Memory Coalition of Excellence

Recommendations for the National Semiconductor Technology Center

Table of contents

Executive summary **3**

Background **3**

Memory industry **6**

Introduction 6

NAND and DRAM scaling 6

Emerging and other memories 7

Future technology trends and challenges 7

Memory-centric technical scope **8**

The memory wall 8

Memory-centric compute 10

Memory Coalition of Excellence **12**

Infrastructure 13

Collaboration framework 15

Conclusion **15**

List of contributors

Micron

Scott DeBoer
Executive Vice President Technology and Products

Gurtej Sandhu
Senior Fellow & VP Technology Pathfinding

Steve Kramer
Principal Engineer Technology Pathfinding

Ameen Akel
SMTS, Vertical Systems Research

Bambi DeLaRosa
SMTS, Artificial Intelligence

Western Digital

Dr. Siva Sivaram
President, Technology & Strategy

Richard New
Vice President, Research

Didier Ryser
Senior Director, Corporate Strategy

Neil Robertson
Senior Director, Research

Tom Boone
Senior Technologist, Research

Executive summary

Microelectronic chips, also known as integrated circuits (ICs), are at the heart of our modern-day society. They are the essential component of electronic devices, enabling advances in communications, computing, healthcare, military systems, transportation, clean energy, and countless other applications critical to U.S. national and economic security. Discrete semiconductor memory and storage (DRAM and NAND) currently account for almost a third of all integrated circuits sales and are growing faster than any other segment in the semiconductor industry. This trend is expected to continue, and currently, memory and storage account for approximately two-thirds of the world's 300 mm semiconductor wafer output.

Semiconductor memory and storage play an increasingly central role in overall computing infrastructure, largely fueled by the data economy and the current "data explosion" era, resulting in an exponential increase in the amount of data generated and stored in the computing ecosystem. As this data growth continues, workloads and applications are forced to migrate towards more memory-heavy architectures. Additionally, advancements in memory and storage set the pace for semiconductor technology development. Semiconductor memory and storage technology iterates at approximately two times the pace of leading-edge logic and requires the world's most advanced manufacturing processes technology and tooling.

The United States' competitiveness in memory faces several challenges compared to other countries, including economies of scale and more limited investment incentives. As the United States looks to bolster investment in the semiconductor industry through the National Semiconductor Technology Center (NSTC), investment in foundational memory technologies and the establishment of a Memory Coalition of Excellence will be critical to ensure continued U.S. competitiveness in the overall microelectronics space. This effort will require many diverse innovations, including ideas for new memory architectures, new materials, device and process technologies, and advances in manufacturing tooling. This document provides an overview of the memory industry, details the competitive challenges faced by the U.S. memory industry, identifies specific technical focus areas that are relevant to the memory domain, and makes recommendations for each area.

Background

Advancements over the last 70 years in semiconductor electronics have enabled and enhanced countless industries, such as telecommunications (radio, television, telephone, internet), commerce, aerospace and defense, and banking. Every facet of our lives is intertwined with semiconductors. As such, semiconductors play a pivotal role in U.S. national economic activity and national security. The cost (capital and operating) of semiconductor industry competitiveness, however, is high compared to other market segments. The U.S. semiconductor industry annually invests about one-fifth of its revenue into R&D ($44 billion in 2020), the second-highest share of any major U.S. industry, only behind pharmaceuticals[1]. Continued advancements critical to fueling U.S. competitiveness in semiconductors require augmented, sustained investment in core research, manufacturing technologies, infrastructure, and ecosystems.

Worldwide semiconductor revenue totaled $595 billion in 2021, an increase of 26.3% from 2020, according to findings from Gartner, Inc[2]. This increase in revenue was fueled by the growing demands of computing infrastructure, as well as by the COVID-19 pandemic. Figure 1 shows the 2021 breakdown by market segment.

