Rebooting Computing.

Conclusion: A Vision of Future IT Technology and Infrastructure

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. The United States needs 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.

Appendix A: Data Explosion Facts

A1. Storage Needs in 2040

Global trends in information storage noted in this report are based on research by Hilbert and Lopez (2011),[41] where a detailed inventory of all storage media was created. A summary of their findings is shown in Figure A1, from which several observations can be made and include: (1) the majority of data was analog before 2002, (2) the analog data reached a maximum around 2000 and steadily decreased afterwards, (3) the amounts of stored analog and digital data became equal around 2002, and (4) after 2007 the vast majority of information became digital—a trend that continues today.

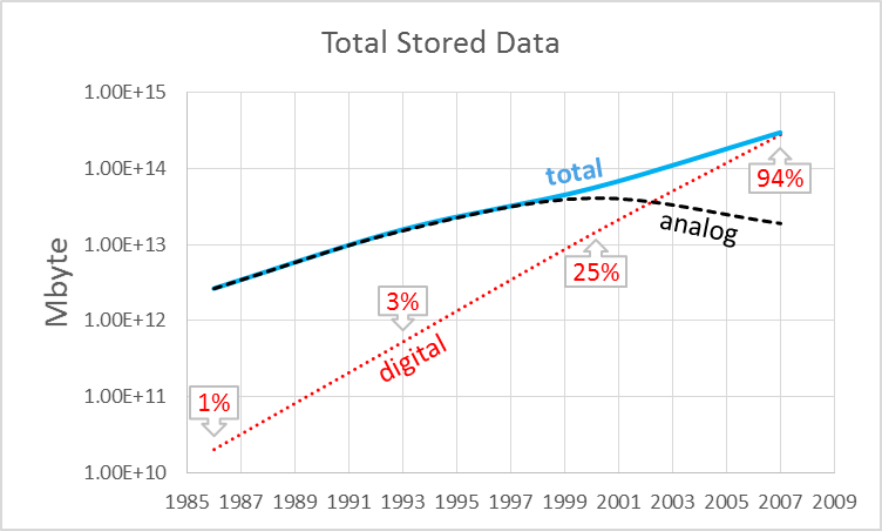

Fig. A1. Timeline of analog and digital data storage including percentages of digital data with time.[2]

Extrapolation of the digital line in Figure A1 provides projections for required global data storage. An important caveat is that the growth rate of digital storage (red line) is considerablly higher than the total growth rate (blue line). This reflects analog data dominating the total storage capacity for the majority of the measurement period. Extrapolation of the digital line in Figure A1 is treated as an upper bound for storage, while the extrapolation of the total storage capacity is used as a conservative estimate.

Figure A2 shows the projections of global memory demand, both a conservative estimate and an upper bound. These extrapolations are compared to independent estimates of the data storage in 2010–2014 (solid red dots) and projected storage needs in 2020 and 2030 (open red dots) in Figure A2. All estimages and projections are within the defined boundries (formed by the “conservative” and “upper bound” lines).

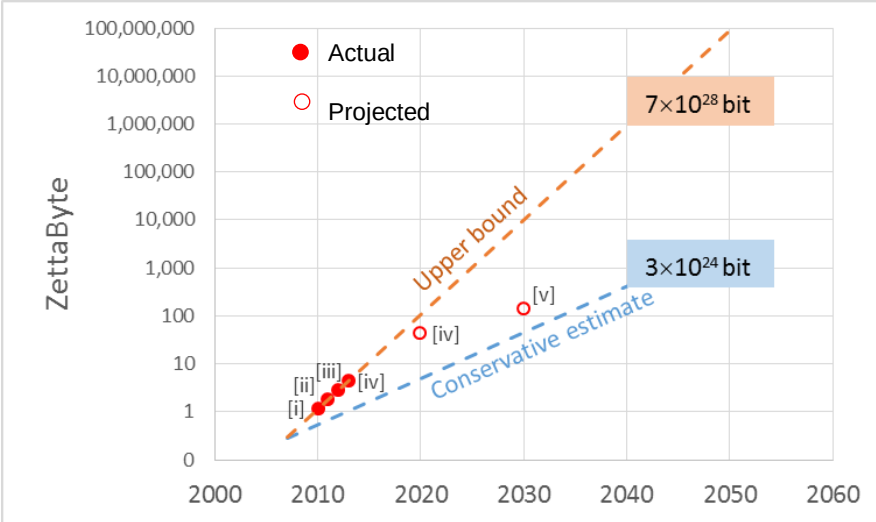

Fig. A2. Estimated and projected global memory demand, including a conservative estimate and an upper bound.

