JUMP 2.0 Research Needs

SRC/DARPA Critical Research Themes:

7 Centers of Excellence for 2030 and Beyond

MISSION STATEMENT

Continuous innovation in semiconductor technologies has spurred one of the most prolific and substantive periods of growth in human history. From early post-WWII transistors to a collective and precise mastery of electron flow, we've seen astonishing leaps in scientific discovery and performance gains. Semiconductor technologies have immeasurably expanded the limits of scientific inquiry and delivered unprecedented levels of prosperity, security, and freedom. Over the coming decade, machine intelligence that seamlessly integrates technology into our daily lives promises revolutionary advances in all aspects of human society.

That promise comes with an asterisk; many future opportunities are unachievable on current trajectories. The underlying hardware is approaching fundamental physical limits. Predictably linear hardware performance improvement and scalability are breaking down, and the current software-driven value paradigm in information and communication technologies (ICT) must be adapted to a new era of microsystems integration and HW/SW codesign.

The SRC/SIA Decadal Plan for Semiconductors calls for industry and government action to address the existential challenges before us. Outlined as five Seismic Shifts, they are:

- **The Analog Data Deluge:** Fundamental breakthroughs in analog hardware are required to generate smarter world-machine interfaces that can sense, perceive, and reason.
- **Memory & Storage:** Growth of memory demands will soon outstrip global silicon supply, presenting opportunities for radically new memory and storage solutions.
- **Communication Capacity vs. Data Generation**: Always-available communication requires new research directions that address the imbalance of communication capacity vs. data generation rates.
- **ICT Security:** Breakthroughs in hardware research are required to address emerging security challenges in highly interconnected systems and Artificial Intelligence.
- **Compute Energy vs. Global Energy Production**: Ever-rising energy demands for computing vs. global energy production are creating new risks; new computing paradigms offer opportunities with dramatically improved energy efficiency.

Preeminent scientists and technologists across 16 industry-leading companies have outlined seven critical research themes to address these rising challenges. Representing both commercial and defense interests, the science advisory team took a holistic system-driven approach, drafting a proposal for high-risk, high-reward investigation by complementary, multidisciplinary Centers. Together, the Centers will drive foundational breakthroughs that address the Decadal Plan's 'Grand Challenges' outlined for 2030 and beyond.

In crafting these themes, the science advisory team was informed by two other key source documents:

- The 2019 IEEE Heterogeneous Integration Roadmap, which seeks to *"stimulate application-driven, pre-competitive collaboration among industry, academia, and government to accelerate progress."*

- DARPA Electronics Resurgence Initiative 2.0, which seeks to forge "*forward-looking collaborations among the commercial electronics community, defense industrial base, university researchers, and the DoD…"* around critical application and technology focus areas.

SEVEN CRITICAL SYSTEMS (S) AND TECHNOLOGY (T) THEMES

1. Cognition (S)
2. Communication and Connectivity (S)
3. Intelligent Sensing to Action (S)
4. Systems and Architectures for Distributed Compute (S)
5. Intelligent Memory and Storage (S)
6. Advanced Monolithic and Heterogeneous Integration (T)
7. High-performance, Energy-Efficient Devices for Digital and Analog Applications (T)

Core principles underlying every theme:

- Microsystems Integration relying on disruptive semiconductor technologies and HW/SW innovations and codesign as key drivers of ICT value.
- Radical improvements and abstractions needed at every stage in the value chain from materials to systems, and related information representation and processing.
- Innovations must demonstrate improvement in at least one area, without compromising overall system performance. System-level modeling and demonstration of proposed improvements and trade-offs is critical.
- Commercial and Defense converge to share application space insights; drive and fund the research agenda; support researchers with guidance, insight, active collaboration, design and prototyping.
- Organic, evolutionary cross-theme collaboration will be a critical element determining the success of the program.