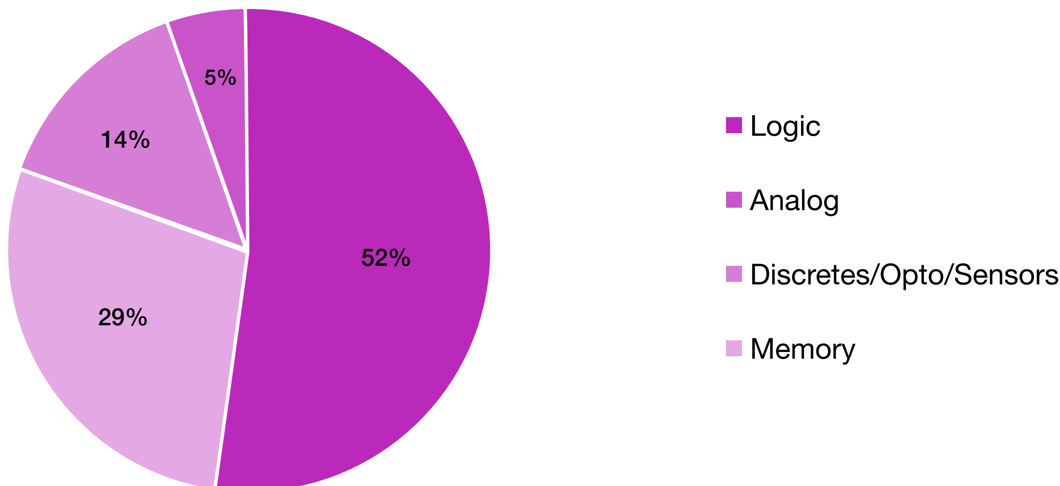


Figure 1: 2021 Semiconductor industry revenue by segment

At a macro level, the semiconductor industry comprises four major segments with common underpinnings in basic solid-state technology, each with its own unique needs and specializations: Logic; Memory/Data Storage; Analog; and Optoelectronics, Sensors and Discrete (OSD) components.

The Logic segment is characterized by integrated circuits that process digital data, whose components require continual scaling for competitiveness in cost, performance, power, and features. Examples include microprocessors, graphics processors, wireless baseband processors, wireless network-on-chip, and microcontrollers, to name a few. The product requirements for the prolific, post-consumer markets underpin telecommunications and the internet (data centers, smartphones, game devices, etc.) advanced semiconductor process technologies. The cost of continually redeveloping and improving logic chips over the past decade has become exorbitant, limiting production capabilities to only a few of the world's largest companies.

The Memory/Data Storage segment is characterized by integrated circuit components that store and retrieve data across a spectrum of performance and retention requirements. This segment is dominated by DRAM and NAND technologies and products and requires some of the most advanced and sophisticated semiconductor process technologies. DRAM and NAND are used, respectively, as working memory and storage for nearly all electronic applications

[1]https://www.semiconductors.org/wp-content/uploads/2020/03/2021_SIA_Industry-Facts_5-19-2021.pdf
[2]Gartner Market Share Analysis, "Market Share Analysis: Semiconductors, Worldwide, 2021."

and systems, including smartphones, PCs, servers, and vehicles. While DRAM and NAND share some similarities, they also have key differences, which are discussed in the following section. The memory segment is unique in that technical innovations are equally critical for both embedded integrated circuit technology, which is produced largely by foundries, and stand-alone products, which are produced in dedicated facilities. To date, the semiconductor industry has experienced a predictable cadence of advancements relying on Moore's Law scaling advantages. However, this cadence is thwarted by technologies approaching atomic scaling limits. The transition to 3D architectural design approaches can extend these scaling advancements, particularly in the case of semiconductor-based memory and storage. Focused advancements in semiconductor-based memory and storage are of utmost importance, given the explosive demand for these technologies, as well as the burgeoning demand for greater performance, improved energy efficiencies, and more advanced functionality.

The Analog segment comprises integrated circuit components that must interface with continuous, non-discrete (non-digital) information, such as information from sensors, electrical equipment, and on-air broadcasts. This segment includes mixed-signal control, where analog signals are converted to digital, and vice versa. This category of semiconductors utilizes specialized process technologies that are tuned for high sensitivity precision requirements and tends to be fabricated using non-leading-edge technologies, referred to as legacy or trailing process nodes.

The last category is a broad catch-all for other semiconductor technologies: Optoelectronics, Sensors, and Discrete (OSD) components. In particular, discrete components perform individual electronic functions such as resistors, transistors, and rectifiers. Like analog components, these chips use trailing technology process nodes, or in the case of some discretes, completely different and less stringent processing.

Given the critical role that semiconductors play across all the segments highlighted in our economy and national defense, it is urgent that U.S. government funded research expenditures reflect the importance of the industry to the country's future security and economic health. While the federal government accounts for 13% of total semiconductor R&D investment in the U.S., this percentage is well below the 22% average across all other technology sectors, see figure 2. The United States is widely recognized for its strong leadership across the semiconductor industry. With the growing importance of memory in enabling next generation compute, it is imperative for U.S. federal investments to prioritize memory and storage R&D.