As indicated by Figure A2, future information and communication technologies are expected to generate enormous amounts of data, far surpassing today's data flows. Global memory demand is estimated to exceed 3×10^{24} bits by 2040. As an illustration, at this conservative estimate, the total mass of silicon wafers required for ultimately scaled flash memory, would exceed the total available silicon supply (see below) and would be equal to 3×10^{24} bits $\times$ 1.63×10^{-12} g/bit = 2.4×10^{9} kg (see Section A5).

A2. Global Silicon Wafer Supply Trend

The data for global silicon wafer supply was acquired from Weingarner[42] and shows a linear growth between 1990 and 2015 (Figure A3). Figure A4 is an extrapolation of this data into the future, which projects that silicon wafer production will be ~2.4×10^{7} kg in 2040.

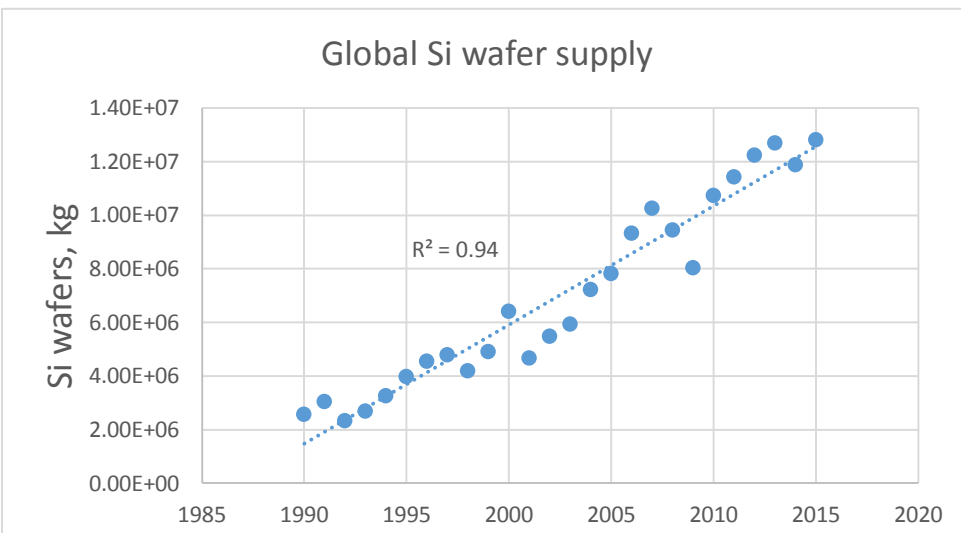

Fig. A3. Global Si wafer supply between 1990 and 2015.[42]

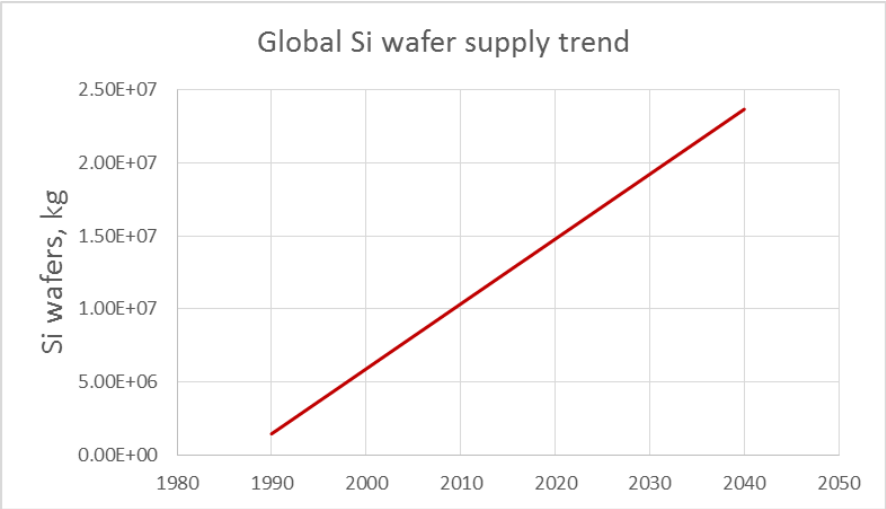

Fig. A4. Projected global Si wafer supply extrapolated from [42]

A3. Total Communication Capacity in 2040

The world's technological effective capacity to broadcast and telecommunicate information is projected to rise to ~ 10^{23} bit/year in 2040 (based on research by Hilbert and Lopez[2]). As can be seen in Figure A5, there is a growing gap between the world's technological information storage needs and communication capacities. For example, whereas in 2014 it would be possible to transmit all world's stored data in less than one year, i.e., in nearly real time, in 2040 it would require at least 20 years (even if the conservative data storage estimate in Figure A2 is assumed).

