Grand Challenge R&D Priorities for U.S. Government Semiconductor Investments

Recommendations from the Semiconductor Industry Association
Chief Technology Officer Committee

To

Deputy Secretary of Commerce Dabbar & the
CHIPS Research and Development Office

October 2, 2025

1. Scaling AI Inference for Next-Generation, High-Performance Computing

High-performance computing (HPC) underpins U.S. technology leadership, enabling national priorities including artificial intelligence (AI), machine learning, and scientific discovery. As highlighted in the 2025 Semiconductor Research Corporation (SRC) Microelectronic and Advanced Packaging Technologies (MAPT) Roadmap, AI workloads have shifted from emerging to dominant, reshaping both data center and edge computing. However, as deployment of AI continues and demand accelerates, global data center power consumption is projected to exceed 160 GW by 2030 (source: PWC, IEA), creating significant pressure on the grid and on AI providers who will faces escalating cost of ownership for their compute centers. To mitigate these pressures while continuing the U.S. buildout of AI infrastructure, there is a critical need to reduce power consumption at every level—from transistors and chip-level interconnects to packaging, system-scale interconnects, and datacenter cooling.

Large-scale AI training and inference now represent a dominant part of data center workloads, driving a wave of co-design from silicon to systems. While training remains an important focus area, **both centralized and edge AI inference has emerged as the critical capability for deploying large-scale AI models**.

Grand Challenge

To maintain establish and maintain leadership in this domain, the U.S. should:

Achieve a 100-1000x increase in per-user inference throughput for advanced, large language models (LLMs, >500B parameters) compared to today's best scale-out inference systems within a similar power envelope.

These improvements would enable more intelligent and capable models, building on the scaling advantages of reasoning-based architectures such as chain-of-thought models, as popularized by leading AI companies.

Technical Foundations

System level improvements of 100-1000x will require transformative innovation across the entire inference stack and between layers of the stack (system technology co-optimization, "STCO"), from memory and advanced packaging to frameworks, runtimes, and models. And, chip design software must evolve in parallel to enable designers to optimize systems that leverage innovations throughout the stack. Leading-edge systems are increasingly constrained by the memory-bound decode phase, with two primary systems bottlenecks:

1. **Memory Bandwidth at Nominal Capacity**
   Performance scales proportionally with memory bandwidth, even in state-of-the-art inference accelerators.

2. **Interconnect Latency and Bandwidth in Scale-out Systems**
   Scaling out increases capacity and bandwidth but introduces latency and

synchronization overhead. Reducing scale-out degree (e.g., with higher bandwidth and capacity memory) can mitigate these issues.

Key Enablers

To meet the grand challenge within a comparable power envelope to today's systems, the semiconductor industry must deliver the following foundational technologies:

1. **Extreme-bandwidth, Capacity-optimized DRAM-based Memory:**
   Target: >10x bandwidth improvement using DRAM or denser alternatives.
   3D-integrated DRAM is a promising path, as its bandwidth scales with accelerator area, unlike 2.5D memories, which are limited by the perimeter ("shoreline") of an accelerator.

2. **High-performance Interconnects**
   Target: >10x better synchronization latency and >10x better bisection bandwidth at realistic accelerator scales. This innovation could include on-package, intra-rack, and inter-rack interconnects.

3. **Memory-centric Architectures**
   Target: >10x end-to-end inference performance using processing-in-memory (PIM), processing-near-memory (PNM), and processing using memory (PUM). These architectures must maintain DRAM-equivalent or greater capacity and require full-stack innovation (from hardware to software) to demonstrate real-world performance gains.

4. **Thermal Performance**
   Target: Ensure proposed solutions maintain similar thermal performance to incumbent systems. Thermal constraints will define the feasibility of 3D integration and high-density packaging. Innovations in cooling and advanced thermal materials are essential to sustain performance scaling.