Comparison of U.S. federal government share in total R&D investment

Across all sectors

22%

Semiconductors

13%

Figure 2: Percent of U.S. federal government investment in total R&D spend, 2018

Memory industry
Introduction

As discussed, by providing foundational capability for AI, 5G, and data centers, memory and storage advancements spur innovation across industries, including healthcare, automotive, communications, and defense. Because of this, and the previously mentioned co-generated "data explosion", memory and storage have grown from 10% of global semiconductor industry revenue in the year 2000 to approximately 30% of the industry's revenue today. This trend will continue as technological advances require increased density, performance, and advanced capabilities. For example, 5G smartphones have 50% more memory (DRAM) and double the storage (NAND) content compared to 4G phones. Today's autonomous vehicles require as much DRAM and NAND storage as found in advanced data center servers. Memory consumption will continue to increase as this technology evolves and proliferates. The International Data Corporation (IDC) predicts that 175 zettabytes (each zettabyte equals one trillion gigabytes) of data will be generated worldwide in 2025. Information storage underpins this data economy, allowing semiconductor memory to permeate nearly every facet of daily life and sets the pace for advancements in the broader semiconductor ecosystem.

The ubiquitous nature of memory in electronic systems means that memory cells make up approximately 85% of the entire device count in semiconductor manufacturing. However, U.S. based manufacturing of memory only accounts for 2% of the global total. This growth will continue, given DRAM's and NAND's essential role in all computing and as a foundation for data-centric infrastructure demand. DRAM and NAND unlock economic opportunity by empowering precision medicine, optimizing smart manufacturing, powering financial services, and helping deliver autonomous transportation. It is vital for the United States to stay at the forefront of memory technology, both because of its significant role in the U.S. economy and the importance of data security. Federal initiatives, such as the proposed NSTC, provide a unique opportunity to support sustained domestic memory technology innovation, thereby bolstering U.S. national and economic security.

NAND and DRAM scaling

While both DRAM and NAND Flash share similar technology elements around basic constructional device formation and back-end metallization, each also drive different unique leading edge semiconductor technology requirements. NAND has several unique requirements, in particular high aspect ratio etch related technologies, which are much more advanced than, in general, logic applications. Similarly, DRAM requires precision deposition of unique materials and leading-edge lithography techniques for high density capacitor structures not needed by other semiconductor segments. For both DRAM and NAND, generational bit growth, cost reductions and ultimately the performance of a wide variety of end products depend on healthy scaling roadmaps.

Emerging and other memories

There are additional memory technologies filling niche application and markets, this includes volatile and nonvolatile memory technologies not readily filled by DRAM and NAND Flash. These include stand-alone SRAM, NOR Flash, and Mask Programmable ROM. The “Emerging Memory” category encompasses developments focused on novel materials and architectures, and are upstarts focused on addressing new tiers in the overall compute paradigm. As well as addressing longer-range scaling limitations of the existing DRAM and NAND roadmaps. These emerging memories include novel materials for the memory storage unit – resistive RAM, phase-change materials (PCM), Magnetic RAM (MRAM), and Ferroelectric materials-based RAM (FeRAM). While ReRAM and PCM have found limited success in niche applications they do not serve as a replacement technology for DRAM and NAND Flash architectures.

Future technology trends and challenges

Continued improvements in data density, bandwidth capability, and power management remain the priorities for the memory and storage industry. These priorities will be enabled by new innovative materials and process technologies allowing for continued technology scaling, in combination with 2.5D and 3D enabled new compute architectures and paradigms, and more advanced system-on-a-chip (SoC) and packaging solutions. With the level of integration of today’s most advanced semiconductor solutions, this R&D effort will also need to include key elements of the technology ecosystem. This ecosystem spans core research across our national labs and academia, equipment vendors for intrinsic process capabilities, heterogeneous packaging innovation to enable product advancements, and cost-effective test methodologies that keep pace with capacity gains.

As DRAM moves into the next phase of development, it faces several challenges as the technology approaches its fundamental physical limits based on currently identified materials and processes. These limitations include very costly extreme ultraviolet (EUV) lithography that requires significant cost per bit scaling. Cutting-edge DRAM in today’s most advanced devices and systems is based on roughly 12 nm to 15 nm minimum features, which as a result of the structure of the DRAM require lithography capability beyond the most advanced logic requirements. As the physical limit for traditional DRAM scaling approaches, there is an opportunity for a disruptive technology transition with significant impacts on industry dynamics. There are efforts in R&D across the globe to disrupt planar DRAM technology by migrating to 3D, similar to the development of NAND. While considerable R&D has explored displacive types of memory technology to supplant DRAM, none have yielded the combination of speed, reliability, and scalability to compete with DRAM.