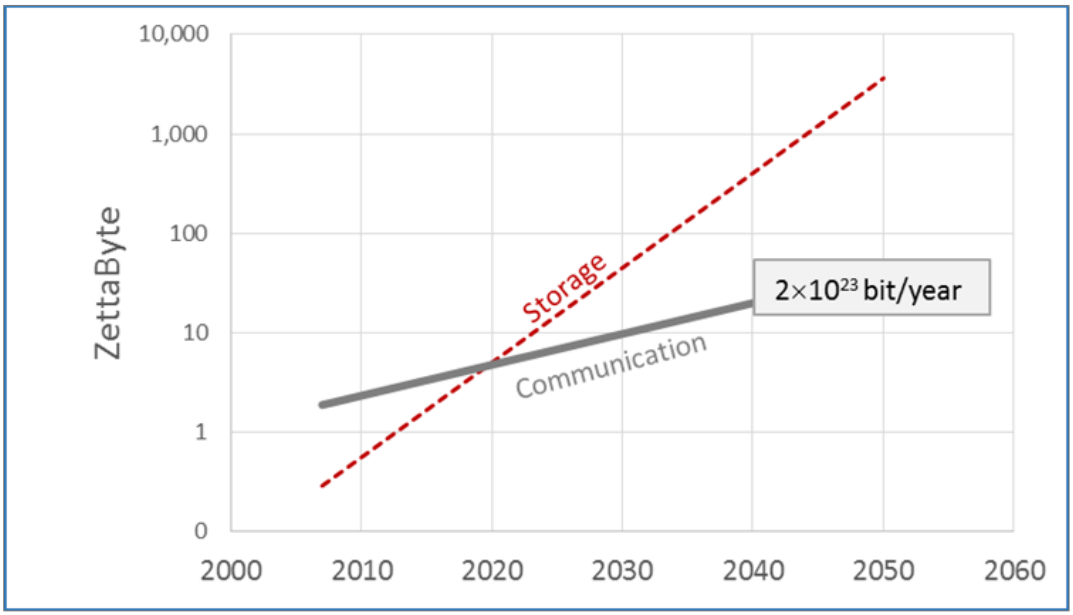

Fig. A5. Trend in world's technological information storage and communication capacities.

A4. Total Effective Capacity to Compute Information in 2040

The overall computational performance of microprocessors is often measured in millions of instructions per second (MIPS) that can be executed across a standard set of benchmarks. As can be seen in Figure A6, the world's technological effective capacity to compute information is projected to rise to ~5×10^{23} MIPS in 2040, including ~5×10^{19} MIPS of general-purpose computing (based on the research by Hilbert and Lopez[2]).

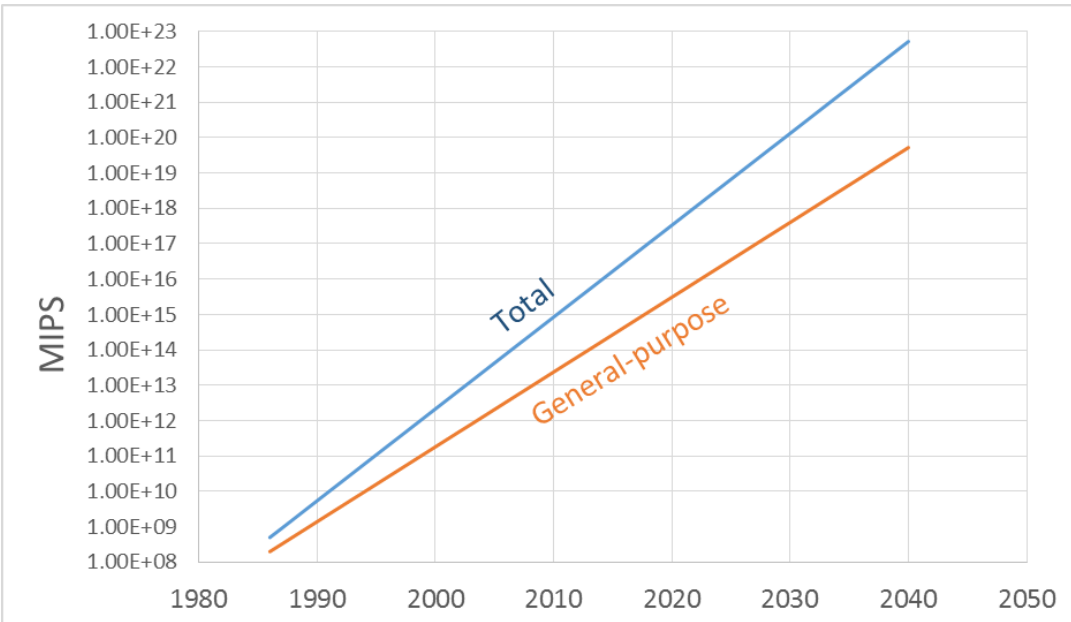

Fig. A6. Trend in world's technological information processing capacity.

