Semiconductor Research Corporation
Semiconductor Industry Association

Decadal Plan for Semiconductors

ABRIDGED REPORT

Published October 2020

Decadal Plan Executive Committee

James Ang, PNNL
Dmytro Apalkov, Samsung
Fari Assaderaghi, Sunrise Memory
Ralph Cavin, SRC
Ramesh Chauhan, Qualcomm
An Chen, IBM
Richard Chow, Intel
Robert Clark, TEL
Maryam Cope, SIA
Debra Delise, Analog Devices
Carlos Diaz, TSMC
Bob Doering, Texas Instruments
Sean Eilert, Micron
Ken Hansen, SRC
Baher Haroun, Texas Instruments
Yeon-Cheol Heo, Samsung
Gilbert Herrera, SNL
Kevin Kemp, NXP
Taffy Kingscott, IBM
Stephen Kosonocky, AMD
Matthew Klusas, Amazon
Steve Kramer, Micron
Donny Kwak, Samsung
Lawrence Loh, MediaTek
Rafic Makki, Mubadala
Matthew Marinella, SNL
Seong-Ho Park, SK hynix
David Pellerin, Amazon
Daniel Rasic, SRC
Ben Rathsak, TEL
Wally Rhines, Mentor Graphics
Heike Riel, IBM
Kirill Rivkin, Western Digital
Gurtej Sandhu, Micron
Ghavam Shahidi, IBM
Steve Son, SK hynix
Mark Somervell, TEL
Gilroy Vandentop, Intel
Jeffrey Vetter, ORNL
Jeffrey Welser, IBM
Jim Wieser, Texas Instruments
Ian Young, Intel
David Yeh, TI & SRC
Victor Zhirnov, SRC
Zoran Zvonar, Analog Devices

The Decadal Plan Workshops that helped drive this report were supported by the U.S. Department of Energy, Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) Research Programs, and the National Nuclear Security Agency, Advanced Simulation and Computing (ASC) Program. SIA and SRC are grateful for their support.

Acronym Definitions

AI	Artificial Intelligence
aJ	attojoule (10^{-18} joules)
bps	bit-per-second
CPU	Central Processing Unit
CMOS	Complementary Metal-Oxide-Semiconductor
DNA	DeoxyriboNucleic Acid
DRAM	Dynamic Random-Access Memory
FPGA	Field-Programmable Gate Array
GPU	Graphics Processing Unit
IoT	Internet of Things
ICT	Information and Communication Technologies
I/O	Input/Output
IP block	semiconductor Intellectual Property core
HW	Hardware
NAND flash	highest-density silicon-based electronic nonvolatile memory
Mbps	megabit-per-second
MIMO	Multiple-Input and Multiple-Output
mm-Wave	millimeter wave
nJ	nanojoule (10^{-9} joules)
NVM	Nonvolatile Memory
R&D	Research and Development
SIA	Semiconductor Industry Association
SRC	Semiconductor Research Corporation
SW	Software
Tbps	terabit-per-second
THz	Terahertz
Zettabyte	10^{21} bytes
ZIPS	10^{21} compute instructions per second

Semiconductors, the tiny and highly advanced chips that power modern electronics, have helped give rise to the greatest period of technological advancement in the history of humankind.

Chip-enabled technology now allows us to analyze DNA sequences to treat disease, model nerve synapses in the brain to help people with mental disorders like Alzheimer's, design and build safer and more reliable cars and passenger jets, improve the energy efficiency of buildings, and perform countless other tasks that improve people's lives.

During the COVID-19 pandemic, the world has come to rely more heavily on semiconductor-enabled technology to work, study, communicate, treat illness, and do innumerable other tasks remotely. And the future holds boundless potential for semiconductor technology, with emerging applications such as artificial intelligence, quantum computing, and advanced wireless technologies like 5G and 6G promising incalculable benefits to society.

Fulfilling that promise, however, will require taking action to address a range of seismic shifts shaping the future of chip technology. These seismic shifts—identified in *The Decadal Plan for Semiconductors* by a broad cross-section of leaders in academia, government, and industry—involve smart sensing, memory and storage, communication, security, and energy efficiency. The federal government, in partnership with private industry, must invest ambitiously in semiconductor research in these areas to sustain the future of chip innovation.

For decades, federal government and private sector investments in semiconductor research and development (R&D) have propelled the rapid pace of innovation in the U.S. semiconductor industry, making it the global leader and spurring tremendous growth throughout the U.S. economy. The U.S. semiconductor industry invests about one-fifth of its revenues each year in R&D, one of the highest shares of any industry. With America facing increasing competition from abroad and mounting costs and challenges associated with maintaining the breakneck pace of innovation, now is the time to maintain and strengthen public-private research partnerships.

As Congress works to refocus America's research ecosystem on maintaining semiconductor innovation and competitiveness, *The Decadal Plan for Semiconductors* outlines semiconductor research priorities across the seismic shifts noted above and recommends an additional federal investment of $3.4 billion annually across these five areas. The interim report is included here, and the full report is scheduled to be released in December 2020.

Working together, we can boost semiconductor technology and keep it strong, competitive, and at the tip of the innovation spear.

Sincerely,

John Neuffer
President & CEO
Semiconductor Industry Association (SIA)

Todd Younkin
President & CEO
Semiconductor Research Corporation (SRC)

Executive Summary

The U.S. semiconductor industry leads the world in innovation, based in large part on aggressive research and development (R&D) spending. The industry invests nearly one-fifth of its annual revenue in R&D each year, second only to the pharmaceuticals sector. In addition, Federal funding of semiconductor R&D serves as the catalyst for private R&D spending. Together, private and Federal semiconductor R&D investments have sustained the pace of innovation in the U.S., enabling it to become the global leader in the semiconductor industry. Those R&D investments have nurtured the development of innovative and commercially viable products, and as a direct result, have led to a significant contribution to the U.S. economy and jobs.

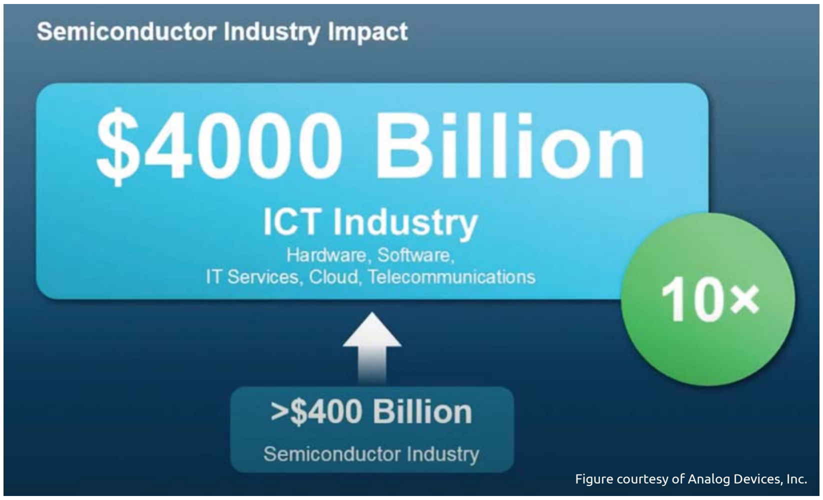


Figure courtesy of Analog Devices, Inc.

The current hardware-software (HW-SW) paradigm in information and communication technologies (ICT) has made computing ubiquitous through sustained innovation in software and algorithms, systems architecture, circuits, devices, materials, and semiconductor process technologies among others. However, ICT is facing unprecedented technological challenges for maintaining its growth rate levels into the next decade. These challenges arise largely from approaching various fundamental limitations in semiconductor technology that taper the otherwise necessary generational improvements in the energy-efficiency with which information is processed, communicated, stored, sensed and actuated on. Long term sustainable ICT growth will rely on breakthroughs in semiconductor technology capabilities that enable holistic solutions to tackle information processing efficiency. Disruptive breakthroughs are needed in the areas of software, systems, architectures, circuits, device structure and the related processes and materials that require timely and well-coordinated multidisciplinary research efforts.

This Decadal Plan for Semiconductors outlines research priorities in information processing, sensing, communication, storage, and security seeking to ensure sustainable growth for semiconductor and ICT industries by:

- informing and supporting the strategic visions of semiconductor companies and government agencies
- guiding a (r)evolution of cooperative academic, industry and government research programs
- placing 'a stake in the ground' to challenge the best and brightest researchers, university faculty and students

The Semiconductor Industry Association (SIA) June 2020 report[1] demonstrates that federal investment in semiconductor R&D spurs U.S. economic growth and job creation and presents a case for a 3x increase in semiconductor-specific federal funding. For every dollar spent on federal semiconductor research has resulted in a $16.50 increase in current GDP.