NAND Flash architecture has already migrated to 3D and each successive new generation of 3D NAND drives increases in areal density of bits by adding more memory layers, which also leads to lateral scaling of the memory array for adding contacts to the memory bits, thereby decreasing each new 3D node’s ability to deliver increasingly cheaper memory. Similar to DRAM, monolithic 3D NAND solutions require tremendous future innovation to continue to enable advancements in performance and cost as processes become more and more complex as the industry progresses to many hundreds, even thousands of layers.

To help ensure the continued pace of bit density scaling and/or bit cost reduction in memory technology, additional research paths must be fortified in “emerging” and new memory concepts based on alternate storage mechanisms. There must also be a concurrent focus on architectural innovations that work to harness the capabilities enabled by new memory technology and maximize system level performance and cost benefits for end applications in the marketplace. These new memory system conceptualizations, or reimagination of the logic-memory hierarchy, can lead to much more efficient systems that sidestep the current limitations by flexibly using memory and logic devices to optimize for substantial system-level performance gains.

In addition, further investments are needed to develop new methods for chip stacking, so-called heterogeneous integration (HI), which requires multi-die bonding and specialized packaging. This technology brings disparate parts of computer architecture that have yet to be integrated homogeneously closer together thus providing for much higher information transfer speed and energy reductions. HI also allows for new architectures to be realized that are too complex and/or impractical for traditional wire and solder ball bonding.

Memory-centric technical scope

The memory wall

Current data handling schemes rely on an architecture in which data storage is separate from data processing. This creates a need to constantly shuttle information to and from memory, which happens at great performance cost, in both time and energy. “The memory wall” refers to this time and energy bottleneck in the system. The huge increases in data quantity driven by advanced analytics, big data, AI, machine learning, and video streaming exacerbate this problem. New memory innovation begins by harnessing methodologies that eliminate costly data movement by making memory more central to compute, creating so called “memory-centric” architectures. Organizations are bringing compute closer to the source of the data, leveraging memory technology innovations like never before, to drastically improve performance and launch a new era of technology transformation. Developing leading-edge memory technology is essential to support this transformation.

Larger gains in efficiency are possible by putting the compute functionality near DRAM, also referred to as near-memory compute. Even greater efficiency is achieved by performing computation directly on the fast memory (like DRAM) for in-memory compute. Analog compute and fully analog accelerators further expand the scope to improve efficiency by providing a large number of possible states for each memory cell and performing compute on a large volume of data in parallel. While this is a promising direction, device characteristics and variability remain key challenges, and a suitable, high-quality analog memory device remains elusive. Certain data-bound workloads are more amenable to certain types of memory-centric processing solutions, or combinations of processing solutions, which is exacerbated by the trend toward domain-specific architectures (DSA). DSAs can achieve higher efficiency by tailoring the architecture to characteristics of the workload, or domain, being addressed.