For general-purpose computers, there is a strong correlation between the overall computational performance measured in MIPS and the system's binary throughput (BTP), measured in bit/s, which represents a characteristic number of "raw" binary transitions in the system needed to implement an instruction:

$$BTP \approx k \times MIPS^{r} \qquad (1)$$

Formula (1) is valid for a variety of microprocessor chips (a selection of 39 chips produced in 1971–2011 by ten different companies; for details see Zhirnov, Cavin, and Gammaitoni 2014[43]), k~3×10^{10} and r~1.56 with a high degree of accuracy (the determination coefficient R^2=0.98). This strong correlation suggests a possible fundamental law behind the empirical observation. Currently, there is no established correlation between MIPS and binary throughput for special-purpose computing (which includes various microcontrollers, graphic processing units, etc.). In the following, in order to at least partially account for the raw bits generated by these systems, result (1) is multiplied by a factor of two. The resulting total number of binary transitions required for computing per year is shown in Figure A7.

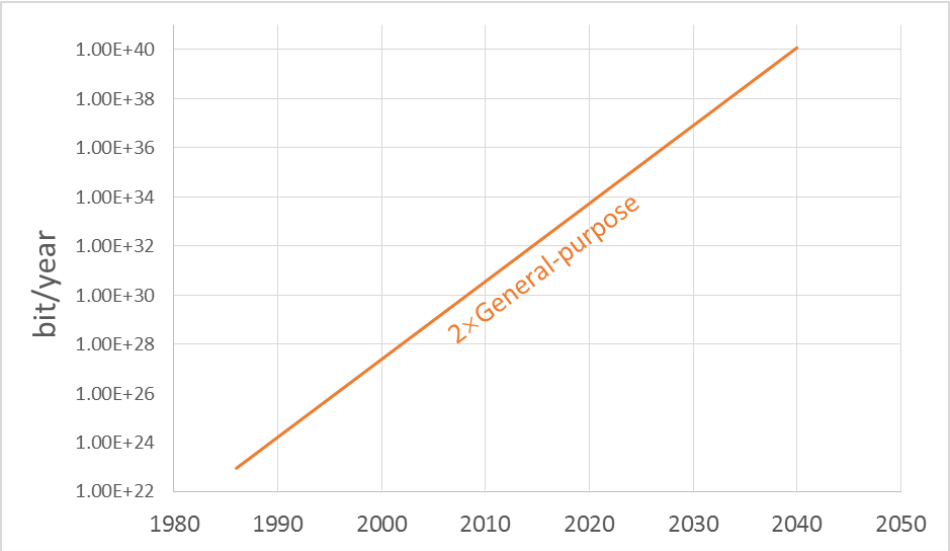

Fig. A7. The total number of binary transitions required for computing.

The total energy expenditure for computing (Figure A8) is directly related to the number of raw bit transitions. The system-level energy per bit operation is a combination of many components, such as logic circuits, memory arrays, interfaces, I/Os, etc (Figure A9), each significantly contributing the total energy budget. In current mainstream systems, the lower-edge system-level energy per one bit transition is ~10^{-14} J, [43] which is referred as the "benchmark" in Figure A8. For this benchmark energy per bit, computing will not be sustainable by 2040, when the energy required for computing will exceed the estimated world's energy production. Thus, radical improvement in the energy efficiency of computing is needed. The physics-based device-level theoretical lower limit is 3×10^{-21} J/bit (known as the Landauer limit for binary switching), and a practical lower limit for a system-level energy consumption can be estimated, based on the analysis in[43], to be ~10^{-17} J/bit, which is referred to as the "target" in Figure A8.

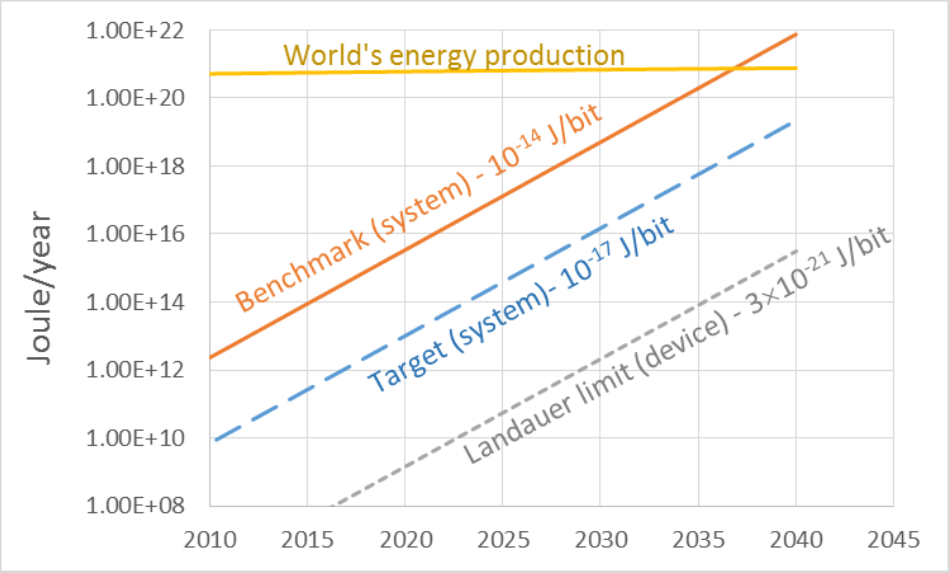

Fig. A8. Total energy of computing.