The Decadal Plan for Semiconductors complements this report and identifies specific goals with quantitative targets. It is expected that the Decadal Plan will have a major impact on the semiconductor industry, similar to the impact of the 1984 10-year SRC Research Goals document that was continued in 1994 as the National Technology Roadmap for Semiconductors, and which later became the International Technology Roadmap for Semiconductors in 1999.

Trends and drivers

Currently information and communication technologies are facing five major seismic shifts:

The Grand Challenge

Information and communication technologies make up over 70% of the semiconductor market share. They continue to grow without bounds dominated by the exponential creation of data that must be moved, stored, computed, communicated, secured and converted to end user information. The recent explosion of artificial intelligence (AI) applications is a clear example, and as an industry we have only begun to scratch the surface.

Having computing systems move into domains with true cognition, i.e., acquiring understanding through experience, reasoning and perception is a new regime. This regime is unachievable with the state-of-the-art semiconductor technologies and traditional gains since the reduction in feature size (i.e., dimensional scaling) to improve performance and reduce costs in semiconductors is reaching its physical limits. As a result, the current paradigm must change to address an information and intelligence-based value proposition with semiconductor technologies as the driver.

[1] Sparking Innovation: How Federal Investment in Semiconductor R&D Spurs U.S. Economic Growth and Job Creation, SIA Report, June 2020

Call to Action: Semiconductor Technology Leadership Initiative

Maintaining and strengthening the leadership of the United States in ICT during this new semiconductor era requires a sustained additional $3.4B federal investment per year throughout this decade (i.e. tripling Federal funding for semiconductor research) to conduct large-scale industry-relevant, fundamental semiconductor research. (The Decadal Plan Executive Committee offered recommendations on allocation of the additional $3.4B investment per year among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies).

The investments through new public-private partnerships must cover a wide breadth of interdependent technical areas (compute, analog, memory/storage, communications, and security) requiring multi-disciplinary teams to maintain U.S. semiconductor technology leadership.

These investments need to be organized and coordinated to support a common set of goals focused on market demand to provide technologies which enable new commercial products and services over the course of the program. The Decadal Plan has identified five seismic paradigm shifts required to accomplish this overarching Grand Challenge.

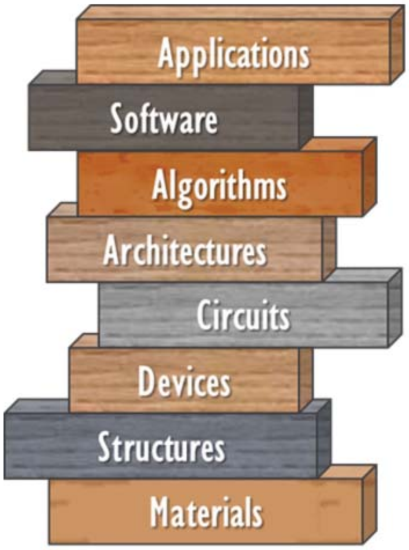
A Note on Funding Strategic National Initiatives:

The Decadal Plan serves as a blueprint for policymakers who recognize this challenge and seek guidance on areas of research emphasis for scientific research agencies and public-private partnerships.

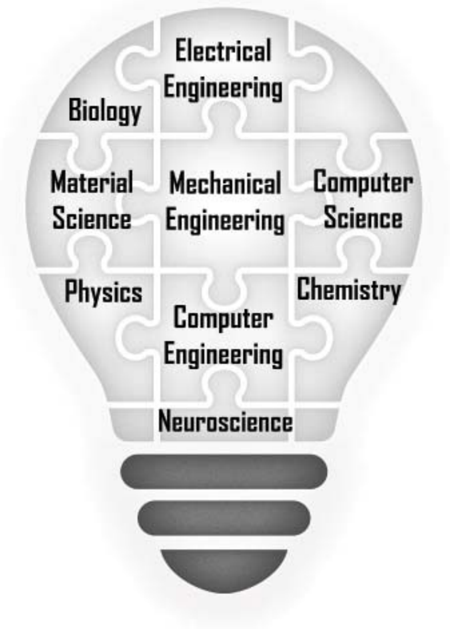
Semiconductor Technology Breakthroughs Rely On

Holistic Optimal Solutions

Driven by Hardware/Software Co-Optimization



Interlocked Multidisciplinary Research



1. Introduction

Currently information and communication technologies are facing five major seismic shifts:

Seismic shift #1 Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces that can sense, perceive and reason.

Seismic shift #2 The growth of memory demands will outstrip global silicon supply presenting opportunities for radically new memory and storage solutions.

Seismic shift #3 Always available communication requires new research directions that address the imbalance of communication capacity vs. data generation rates.

Seismic shift #4 Breakthroughs in hardware research are needed to address emerging security challenges in highly interconnected systems and Artificial Intelligence.

Seismic shift #5 Ever rising energy demands for computing vs. global energy production is creating new risk, and new computing paradigms offer opportunities with dramatically improved energy efficiency.

To support the Decadal Plan development, an international series of five face-to-face workshops has been conducted to assess quantitatively each seismic shift, assign targets and suggest initial research directions. Participants and contributors to these workshops included academic, government and industrial domain experts. The output of these workshops has guided the recommendations in the 2020 Decadal Plan for Semiconductors.

These workshops provided highly interactive forums where key research leaders evaluated the status of nanoelectronics research and application drivers, discussed key scientific issues, and defined promising future research directions. This is instrumental for the 2020 version of the Decadal Plan for Semiconductors to reflect an informed view on key scientific and technical challenges related to revolutionary information and communication technologies, based on new quantitative analyses and projections.

The primary objectives of the Decadal Plan include

- Identify significant trends and applications that are driving Information and Communication Technologies and the associated roadblocks/challenges.
- Assess quantitatively the potential and status of the five seismic shifts that will impact future ICT.
- Identify fundamental goals and targets to alter the current trajectory of semiconductor technology.

The Decadal Plan provides an executive overview of the global drivers and constraints for the future ICT industry, rather than to offer/discuss specific solutions: **The document identifies the what, not the how.** In doing so, **it focuses and organizes the best of our energies and skills to the key challenges** in a quantitative manner about which creative solutions can be imagined and their impact measured.

Seismic shift #1

Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces that can sense, perceive and reason.

Analog electronics deals with real-world continuously variable signals of multiple shapes (in contrast to digital electronics where signals are usually of standard shape taking only two levels, ones or zeros). The analog electronics domain encompasses multiple dimensions as shown in Figure 1. Also, all inputs human can perceive are analog, which calls for bio-inspired solutions for world-machine interfaces that can sense, perceive and reason based on ultra-compressed sensing capability and low operation power (Figure 2).

Figure 1: The Dimensions of Analog Electronics

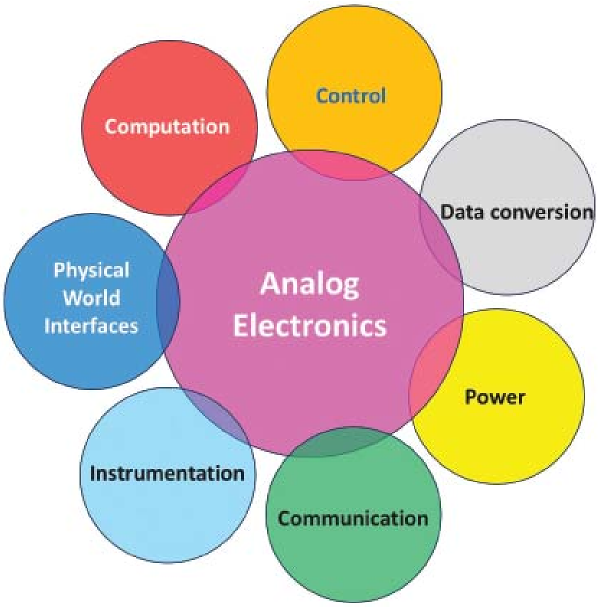
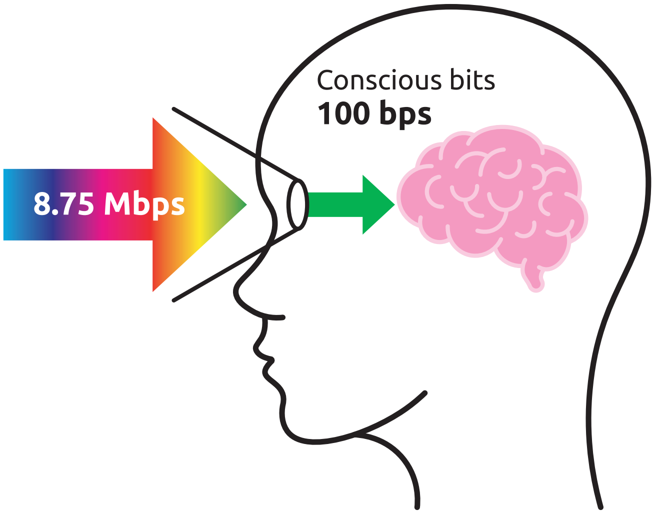