Conclusion

Delivering a 100-1000x improvement in AI inference throughput necessitates breakthroughs in effective memory bandwidth and architectures, interconnect performance, and thermal performance. These advancements will unlock the full potential of existing logic capabilities, enabling scalable, energy-efficient systems for next-generation AI inference. By addressing these systems bottlenecks, the U.S. will ensure leadership in high-performance AI infrastructure.

2. Low-Power & Edge Compute

As systems become increasingly intelligent and as data becomes more plentiful at the edge, demand for low power and edge compute will expand significantly. Power efficient compute solutions in edge systems will enable increasing functionality in these products, often deploying capabilities that are currently only possible in a large-scale data center. Localizing processing will also alleviate capacity demand on capacity constrained long-range communications infrastructure to transmit unprecedented data volumes to centralized compute resources. Further, edge processing is critical to drive down latency for real-time actions in end markets like autonomous systems and industrial robotics. Finally, while AI models will continue to be trained in data centers, fully trained models will commonly be deployed on intelligent edge devices, often exploiting "federated learning" architectures.

The term "edge" is relative to large-scale, highly centralized systems and can refer to varying levels of decentralization (even encompassing some smaller data centers or "on prem" systems). For the purposes of this document, the "edge" refers to systems that are not line-powered. Rather, they are commonly powered by a battery (i.e., EV battery bank, coin-cell battery, etc.) or in some cases a remote generation source (i.e., on-board solar cells or other energy harvesting methods). Edge systems are also often mobile (i.e., vehicles, robots, mobile phones, wearables, etc.). For these reasons, low-power, small form-factor solutions are essential to advance compute performance in a constrained power window.

Grand Challenge

A key challenge facing low-power and edge systems is the sheer diversity of use cases. In general, compute performance and efficiency can be advanced by customizing hardware to its workload. However, designing entirely bespoke systems for each edge application is not a scalable strategy. Rather, **plug-and-play chiplet platforms/ecosystems enabling mixed material and heterogeneity will enable more facile customization** to meet a wide variety of requirements at the edge. Moreover, these solutions must deliver efficiency early on—**targeting 100-1000x efficiency improvements over current state of the art—and ensure security as the chiplet marketplace grows and diversifies.**

Technical Foundations

To propel low-power and edge compute across commercial and military applications, chiplet-based memory and storage systems must generate compute paradigms that minimize data movement and improve the energy efficiency of data movement where required. Advanced packaging of system components must also enable dramatically improved interconnect power efficiency. Technological advancements focused on power efficiency must deliver cost, security, and functional safety to promote ownership and usage. Key applications for innovation include:

1. **Mobile phones and laptops**—Ultra-low-power needs in personal electronics is a key driver for low-power/edge innovation. Given the unrivaled market penetration of mobile phones and laptops worldwide (and consumer propensity to replace these devices on a periodic cadence), developing low-power and edge compute for these markets allows providers to achieve scale rapidly. As phones incorporate increasing functionality, demand for **longer battery life**, **always-on connectivity**, and advanced features like **on-board AI, high-resolution displays, and real-time sensing** has intensified the need for **ultra-low-power consumption**. This requirement drives silicon design across multiple domains, such as imaging sensors and display drivers, to optimize power by pushing toward more advanced node processing. In addition, power management ICs must also become more efficient while keeping within a smaller form factor, further driving products to advanced nodes.

2. **Autonomous systems**—Self-driving vehicles, robots, drones, etc.— generate large data volumes from sensors that must then be stored and processed using local memory and compute to guide autonomous decision-making in real-time. However, sophisticated AI inferencing requires much greater compute and memory resources than current technologies and power budgets allow. Aggressive goals for autonomous systems—Level 5 autonomous driving, large-scale deployment of robotics for industrial and home services, and Beyond Visual Line-of-Sight operations for drones—will drive American innovation. Advances in sensors for navigation, perception, positioning, and sensor fusion required for safe and reliable operations must develop while simultaneously improving power efficiency.

3. **Wearable technology**—Innovation in augmented reality glasses, fitness trackers, health care, and other wearables present challenges primarily stemming from the requirements of small form factor and special packaging needs. The small form factor of these devices limits the amount of energy storage available, thus requiring nanopower operations. This calls for new solutions for lowering the processing power—today, industry is far from incorporating all desirable compute in the user wearable device. Wireless communication also represents a significant portion of the energy/power budget, pushing the need for ultra-low-power wireless communications mechanisms. Another emerging area important to wearables and medical applications is flexible packaging. Key issues are the packaging of components made with disparate fabrication technologies combining flexible and rigid sensors on flexible substrates that provide for the unique packaging needs of specific sensor architectures.

4. **Infrastructure needs—**Adding intelligence to infrastructure complements the proliferation of autonomous/mobile devices. “Smart” cities, “smart buildings/factories,” “smart grids,” and “smart highways” will include sensing for security, safety, and resilience across many infrastructure applications, such as bridge integrity, highway flow/management and safety, grid security and management, and intrusion and monitoring security, as well as industrial manufacturing where remote sensors are needed to intelligently manage the process being monitored and controlled. As these uses are often in remote and challenging environments, remote or self-sufficient power is required via batteries or energy harvesting, with the need for 10+ years of service without the need to replace the power source. This drives need for both high efficiency remote sensor function/processing and power management/conversion.

Key Enablers

As previously noted, an important characteristic of “the edge” is the sheer diversity of applications, technologies, and developers that must combine to realize commercially viable products. Creating a pre-production flow to build 1,000-10,000 prototypes (to allow field trial/market validation) is oftentimes an insurmountable obstacle, particularly when package innovation is required (frequently needed for new sensor technology or small form factor requirements). Facilitating prototyping across communities of developers will be particularly valuable for this sector. A platform approach will greatly accelerate such prototypes.

For low-power/edge devices, advanced hetero-integration capabilities to integrate diverse technologies from multiple vendors (including sensors/actuators, processors, memory, radios, etc.) could be a key enabler for substantially accelerating innovation over the next 5 years. This should include evolution of de facto interface standards for inter-chip connectivity.

Conclusion

As we push the boundaries of artificial intelligence and advanced computing, advances in the “core” and “edge” capabilities will complement one another: more data enables greater intelligence, which enables greater autonomy, which can generate more data. Breakthroughs at the edge are as important as leadership in data center compute.

3. Advanced Communications

Advanced communications technologies offer immense promise to enable industries of the future for connected devices, smart cities, and autonomous robotics. Moreover, this segment spans multiple technologies from Bluetooth, to Wi-Fi, to 6G—each with their own respective bottlenecks and challenges. Yet, catalyzing U.S. leadership in this critical sector will not only involve transformational R&D, but it will also require very significant infrastructure buildouts and ecosystem development to ensure that technologies can be successful in the market. The U.S. presently enjoys a strong technical leadership share in semiconductors for advanced communication, but its position in the communications equipment space has slipped in recent decades.

For the purpose of this document, “advanced communications” refers to wireless applications and technologies. While the “full stack” of wireless is critical, the primary critical technologies which impact ubiquitous connectivity are 6G and fast-growing satellite applications. For the last 5 years, U.S. data usage has increased by roughly one-third each year with 2024 showing the largest year-over-year increase ever to reach 132.5 trillion MB. Despite the incredible advancement brought by 5G, a move to 6G will be critical in the not-so-distant feature. As AI continues to be integrated into new products and services, 5G networks are straining. AI applications are expected to increase mobile data traffic beyond the capacity of current 5G networks before the end of the decade. A move to 6G will be essential to support the most advanced AI applications and fully unleash the AI potential and allow the U.S. to be at the forefront of technological innovation.

Grand Challenge

Continued exponential expansion of connected devices and the amount of data traffic places staggering demand on communication networks. Useable spectrum, available power, and equipment cost are all constrained—or severely constrained—resources.

Energy efficiency is key for both infrastructure and terminal devices as each has limitations—whether thermal in infrastructure arrays or thermal and battery energy in mobile “terminals.” Technical targets should include **< 1nJ/bit transferred and latency in <1ms range “guaranteed” for mission critical communications such as robotics, drones, EV, and industrial control.**

Technical Foundations

1. **Increasing bandwidth:** Increased bandwidth requirements naturally drive networks to higher frequencies to deliver higher capacity with more cost-effective arrays. Interference mitigation is required to increase capacity via full-duplex wireless and for robust/resilient and secure communications.

2. **Capacity and coverage:** Each network faces the multi-dimensional challenge of handling enormous traffic density in high demand areas (such as large cities) as well as providing comprehensive coverage to “the long tail” of less densely populated/remote areas. Often,

the optimal solution will require a hybrid approach that exploits a blend of technologies to meet the overall system requirement (e.g., satellite coverage for remote rural access).

3. **Merging communications and sensing:** Integrating sensing with communications provides more effective and robust use of wireless spectrum along with dynamic allocation, effectively boosting capacity. This will be especially critical for sub-mmWave frequencies which enable the longest reach communications. Additionally, in the mmWave frequency bands, technology can be leveraged for other sensing applications such as radar and high-resolution industrial imaging. Sensing will also enable more effective spectrum use/sharing to further enhance robustness and capacity.

4. **Low Latency:** Low latency connectivity is necessary for applications that are "response time critical."

5. **Security/resilience:** In increasingly crowded/hostile environments, avoiding interference/jamming (either incidental or malicious) is a critical system criterion.

Key Enablers

Large, very large, and massive element arrays are key to operating effectively in higher frequency bands and crowded environments. This trend drives several important technology innovations:

1. **Heterogenous integration** advances will be required to integrate diverse technical elements (including antenna functionality) into extremely small form factors.

2. **Radio elements** must become smaller, cheaper, and operate with much greater power efficiency. In some cases, this may involve exploiting specialized semiconductor process technologies.

3. **Beam forming, linearization, interference cancellation, system security, and other advanced radio functionality** can be computationally intense. Advances required in the high-performance compute sector will need to be exploited in wireless systems to achieve these aims.

4. **Standardization** is essential for effective communications solutions. Not only is this required for fundamentals of communications (i.e., frequencies, BW, modulation, protocol, etc.) but also for spectrum use/prioritization and security. Many of the important standards are global.

Conclusion

Advanced wireless communications semiconductor technology is key to advancing society and national security with key opportunity for American leadership by addressing the key enablers identified above. Specific focus on 6G and satellite applications for coverage, capacity, robustness and security is highlighted. Heterogeneous integration is required due to the wide range of frequencies and fundamental technologies enabling efficient solutions.

4. Semiconductor Manufacturing

Semiconductor manufacturing is among the most capital-intensive ventures in the modern economy, with leading-edge fabs exceeding $20 billion in capital expenditures. And, as design and manufacturing cycles continue to get longer, costs continue to escalate. **Reducing the time and cost of manufacturing is key to driving U.S. leadership in the next generation of semiconductor manufacturing and for establishing economically viable pathways to domestic production.**

Technical Foundations

1. **Digital Twins:** Digital twins will accelerate process development, dynamic manufacturing flows, and adaptive processing, and in doing so they will enhance yield and performance. Real-time digital twins connected with physical systems will transform semiconductor design and manufacturing across a broad spectrum of tasks—including new materials discovery and process integration, wafer fault detection, advanced process control, virtual metrology, semiconductor design domains, predictive tool maintenance, etc. In total, full-scale implementation of digital twins holds potential to reduce U.S. chip development and manufacturing costs by >40% and reduce manufacturing development cycle times by >35%. Furthermore, digital twins will accelerate innovation for topics ranging from advanced packaging, to materials discovery, to supply chain risk mitigation. The Semiconductor Manufacturing and Advanced Research with Twins USA Institute (SMART USA) is working in close partnership with industry and the Department of Commerce to reclaim global leadership in semiconductor manufacturing through the development and deployment of digital twin technology.