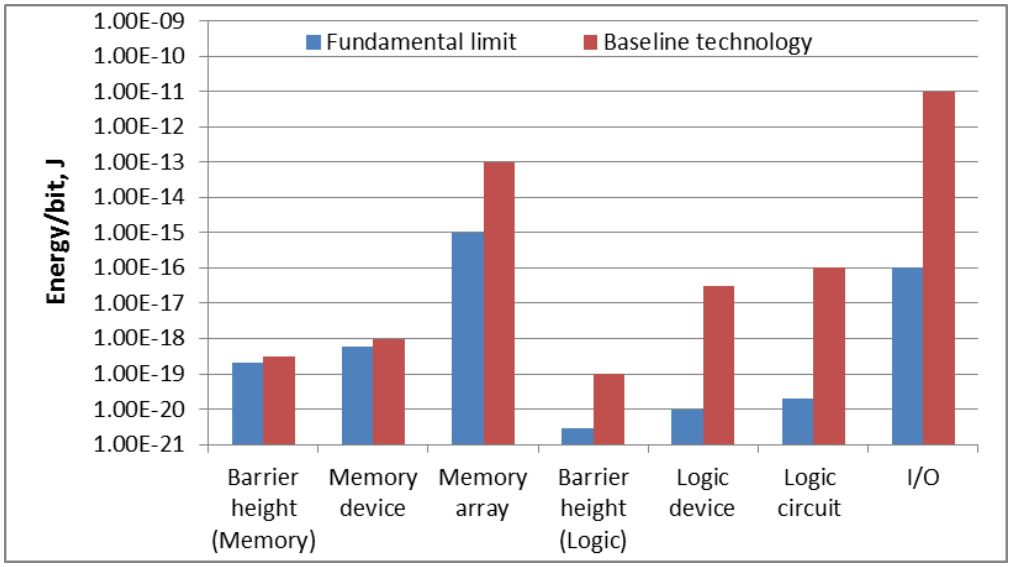

Fig. A9. Energy consumption per bit in ICT: A summary

A5. The Weight of a Flash Bit

Flash memory is organized into dense arrays of cells that are fabricated on a silicon wafer. The array capacity is constrained, in part, by the mass of the silicon. The thickness of the wafer is diameter-dependent (Table A1) because of the mechanical properties of silicon. For example, the initial wafer must be thick enough to avoid cracking during high-volume and speed handling.

Table A1. Diameter and thickness of standard Si wafers

Wafer diameter (mm)	Wafer thickness (mm)
100	0.525
150	0.625
200	0.725
300	0.775
450*	0.925

* Preproduction wafer diameter under development.

The weight of silicon needed for one flash memory bit, m_{flash}, is determined by the thickness of the wafer, h, and the area of the flash cell, A_{cell}:

$$m_{flash} = r_{Si} \times A_{cell} \times h \qquad (2),$$

where ρ_{Si} = 2.33 g/cm^3 is the density of silicon.

The area of one NAND flash cell is $A_{cell} = 4F^2$, where F is its smallest dimension, based on the physics of operation of flash memory,[44] F is 10–15 nm. For practical scaling reasons, F = 15 nm has been chosen for this analysis, yielding $A_{cell} = 9\times10^{-12}$ cm^2. Assuming h = 0.775 mm for a 300 mm wafer, Eq. 1 results an $m_{flash} = 1.63\times10^{-12}$ g/bit. (It should be noted that the thickness of an actual silicon die in a flash package is considerably smaller than the thickness of the initial wafer, as the die thickness is reduced to 20-50 µm

by back grinding and polishing steps. However the weight (2) is representative as the total amount of silicon required to build a flash bit. Flash memory can also be organized in 3-dimensional (3D) stacks of several flash memory dies in one package. The number of stacks and therefore the 3D density depends in the die thickness (see above). If a hypothetical ultimate 3D stacking of the flash is considered, where the die thickness is equal to the smallest cell's spatial dimension, *F,* the 3D packing density of memory cells is $n_{3D}=1/32F^2$.[9] For F=15 nm, a total of 10^{16} bit of flash could fit the volume of a 1-cm cube.