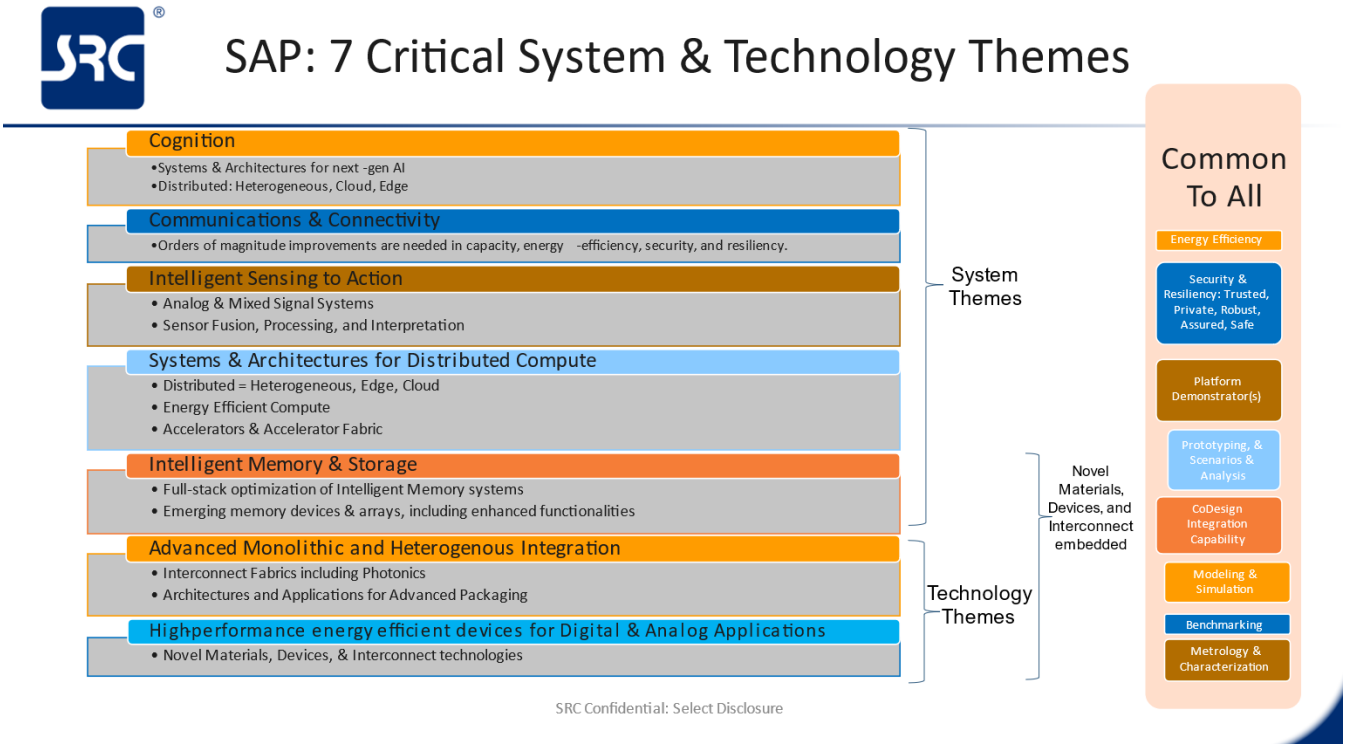

Figure 1 Critical Research Needs

THEME RESEARCH TRAJECTORIES

1. Cognition (S)

Goal: Create cognitive computing systems that continuously learn at scale, securely and efficiently perform reasoning and decision making, operate with purpose, resilience, autonomy, and interact with humans, the environment, and other intelligent systems naturally and in real time. A key goal is to create systems that, without explicit objectives, operate in the natural world on their own by forming and extending models of the world they perceive around them, and by interacting with local human decision makers and global distributed intelligent networks in performing actions to achieve useful yet complex goals. It is proposed that research would span 5 key areas:

Novel Compute Approaches
To successfully build machine intelligent systems with both cognitive and autonomous characteristics, this theme will explore multiple approaches that go beyond traditional deep learning. These novel systems may significantly leverage nontraditional computing methods, relying on new information representation and processing paradigms such as hyper-dimensional representation, analog computing, stochastic computing, Shannon-inspired computing, approximate computing, and bio/brain-inspired models including neuromorphic computing. Such systems can be solely nontraditional, solely von Neumann or a combination of both elements while enabling heterogeneity of multiple architectures (CPUs, GPUs, FPGAs, accelerators, compute-in-memory, etc.) and scalability to large systems. They should seek disruptive benefits in energy-efficiency, performance, and scalability while supporting enhanced functionalities and cognitive capabilities.

Beyond von Neumann Innovation
The von Neumann processing approach of separating memory and processing limits further advances in computation, driving unacceptable latency, thermal, and power inefficiencies. Fundamental improvements in performance and energy efficiency require breakthroughs in novel information representation and processing, programming paradigms, algorithms, architectures, human-machine interfaces, circuits, and materials and device technologies.

Systems Approach to New Compute Paradigms
Exploration of the related system architecture are critical in seeking compute paradigms that are fundamentally error-resilient, energy-efficient, support continuous learning, enhance cognitive capabilities, support better scalability with problem size, and are inherently secure end-to-end, explainable, and trustworthy. The emergence of in-memory computation (i.e., novel accelerators tightly coupled to high-capacity memory) and highly heterogeneous systems will continue to provide challenges to system architects and developers. Where cognitive-level performance is required, regardless of type, the design and management of the flow of data and information will need to be optimized in the systems, materials, devices, algorithm and software programming domains, and balanced against the benefit provided from the new processing approaches.

Cross-Disciplinary Codesign
This research must span the gamut of theory, testing, validation, and demonstration of a complete scalable compute system, requiring a cross-disciplinary effort embracing systems software, programming abstractions, algorithms, architectures, circuits, devices, materials, tools and methodologies. This type of cross-disciplinary codesign will be necessary to enable the efficient deployment of Artificial Neural Networks and other cognitive models from datacenter to the real-world edge.

Co-simulation of and codesign of heterogeneous semiconductor elements, as well as co-optimization of software stacks and accelerators, will be a necessary consideration.

Modeling and Benchmarking
There is still no clear sense of a universal path beyond the current Von Neumann trends, despite much research evaluating various options and evaluating high-level performance in the lab. It is believed that a legitimate and complex research task remains to conduct high level trade-off analysis & benchmarking of the emerging solutions in terms of system performance, size, and power. From an architectural standpoint there are many other quality attributes that might be inspected in a research environment, such as programmability, time to develop solutions, reliability, and performance over a range of environmental conditions. While less germane to technology three to five years out, applied to future technology needs in 2030 and beyond, they can greatly inform the transferability of the research into the real world.

Specific need areas include, but are not limited to:

1. Theoretical underpinnings, algorithms, and foundational techniques beyond deep learning
2. Novel architectures, systems, and programming models
3. Demonstrate feasibility in real-world settings
4. Full system energy efficiency, performance benefits using industry relevant models with industry relevant datasets
5. Proof of concepts that demonstrate at-scale capabilities in one or more domains
6. Methodology to validate reliable, trustworthy, secure cognition systems
7. Cognitive system- and architecture-driven devices, circuits, and materials research in collaboration with other Centers, especially themes 5-7. Envisioned is a fundamental examination/re-engineering of compute stack via codesign of materials/device/circuit/ architecture/algorithm/programmability.