Figure 2: The brain's ability to perceive and reason is based on ultra-compressed sensing capabilities with 100,000 data reduction and a low operation power

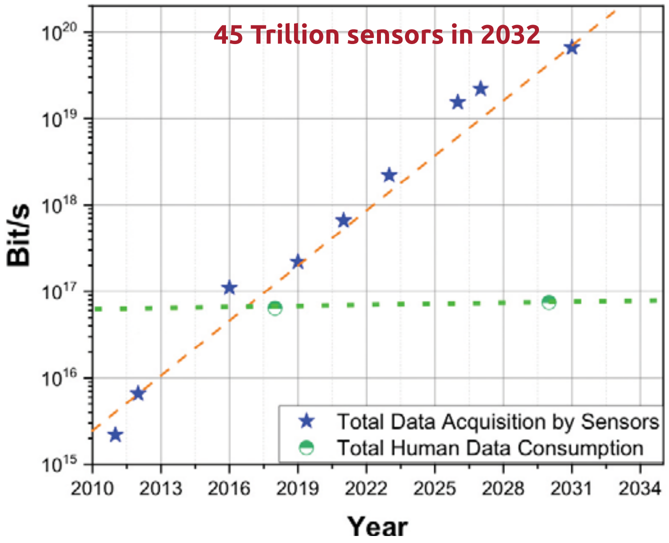


The physical world is inherently analog and the "digital society" places an increasing demand for advanced analog electronics to enable interaction between the physical and computer "worlds."

Sensing the environment around us is fundamental to the next generation of AI where devices will be capable of perception and reasoning. The world-machine interface lies at the heart of the *current information-centric economy*. As one example, the next wave of the advanced manufacturing revolution is expected to come from next-generation analog-driven industrial electronics, that includes sensing, robotics, industrial, automotive, medical etc. For mission critical applications, the reliability of electronic components is a priority. Today, for example, analog chips constitute 80% of failures in automotive electronics, which is ten times worse than digital chips.

The estimated total analog information generated from the physical world is equivalent to ~10^{34} bit/s. As a reference, the total collective human sensory throughput pales at ~10^{17} bit/s (Figure 3). Thus, our ability to perceive the physical world is significantly limited. There are tremendous opportunities for future analog electronics that augment the human sensory system, which is expected to have significant economic and social consequences. Examples include, but are not limited to, creating multimedia that specifically targets human sensory and cognitive systems, including nervous system interfaces and communications. This can result in new human-centric technologies such as multi-sensing-based medical diagnostics and therapy, complete virtual reality with virtual aroma synthesizers, or active odor cancellation based on indoor air quality.

Figure 3: Trend in world's installed sensing capacities



Today the ability to generate analog data is growing faster than our ability to intelligently use the data. This situation will become even more serious in the near future, when data from our lives as well as from IoT sensors may create an analog data deluge that will obscure valuable information when we need it the most. Sensor technologies are experiencing exponential growth with forecasts of ~45 trillion sensors in 2032 that will generate >1 million zettabytes (10^{27} bytes) of data per year. This is equivalent to ~10^{20} bit/s, thereby **surpassing the collective human sensing throughput**. Therefore, a significant paradigm shift towards extracting key information in the predicted data deluge and applying it in an appropriate way is key to harnessing the data revolution. Thus, the *Analog Grand Goal* is for revolutionary technologies to increase useful/actionable information with less energy and data bits e.g. sensing-to-analog-to-information reduction with a practical compression/reduction ratio of 10^5:1.

For many real-time applications, the value of sensory data is brief, sometimes only a few milliseconds. The data must be utilized within that time frame and in many cases locally for latency and security considerations. Therefore, pursuing breakthrough advances in information processing technologies such as developing hierarchical perception algorithms that enable understanding of the environment from raw sensor data is a fundamental requirement. New computing models such as analog "approximate computing" are required. This is aligned with the Grand Goal #5 of discovering a radically new 'computing trajectory', outlined later. New analog technologies can also offer great advancements in communication technologies. Even in computer-to-computer communication, analog interfaces are required at long distances. The ability to collect, process and communicate the analog data at the input/output (I/O) boundaries is critical to the future world of IoT and Big Data. Analog technology advancements to the THz regime will be required for both the sensing and communication needs of the future.

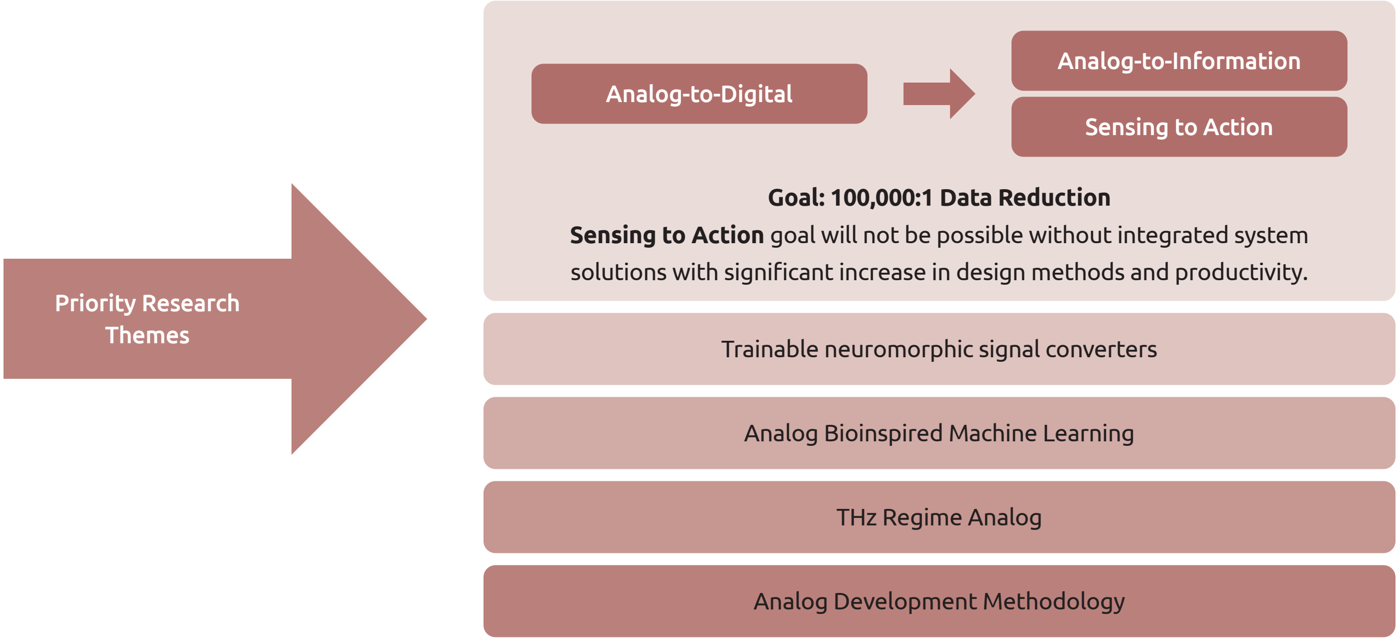
Grand Goal #1:

Analog-to-information compression/reduction with a practical compression/reduction ratio of 10^5:1 that drives to a practical use of information versus "data" in a way that is more analogous to the human brain.

Call for Action

The analog interface bridges the physical and digital worlds. Our collective ability to access the information of the physical world through analog signals is 10,000 trillion times below what is available, and radical advances in analog electronics will be required soon. New approaches to sensing such as *sensing to action*, analog “artificial intelligence” (AI) platforms, brain inspired/neuromorphic and hierarchical computation, or other solutions will be necessary. Breakthrough advances in information processing technologies, such as developing perception algorithms to enable understanding of the environment from raw sensor data, are a fundamental requirement. New computing models such as analog “approximate computing,” which can trade energy and computing time with accuracy of output (presumably how the brain does) are required. New analog technologies will offer great advancements in communication technologies. The ability to collect, process and communicate the analog data at the input/output boundaries is critical to the future world of IoT and Big Data. Additionally, analog development methodologies require a step increase (10x or greater) in productivity to address the application explosion in a timely manner. Altogether, collaborative research to establish revolutionary paradigms for future energy-efficient analog integrated circuits for the vast range of future data types, workloads and applications is needed.