Appendix B. Rebooting the IT Revolution Workshop

Workshop Steering Committee

Avram Bar-Cohen, DARPA
Ralph Cavin, Semiconductor Research Corporation
Scott DeBoer, Micron Technologies
Bob Doering, Texas Instruments
Sam Fuller, Analog Devices Inc.
Steve Hillenius, Semiconductor Research Corporation
Dale Ibbotson, Altera
David Isaacs, Semiconductor Industry Association
Steven Johnston, Intel Corporation
Taffy Kingscott, IBM
Rafic Makki, GLOBALFOUNDRIES
Russell Meyer, Micron Technologies
Celia Merzbacher, Semiconductor Research Corporation
Tak Ning, IBM
Mike Roco, National Science Foundation
David Seiler, National Institute of Standards and Technology
Hans Stork, ON Semiconductor
Tom Theis, Semiconductor Research Corporation
Dustin Todd, Semiconductor Industry Association
Gilroy Vandentop, Semiconductor Research Corporation
Nicholas Yu, Qualcomm
Victor Zhirnov, Semiconductor Research Corporation

The workshop was sponsored by the Semiconductor Industry Association, the Semiconductor Research Corporation and supported by the National Science Foundation.

The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; [and] to secure the national defense…" With a budget of $7.3 billion in FY 2015, NSF is the funding source for approximately 24 percent of all federally supported basic research conducted by America's colleges and universities. More information about NSF is available at http://www.nsf.gov/about/.

The Semiconductor Industry Association (SIA) is the trade association representing leading U.S. companies engaged in the design and manufacture of semiconductors. Semiconductors are the fundamental enabling technology of modern electronics that has transformed virtually all aspects of our economy, ranging from information technology, telecommunications, healthcare, and transportation, to energy and national defense. The United States is the global leader in the semiconductor industry, and continued U.S. leadership in semiconductor technology is essential to America's continued global economic leadership. More information about SIA is available at http://www.semiconductors.org.

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 seeks to broaden the university base for such research and to increase the supply of qualified personnel for the industry's future workforce. SRC's mission is to assess industry's needs for research,

develop strategies to meet these needs, fund research at academic institutions consistent with these strategies, facilitate structured engagement of industry experts with the researchers, and proactively transfer research results to its members. 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 is a program jointly supported by MARCO (a subsidiary of SRC) and the Defense Advanced Research Projects Agency (DARPA), and supports 6 multi-university centers that involve 42 universities across the country. SRC's core program, Global Research Collaboration, has partnered with NSF on a number of joint programs. 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/.

Workshop Agenda

Workshop on Rebooting the IT Revolution
March 30-31, 2015
IBM Conference Center, 600 14th Street NW, Suite 300, Washington, DC

Monday, March 30, 2015		
12:00P – 1:00P	Registration / Networking / Light Lunch	
1:00P – 1:15P	Welcoming Remarks	Taffy Kingscott / IBM John Neuffer / SIA Steve Hillenius / SRC Mike Roco / NSF Avram Bar-Cohen / DARPA Dave Seiler / NIST
1:15P - 2:15P	VISION: Rebooting Semiconductor Research for the Next Era of Information Technology	Zach Lemnios / IBM
Session I - Always Aware & Secure: A Vision for Ultimate Intelligent Infrastructure (Facilitator: Linton G. Salmon / DARPA)		
2:15P - 4:15P	**Keynote:** Getting Closer to the Customer	Ken Gabriel / Draper Laboratory
	Roundtable / Open Mic Discussions: Matt Scholl / NIST Edward Lee / UC Berkeley Rafic Makki / GLOBALFOUNDRIES Kevin Kemp / Freescale	
4:15P – 4:30P	Break	
Session II - Always Connected: Billions of Connected Nanosystems (Facilitator: Gilroy Vandentop / SRC & Intel)		
4:30P – 6:30P	**Keynote**: Ultimate Connectivity: The TerraSwarm Vision	Jan Rabaey / UC Berkeley
	Roundtable / Open Mic Discussions: Ajith Amerasekera / Texas Instruments Geng Wu / Intel Sam Fuller / Analog Devices Josep Jornet / SUNY Buffalo	
6:30P – 8:00P	Reception	
Tuesday, March 31, 2015		
8:15A – 8:30A	Registration / Light Continental Breakfast	
Session III - Intelligent Sensor Nodes (Facilitator: An Chen / GLOBALFOUNDRIES)		
8:30A – 10:30A	**Keynote:** Cubic-Millimeter Sensor Nodes	Dennis Sylvester / Univ. of Michigan
	Roundtable / Open Mic Discussions: Hans Stork / ON Semiconductor Nigel Paver / ARM Manos Tentzeris / Georgia Tech Bob Doering / Texas Instruments	
10:30A – 10:45A	Break	
Session IV - Semiconductor Technologies for Big Data (Facilitator: Jeff Welser / IBM)		
10:45A – 12:45P	**Keynote:** Extreme Computing	Daniel Reed / Univ. of Iowa
	Roundtable / Open Mic Discussions: Randal E. Bryant / White House OSTP Gurtej Sandhu / Micron Steve Trimberger / Xilinx Avram Bar-Cohen / DARPA	
12:45P – 2:00P	Lunch	
Session V - Research Needs and Potential Responses (Facilitator: Ralph Cavin / SRC)		
2:00P – 4:00P	Examine technology gaps and gating research issues critical to enabling the future IT Revolution Infrastructure	All
4:00P – 4:30P	Closing comments / Adjourn	