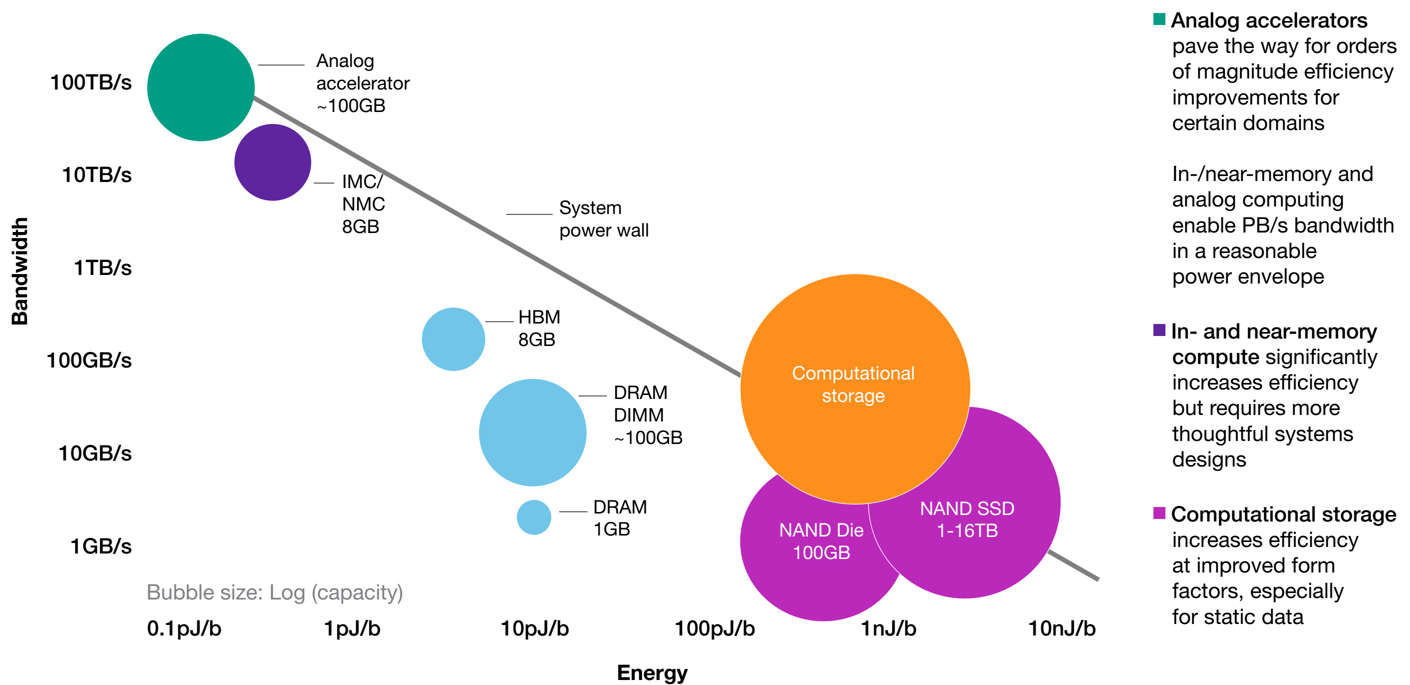


Figure 3: Domain-specific architectures scaling through increased memory efficiency

Each computing system will have a maximum allowable power, which in Figure 3 is labeled the "system power wall." It is a maximum combination of data bit transfers (bandwidth in GB/s) and cost of data movement (pJ/b). The highest efficiency lies with an analog accelerator, but only certain workloads are amenable to this computational method. Therefore, DSAs will dictate how best to distribute workloads between novel and traditional architectures.

We have begun to see integrated circuit infrastructure transform and memory evolve, to support the demands of an increasingly data-driven lifestyle. Memory and storage chip technologies have transitioned into a post-Moore's law regime: the 3D scaling paradigm. This transition has driven a major shift in the development of next-generation technology solutions for materials, unit processes, devices, circuits, and architectures. Using a first principles-based approach, an efficient way to explore and evaluate new materials and devices for successive memory chip solutions is needed. The ecosystem to support this transition, including tools for materials, processes, complex 3D structures, and platforms for materials, structure and TCAD modeling, needs to be developed and matured on a path to support sustained 3D scaling.

Continued progress on this path requires a complete reimagination of the interaction between compute, memory, and storage. Optimal solutions emerge from a holistic view of all components as one, including materials, novel devices, circuit design, architecture, and heterogeneous (3D) packaging, while folding in frameworks, operating systems, software, and application optimizations and still addressing security requirements and needs for new metrology.

Memory-centric compute

Memory-centric compute is the logical path to perform advanced computing at low energy and high performance for memory-bound workloads, including AI inference and training. Any serious move toward memory-centric compute requires integrated innovation from applications to the storage bits, including architectures, frameworks, operating systems, and memory systems. All items in the compute stack must evolve together in a system-level driven memory-centric paradigm.

Embracing memory-centric architectures creates immense potential for radically faster and more efficient compute systems. However, trade-offs in bandwidth and energy will be required depending on the structure of data-centric workloads. To realize this, systems must embrace heterogeneous design, shown in Figure 4, that: 1) evolves from current, general purpose, compute-centric architectures; 2) makes use of new, special-purpose, accelerator-aware designs; and 3) pushes more compute toward memory with novel memory-centric and domain-specific architectures, with data-movement-aware programming models, and tightly coupled memory and compute architectures.