Outcomes: Benchmarked to traditional computing approaches (on state of the art COTS hardware), novel cognition systems should achieve 100X better on at least one key metric (energy, speed, area of the chip, complexity, reliability, etc.) without compromising other key metrics, enabling new killer applications.

2. Communication and Connectivity (S)

Goal: Uncover new technology approaches that enable high-bandwidth, energy-efficient connectivity for a range of future applications for: (A) wireless communication; (B) system-to-system communication, and (C) chip-to-chip communication. Future wired, wireless, and optical communication systems operating from GHz to THz frequencies face an imbalance between communication capacity vs. data generation rates. Data generation and movement will be at the heart of the always-available, deeply integrated, and immersive systems of the future; bandwidth for machine-to-machine traffic will surpass human-to-machine, presenting new challenges and opportunities for systems and networks. Future compute and memory system improvements must be accompanied by communication and connectivity breakthroughs and novel communications technologies. Orders of magnitude improvements are needed in capacity, performance, energy efficiency, security, and resiliency.

The theme will address enabling devices, component demonstrations, and integration approaches that will demonstrate compelling improvements in performance, connectivity capacity, and energy efficiency (i.e., pJ/bit) across the example applications. Peak rate transmission must support 10x

improvement in zettabyte-scale data movement. Networks must provide robust service, intelligently use bandwidth to support future workloads, and maximize network capacity.

To achieve the necessary improvements, vertically integrated breakthrough research efforts are needed in five key areas:

1. Application-driven, software-programmable architectures and demonstrations
2. Components and integrated subsystem solutions to enable communication and connectivity architectures
3. Enabling devices and components to achieve system-level performance
4. Heterogeneous integration and unique packaging approaches
5. Security, data integrity, and trust with a focus on moving data across challenging and typically lossy channels (wired, wireless and optical) requiring innovative technology and signal processing.

Ever-increasing data center compute and wireless backhaul will require growing data rates with varying reach and require advances in optical communication density, efficiency, and integration. Additionally, wireless communications capacity, data rate with increasingly crowded spectrum with exponentially growing IoT, will require more selective and adaptive solutions with steerable beam arrays. For extreme operating conditions (electro, thermal, mechanical), current modeling and metrology techniques break down; peak junction temperature measurement, for example, is virtually impossible for high-power arrays. Wired and optical communications require significant improvement in interface and modulation efficiencies, such as electrical-optical-electrical conversions and packaging interfaces with low loss at >100Gbps data rates.

Specific need areas include, but are not limited to:

1. **Devices and Materials** that will enable RF-THz communication signals for wired, wireless and optical front ends (e.g., power amplifiers for THz frequencies). Integrated photonics modulators, detectors, and processing, as well as novel low-loss interconnect (optical and electrical) will enable advanced efficient communication solutions. Modeling and metrology techniques will be necessary to design and predict performance.
2. **Array Technology and Architectures** with active sensing and beam forming/tracking processing at very high efficiency. Sensing spectrum and dynamically adapting communications links to best efficient channels while mitigating interference will be required for increased capacity and efficiency.
3. **High-efficiency data modulation, adaptive processing, and data recovery** for advancing data rates of challenging wired and optical links (including chip to chip) driving to 0.1pJ/bit at Tbps rates. These solutions will specifically address difficult channels leveraging advanced equalization, crosstalk mitigation and full duplex operation with associated signal processing/conditioning.
4. **Packaging for Communication and Heterogeneous Integration** solutions, including integrated antenna arrays and low loss in the mmWave-THz regime. Heterogeneous technologies include CMOS, III-V (i.e., GaN, InP, SiGe), photonics, passives, and low loss interconnects.
5. **Communication Solutions, Processing and Architectures** at the system level with holistic optimization will provide the optimum partitioning and effective solution. The needs for 5G and beyond will require novel and new signal processing to address capacity and efficiency. As overall communication bandwidth increases, chip-to-chip communication will need to be addressed to avoid a bottleneck with the Tbps required at that level. Finally, new RF/THz generation techniques for both reach and bandwidth will be required for wired, optical and

wireless systems. The holistic optimization needs to include efficiency, capacity, energy as well as robustness and security.

Outcomes: The broad research agenda defined will address the world need to move data and information in a timely manner and keep up with the exponential data growth. The capacity and capability will enable radical improvements in our communication infrastructure and devices for a range of applications and systems. It would enhance our ability to communicate insight via timely delivery of key data and information, significantly increase capacity and energy efficiency, reduce latency, and enable ubiquitous worldwide access to world information in real time.

3. Intelligent Sensing to Action (S)

Goal: This theme seeks fundamental breakthroughs in analog hardware to generate smarter world-machine interfaces that can sense, perceive, and reason. These qualities are essential to the next generation of AI and edge devices. Application demands in both civilian and defense spheres rely on the rapid expansion of the sensor interface surface, and sensor processing to support actuation and machine intelligence. However, the exponential growth of analog data— 'the analog data deluge'— cannot be effectively utilized on current technology trajectories, and existing approaches to sensor data processing will strain communication, storage, and compute infrastructure. Future sensing systems must be highly efficient, reliable, autonomous, and secure. They must take a holistic, application-specific approach to utilize and address the growing deluge of sensed data, with efficient devices and cognitive signal processing to synthesize data, make decisions, and take timely action.

To address these solutions across the indicated applications, research and innovation are required across the sensing-to-action "loop". Application examples include medical, robotics/autonomous systems, infrastructure, national security and defense, and energy management. All fundamentally rely on the sensed domain to deliver future workloads.

Specific need areas include, but are not limited to:

1. Devices including novel sensors themselves, which greatly reduce subsequent processing
2. RF and THz devices and components, which extend sensing modalities
3. Photonics and other modalities of sensors (can all reduce downstream signal processing)
4. Array enabling technologies and signal processing including mmWave/THz, multi-sensor/sensor fusion and time-sensitive networked sensor systems (will add intelligence and robustness in decision making)
5. Packaging, heterogeneous integration and energy efficiency will all be elements enabling an intelligent sensing solution
6. Power management and delivery will also be critical analog elements of future systems which are increasingly power density limited and will require technology breakthroughs. A holistic system view and development methodology will be necessary to explore and optimize the intelligent sensing solution and define individual blocks.
7. Careful threat analysis of the novel sensors must be included and the implications on the end to end system must be understood. For example, sensors data should not be blindly trusted by downstream processes in the sensing-to-action "loop" and algorithms to determine the reasonableness of sensor data and detect sudden or unusual changes must be designed into the overall system.

Outcomes: Sensing and actuating systems which effectively reduce the expected data deluge by multiple orders of magnitude while providing robust and timely action for rapid response. New sensing capabilities in the THz regime enabled by devices, circuits and systems will provide higher levels of intelligence and insight into medical, infrastructure and defense. Robotics and industrial control systems will be better able to react to changes on a local basis with reduced communication needs. Coupled with improvements in other themes, defense applications will have greater reach and capability to respond. Overall, there will be new physical insights and management for world health, safety, and energy.

4. Systems and Architectures for Distributed Compute (S)

Goal: Breakthrough advances in distributed, energy-efficient general-purpose computing to enable sustainable, scalable heterogeneous 'systems of systems' of specialized accelerators, accelerator fabrics, and architectures.

Future distributed systems will need exponential performance improvements, energy efficiency, resilience, confidential computing (minimal trusted hardware and software footprints, i.e., small secure hardware enclaves and assumption of untrusted hypervisor and OS), and connectivity at any time and place. To achieve necessary efficiencies for emergent workloads, integration of a broad fabric of novel non-von Neumann system architectures, data-driven computational approaches, and in-memory based computing systems will be required. The scope envisioned is the entirety of the distributed computing infrastructure, across a broad range of applications from consumer to defense.

Full-stack, Application-Driven Integration
The systems developed should take a full-stack, holistic view of materials, devices, architectures, and software to meet the computing needs of emerging applications, including challenging, compute-intensive applications as well as highly constrained (energy, bandwidth, storage etc.) applications. Novel architectures and application/workload specific accelerators promise performance and efficiency, but introduce complexity and cost for computer architects, system designers, and developers of system software, run-times, and applications.

Application-driven research is needed to comprehend and quantify projected end-to-end benefits, efficiency, and scalability for target workloads. Simulation and modeling to support sensitivity analysis and trade-offs is a desired result, to project scaling efficiencies for real-world applications across the system stack over otherwise state-of-the-art evolutionary solutions. Desired insights include corresponding novel specialized accelerators, design complexity and cost, tooling, implementation complexity, programmability, software stack and integration, etc.

Emergent Heterogeneous Systems
Research should encompass a highly heterogeneous system architecture of both traditional and novel architectures and devices dynamically evolving over time, with components of different age, compute and sensing capabilities, power consumption requirements, and environmental limits. Heterogeneous systems should be composable with field-replaceable components (GPUs, CPUs, FPGAs, accelerators, and in-memory compute devices) and allow dynamic system performance optimization across a broad range of workloads and compute scale. 'Micro data centers' should deliver machine intelligence where most efficient. The distributed system infrastructure must be secure, resilient, serviceable, and infinitely scalable, including seamless inclusion and support of billions of edge devices. Collectively, system elements and devices must work collaboratively for

greater overall effectiveness, efficiency, and resiliency, and dynamically adapt to continuously deliver predictable, optimized performance and scale.

Cross-Disciplinary Codesign

Codesign will be necessary to enable energy efficiency at any scale. The design and management of the flow of data and information through the system will need to be optimized in the materials, devices, systems, algorithm, and software programming domains, and balanced against the benefit provided from the new processing approaches. The proposed research should yield orders of magnitude improvements in scale, capability, and efficiencies over the 10-year timeframe.

Specific need areas include, but are not limited to:

1. Architecture:
    - Breakthrough energy efficiency at all levels, >100x power reduction; SWAP constrained edge computing including ultra-light nodes.
    - Disaggregated, composable architectures supporting dynamic allocation and de-allocation of virtualized hardware resources based on application demands
    - Robust architectures providing predictable performance across a range of operating environments, and in the presence of connectivity interruptions and component failures (i.e., self-repair, self-recovery, multi-node resiliency).
    - Adaptive, dynamic methods to realize ambient computing
    - HW/SW co-optimization to optimally partition computation within tiers and across tiers
    - Reducing effort, time, and cost for: (1) development and deployment of the accelerator ASIC, (2) the tool chain (compilers libraries, etc.) for the ASIC, and (3) for the development, deployment, and maintenance of applications (application life-cycle management) that leverage the accelerators.
    - Interconnects: as cores per node increase, innovations are required for the interconnects between cores to replace buses.
    - Memory:
        - Rethink of virtual memory and cache hierarchy
        - Large global address space
        - Exploitation of NVM density and power characteristics—architectural manifestation, address space management
    - New consistency and coherence models to offer better performance where traditional constraints can be relaxed, e.g., relaxation algorithms, possibly DNN training, etc.
2. Intelligence and Systems Management:
    - Intelligent systems management to collaboratively provide a desired solution
    - Focus on self-management methods for monitoring, dynamic optimization of resource utilization, and dynamic application optimization
    - Distributed system management tools, including modeling and characterization
    - Effective execution for distributed learning and inference; middleware services for AI including multi-modal distributed learning
    - "Intelligence as a Service" delivery: functions at the edge should leverage compute capacity of the cloud to implement state-of-the-art AI for a broad spectrum of applications
3. Security and Privacy:
    - Integrated security: i.e., trusted hardware footprint (enclave, trust zone) on chip
    - Evolve confidential computing to comprehend heterogenous HW and resource disaggregation
    - Ability to include non-trusted networks while delivering overall security, privacy, and resiliency to failures and attacks

- End-to-end comprehension of security, including policy, metrics and auditability, enable security orchestration
- New algorithms and programming models focused on security and privacy
- Confidential computing with zero software trust zone, i.e., untrusted hypervisors, OS and system software stack

4. Software:
    - Enable rapid, lower cost design of components such as specialized devices and accelerators and reduce cost/complexity to integrate new compute elements
    - Making new technologies usable by system/software designers - programmability, systems and dynamic performance management, composability, virtualization, etc.
        - Programming language constructs, Architectural abstractions, microarchitectural innovations
    - Development tools for distributed systems design
    - Future-proof compute infrastructure, such that newly installed devices support dynamic optimization of system/platform stack
    - Address emerging workloads (applications) in architecture, system management, and data privacy, e.g., graph neural networks (GNN), post-quantum cryptography (PQC)

Outcomes: Scalable, energy-efficient distributed systems that deliver optimum levels of heterogeneity for general purpose compute workloads. Leveraging of key innovations, collaborations, and results from the Communication, Compute, and Memory and Storage themes, to achieve the overall distributed computing and edge intelligence goals.

5. Intelligent Memory and Storage (S)

Goal: Innovation in system and memory subsystem architectures to achieve >100x overall improvement in application performance. Improvements in memory- and storage-constrained application performance should be achieved through an integrated approach to power, performance (bandwidth), area and cost (PPAC) while addressing necessary RAS and security requirements.

As projected in the SRC Decadal Plan, the exponential growth in global demand for memory bits and interdependent computing power - driven largely by data movement to and from memory – is unsustainable. Continued and successive system improvements will require fundamental changes in memory and storage paradigms and technologies.

Full-stack Memory and Storage-Centric Approach

As the cost of data movement is a key problem, increasingly, systems will need to be optimized for memory and storage. Research is needed to investigate the necessary changes in memory and storage as a vertical system. To achieve optimal PPAC, it may be necessary to co-architect and codesign memory and processing elements. Both conventional and novel compute components must move closer to memory, and even into memory in opportune instances; in many cases these ASICS will be designed to perform specialized domain-specific functions as part of a heterogeneous architecture. Innovation should support a robust component ecosystem, and deep technical relationships between memory and ASIC suppliers, and perhaps third-party foundries. These changes are expected to permeate all layers of the stack from the application layer, through the operating system, and the underlying system and memory subsystem architectures, including the various abstractions along the data path. Holistic system consideration will encompass necessary enhancing components such as hypervisors and virtual memory systems, as well as data layout, security, and management.

Architecture and Algorithms
The lines between traditional storage and memory are becoming blurred as applications increasingly use sparse access to extremely large pools of data. The need to efficiently search large pools of data calls for the introduction of large-scale content addressable memories. Algorithms with low operational intensity are becoming widespread, calling for system architecture changes that can facilitate the introduction of near-data-computing concepts. Compute-in-memory/storage requires architectural changes for data layout in memory systems, new methods for managing cache coherence, memory management for distributed compute, data transform, and alternate methods of data encryption.

Virtualization and virtual addressing are complex topics in the context of in-memory compute and need to be considered as well. In-memory compute must coexist with the rest of the developed memory system: it must gracefully, securely, and performantly understand and comprehend virtual addressing. Datacenters often virtualize their hardware by multiplex applications to serve multiple users on the same hardware to maximize efficiency; in-memory compute hardware will benefit similarly. Today, memory sub-systems across many classes of systems make heavy use of virtual memory to facilitate programmability and security. Pioneering research in this vein must expand this processor-centric concept to in- or near-memory accelerators without hindering developed performance advantages.

Device/Circuit/Package-Level Innovation and Integration
To achieve optimal PPAC, an architecture-driven approach is required to identify and incorporate novel ideas and advances in memory materials, memory devices, and access device innovations. Bottlenecks and power/thermal design challenges must be addressed by the overall memory system design and underlying technologies.

Electrical interfaces between and amongst system components will be especially important in heterogeneous system architectures, so significant innovation in algorithms and in signaling will be paramount. It is expected that new memory-centric integration methodologies will require a wholly new "heterogeneous architecture," i.e., a re-engineering of the entire memory/compute system. Coincident with new architectures, it is expected that new concepts for advanced package-level integration and 3D device layer stacking be explored.

Collaboration within and across Centers will be required. A successful research agenda will ensure necessary core expertise in memory systems and underlying technologies, and effectively leverage complementary integration and device research in other themes.

Demonstrator Platform Guidance
Realizing the true benefit and gains of architecting and optimizing the system around *new memory and storage paradigms*, will require exploration of existing applications beyond the scale possible today. It is expected that the Center will need to co-develop a "Grand Challenge" application (or complementary set) to explore and demonstrate new system and memory subsystems and new compute paradigms. Mapping to extant or emerging systems may not deliver the necessary degree of innovation. Flexibility and programmability of the computing components for the Grand Challenge(s) is also paramount, such that general applicability may be achieved, and it may be necessary to explore a run-time resource management system to dynamically tune workloads to the current state of the environment.

Benchmarks and Modeling
Given the difficulties in prototyping this new system architecture, it is expected that full-stack modeling will play a very important role in successful innovation within this theme. Modeling should encompass system, architecture and circuits, and the impact of materials and device attributes. It is

expected that modeling will provide new approaches to memory/compute system characterization and metrology, new methods for efficiency gains, and augmentations by AI and ML methods.

With the emergence of in-memory acceleration or domain-specific acceleration, Amdahl's law tells us that another portion of the workload will become a bottleneck. With memory acceleration as an example, the memory bottleneck may be removed, but this will leave storage, network, or compute as the limiting portion of the system architecture. Thus, this field of research requires a forward-looking mindset to identify and tackle new bottlenecks. Extension of current system-level benchmarking will be required to analyze impact of emergent bottlenecks, e.g., storage and networking.

Specific need areas include, but are not limited to:

1. Architectural solutions for >100Ximprovement in PPAC, while addressing RAS and security of memory and storage
    a. Memory and data storage paradigms
        i. Identify and develop killer apps for in-memory computing at different levels of memory size and memory types
        ii. Access paradigms
        iii. Content addressable associative memories
        iv. New methodologies for storage and memory
    b. Near-Data computing:
        i. Compute in or near memory
        ii. Compute in storage
        iii. Merged memory and logic through heterogeneous integration
    c. Memory controller architectures
        i. For heterogeneous domain specific computing or distributed computing
        ii. In the context of processing near or in memory
        iii. Facilitating sparse accesses
    d. Memory sub-system architectures that embrace novel memory electrical interfaces and innovative packaging strategies, e.g., high bandwidth, low energy per bit, low latency
2. Enable new applications that utilize architectural solution above, e.g., memory/storage architectures for
    a. Advanced algorithms (e.g., hypergraph for AI/ML)
    b. Cognitive computing
    c. Distributed computing, e.g., combined disaggregated memory & storage architecture to minimize the impact of data movement
    d. Applications for heterogeneous systems
    e. "Grand Challenge" applications that take advantage of tight coupling of memory, compute and domain-specific accelerators
3. Overhaul current modeling methodologies to achieve powerful, new, holistic methodologies that better and more efficiently simulate the upper system hierarchy—system architecture, circuits, cells, down to memory elements.
    a. System modeling supporting architecture analysis
    b. Circuit design and modeling
    c. Impact of device and materials attributes against target workload(s)*
    d. Data science and ML methodologies to augment modeling
4. Innovative metrology, characterization, and test platforms for memory systems
    a. Novel memory systems figures of merit

5. Novel memory device concepts and related materials and processes that can be integrated with CMOS and/or novel select devices and bring in disruptive benefits in PPA, functionality, and complexity in at least one level of the memory hierarchy.*
    a. Materials, processes, and device architectures for volatile and NVM memory cells: addressing physical origin of switching, device non-idealities (variability, endurance, drift, TDDB, etc.), BEOL compatibility, scaling, and ultrahigh energy efficiency.
    b. Memory materials (and compatible select device materials for 3D memory/storage: materials, deposition and deep etching methods enabling stacked-layer configuration and/or 3D-NAND type embodiments of NVM, with an eye toward BEOL compatibility.
    c. Atomic-scale memory devices, in pursuit of ultimate scalability
    d. Front-end devices
6. New Select devices for 3D and Stacked memory cells
7. Fundamental examination/re-engineering of in-memory compute stack requirements and the implications on codesign of materials/device/circuit/architecture/algorithm/programmability:
    a. Device, materials, and architecture to enable small, low-cost compute + memory + sensor systems
    b. Compute/processing-in-memory (CIM/PIM), e.g., digital and analog embedded NVM devices
    c. Compute-near-memory, e.g., digital embedded NVM devices
    e. Hardware for secure computing, e.g., homomorphic encryption, physically unclonable functions, random number generators
    f. Deep learning acceleration, e.g., multi-state and analog synaptic cells with optimal stability, noise, etc. for inference, and update behavior, endurance, etc. for training
    g. 3D deep learning architectures—logic, memory
    h. New device concepts and reimagined technologies that provide for deterministic and large resistive ratio memory devices that are ideally suited for multi-state, toward analog-like performance (may include, but not confined to magnetic, FE, resistive, and PC materials).
        i. Memory technologies for analog and/or brain-inspired compute -> device / novel materials that can reliably achieve deterministic, multi-state functionality.

*Suggested collaboration with Theme 7: i.e., device center to identify top materials/device with said attributes, with Systems modeling of impact of the identified devices across a broad workload.

Outcomes: Successful innovation through research in this area will result in monumental improvements in the efficiency and throughput of algorithms operating on large data sets. In-memory-compute alone offers the potential of orders-of-magnitude reduction in energy per operation, again, due to the inherent reduction of energy required to move data.

6. Advanced Monolithic and Heterogeneous Integration (T)

Goal: Radical improvements in 3D monolithic and heterogeneous integration and packaging system integration to deliver >100x performance density and efficiency for future compute platforms. To achieve the performance density and efficiency required for future architectures and devices, research must drive fundamental technology breakthroughs for new logic and memory tiers, interconnects, and power and thermal infrastructure. Research must also clearly ‘connect the dots’

between architecture, devices, and manufacturing, enabling breakthrough system-level architectures that scale to manufacturing and assembly. This will result in not only higher performance and more power efficient solutions but also enable smaller, more flexible, and more portable solutions with a faster time to market and lower cost.

The scope of this theme is threefold:

1. Address the challenges of ever-increasing system-level integration density requirements of heterogeneous technologies as an integrated packaging solution with multiple dies, and / or as a monolithic integrated single die/substrate solution—each where the resulting solution is viable and most cost effective. Objective is to achieve >100X power efficiency versus other methods where size and power value justify cost of integration.
2. Devise novel disruptive interconnect fabrics including but not limited to ultralow resistance interconnects both at the chip-level and package-level, and silicon photonics fabrics.
3. Explore the related materials and process enabling platform-capable solutions for 1 and 2 above.

Challenges include the codesign/development of the integrated solutions rather than individual components, as well how these functions will be interconnected, physically packaged, powered, and thermally managed. Additionally, monolithic integration requires research regarding processing multiple layers of technology on a single substrate and process compatibility (materials and thermal cycles) along with the needed interconnects. Packaging solutions will require research into methods of stacking, attaching, interconnecting multiple die (i.e., “Chiplets” or other) in a single package integrated solution with die attach and die-to-die bonding, interface, and co-development for a “system” solution.

Semiconductor packaging materials that are both environmentally friendly and reduce waste in their production and use need to be developed. Heterogeneous integration and thin die handling require innovations in Wafer Support systems and in improving die strength, these challenges must be solved through both new materials including high temperature temporary adhesives or bonding materials, as well as improved die handling systems and high accuracy die placement or self-aligned or alignment tolerant structures. The investigation and application of nano materials in packaging must be advanced through practical implementation methods for some of these monolayer and 2D materials. Modeling, EDA and verification tools to integrate different components, circuits and integrated chiplets will require development of new smart AI based layout and verification tools that tie together the ability to simulate and verify circuit up to package and Signal Integrity and Power Delivery networks.

Fundamental packaging research should include thermal, magnetic, low-loss dielectrics, interfaces & interconnect transitions (metallics/photonics/wireless) and encompass modeling and metrology for the widest possible range of operating characteristics to cover the scope of applications including high performance computing, IoT, automotive, medical, space and defense environments. At the component level, it is expected that performance will achieve near-equivalence to monolithic integration.

Key packaging elements may include scalable non-volatile memory that can be integrated with CMOS; new device/memory/analog/sensor tiers needed in compute-in-memory, near memory computing, THz communication, and large-scale cognitive computing; novel on-chip 2D and 3D interconnect with high density, including efficient power distribution and revolutionary thermal management technologies to support >2X increase in power density. It is viewed that new architectures will drive materials and devices for 3D monolithic integration that will lead also to new system implementations. Key enablers may include bottom-up integration or layer transfer, low thermal

budget processing, high reliability, improved thermal management (architecture and materials), novel and efficient test and validation methods, design, and fabrication compatibility with 2.5D/3D Packaging solutions, and advanced modeling that encompasses new materials through system-level modeling and PPA assessment to narrow design space and accelerate prototyping.

This theme should support and enable prior themes by addressing the integration of multiple different technologies (i.e., heterogeneous). Advanced packaging of heterogeneous technologies will be required to demonstrate feasibility, performance, and energy efficiency for future commercial and defense systems, but may present significant challenges for Systems Centers.

Specific need areas include, but are not limited to:

4. Exploratory research into materials, processes, and manufacturing methods for dense, low-latency, low-power electrical interconnects as well as silicon photonics with CMOS-technology compatibility
    a. Disruptive monolithic or heterogeneous integration concepts for active and passive devices and interconnect fabrics
    b. Synthesis of custom materials, devices, and interconnects and predictive modeling
    c. Novel power delivery and heat removal scalable solutions
5. Modeling of technologies and designs for application requirements
    a. Predictive materials modeling and nanomechanical structural modeling including interfaces
    b. Full-stack modeling—process relation to structures and interfaces throughout manufacturing, including thermal effect (e.g., self-heating effect, heat dissipation scheme, etc.)
    c. Enabling data science/ML for augmenting materials and structural modeling
    d. Benchmarking of technologies and designs for application requirements
6. Innovative metrology, characterization, test platforms, and from devices to systems.
    a. Thermal characterization
7. Advanced Manufacturing Technology for Integration, including:
    a. Methodologies for achieving high connectivity, low defect, 3D stacking of chips and device layers
    b. High-aspect ratio etching
    c. Techniques to address misalignment
8. Enabling the fundamental examination/re-engineering of compute stack via codesign of materials/device/circuit/architecture/algorithm/programmability and conceptualizing disruptive monolithic or heterogeneous integration scalable solutions. Leveraging and enabling novel concepts from all other themes is expected.

7. High-Performance, Energy-Efficient Devices for Digital and Analog Applications (T)

To achieve the energy-efficient performance and scale outlined as a goal in the Decadal Plan, this theme seeks disruptive innovations in advanced active and passive devices for FEOL or BEOL integration for orders of magnitude improvements in scaling, energy efficiency, density, performance, throughput, latency, and novel functionalities. Breakthroughs in the foundational technology building blocks from novel materials and processes to devices play a key role in enabling ultra-scalable distributed compute, communication, sensing, networking, memory, and storage systems to meet the demands of future workloads.

A broad, horizontal research agenda is needed to showcase novel materials with new functionalities and properties that can augment and/or surpass evolutionary improvements in state of the art conventional semiconductor technologies for both monolithic and heterogenous interconnects. For devices and materials targeting novel algorithms for unconventional computing, codesign with circuit and architectural considerations is essential, with results supporting a wide range of different figures of merit and operating characteristics – i.e., thermal and power budgets, etc.

Materials development, innovative metrologies, and device demonstration of viable process integration are within the scope, providing guidance to future manufacturing. Along with experimental demonstrations, benchmarking and multi-scaled physics-based modeling are expected.

Discovery of new classes of materials are likely needed to achieve the orders of magnitude improvement in device, array, chip, and system properties that form the goals in the other Themes and the Decadal Plan. The limited pace, and lack of systematic method to synthesize and characterize the billions of possible material combinations, significantly constrain the materials search space and limit the research output to evolution of known materials. This theme encourages exploration and implementation of a systematic and deliberate approach to enable materials screening at >100X the conventional pace, by leveraging parallelism in materials synthesis and characterizations.

Specific need areas include, but are not limited to:

1. High-performance, energy-efficient devices for logic, memory, analog computing, power and sensing
    a. FEOL high-performance devices beyond Si
    b. BEOL compatible high-performance devices
    c. Nonvolatile logic
    d. Low-thermal budget logic devices for 3D
    e. Low power logic
    f. Bio-inspired electronics
    g. Integrated photonics for sensing and computing
    h. Integrated very high efficiency power conversion and delivery devices
2. High-performance passive components
    a. High-density capacitors for power and >300GHz performance
    b. Magnetics/inductors for integrated RF and power
    c. Low-loss interconnect: metallic, optical, new material
3. Novel memory (e.g., scalable NVM that can be integrated with CMOS) and exploratory memory element concepts
    a. Memory materials and selectors for 3D memory/storage: materials, deposition and etching methods enabling stacked-layer configuration and/or 3D-NAND type embodiments of NVM
4. Devices enabling disruptive memory applications including but not limited to BEOL compatible transistors for dense memory arrays and innovative Selector devices
    a. Transistors supporting Ultra-low leakage (< atto-amp) and High on/off ration (>10^6) for memory-centric applications
5. Modeling of technologies and designs for application requirements for synthesis
    a. Predictive materials modeling and nanomechanical structural modeling including interfaces
    b. Full-stack modeling—process relation to structures and interfaces throughout manufacturing, including thermal effect (e.g., self-heating effect, heat dissipation scheme, etc.)

    c. Enabling data science/ML for augmenting Materials & Structural modeling
    d. Benchmarking of technologies and designs for application requirements
6. Advanced Manufacturing Technology & Integration, including advanced patterning
    a. Selective processing (ASD, ALD, ALE, MLD)
    b. Atomically precise processing and interfaces
    c. Ability to control bottom-up process
    d. Processing techniques that do not create surface damage
    e. High-throughput atomic lithography capabilities
7. Innovative metrology, physical-electrical-thermal characterization, test platforms, and from devices to memory systems
8. Material development beyond logic, memory, and interconnect needs:
    a. Materials for > 300GHz RF and analog devices
    b. Materials for photonics: laser integration, photonics processing, interconnect
    c. Thermal materials and materials for extreme environments (thermal shock, high E-field, radiation hard, etc.)
9. Enhanced Materials Discovery. HTE approaches that enable coupling between
    a. High throughput synthesis,
    b. Rapid measurement of the most important 1-2 physical properties, and
    c. AI/ML guided algorithm to learn from the measured data and intelligently identify more promising material compositions for further synthesis, are encouraged.

    *Examples can be found in the state-of-the-art research in clean energy and catalysis, and recently initiated thermoelectric material screening. Detailed description and further examples of HTE applications in electronic, magnetic, optical, and energy-related materials can be found in Green et al, Journal of Applied Physics 113. 231101 (2013), https://aip.scitation.org/doi/10.1063/1.4803530).

10. Enabling the fundamental examination/re-engineering of compute stack via codesign of materials/device/circuit/ architecture/algorithm/programmability. List below provides examples from 6 themes above as a guide; not inclusive or prescriptive.
    a. Optimized device, materials, and architecture to enable for novel algorithms and nontraditional compute, e.g., stochastic computing, high-/hyper-dimensional computing, spiking NN
    b. Device, materials, and architecture to enable small, low-cost compute + memory + sensor systems
    c. Compute/processing-in-memory (CIM/PIM), e.g., digital and analog embedded NVM devices
    d. Compute-near-memory, e.g., digital embedded NVM devices
    e. Hardware for secure computing, e.g., homomorphic encryption, physically unclonable functions, random number generators
    f. Deep learning acceleration, e.g., multi-state and analog synaptic cells with optimal stability, noise, etc. for inference, and update behavior, endurance, etc. for training
    g. 3D deep learning architectures—logic, memory
    h. New device concepts and reimagined technologies that provide for deterministic and large resistive ratio memory devices that are ideally suited for multi-state, toward analog-like performance (may include, but not confined to magnetic, FE, resistive, and PC materials).
        i. Memory technologies for analog and/or brain-inspired compute -> device / novel materials that can reliably achieve deterministic, multi-state functionality

i. Components and architecture for stochastic computing-based Boltzmann Machines and deep learning network:
    i. Low-cost, high-performance, massively parallel, scalable, tunable stochastic components and circuit implementation concepts
    ii. Architecture- and circuit-level feasibility study for demonstrating orders of magnitude advantage beyond von Neumann-based architecture.