Workshop Participants

Amerasekera, Ajith / Texas Instruments

Atwood, Greg / Micron

Bar-Cohen, Avram / DARPA

Basu, Sankar / NSF

Bryant, Randal E. / OSTP

Cavin, Ralph / SRC

Chen, An / GLOBALFOUNDRIES

Doering, Robert R. / Texas Instruments

Felbinger, Jonathan / DOD

Friedersdorf, Lisa / NNCO

Fuller, Samuel / Analog Devices

Gabriel, Kaigham / Draper Laboratory

Goldberg, Lawrence S. / NSF

Hillenius, Steve / SRC

Isaacs, David / SIA

Johnston, Steven / Intel

Jornet, Josep M. / SUNY Buffalo

Kemp, Kevin G. / Freescale

Kingscott, Kathleen N. / IBM

Lee, Edward A. / UC/Berkeley

Lemnios, Zachary / IBM

Makki, Rafic / GLOBALFOUNDRIES

McCants, Carl / IARPA

Meisner, Bob / US DOE

Merzbacher, Celia / SRC

Neuffer, John / SIA

Paver, Nigel / ARM Inc.

Pomrenke, Gernot / AFRL

Rabaey, Jan / UC Berkeley

Reed, Daniel A. / Univ. of Iowa

Richter, Curt A. / NIST

Roco, Mihail C. / NSF

Salmon, Linton G. / DARPA

Sandhu, Gurtej S. / Micron

Scholl, Matthew / NIST

Seiler, David G. / NIST

Staffin, Robin / DOD

Stork, Hans / ON Semiconductor

Sylvester, Dennis M. / Univ. of Michigan

Tentzeris, Emmanouil / Georgia Tech

Theis, Thomas N. / IBM;

Todd, Dustin / SIA

Vandentop, Gilroy J. / Intel

Trimberger, Stephen / Xilinx

Welser, Jeffrey J. / IBM

Whitman, Lloyd / OSTP

Wu, Geng / Intel

Ying, Charles / NSF

Zhirnov, Victor / SRC

References

[1] IDC and EMC2,"The Digital Universe in 2020," IDC's sixth annual study of the digital universe, 2012. Website: http://www.emc.com/leadership/digital-universe/2012iview/index.htm

[2] M. Hilbert and P. Lopez, "The world's technological capacity to store, communicate, and compute information," *Science* 332 (2011) 60-65.

[3] IDC and EMC Corp., *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. 7th annual study of the digital universe. Executive Summary available online: http://idcdocserv.com/1678, p. 3.

[4] *TSensors Roadmap* v. 1.1, TSensors Summit, Stanford University, October 23–25, 2013. Available online: http://tsensorssummit.org/Resources/TSensors%20Roadmap%20v1.pdf

[5] C. G. Bell, R. Chen, and S. Rege, "The effect of technology on near term computer structures," *Computer* 2(5): 29-38 (March/April 1972).

[6] G. Bell, "Bell's law for the birth and death of computer classes," *Commun. ACM* 51 86-94 (2008).

[7] D. Sylvester (Univ. Michigan), "Cubic-millimeter sensor nodes," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[8] T. Nakagawa, M. Miyazaki, G. Ono, R. Fujiwara, T. Norimatsu, T. Terada, A. Maeki, Y. Ogata, S. Koyayashi, N. Koshizuka, and K. Sakamura, "1-cc computer using UWB-IR for wireless sensor network," *IEEE J. Solid-State Circuits* 43(4): 964-973 (2008). DOI: 10.1109/JSSC.2008.917543.

[9] V. V. Zhirnov and R. K. Cavin, *Microsystems for Bioelectronics: Scaling and Performance Limits* (Elsevier 2015).

[10] Nigel Paver (ARM), "Sensor challenges," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[11] M. Tentzeris (Georgia Tech), "Ink-jet-printed, nanotechnology-enabled 'zero-power' wireless sensor nodes for 'smart skin' applications," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[12] J.-H. Seo, T.-H. Chang, J. Lee, R. Sabo, W. Zhou, Z. Cai, S. Gong, and Z. Ma, Microwave flexible transistors on cellulose nanofibrillated fiber substrates, *Appl. Phys. Lett*. 106: 262101 (2015); DOI: 10.1063/1.4921077.