Invest $600M annually throughout this decade in new trajectories for analog electronics. Selected priority research themes are outlined below.[2]



Priority Research Themes

Analog-to-Digital

Analog-to-Information

Sensing to Action

Goal: 100,000:1 Data Reduction

Sensing to Action goal will not be possible without integrated system solutions with significant increase in design methods and productivity.

Trainable neuromorphic signal converters

Analog Bioinspired Machine Learning

THz Regime Analog

Analog Development Methodology

[2] The Decadal Plan Executive Committee offered recommendations on allocation of the additional $3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

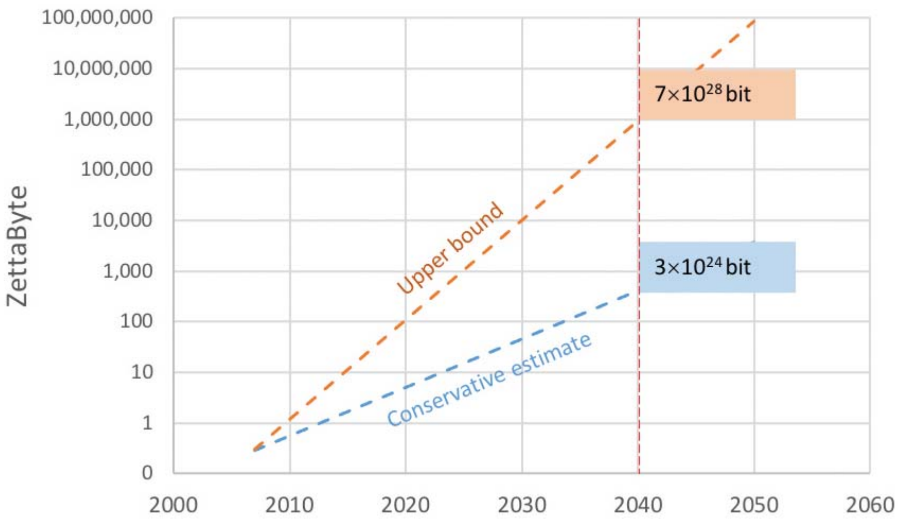
Seismic shift #2

The growth of memory demands will outstrip global silicon supply presenting opportunities for radically new memory and storage solutions.

Radical new solutions in Memory and Storage technologies will be needed for future ICT with major innovations in devices, circuits and architectures. By end of this decade, the continuing improvements in ICT energy efficiency and performance will stall as the underlying memory, and storage technologies will meet scaling limitations. At the same time, training data for AI applications is exploding with no limit in sight. It is becoming increasingly clear that in future information processing applications, synergistic innovations from materials and devices to circuits and system-level functions, likely using unexplored physical principles, will be a key to achieving new levels of bit density, energy-efficiency and performance.

Global demand for data storage grows exponentially, and today's storage technologies will not be sustainable in near future due to excessive material resources needed to support the ongoing Data Explosion. Thus, new radical solutions for data/information storage technologies and methods are required. Figure 4 shows the projections of global data storage demand—both a conservative estimate and an upper bound. As indicated by Figure 4, future information and communication technologies are expected to generate enormous amounts of data, far surpassing today's data flows. Currently, the production and use of information has been growing exponentially, and by 2040 the estimates for the worldwide amount of stored data are between 10^{24} and 10^{28} bits as shown in Figure 4. Furthermore, as a result, while the *weight of a single bit* in the case of ultimately scaled NAND flash memory is 1 picogram (10^{-12} g), the *total mass* of silicon wafers required to store 10^{26} bits would be approximately 10^{10} kg, which would exceed the world's total available silicon supply (Figure 5).

Figure 4: Global demand for memory and storage (utilizing silicon wafers), is projected to exceed the amount of global silicon that can be converted into wafers.



Challenge

Global demand for conventional silicon-based memory/storage is growing exponentially (Figure 4), while silicon production is growing only linearly (Figure 5). This disparity guarantees that silicon-based memory will become prohibitively expensive for Zetta-scale "big data" deployments within two decades.

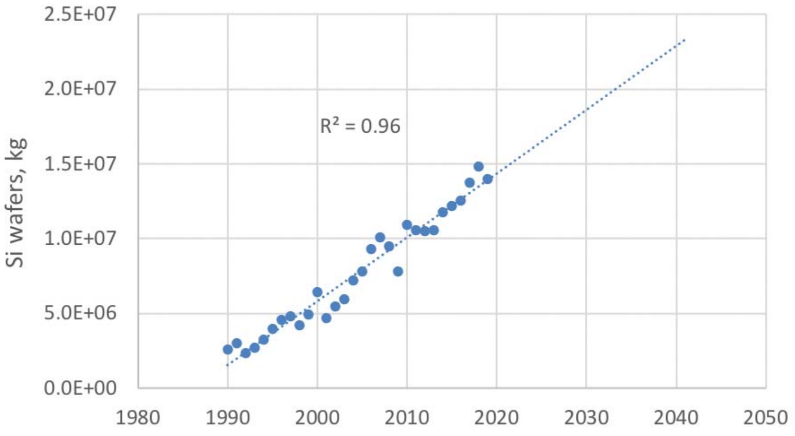
Grand Goal #2a:

Develop emerging memories and memory fabrics with >10-100X density and energy efficiency improvement for each level of the memory hierarchy.

Grand Goal #2b:

Grand Goal #3b: Discover storage technologies with >100x storage density capability and new storage systems that can leverage these new technologies.

Figure 5: Global Si wafer supply: 1990-2020 data and future trend.

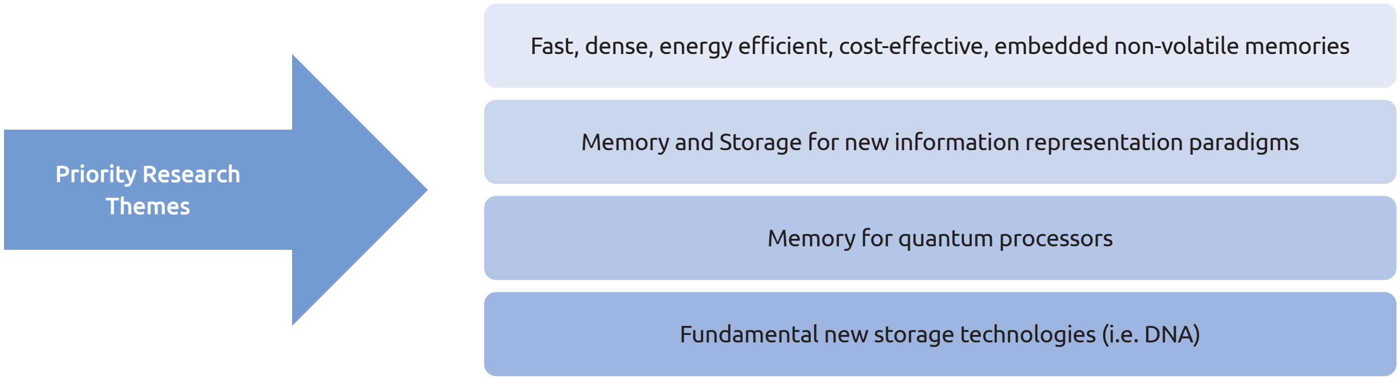


In addition, memory, such as DRAM, is an essential component of computers and further advances in computing are impossible without 'reinventing' the compute memory system including device physics, memory hierarchy architecture and physical implementation. For example, traditional embedded nonvolatile memory can no longer be scaled below 28nm, thus alternatives are needed. Finally, new memory solutions must be able to support multiple emerging applications, such as, e.g. artificial intelligence, large-scale heterogeneous high-performance and data-center computing, and various mobile applications that also meet the rugged environmental requirements of the automotive market, etc.

Call for Action

Radical advances in memory and data storage are required soon. Collaborative research 'from materials to devices to circuits to architecture to processing and solutions' for future high-capacity energy-efficient memory and data/information storage solutions for the vast range of future applications is needed.