2. **Advanced Packaging**: 3D heterogeneous integration (3DHI) and the “chiplet ecosystem” represents one of the most significant areas of opportunity in the semiconductor industry across a diversity of end markets—including large chips required to drive unprecedented AI workloads and low-power compute for AI inferencing at the edge. Fabbing system components independently simplifies and shortens front-end fabrication flows, increases yield tolerances, and reduces upfront capital expenditures—effectively reducing barriers to entry for new chiplet providers and unlocking broader innovation. Realizing this opportunity requires new manufacturing and tooling developments (wafer/panel level) that the domestic ecosystem will struggle to make on its own at a pace sufficient to capture technology leadership. Driving domestic leadership in advanced packaging will require a suite of technology developments spanning design/EDA, thermal management, photonics/optical connectors, new manufacturing tooling resembling front-end production, materials, etc. Critically, establishing the U.S. as an epicenter advanced packaging will also require enablement of full-scale/industry-wide standards to accelerate the chiplet ecosystem.

Key Enablers

1. **New Materials to enable CMOS+X (heterogeneous packages):** Materials innovations are vital for next-generation transistor and packaging innovations. At the device level, new materials offer potential for superior electrical performance (e.g., operating voltages,

switching speeds, low loss, etc.), denser memory technologies, and photonic integration. At the packaging level, new materials are essential for substrates and interposers, interconnect materials, bonding, thermal dissipation, etc. Additionally, new materials hold potential for improved mechanical performance, processability, and alleviating reliance on materials with supply chain risk. Given the immense challenge and risk of developing and incorporating new materials for fab integration, digital twin infrastructure can significantly accelerate new materials discovery, process development, and supply chain exploration.

2. **Extreme Ultraviolet Lithography:** EUV lithography—and subsequent iterations including "high NA EUV" and "super high NA EUV"—is essential for advanced node processing. But in the semiconductor industry, innovations seldom operate in isolation of one another, and EUV lithography is no exception. In addition to driving advancement in the coming generations of EUV lithography (e.g., wavelength scaling, power scaling, etc.), the industry must also advance innovation in photoresists that can deliver more chemistry per incident photon and can deliver those changes at lower energies of incident photons. Moreover, these photoresists must be applied on top surfaces at thinner and thinner dimensions to increase numerical aperture, which will require new deposition techniques that do not trade off uniformity. Finally, high throughput measurement techniques to ensure thickness and uniformity for wafer dispositioning will also be needed. Notably, computational lithography software platforms (including digital twins) can accelerate R&D of cutting-edge lithography capabilities. As transistor geometries become smaller and more three-dimensional—requiring longer processing flows to build complex geometries—reliance on EUV stands to increase. Given that EUV lithography modules are the most expensive tools in the fab, investments should support research that alleviates requirements for additional EUV processing, thereby helping companies to manage escalating costs of manufacturing. Additionally, fundamentally new techniques to compete with EUV (e.g., free electron lasers) hold potential to challenge exclusive reliance on EUV where 2D scaling remains essential.

3. **Equipment:** While semiconductor factory tooling is known for its formidable upfront capital expenditures, tools also entail high operating costs due to energy requirements, use of expensive processing chemical and precious materials, etc. Increasingly, tool providers are facing demand from device manufacturers for solutions that drive down operating expenses. These solutions span energy efficiency, rapid switching between idle and production modes (to reduce energy consumption during non-processing time), use of collaborative robots ("co-bots") to perform simple maintenance tasks without need to open a production tool, integrated in-situ wafer-level metrology to limit need for stand-alone process monitoring (and enable adaptive processing), and processing technologies that maximize atomic efficiency and minimize waste of chemicals and materials.

Conclusion

Advancing U.S. competitiveness in semiconductor fabrication across an increasing array of products and end markets relies not only on our ability to push the boundaries of physics; it also requires investments that are specifically aimed at controlling costs and challenging current innovation and process development practices.