[13] Hans Stork (ON Semiconductor), "Intelligent Sensor Nodes: Fabrication and Design", Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015

[14] Bob Doering (Texas Instruments), "Sensors--A Vast Multidimensional Space", Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[15] Ajith Amerasekera (Texas Instruments), "Deeply interconnected systems," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[16] Geng Wu (Intel), "Always Connected: Billions of Connected Nanosystems," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[17] Samuel Fuller (Analog Devices), "Always connected: Billions of connected nanosystems," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[18] J. Cubo, A. Nieto, and E. Pimentel, "A Cloud-based Internet of Things platform for ambient assisted living", *Sensors* 14 (2014) 14070-14105.

[19] Cisco, *Cisco Visual Networking Index (VNI): Global Mobile Data Traffic Forecast Update, 2014–2019* (2015). Available online: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

[20] M. Meeker (Morgan Stanley), "Economy and Internet trends" (presentation) Web 2.0 Summit, San Francisco, Oct. 2009: http://www.slideshare.net/ahawkinson/mary-meeker-2009-web-20-summit-presentation.

[21] Josep Jornet (SUNY Buffalo), "Which wireless technology will support the IT revolution? A communications view," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[22] J. Rabaey (UC Berkeley), "Providing ultimate connectivity," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[23] I. F. Akyildiz, J. M. Jornet, and C. Han, "Terahertz band: Next frontier for wireless communications," *Physical Communication Journal* (Elsevier) 12:16-32 (2014).

[24] Steve Trimberger (Xilinx), "A new Moore's Law," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[25] Square Kilometer Array website, https://www.skatelescope.org/

[26] Daniel Reed (Univ. Iowa), "Extreme scale lessons," presentations at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[27] Avram Bar-Cohen (DARPA), "Cognitive computing, Towards the electronic brain," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[28] Wikipedia, "Theory of Computation," https://en.wikipedia.org/wiki/Theory_of_computation.

[29] ZDNet, "Cheat Sheet: Microservers," http://www.zdnet.com/article/cheat-sheet-microservers/.

[30] D. Linthicum, "Cloud appliances: The cloud that isn't the cloud," InfoWorld (2012), http://www.infoworld.com/article/2619260/iaas/cloud-appliances--the-cloud-that-isn-t-the-cloud.html.

[31] R. Bryant (White House Office of Science and Technology Policy), "Setting a 1 EB target for 2025," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[32] G. Sandhu (Micron), "Semiconductor technologies for big data," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[33] P. Dlugosch, D. Brown, P. Glendenning, M. Leventhal, H. Noyes, "An efficient and scalable semiconductor architecture for parallel automata processing", IEEE Trans. Parallel and Distr. Syst. 25 (2014) 3088-3098.

[34] National Science Foundation, "Cyber-Physical Systems (CPS)", Directorate for Computer & Information Science & Engineering, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503286.

[35] Edward Lee (UC Berkeley), "Cyber-Physical Systems," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[36] Kevin Kemp (Freescale), "The Internet of Things," presentation at the Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[37] Rafic Makki (GLOBALFOUNDRIES), "Semiconductor Industry Challenges", Workshop on Rebooting the IT Revolution, Washington, DC, March 30 & 31, 2015.

[38] Executive Order Creating a National Strategic Computing Initiative, https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative.

[39] Federal Register Notice of Request for Information, "Nanotechnology-Inspired Grand Challenges for the Next Decade", https://www.federalregister.gov/articles/2015/06/17/2015-14914/nanotechnology-inspired-grand-challenges-for-the-next-decade

[40] https://www.src.org/calendar/e005817/.

[41] M. Hilbert and P. Lopez, "The world's technological capacity to store, communicate, and compute information," *Science* 332 (2011) 60-65.

[42] R.M. Winegarner, *Silicon Industry 2011*, Healdsburg, CA: Sage Concepts (2011), http://sageconceptsonline.com/.

[43] V. Zhirnov, R. Cavin, and L. Gammaitoni, "Minimum energy of computing, fundamental considerations" (Chapter 7), In *ICT - Energy - Concepts Towards Zero - Power Information and Communication Technology*, Giorgos Fagas (Ed.), InTech 2014, http://intechopen.com/books/ict-energy-concepts-towards-zero-power-information-and-communication-technology/minimum-energy-of-computing-fundamental-considerations.

[44] V. V. Zhirnov and M. J. Marinella, "Memory Technologies: Status and Perspectives", in *Emerging Nanoelectronic Devices*, A. Chen et al. (Eds.), Wiley 2015.