Invest $750M annually throughout this decade in new trajectories for memory and storage. Selected priority research themes are outlined below.[3]



Priority Research Themes

Fast, dense, energy efficient, cost-effective, embedded non-volatile memories

Memory and Storage for new information representation paradigms

Memory for quantum processors

Fundamental new storage technologies (i.e. DNA)

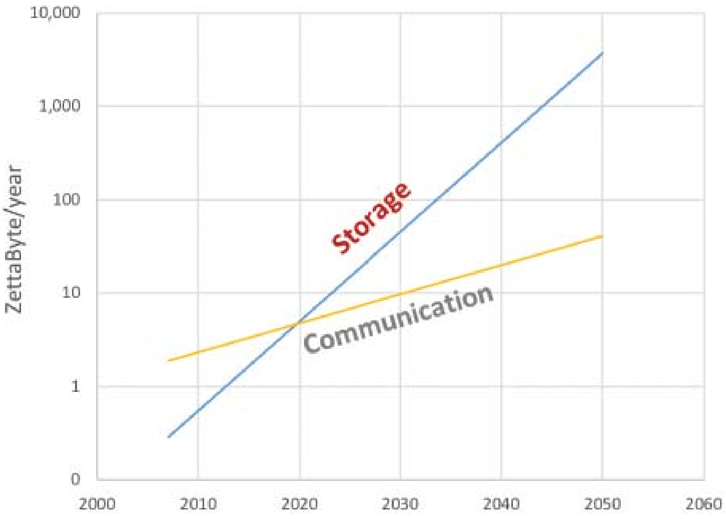
[3] The Decadal Plan Executive Committee offered recommendations on allocation of the additional $3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

Seismic shift #3

Always available communication requires new research directions that address the imbalance of communication capacity vs. data generation rates.

The current state of the developed world is characterized by (almost) *always-available* communication and connectivity, which has a tremendous impact on all aspects of life. A manifestation of this is Cloud Storage and Computing. The ability to get data from anywhere and send it to everywhere has transformed both the way we do business and our personal habits and lifestyle. Social networks are an example. However, the main concept of the cloud is based on the assumption of constant connectivity, which is not guaranteed. Furthermore, the demand grows daily for communication to become more ubiquitous as we become more connected. An alarming trend is a growing gap between the world's technological information storage need, as just discussed, and communication capacities shown in Figure 6. For example, while currently it is possible to transmit all world's stored data in less than one year, in 2040 it is predicted to require at least 20 years for the transmission. A global storage-communication cross-over is expected to happen around 2022 which may have a tremendous impact on ICT. Even with growing trend on edge computing for AI systems to cater for privacy and faster response time, the explosion of information generated and stored will require tremendous growth on cloud storage and communication infra-structure.

Figure 6: The Global Communication Data Generation Crossover occurs when the data generated exceeds the world's technological information storage and communication capacities, creating limitations to transmission of data.



Grand Goal #3a:

Advance communication technologies to enable moving around all stored data of 100-1000 zettabyte/year at the peak rate of 1Tbps@<0.1nJ/bit.

Grand Goal #3b:

Develop intelligent and agile networks that effectively utilize bandwidth to maximize network capacity.

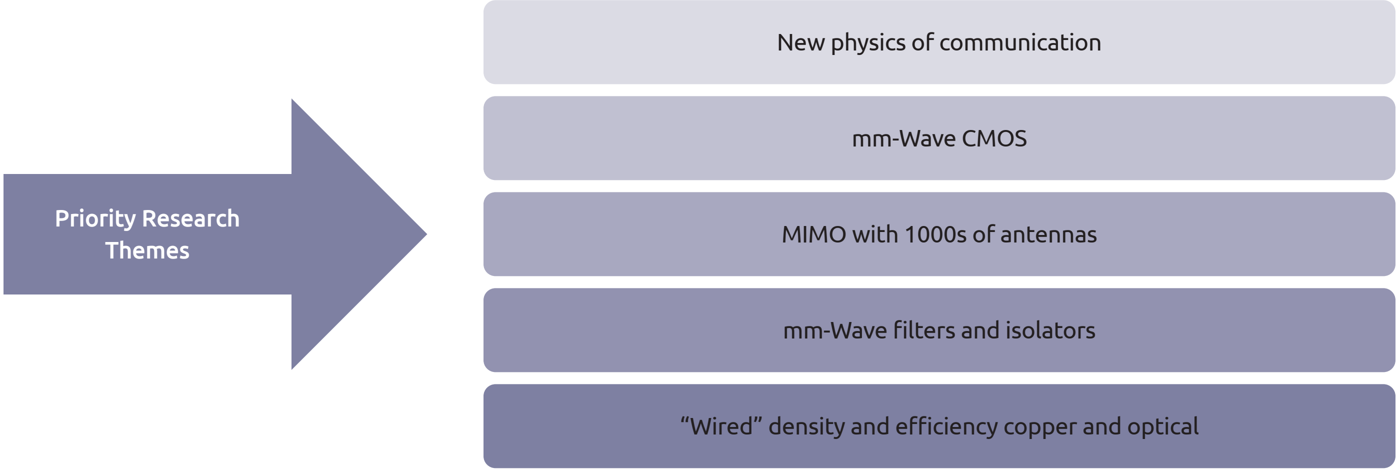
Call for Action

Radical advances in communication will be required to address growing demand. For example, the cloud technologies may undergo substantial changes with emphasis shifting toward edge computing and local data storage.

Broadband communications will expand beyond smart phones to immersive augmented reality, virtual meetings and smart office settings. New capabilities will enrich user experiences through new use cases and new vertical markets. This requires collaborative research spanning a broad agenda aiming at establishing revolutionary paradigms to support future high-capacity, energy-efficient communication for the vast range of future applications. The DOE Office of Science published a report in March 2020 to identify the potential opportunities and explore the scientific challenges of advanced wireless technologies.[4]

Challenges would include wireless communication techniques expanding to THz region, wireless and wireline technologies interplay, new approaches to network densification, increasing importance of security, new architectures for mm-wave, device technology to sustain bandwidth and power requirements, packaging and thermal control.

Invest $700M annually throughout this decade in new trajectories for communication. Selected priority research themes are outlined below.[5]



Priority Research Themes

New physics of communication

mm-Wave CMOS

MIMO with 1000s of antennas

mm-Wave filters and isolators

“Wired” density and efficiency copper and optical

[4] https://www.osti.gov/servlets/purl/1606539

[5] The Decadal Plan Executive Committee offered recommendations on allocation of the additional $3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