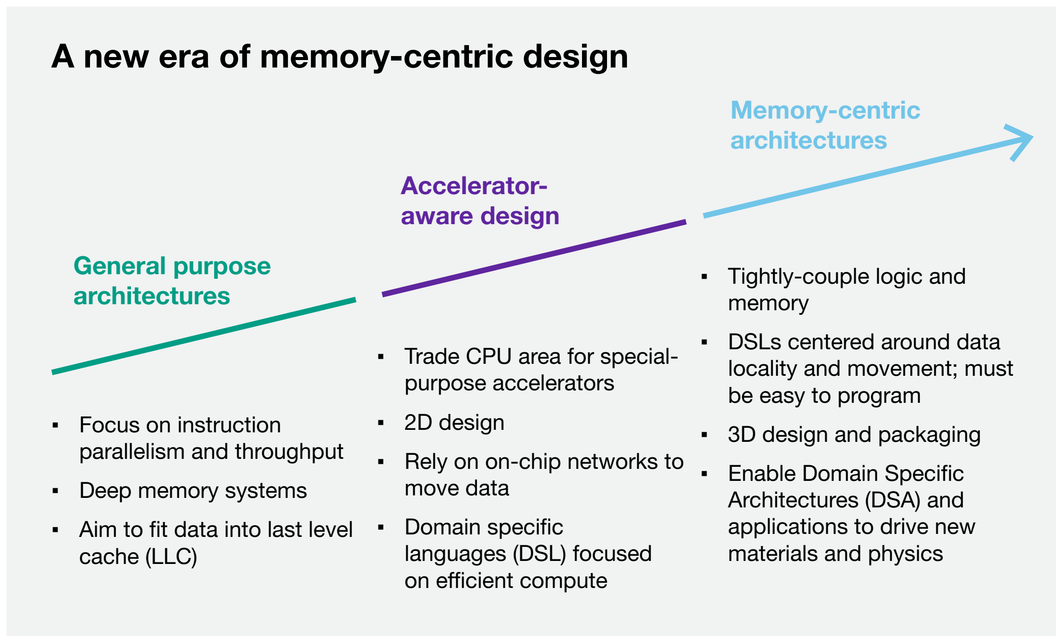


Figure 4: A new era of memory-centric design.

Figure 5 presents the building blocks for this new paradigm of memory-centric compute, both current and futuristic. Bringing workloads closer to memory has begun with stacking memory chips in a 3D fashion, termed high-bandwidth memory (HBM), and integrating these stacks with systems in a 2.5D manner. New memory architectures are now being explored that insert logic functionality into memory chips at the silicon level, both alongside and within the memory array, for deep, in-memory capabilities. Taking this memory and compute synergy one step further, one can imagine the complete merging of memory and logic where analog memory functions are arranged to provide concurrent compute capability.

Building blocks in a memory-centric world

Goal: Explore architectures, frameworks, and operating systems for future memory and storage building blocks

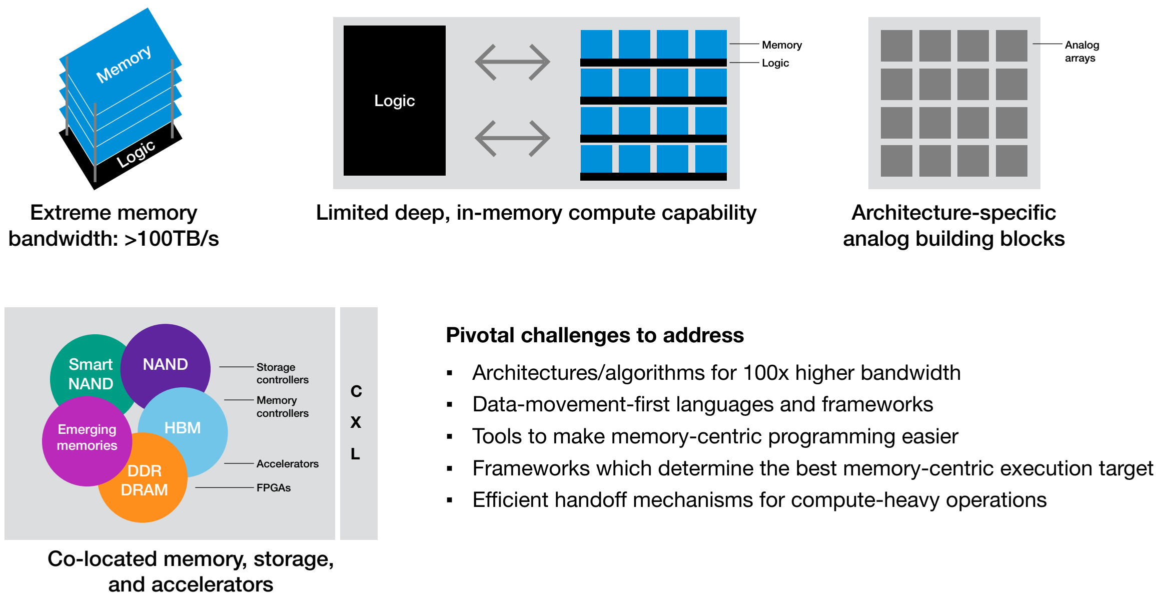
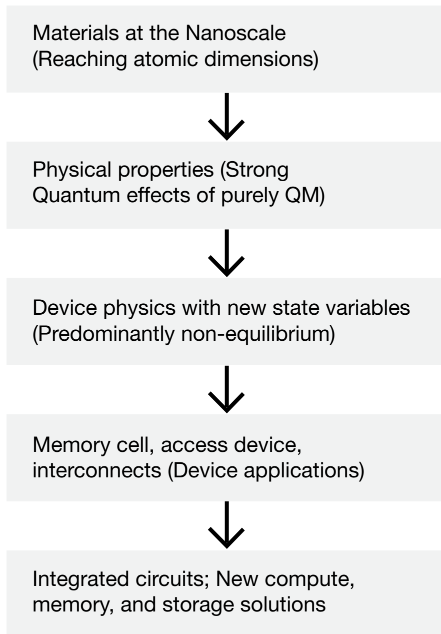
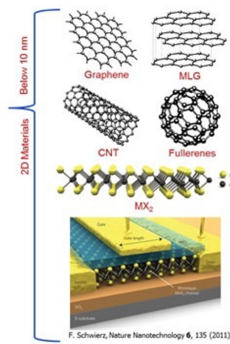


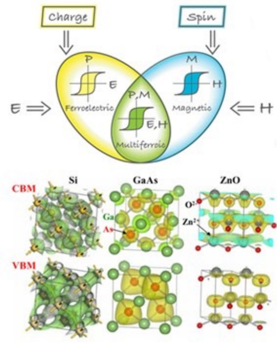
Figure 5: Building blocks of memory-centric compute.

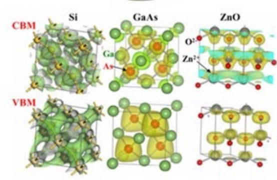
For paradigmatic advancements in memory technology, novel concepts and related materials that can be integrated with traditional devices are needed. To be competitive against current DRAM and NAND technology, any new memory or select device discoveries must provide disruptive benefits in many, if not all, of the key device metrics of performance, power, area, functionality, cost and complexity. The benefit assessment must consider overall system level needs and benchmark across the full stack from materials, processes, device and circuit, and system architectures. Scaling considerations have been pushed to the point that may require device solutions that take advantage of inherently 2D and 1D materials, and possibly concepts that work at near atomic resolutions. It will be important to understand the underlying device mechanisms, and certain device concept pitfalls that, for example, may involve atomic motion in the switching mechanism. This generally results in device level variability or stochasticity. For use in large array implementations, any device performance variability must be near wholly eliminated.

Discover new materials with novel mechanisms and physics for system-level enablement









Needs: New materials, novel physics, and improved materials quality

Memory cell scaling

- Low I_{off} (~attoamp) semiconductor materials
  - Wide band gap (2+ eV)
  - Cycling stability (>1E8)
  - Novel materials (e.g., 2D) for physical scaling
  - Operational at 100C

Neuromorphic compute

- Multi-Level Cell (MLC) with deterministic programmability
  - Compatible with high performance CMOS
  - Low T processing
  - Large programming window
  - Low variability (highly stable)

Unearth novel mechanisms and physics in new material vs. rehashing existing device concepts via the periodic table. Completed in lockstep with first principles modeling driven by system-level benefits.

Figure 6: Novel memory technology R&D. First principles modeling driven by system-level benefits.

Memory Coalition of Excellence

In order to secure U.S. leadership in the critical area of semiconductor memory and storage technology, the NSTC should develop and articulate a long-term (>5 years) vision and roadmap for the enablement of the next generation of these technologies.

A Memory Coalition of Excellence (MCOE) will support this era of transformation and new technological innovations required. This MCOE should be a focused effort across industry, academia and government with clearly defined objectives related to overcoming the challenges outlined in this paper and should be aligned with other key coalitions of excellence (COEs) to support the overall objectives of the NSTC.

The MCOE should focus on pre-competitive research in materials, process, 3D structures and manufacturing technologies for memory and collaborate with other COE's on packaging and interconnect technologies to enable next generation of energy-efficient computing and domain-specific accelerators. The activities should include development of 3D design automation and modeling tools/methods. The MCOE should also identify a set of nation-wide grand challenges facing memory performance scaling that encourage large-scale collaboration across the U.S. semiconductor ecosystem.

Key activities for this MCOE to realize next generation solutions can include for example:

- Advanced research and development for materials, foundational process/metrology technologies, and state of the art analysis techniques
- Modeling methodologies and tools for rapid development and co-optimization of complex technologies and systems
- Next generation 3D memory technology and supporting vectors development
- Heterogeneous integration (functional and/or physical) at wafer and chip level
- X-point array integrated with advanced CMOS for new concept validation
- Advanced packaging to address challenges related to stacked memory chips

The roadmap for the Memory Coalition of Excellence should focus on memory-centric computing architectures using new concepts such as Near-Memory compute, In-Memory compute, and Analog compute with the aim of accelerating pervasive data intensive workloads such as AI inference and training.