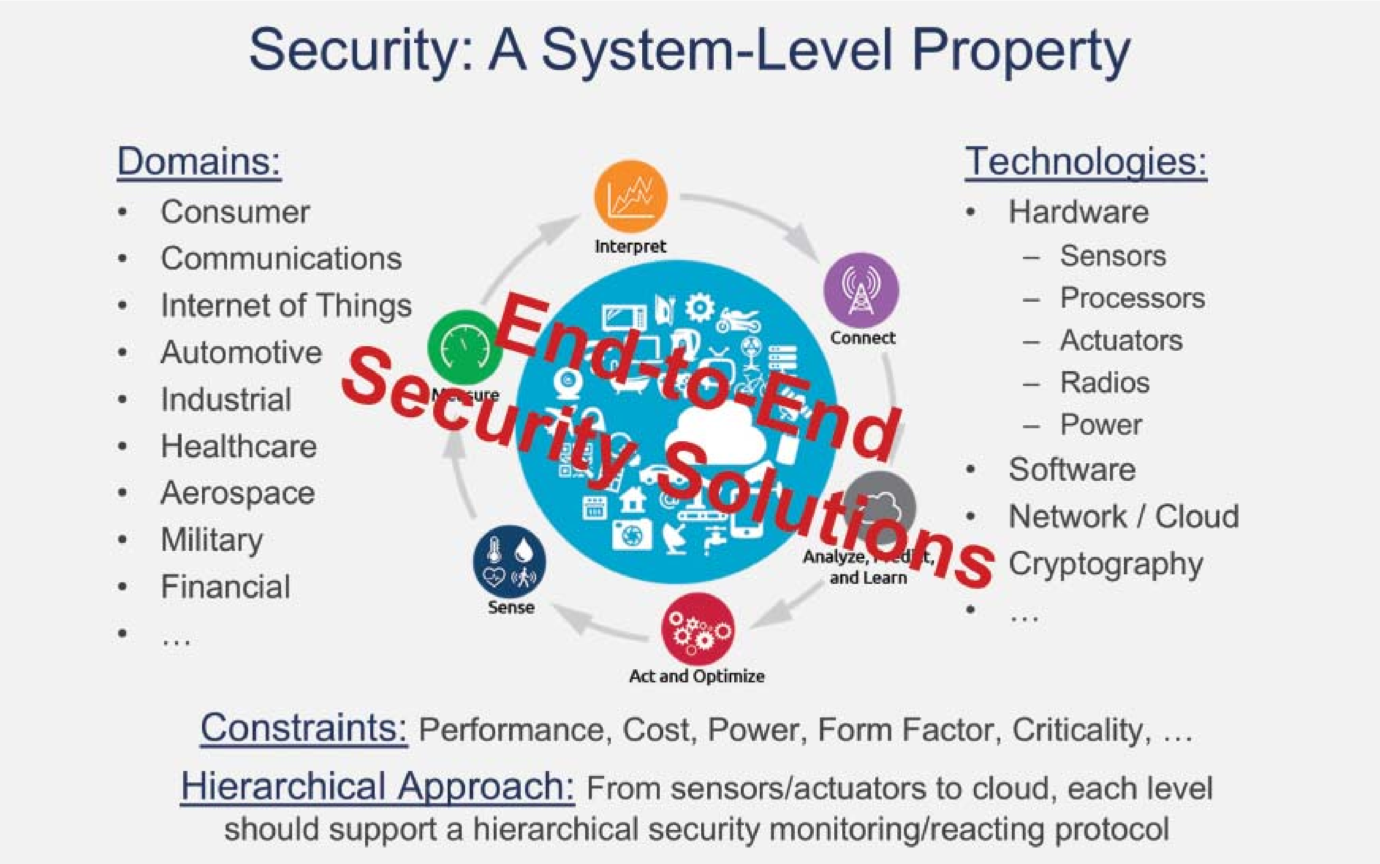
Seismic shift #4

Breakthroughs in hardware research are needed to address emerging security challenges in highly interconnected systems and Artificial Intelligence.

Today's highly interconnected systems and applications require security and privacy for proper operation (Figure 7). Corporate networks, social-networking and autonomous systems are all built on the assumption of reliable and secure communication but are exposed to various threats and attacks ranging from exposure of sensitive data to denial of service. The field of security and privacy is undergoing rapid flux these days as new use cases, new threats, and new platforms emerge. For instance, new threat vectors through the emergence of quantum computing will create vulnerabilities in current cryptographic methods. Thus, new encryption standards resistant to quantum attack must be developed, with consideration given to the impact of these standards on system performance. Also, privacy has emerged as a major policy issue drawing increased attention by consumers and policy makers across the globe. Technical approaches to enhancing privacy include obfuscating or encrypting data at the time of collection or release.

In another direction, devices have permeated the physical world, and thus trust in these devices becomes a matter of safety. Security has thus never been more important. Safety

Figure 7: Systems view of security (courtesy of Yiorgos Makris / University of Texas at Dallas)



Security: A System-Level Property

Domains:
- Consumer
- Communications
- Internet of Things
- Automotive
- Industrial
- Healthcare
- Aerospace
- Military
- Financial
- …

Technologies:
- Hardware
  - Sensors
  - Processors
  - Actuators
  - Radios
  - Power
- Software
- Network / Cloud
- Cryptography
- …

Constraints: Performance, Cost, Power, Form Factor, Criticality, …

Hierarchical Approach: From sensors/actuators to cloud, each level should support a hierarchical security monitoring/reacting protocol and reliability of systems need to consider malicious attacks in addition to the traditional concerns of random failures and degradation of physical-world systems. Security of cyber-physical systems needs to consider how to continue to function or fail gracefully even after attacks. We need intelligent algorithms that sift through contextual data to evaluate trust, to do secure sensor fusion over time. This is a difficult problem as contextual data has tremendous variety and quantity—the systems of the future are actually systems of systems with limitless possibilities for communication and signaling. For instance, cars can communicate with each other and also with roadside infrastructure. Like humans, we need to augment systems with the intelligence to trust or not trust all they perceive.

Our hardware is also changing. Complexity is the enemy of security, and today's hardware platforms are highly complex due to the drivers of performance and energy-efficiency. Modern system-on-chip designs incorporate an array of special-purpose accelerators and intellectual property (IP) blocks. The security architecture of these systems is complex, as these systems are now tiny distributed systems where we must build distributed security models with different trust assumptions for each component. Furthermore, these components are often sourced from third-parties, implying the need for trust in the hardware supply chain. The pursuit of performance has also led to subtle issues in microarchitecture. For instance, many existing hardware platforms are vulnerable to speculative execution side-channel issues, famously exposed by Spectre and Meltdown. Driven by these problems and others, the future requires fundamentally new hardware designs.

The major workload of today is AI. Many security systems, for instance, use anomaly detection to identify attacks or employ feature analysis for contextual authentication. AI's capabilities continue to increase, and applications for these trusted systems continue to grow. However, the trustworthiness of the AI for these systems is unclear. This is a problem not just for security systems but even for general systems with implicit trust assumptions, for instance, visual object detection in autonomous vehicles. Researchers have shown that small perturbations to an image can sway neural network models into the wrong conclusion. A well-placed small sticker on a stop sign can make a model classify it as a Speed Limit 45 sign.[6] Other applications of deep learning systems have similar trust issues: the output of speech recognition might be manipulated with imperceptible audio changes, or malware might go undetected with small changes to the binary. The brittleness of deep learning models is related to their famous inscrutability. Neural networks are black boxes with no explanation for their decisions. Other important problems with neural networks are algorithm bias and fairness. Approaches are needed to make deep learning systems more trusted, explainable, and fair.

Finally, in the last decade, the systems that we must secure have become immeasurably more complex. The cloud has become the standard for outsourcing computation and storage, while maintaining control. We are still grappling with security challenges arising from the cloud—multi-tenancy, provider assurance, and privacy—while cloud offerings continue to increase in complexity. Clouds now offer trusted execution environments and specialized, shared hardware and software. At the same time, interest in edge computing is growing as we realize clouds lack the performance and privacy guarantees of nearby compute infrastructure. The heterogeneous nature of the edge implies trust in edge service providers is a major issue and, of course, the security of IoT devices has plagued us for years. Developing security must be made easier for resource-constrained, often low-cost devices. Even if care is taken in the security design, difficulties arise from extreme environments, such as medical implants. To compound the problem, systems have become more complicated at every level—modern system-on-chip designs incorporate an array of special-purpose accelerators and IP blocks, basically tiny distributed systems where we must build distributed security models with different trust assumptions for each component.

[6] "What is adversarial machine learning?" by Ben Dickson–July 15, 2020 https://bdtechtalks.com/2020/07/15/machine-learning-adversarial-examples/

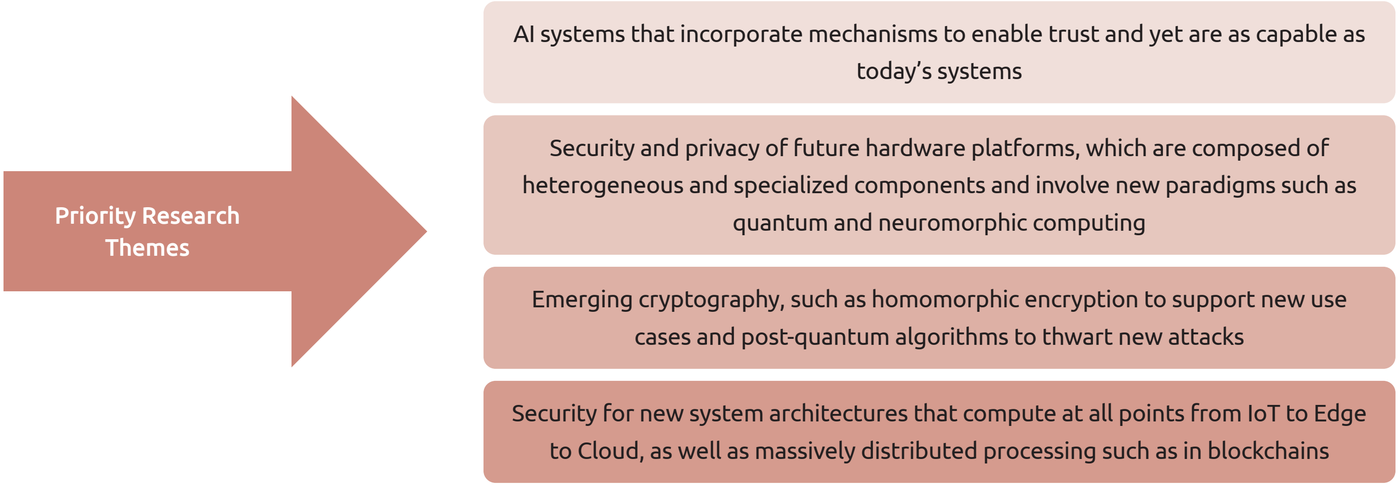
Call for Action

The pace at which today's systems are increasing in intelligence and ubiquity is astounding. At the same time, the increased scale and complexity of these systems have forced hardware specialization and optimization to address performance challenges. All these advances in capability must go hand-in-hand with advances in the security and privacy. Examples include securing weaknesses in the machine-learning or conventional cryptography, protecting privacy of personal data, and addressing vulnerabilities in the supply-chain or hardware.

Grand Goal #4:

Develop security and privacy advances that keep pace with technology, new threats, and new use cases, for example, trustworthy and safe autonomous and intelligent systems, secure future hardware platforms, and emerging post-quantum and distributed cryptographic algorithms.

Invest $600M annually throughout this decade in new trajectories for ICT security. Selected priority research themes are outlined below.[7]



Priority Research Themes

AI systems that incorporate mechanisms to enable trust and yet are as capable as today's systems

Security and privacy of future hardware platforms, which are composed of heterogeneous and specialized components and involve new paradigms such as quantum and neuromorphic computing

Emerging cryptography, such as homomorphic encryption to support new use cases and post-quantum algorithms to thwart new attacks

Security for new system architectures that compute at all points from IoT to Edge to Cloud, as well as massively distributed processing such as in blockchains

[7] The Decadal Plan Executive Committee offered recommendations on allocation of the additional $3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

Seismic shift #5

Ever rising energy demands for computing vs. global energy production is creating new risk, and new computing paradigms offer opportunities with dramatically improved energy efficiency.

Rapid advances in computing have provided increased performance and enhanced features in each new generation of products in nearly every market segment whether it be servers, PCs, communications, mobile, automotive, and entertainment among others. These advances have been enabled by decades of R&D investments by both the private sector and the government yielding exponential growth in compute speed, energy efficiency, circuit density, and cost-effective production capacity. Sustained innovation in software and algorithms, systems architecture, circuits, devices, materials, and semiconductor process technologies, have been foundational to that growth pace. Although this trend has persisted for decades by successfully overcoming many technological challenges, it is now recognized that conventional computing is approaching fundamental limits in the energy efficiency and therefore presenting challenges that are much harder to surmount. Consequently, disruptive innovations in information representation, information processing, communication, and information storage are all pressing and critical to sustainable economic growth and the United States technological leadership.

As the computations per year increases, the number of bits used to support these computations also increases. It is projected that in 2050 we will be dealing with nearly 1044 bits. As shown in Figure 8a, the total energy consumption by general-purpose computing continues to grow exponentially and is *doubling approximately every 3 years* while the world's energy production is growing only linearly, by approximately 2% a year. The rising global compute energy is driven by ever growing demands for computation (Figure 8b) and this is in spite of the fact that the chip-level energy per one-bit transition in compute processor units (e.g. CPU, GPU, FPGA) has been decreasing over last 40 years (as manifested by the Moore's law), and is ~10 aJ or 10^{-17} J in current processors.

However, the demand for computation growth is outpacing the progress realized by Moore's law. In addition, Moore's law is currently slowing down as device scaling is approaching fundamental physical limits. If the exponential growth in compute energy is left unchecked, market dynamics will limit the growth of the computational capacity which would cause a flattening out the energy curve (the 'market dynamics limiting' scenario in Figure 8a). Thus, the radical improvement in energy efficiency of computing is required to avoid the 'limiting' scenario.

The underlying difficult problem is *bit utilization efficiency* in computation, i.e. the number of single bit transitions needed to implement a compute instruction. The current CPU compute trajectory is described by a power formula (shown as inlet in Figure 9) with an exponent bounded by p~ ⅔. The theoretical basis for the observed trajectory and for the value of the exponent is not clearly understood and thus the theoretical basis for computation needs to be further developed. As an observation, if it could be possible to increase the exponent in the formula by only ~30%, the compute efficiency and thus energy consumption would have a 1000,000x improvement. This is illustrated as "new trajectories" in Figure 9.

Grand Goal #5:

Discover computing paradigms/architectures with a radically new 'computing trajectory' demonstrating >1,000,000x improvement in energy efficiency. Changing the trajectory not only provides immediate improvements but also provides many decades of buffer (as shown in Figure 8). This would be much more cost effective than attempting to increase the world's energy supply dramatically.

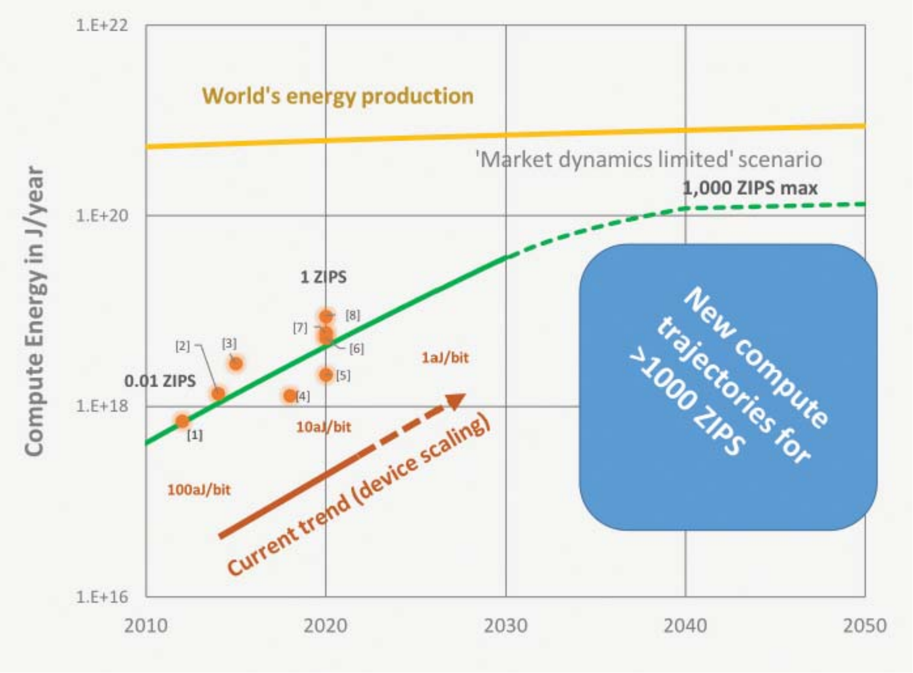


Figure 8 (a): Total energy of computing: The solid green line indicates continuing the current computing trajectory while improving the device energy performance. The dashed green line indicates a 'market dynamics limited' scenario stopping further increase in the world's computing capacity and resulting in a flattening out the energy curve. The blue box indicates a scenario where a radically new computing trajectory is discovered. The Decadal Plan model (green line) is compared to independent data by different groups (circled dots):

[1] W. Van Heddeghem et al (U Ghent), "Trends in worldwide ICT electricity consumption from 2007 to 2012", Computer Communications 50 (2014) 64

[2] IEA (2017), Digitalization and energy, IEA Publications, Cedex, Paris

[3] European Commission (2015), Eco-design preparatory study on enterprise servers and data equipment. Luxembourg: Publications Office of the European Union

[4] E. Masanet, et al (Northwestern U, LBNL, Koomey Analytics), "Recalibrating global data center energy-use estimates", Science 367 (2020) 984

[5] H. Fuchs et al (LBNL),"Comparing datasets of volume servers to illuminate their energy use in data centers", Energy Efficiency 13 (2020) 379

[6] Malmodin et al. (Ericsson, Telia Company), "The future carbon footprint of the ICT and E&M sectors Proceedings of the 1st International Conference on Information and Communication Technologies for Sustainability ETH Zurich, February 14-16, 2013 pp. 12-20

[7] A. S. G. Andrae and T. Edler (Huawei – Sweden), "On global electricity usage of communication technology: Trends to 2030", Challenges 6 (2015) 117-157

[8] L. Belkhir and A. Elmeligi (McMaster U, Canada), "Assessing ICT global emissions footprint: Trends to 2040 & recommendations", J. Cleaner Production 177 (2018) 448

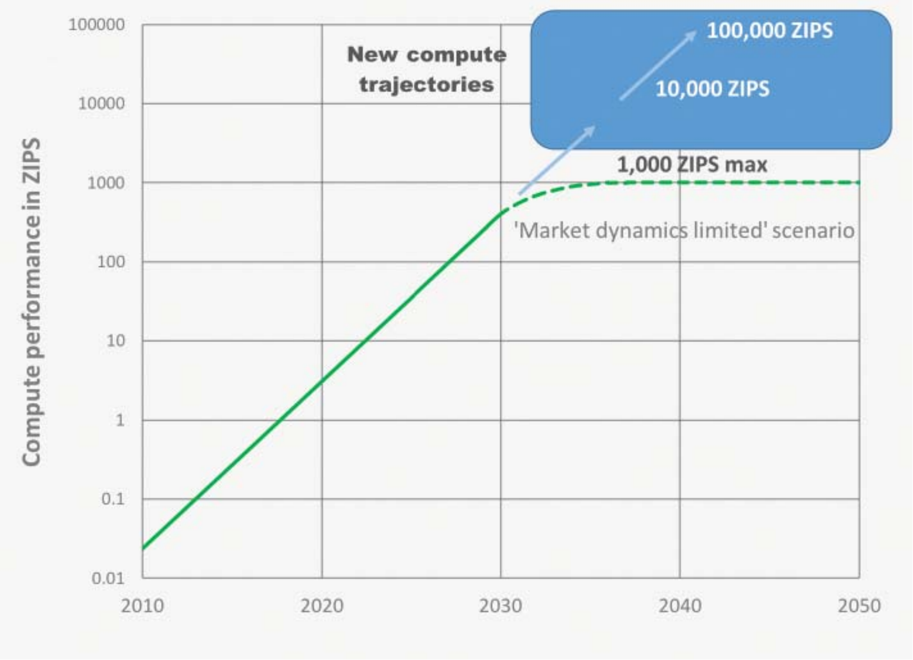
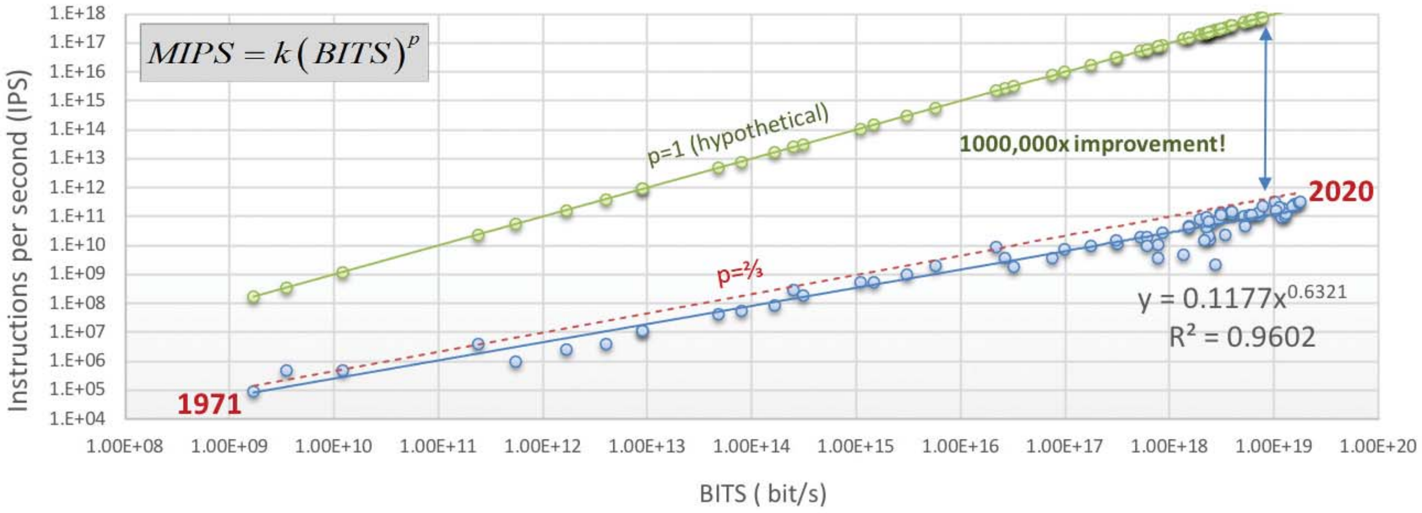


Figure 8 (b): World's technological installed capacity to compute information, in ZIPS, for 2010-2050. The solid green line indicates the current trend (based on research by Hilbert and Lopez[4]). The dashed green line indicates a 'market dynamics limited' scenario stopping further increase in the world's computing capacity due to limited energy envelope. The blue box indicates a scenario where a radically new computing trajectory is discovered.

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

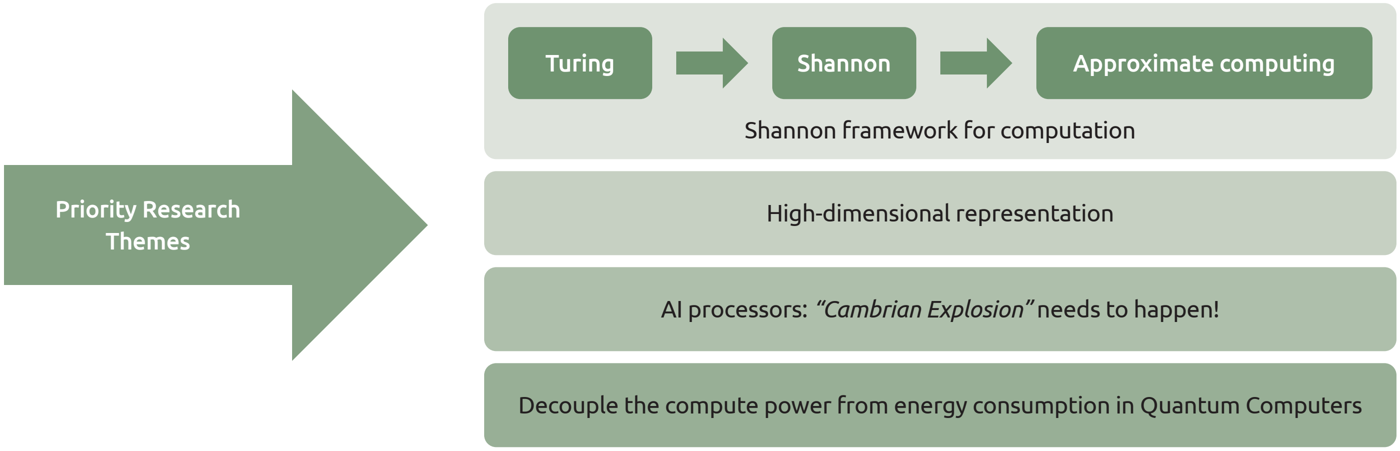
Figure 9: The current CPU compute trajectory



Call for Action

Revolutionary changes to computing will be required soon. Computational loads continue to grow exponentially as evidenced by the growth in "artificial intelligence" (AI) applications and training demands. New approaches to computing such as in-memory compute, special purpose compute engines, different AI platforms, brain inspired/ neuromorphic computation, quantum computing, or other solutions will be necessary and will need to be combined in a heterogeneous manner. The scope of potential heterogeneous computing architectures are described in a recent National Science and Technology Council (NSTC) report on the Future of Computing.[9] This research will require a cross-disciplinary, cross-functional approach to realize commercially viable and manufacturable solutions with multi-decade longevity to replace the mainstream digital approach. This document is intended to stimulate collaborative research 'from materials to architecture and algorithms' to establish revolutionary paradigms that support future energy-efficient computing for the vast range of future data types, workloads and applications. For additional background, see the DOE Office of Science, Basic Research Needs for Microelectronics workshop report.[10]

Invest $750M annually throughout this decade to alter the compute trajectory. Selected priority research themes are outlined below.[11]



Priority Research Themes

Turing → Shannon → Approximate computing

Shannon framework for computation

High-dimensional representation

AI processors: *"Cambrian Explosion"* needs to happen!

Decouple the compute power from energy consumption in Quantum Computers

[9] https://www.nitrd.gov/pubs/National-Strategic-Computing-Initiative-Update-2019.pdf

[10] https://science.osti.gov/-/media/bes/pdf/reports/2019/BRN_Microelectronics_rpt.pdf

[11] The Decadal Plan Executive Committee offered recommendations on allocation of the additional $3.4B investment among the five seismic shifts identified in the Decadal Plan. The basis of allocation is the market share trend and our analysis of the R&D requirements for different semiconductor and ICT technologies.

Semiconductor
Research
Corporation

4819 Emperor Blvd
Suite 300
Durham, NC 27703

919.941.9400
www.src.org