Infrastructure

NSTC infrastructure should enable the development of key capabilities for building prototypes to demonstrate game-changing improvements for the next generation of microelectronics applications. The facilities and infrastructure should provide advanced memory/storage, logic, and analog system prototyping with enablement of supporting materials, devices, and packaging.

The envisioned infrastructure for such goals includes state-of-the-art, 300 mm clean room space with leading-edge semiconductor tooling capabilities to fabricate full-flow, concept memory chip prototypes, components and modules, as well as a dedicated systems lab for verification and testing. To ensure a fast ramp from lab to fab, materials and equipment should be co-developed in tandem with process and integration technology in each of the COEs. The infrastructure should include:

- Technology development
  - Process/tool hardware development to build complex 3D nanostructures
  - Tools/materials/masks for advanced mask and wafer patterning (EUV) solutions
  - Platform for development of next generation advanced modeling methodologies and tools (physics, materials, structural, TCAD, design, etc.)
  - Combined materials/device approach for accelerated development of next generation solutions
  - Advanced metrology and materials analysis/characterization tools

- Heterogeneous integration
  - Drive technology vectors for advanced memory/logic heterogeneous integration in collaboration with and leverage other advanced R&D related to Packaging/HI in NSTC
  - Develop advanced wafer-to-wafer and chip-to-wafer bonding technologies with extremely high alignment accuracy and low defectivity
  - Drive advanced structure, stress, materials, and electronic design automation (EDA) modeling solutions
- Memory chip vehicle
  - All required manufacturing tools for full-flow processing and metrology steps in with dedicated 300 mm cleanroom space
  - Advanced metro/testing/characterization resources
  - Memory test chip prototyping support

To realize the greatest value from the NSTC investments, maximize execution speed and synergies with existing infrastructure, and provide access to the best expertise, we propose that portions of the Memory Coalition of Excellence be built in an adjacent annex to existing leading-edge facilities. The MCOE will be configured to facilitate expanded interaction opportunities between industry and university researchers and students. In addition, the MCOE will provide an advanced facility with state-of-the-art tools and a mentoring environment for workforce development. The infrastructure will be designed to ensure easy access for NSTC staff. To facilitate a fast ramp from lab-to-fab, the MCOE will be equipped to support projects for proving out new concepts proposed by start-ups and small companies requiring leading edge semiconductor technology prototyping. As a hub of a distributed network of innovative facilities, the Memory Coalition will serve as an anchor for a collaborative ecosystem encompassing academia, National Labs, startups, and industry participants, including tool and software vendors. This ecosystem will provide infrastructure to support both vertical and horizontal integration, with the Memory Coalition set to work in alignment with other Coalitions of Excellence to deliver leading-edge innovation integrating new advances from across the semiconductor ecosystem.

Collaboration framework

Expanded collaboration in semiconductor concept prototyping is vital to bring memory, storage, logic, and analog together. This would then support a greater focus on heterogeneous solutions and facilitate the combination of new concepts from multiple technology vectors, such as materials, device, circuits, architecture, software, and modeling. A collective, collaborative, Memory Coalition framework with on-site, state-of-the art fabrication facilities will create an environment that attracts the best researchers from government, academia, and industry.

To attract the required talent, the NSTC will need to develop a compelling technological roadmap, provide access to state-of-the-art infrastructure, and demonstrate its ability to provide unparalleled opportunities for interdisciplinary collaboration.

Conclusion

The National Semiconductor Technology Center (NSTC) can play a pivotal role in driving U.S. technological innovation and leadership over the long term. Given the intense pace of innovation in the semiconductor industry and the increasing and expanding demand for memory and storage, memory must be a key focus for the NSTC. A concerted drive by the NSTC can accelerate innovation in memory and storage by enabling next-generation memory-centric design architecture, 3D memory structure technology development, and heterogeneous integration. The NSTC should create a Memory Coalition of Excellence to support focused attention on the aforementioned memory-centric innovations needed for compute infrastructure of the future. Investment in memory advancements will prevent semiconductor-based technology from stagnating and secure the continued cadence of technological advancements, thereby ensuring continued U.S. economic and national security.

©2022 Micron Technology, Inc. All rights reserved. Micron, the Micron logo, the Micron symbol and Intelligence Accelerated™ are trademarks of Micron Technology, Inc. All other trademarks are the property of their respective owners.

©2022 Western Digital Corporation or its affiliates. All rights reserved. Western Digital, the Western Digital design, and the Western Digital logo are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries.