Chapter 10
Digital Processing

Scope
The roadmap for digital processing examines the current and emerging drivers for digital processing and the required innovations in the processing paradigms. These requirements dictate the necessary technologies and architectures of processing, memory and support chiplets, interconnection and the overall system architecture. The chiplet and interconnect architectures, in turn, drive the required device and interconnection solutions as well as the development of the necessary materials and processing/manufacturing technologies. The roadmap also specifies ancillary requirements and solutions that are needed for system-level integration of digital processing systems, and related system-level considerations for overall security, power conversion/delivery, system reliability and run-time management needs, which are discussed in other chapters of the roadmap.

Major Impediments and Challenges
This Digital Processing chapter of the MAPT Roadmap identifies several impediments/challenges in realizing heterogeneously-integrated digital processing systems including the need to:

- Address the cost of data movement in terms of performance (latency and bandwidth), end-to-end energy expended per bit transported as the volume and rate of data increase exponentially.
- Limit the overall system scale energy consumption and the need to dramatically improve the energy-efficiency of the processing system as a whole to deal with the data deluge and the necessary processing of such data.
- Address scaling limitations inherent in the architectures in use and for Systems-in-a-Package (SiPs) and those imposed by total power, power distribution, interposers and interconnections pose serious challenges.
- Address the increasing need for security and the reliability of the heterogeneously-integrated digital processing systems, including the need to monitor and interpret all required information to ensure both secure and reliable operations.
- Address and improve end-to-end sustainability, encompassing pre-design, design, manufacturing, use and eventual disposal/recycling.
- Provide high-level design tools that permit function disaggregation into a multi-chiplet architecture, optimizing simultaneously across several parameters are needed; the large design space that needs to be explored poses challenges that require machine learning (ML) based solutions.

Figure 10.1 depicts the overall theme for this chapter. The roadmap for Chapter 10 is organized into three subareas:
1. **The Drivers: Processing Paradigms & HW-SW Co-Design** (Sec 10.1)
2. **System-Level Architecture:** Processing and Memory Architecture/Chiplets, Interconnection Architectures for Heterogeneously Integrated Solutions (Sec. 10.2)
3. **Devices, Interconnects, Materials and Processing** (Sec. 10.3)
The roadmap for the processing paradigms, system-level architecture, and devices/materials for digital processing are shown below.

**Figure 10.1.** Overall theme of this chapter

The roadmap for the processing paradigms, system-level architecture, and devices/materials for digital processing are shown below.

<table>
<thead>
<tr>
<th>Period →</th>
<th>Current</th>
<th>5 years</th>
<th>10 years</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processing Paradigm Roadmap: Development and Maturation Periods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compute Centric, Memory-Centric Data driven Architecture</td>
<td>Accelerator based architectures and memory centric architectures are expected to provide significant benefits in near term.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stochastic Computing Cognitive Computing</td>
<td>Non-numeric computing like probabilistic &amp; cognitive computing will continue to evolve in near and mid-term.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuromorphic computing &amp; Quantum Annealing, Quantum Computing</td>
<td>Neuromorphic and Quantum system are in early stages of development but demonstrate potential for significant power performance advantage for specific applications and will continue to evolve in mid and long term.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneous Integration Roadmap for SiP Architectures for Digital Processing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Integrated Chiplet Types &amp; Count</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-End: Multi-core chiplets, GPU, HBM, small NVM, co-packaged photonics at high end, wider use of interconnection standards, 3D SiPs with 2 or 3 layers + HBM2/HBM3, SRAM stacks. Use of liquid cooling increases with package TDPs to 500W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium-End: similar to above but with GDDR6+ DRAM instead of HBM, low-end GPUs Embedded/IoT: larger NVM, SRAM, low-power accelerator integration optionally with RF chiplets, use of processing data at acquisition point for images Chiplet count within package: 4 to 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High and Medium end: Analog AI/ML chiplets in production SiPs, Neuromorphic accelerator and PIM chiplets appear, use of NVM in memory hierarchy, UVM standards emerge. Increasing use of 3D designs analog/low-power chiplets. Security and run-time monitoring/management chiplets appear. Package TDPs near 1KW mark in extreme SiPs. Embedded/IoT: 5y trends scale up with specialized chiplets, 5G/6G chiplets for IoTs, data consumption at edge goes beyond automotive and specialized markets Chiplet Count: 10 to 30, higher counts realized through 3D stacking in few chiplets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serious challengers to DRAM memory emerge; Co-packed photonics chiplets for IO; NVM memory chiplets see wide use in memory hierarchies; analog ML accelerators permeate, Other specialized chiplets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Accelerator Chiplets and Specialized SiPs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU (full and variable precision) Early memory-centric chiplets Early analog accelerators – low lifetime built with memristors; SiPs supporting quantum annealing and quantum computing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analog NN accelerators using NVM and analog memory technologies with low-power ADCs; Tighter integration of photonics for quantum computing acceleration, mature accelerators for data-centric computing and data-driven computing of more general nature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuromorphic accelerators or SiPs, Small form factor, general quantum computing chips integrating qubits, quantum gates, measurement and control logic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Memory Technologies, Memory Hierarchy and Memory Coherence in SiPs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBMs, Stacked SRAM Caches, Cache chiplet on logic, conventional and extended hierarchy with large L4 cache, standard coherence protocols, some based on CXL, limited proprietary unified virtual memory (UVM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVRAMs across memory hierarchy with non-traditional uses, Open UVM architectures and standards, Advanced HBM generations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New memory chiplet competing with DRAMs, energy-efficient NVMs competing with SRAM in speed and comparable to DRAM densities, UVMs supporting multi-granular coherency.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SIP Interconnections</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silicon bridges with ~1000 to 2000 lines, limited face-to-face bonded 3D connections, interposer with 2 to 3 metal layers, co-packed photonics IO with 2 to 8 wavelength WDM links at high end, back-end vias for power delivery, PAM4 encoding at high-end</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silicon bridges with up to 4000 lines, via and bump pitch down to 5 to 10um in large volume production 3D SiPs, interposers with up to 5 layers, PAM 4 encoding use increases, PAM8 encoding appears at high-end, small form factor photonics IO in production, dense WDM in wider use for SiP IO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silicon bridges with up to 8000 wires, PAM8 encoding appears at high end, nano-scale face-to-face bonding at high end, Up to 7 layers in interposer, dense WDM and package-internal photonics chiplets for IO enter production for high end SiPs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Devices and Interconnections</th>
</tr>
</thead>
<tbody>
<tr>
<td><em><em>Transistor density</em> (#/cm²)</em>* [Synopsis TCAD]</td>
</tr>
<tr>
<td><em><em>Transistor power</em> density (W/cm²)</em>*</td>
</tr>
</tbody>
</table>
10.1 Applications, Processing Paradigms & HW-SW Co-Design

This section identifies the most demanding and prominent applications driving the markets and their requirements. Some of the application requirements also point to shortcoming in current processing architectures and hence we discuss evolution in different processing paradigm. Lastly HW-SW codesign highlights the need for a coordinated effort at different levels of system design for addressing power and performance requirements.

10.1.1 Applications- covered comprehensively in Chapter 8 so not included here.

10.1.2 Processing Paradigms

When the CPU based processing platforms are unable to meet compute/energy/storage efficiency requirements for specific applications the typical response is to go from standard compute to custom compute solution (either as custom processors or hardware accelerators).

In order to improve the energy efficiency and performance of the compute for these applications there is a need to explore alternative compute paradigms beyond the prevalent compute centric paradigm.

The following compute paradigms are expected to grow and influence specific types of applications.

- **Compute Centric**: Most current compute architectures are built focusing on computation units and memory hierarchy & interconnects designed to support the compute fabric. Efficiency in application processing is addressed by exploiting characteristics of compute in terms of instructions (CPU, DSP, ASIP), instruction-level parallelism (Scalar, super-scalar, VLIW) and task level parallelism (multi-cores, heterogenous architectures).

- **Data Driven Architectures**: A large number of applications involve massive amounts of data processing which is beginning to move focus from compute to memory as evident from significant amount of energy spent in memory and interconnect. New memory devices and architectures are evolving for overall efficiency in performance and energy for these applications.

- **Memory Centric**: As the relative cost (performance & power) of compute v/s memory for data intensive applications continues to lean towards memory (called memory wall), there is a growing trend towards architectures that place compute resources closer to the data as compute-in-memory and compute near memory.

- **Stochastic Computing**: The prevailing compute paradigms process and store information as numbers with precision (8, 32, 64bits). Processing of numeric data has it’s associated arithmetic computation cost in performance and energy. A large number of applications that do not impose stringent numeric accuracy requirements, stochastic computing that stores information as probabilistic data (and associated simplified compute) can provide significant energy efficiency improvements.
Cognitive Computing: Similar to Stochastic computing information can also be stored and processed with large random vectors in place of numeric data. This is exploited by hyper-dimensional computing which offers a variety of advantages for AI systems in terms of network architectures, energy efficiency for learning as well as inference and noise/error tolerance. Symbolic data representation with large random vectors also enables multiple layers of cognitive computing.

Neuromorphic Computing: This is Brain-inspired compute paradigm including integrated memory & compute architecture with neurons & synapses. Energy efficiency is also improved through event-driven/asynchronous compute & communication as in spiking-neural networks.

Quantum Computing: While traditional computing stores information as bits and computes with arithmetic, Quantum computing stores information in multi-dimensional states (qubits) and quantum processing exploits superposition and interference of the qubits to provides exponential growth in compute capabilities for large complex problems.

10.1.3 HW-SW Codesign

Over the past two decades, the end of Dennard scaling (in the last 20 years) and slowing of Moore’s law (in the last decade) coupled with a focus on increased performance have led to passing the PetaFLOP milestone with the Roadrunner system (2.3MW) at Los Alamos National Laboratory, and 14 years later with the Frontier system (21.1MW) at Oak Ridge National Laboratory. The Exascale Computing Project established architecture R&D projects with processor and system companies to support the codesign collaborations that led to new capabilities for hybrid CPU-GPU HPC, and extensions to support scientific ML. While energy efficient performance was an important design goal, the path from PetaFLOPS to ExaFLOPS still required an increase in the power requirement by 9x. Future progress in HPC and computing performance at all scales will require dramatic increases in energy efficiency and a focus on fundamental application bottlenecks, not simply floating-point performance, the memory wall, and data movement in particular.

The confluence of this demand alongside the CHIPS Act goals of the democratization of ASIC design and the accessibility of fabrication has led to a growing number of tailored computing technologies within the marketplace. These demands and market trends will continue and will necessitate a broader and deeper ability to consider the system software stack and hardware concurrently. This is deep codesign, in which algorithms and architectures are designed together to form an optimal solution. Over the next decade many computing areas will require an ability to custom tailor processors, memory technologies, accelerators, and discrete functional units into heterogeneous computing designs. The advent of advanced packaging and 3D integration with chiplet technology standards will support their silicon in package (SiP) arrangements in an agile way. Tools to integrate these tailored technologies will be critical components to this capability and will span from circuit design, floor planning, all the way up to full-system simulation. These hardware design tools will need to be applied in tandem with profiling tools to better understand applications and their algorithms, enabling rapid exploration and assessment of hardware technology designs and changes in applications to achieve an optimized solution across complex tradeoffs (performance, power efficiency, ease of porting, design and fabrication cost, total cost of ownership). AI will undoubtedly play a role in this space as witnessed by multiple academic and industry initiatives which have already incorporated AI into their design and optimization cycles.
Alongside these tools for reasoning about the integrated design space there is the need to efficiently shape future technologies from software to integrated circuits. Heterogeneous computing system software will necessitate continued advancements in software frameworks, compiler technologies, runtime and operating systems. From the perspective of hardware technologies, this will necessitate higher productivity tools for ASIC design, SoC design, memory subsystem design, and the ability for compositions of multiple heterogeneous chiplets to interoperate. Individual components manufactured with different process technologies will need to be seamlessly integrated at low cost, low power, and with high yield, necessitating significant advances in heterogeneous packaging technologies.

Coarse-grained reconfigurable architectures (CGRAs) and the ability to dynamically compose discrete components in disaggregated architectures will further expand the possible codesign space. This will necessitate tooling that is capable of optimization at application runtime and which incorporates optimization criteria across concurrently running workloads. While some cloud computing environments are currently faced with aspects of this complexity, the next decade will see reconfigurability along a much larger set of dimensions and application workloads.

Research and development of technologies that enable this deeper level of codesign holds the key to continued exponential gains in performance and power efficiency for our nation’s strategic computing workloads.

10.2 System-level Architectures

As noted in the SRC Decadal report, data volume and data movement costs dominate. At the chip scale, the fraction of the total chiplet energy spent in the interconnections has grown tremendously at advanced nodes, as seen in Figure 10.2. Systems-in-a-package have to address the cost of data movement at all scales, both in terms of absolute power and energy expended per bit in the data transport process.

The system-scale architectural techniques addressed in this chapter that appear promising in mitigating the data movement costs are summarized below:

- Processing near memory.
- Processing in memory, which encompasses development of suitable memory devices with embedded logic.

![Figure 10.2. Chiplet-scale power trend](image)
- Processing at the interconnection interface, within the interconnection (for example, within on-chip routers as data is moved)
- 3D chiplet stacking architectures which reduce the reach of inter-chiplet interconnections dramatically.
- Architectural solutions that avoid system-scale data movements, such as in broadcasting.
- Since clock signals are distributed globally, the associated interconnection power (using H-trees or similar interconnection topologies) needs to be addressed.

At the level of the interconnections themselves, several solutions offer the potential of reducing interconnection energy. These include:
- Use of redistribution layers (RDLs) offered in many chiplet integration technologies, such as CoWoS.
- Use of clock forwarded links for short reaches, that avoid the need for area- and energy-hogging PLLs and DLLs for clock/data recovery.
- Use of advanced symbol encoding going beyond PAM 4, the current state-of-the-art.
- Use of energy-managed links, supporting the analog of DVFS and other techniques.
- Use of photonics for package-level IO. In the longer term, this could be a solution for interposer-scale connections as well but significant developments are needed for non-IO deployment.

Certainly, the evolution of interconnection standards to ease chiplet integration to achieve the full potential of HI is a necessity. Standards are evolving (BoW and other ODSA efforts, UCIe etc.). Others are likely to evolve or be built on the dominant standards, including standards for power distribution, reliability monitoring, security monitoring and test.

The overall trend in system power was shown in the SRC Decadal Plan to exceed the total energy generated globally, if the current trend of chip/system designs continued with projected deployment rates. This grand challenge has to be addressed using a variety of solutions in the coming years and include:
- Innovations in memory technologies including non-volatile memory devices, analog memory devices, and others.
- Making general-purpose multicore architectures more energy efficient.
- Replacing or augmenting general-purpose processing solutions with the more energy-efficient specialized function accelerators.
- Preferred processing engines for AI/ML shifting from conventional digital ML accelerators like GPU and GPUs to alternatives that are significantly more energy-efficient, in particular analog AI/ML accelerators.
- Neuromorphic computing will provide extreme performance per unit energy, exceeding the capabilities of conventional transistor-based solutions by several orders of magnitude.
- Through system architecture innovations that reduce the cost of data movement.

10.2.1. Memory Technologies

Goals/Needs: Stand-Alone Memory Materials and Devices

Methods are needed to extend current incumbent, standalone memory, DRAM and NAND Flash technologies, as well as enabling “emerging” memory, and invention/discovery of new/novel memory possibilities
**Roadblocks/Challenges:** Further DRAM scaling is hampered by the large relative area needed for the capacitor, so for continued density improvement 3D device stacking through layering (homogenous integration), similar to what has been done for 3D NAND is attractive. A major hurdle for 3D homogeneous DRAM is that it requires BEOL-compatible select devices with appropriate drive characteristics and very low leakage which have yet to be achieved. Continued layer stacking in 3D NAND requires etch technology advances and thinner materials (toward 2D), especially for the channel. For both DRAM and NAND, new packaging architectural schemes and supporting technology for chip stacking (3D HI) beyond current HBM methods are needed.

Emerging memory is still “emerging,” with nothing current having a combined set of properties necessary to supplant the mainstream, incumbent technologies of DRAM and NAND Flash in large-scale, stand-alone implementations. Emerging memories must overcome their known, respective shortcomings to rival these incumbents in cost, scaling, reliability, variability, repeatability, cyclability, and other metrics. Much has been done to address the deficiencies of emerging memories (e.g., the various flavors of ReRAM, PCRAM MRAM, & FERAM) but a collective win has not been forthcoming. Without such overall device competitiveness, these technologies will be relegated to smaller-scale, niche implementations like MRAM as a non-volatile SRAM replacement and FERAM as a relatively fast and low voltage, non-volatile memory.

For the whole memory space, evolutionary new application platforms, yet to be established or even envisioned, like new AI architectures or novel consumer devices, could enable emerging memories (or even DRAM and/or NAND in a new space), so new technology opportunities and overlaps should continue receiving attention. Unforeseen opportunities in the stand-alone memory space may unfold through imagined new architectures that go toward clearing the power-sapping, compute to memory interconnect constriction, the so-called “memory wall,” by enabling widespread near- and in-memory compute.

**Possible Solutions:** Material and device solutions for any memory and/or select device must be approached with an overall knowledge of the many target performance metrics and integration challenges that must be met for success, those that enable the device as a possible replacement or supplement to current memory. Development must go in lockstep with modeling (first principles, transport, multi-physics, etc.) to guide development and implementation. New architectures and technology space that helps eliminate current roadmap issues may open new opportunities for established and/or emerging memory technologies.

**Incumbent Mainstream Memory**

- **3D NAND**—scaling requires thinner channel material (2D) with reasonably high mobility (>20 cm^2/Vs) and concurrent 3D deposition technique (ALD) compatible with BEOL processing window. Very high k through use of ferroelectrics.
- **3D DRAM**—stacking requires BEOL compatible select device with high drive current (>10 MA/cm^2) at ~+/−2V, and extremely low leakage (<<10^-15 A). Toward 2D, as well.
- Enabling 3D heterogeneous packaging architectures and methodologies with supporting technologies for DRAM and NAND, and any emerging memory applications of interest.
**Emerging Memory**

- ReRAM—toward deterministic control of resistance states for low device variability and cycling repeatability, with minimal drift and high stability.
- PCRAM—drift and atomic segregation resistant material with lower drive current requirement
- MRAM—MTJ device or analog with 10-100x lower critical current or voltage-switching, and 10-100x higher resistance ratio.
- FERAM (incl FeFET & FTJ)—high and uniform remnant polarization at scale in X-Y and Z with high retention and resistance to imprint and fatigue.

**10.2.2 Conventional Processing**

General purpose computing IC’s will continue to remain a workhorse. These systems have been optimized and used in applications ranging from HPC and all the way down to embedded/IoT applications.

**Roadblocks/Challenges:** SiPs implementing or incorporating significant conventional processing elements face many challenges:

- The realization of large core counts is impeded by the high cost and low yield of large multicore chiplets.
- Unnecessary power dissipation occurs when cores have low utilization.
- Performance of the memory hierarchy and maintaining memory coherence for SiPs with large core counts remain a performance bottleneck that prevents system scale-up.
- For long-running HPC applications, checkpointing the computation state is a scaling bottleneck.
- Interconnection bottlenecks in multi-chiplet implementations of high core count systems can easily diminish any benefits of high core counts.
- When integrated with accelerator chiplets, memory coherence and memory sharing generally suffer from a high performance overhead as the system scales up.

**Possible Solutions:** The solutions, many in current use, for mitigating the challenges listed above are:

- Integration of multiple, smaller multicore chiplets in a 2.5D configuration, taking advantage of the dense interconnections available on a chiplet to provide performance close to that of a large, single chiplet implementation.
- Non-homogeneous multicore chiplets or chiplets with non-homogeneous cores (called the “big-little” configuration) offer a way to mitigate the power wasted at low utilization levels; the system governor and the OS scheduler can make effective use of the non-homogeneous cores.
- Stacked SRAM chiplets can enhance the capacity of the lower level cache or enable the use of a large last level cache outside the core chiplet or outside the core chiplet area on a larger die. Similarly, emerging non-volatile memory technologies that approach the performance of DRAMs or large SRAMs, have similar write endurance and comparable of better energy efficiency with similar capacities as the current incumbents can be used creatively in the memory hierarchy to support checkpointing, early transaction commitment and other benefits.
- In scaling up general purpose computing systems, multi-chiplet implementation of the processor, along with memory, specialized accelerators and other chiplets is clearly a solution, but the techniques of shared memory coherence at larger scales need to be developed. Techniques for Cache
Coherence scaling to hundreds of large caches with increasing distance between them have to be used, leading to the concept of dynamically partitionable NUMA domains.

- As different types of processing chiplets are integrated, architectures for Unified Virtual Memory (UVM) have to be matured to support efficient coherent memory and avoid inefficiencies and copying overhead. Memory coherency techniques that simultaneously support coherency mechanisms at different, chiplet-specific granularities will be very useful in this respect.
- In-network acceleration for multicast, collectives, barriers and other synchronization features to mitigate the increasing core-to-core communication latency.
- Adaptive routing to reduce congestion and unlock the available bisection bandwidth of the interconnections.
- Interconnection bottlenecks can be alleviated in many ways in SiPs – these includes the use of wide bridges in 2.5 D configurations, stacking of low-power cores and/or IO processors or SRAM caches on high power chiplet cores. Finally, having multiple metal layers within the interposer will permit significant connectivity improvements through parallelism in the network.

10.2.3. Processing in Memory

Goals/Needs: Applications such as Deep Neural Networks (DNN) and Homomorphic Encryption (HE) require frequent data movement among different levels of the memory hierarchy. The limited bandwidth and high data movement energy cost degenerate the system performance and increase the energy consumption of the processing. However, one of the operands of such operations is stationary, e.g., the neural network weights in the DNN. Processing-In-Memory (PIM) is a promising solution for accelerating such data-intensive workloads as it enables in-situ computation directly in the memory component. With processing capability at the cache, main memory, and/or storage devices, we aim to improve the system level performance and energy efficiency for applications from data center scale to edge devices.

Processing-in-memory approaches can also be extended to support applications such as transaction processing, database and search applications, accelerate bit-level operations and accelerate specific bioinformatics applications.

Roadblocks/Challenges: PIM chiplet realizations for SiPs face several challenges:
- The architecture design and computational precision of PIM accelerators often relies on overidealized device/circuit parameters, which is difficult to achieve in the actual manufacturing process.
- Due to the limited internal bus bandwidth, the workload that requires flexible data access patterns (e.g., non-local access or collective operations) still suffers from the data movement bottleneck.
- Due to the limited on-chip memory capacity with fast-growing size of the models, the performance of PIM-based accelerator is degraded by the data movement incurred by the data replacement of the PIM.
- There is a lack of support in the software stack. Most compilers/libraries are not aware of the special dataflow in the PIM accelerator and between the PIM accelerator and the host. Hence, they cannot fully utilize the advantage of PIM design.
- The existing PIM design lacks flexibility for the implementation of an entire application. Thus, it requires memory sharing or fine-grained synchronization with the processor, which could lead to performance degradation compared with that demonstrated in the isolated evaluation.
There is no complete solution to integrate PIM into the existing system. No solution has been verified to be compatible with cache coherency, OS memory management, existing memory models in programming languages, etc.

**Possible Solutions:** Heterogeneous integration allows for specialized PIM chiplets to be integrated with other chiplets and high-speed interconnections and IO chiplets in general for useful deployments. PIM integration at different levels of the memory hierarchy is also enabled by this but appropriate APIs need to be provided to permit workload scheduling, data staging and other needs.

Algorithm-hardware co-design is needed to architect and exploit the full potential PIM-based systems. Trace-driven, high-level simulation tools are needed to identify the hardware primitives to establish a basis for designing the associated software stack, libraries and run-time system. Without a software base, large scale deployment is necessary.

**10.2.4 Analog AI Accelerators**

**Goals/Needs:** The dominance of AI/machine learning applications, which is expected to continue in the coming 5 to 10 years, will see deployment at multiple scales from edge devices and mobile platforms (such as autonomous vehicles) to large data centers. A significant set of these applications have relied on neural network variants (CNNs, DNNs etc.) and current product offerings testify to the wide variety of existing and future needs. In any of these systems, timely response and high throughput are critical. The data sets used for training the AI/ML accelerators have also grown significantly. It is therefore imperative to improve the energy efficiency of AI/ML acceleration subsystems dramatically to scale up with the data sets and application scale.

**Roadblocks/Challenges:** Conventional NN-based accelerators are implemented with digital logic, relying on an array of multiply-and-accumulate (MAC) logic. Improved device technologies, use of variable precision support etc. have been traditionally used to scale up the capacities and also improve the energy efficiency of these accelerators. Recent years have also seen the emergence of analog multiplier-based MACs that have relied on mersisters and similar technologies, which are particularly suited for low-power applications where some lack of precision can be tolerated. Several challenges remain in scaling up the accelerators: the data transport overhead have to be addressed, the power dissipation of the combined systems need to be reduced significantly to avoid thermal and associated reliability issues to permit deployment at scale across all application classes.

**Possible Solutions:** Potential solutions for realizing analog AI/ML acceleration chiplets are:

- Dense analog AI accelerator chiplet with long lifetimes and extended precision that can support dynamic or configurable precision adjustment. This can improve on existing technologies, which tend to have lower lifetimes. These, in turn, translate to needs for innovations in materials, low-noise and stable analog voltage regulators, low-power, low footprint ADCs etc. Examples of candidate analog AI accelerators include the Phase Change Memory (PCM)-based crossbar array analog neural network accelerators, analog NN accelerators based on other in-memory architectures.

- Alternative MAC designs implemented with new device technologies that are inherently energy efficient (such as CNTs – carbon nanotube transistors), but reliability/lifetime improvement is a necessity.
● SONOS Flash Analog memory or MTJ memory devices that have SRAM like performance and high endurance for weights etc.
● These can be combined in a 3D chiplet stack, which, in turn, demand high nanobump/microbump densities, new power delivery/conversion strategies etc.

10.2.5. Other Aspects of SiP Scaleup

Scaling system configurations up to support further performance improvements through higher component integration introduce multiple architectural and microarchitectural challenges, beyond the physical and thermal challenges discussed for far. Many operations which take place across cores involve broadcast mechanisms which incur increasing completion latency as core-to-core communication distance increases. Also, despite 2.5D and 3D integration allowing a significant increase in the bi-section BW of future on-chip interconnects, this may necessarily lead to increased effective BW due to interconnect congestion, which can be especially severe when data flows from/to PE across the SiP towards IO and memory controllers in the SiP periphery. Furthermore, interconnect area currently accounts for 20%-30% the total SiP area and is responsible for 5%-10% of the TDP, thus further increasing interconnect density to address such bandwidth issues would require expensive tradeoffs.

A fundamental limitation in system scaleup is the power wall. Since interconnection power in a chiplet forms the major component of its power dissipation, heterogeneous integration offers a way to reduce the overall power by integrating chiplets inside a package with wide and short interconnections to reduce the IO power that would have been otherwise needed in traversing across package boundaries. This is true with 2.5D integration with bridge-connected chiplets and particularly with 3D stacked chiplet architectures. Unfortunately, 3D chiplet integration has to deal with the issue of heat dissipation, power delivery and yield issues. These issues are addressed elsewhere in the roadmap. In the short term and at present, 3D DRAM (such as HBM), one-layer of high-power logic in a stack with low power chiplets offer a practical solution to memory scaling and system scaling. To scale the system size requires the overall energy-efficiency of the system to scale up and the ways to do this requires one or more of: (a) innovations in system architecture; (b) energy-efficiency improvements through device and interconnection innovations, and, dramatic innovations in cooling technologies.

Other Possible Solutions to Support SiP Scaleup: These solutions, complementing and extending the ones discussed in Section 10.2.2, are:
● QoS-aware routing and congestion management of bandwidth-sensitive (e.g., GPUs) vs latency-sensitive (e.g., CPUs) PEs.
● Fine-grained (e.g., router-level) DVFS support for power management.
● Data compression techniques to increase bandwidth and allow narrower interconnect links.
● Integrated photonics for chiplet-to-chiplet links.
● Hardware support for sparsity is needed to improve effective utilization of large 3D caches and vast compute resources in package.
● Co-designed interconnect topologies tailored to the data movement patterns

The architectural studies of such large systems will also require novel simulation techniques that allow simulating multiple levels of abstraction while being able to reason about microarchitectural details within a chiplet, and all the way to overall performance, power and thermal analysis of the whole system across chiplets, different types of processing elements and 3D layers.
In addition to chiplet integration on an interposer in two or three dimensions, larger wafer-scale systems for customized applications can also enable system scale-up, but widespread deployments of such products are unlikely.

10.2.6. Neuromorphic Computing (This section still under development)

10.2.7. Quantum Annealing and General-Purpose Quantum Computing (This section still under development)

10.2.8. Power Distribution and Conversion

Power distribution for heterogeneously-integrated systems introduce challenges that are unique to such systems. The challenges include:

- Routing considerations for signal and power connections within the interposer that can affect connection length and introduce unintended noise and/or crosstalk.
- Line drops due to high current in systems with higher power envelopes.
- Need to address covert challenges where information can be leaked through chiplet-shared shared power converters and/or power connections.
- The need for high-quality power for low-voltage analog chiplets that are potentially integrated with digital chiplets.
- 3D chiplet stacks power and signal routing connection constraints that introduce line sag, noise, crosstalk etc.

Potential solutions that exist (and will continue to be used) include:

- Backside power delivery (aka power vias), that reduce conflicts between signal and power routing and eliminate covert channels centered on power distribution. New material development may be needed to realize appropriately-scaled decoupling components and magnetics, in general, that can enhance this particular solution for use with high-current chiplets. New material development will also enable the use of in-line inductors for the backside power connections.
- Interposer-embedded power converters.
- Use of point-of-load (POL) converter chiplets inside the package, which takes a step beyond the use of POL converters adjacent to the package, resulting in a distributed power conversion inside the package. This solution requires the development of process techniques that integrate GaN devices into the CMOS-dominated process flow, new material development for magnetics and decoupling power capacitors.
- Use of higher voltage supply to high-power SiPs and their POL conversion inside the package to reduce the I^2R losses with high current feeds before they are converted to chiplet-level voltages. This also requires the development of magnetics and decoupling components that can be physically scaled for in-package use with low-losses, the development of new converter designs and topologies to promote wide efficiency across a wide load current range, capable of supporting fast slew rate and capable of using lower inductance and capacitors inside the package for the in-package filtering circuitry.
- For 3D systems, solutions need to be developed for power routing within the 3D stacks; these solutions also need to avoid the setup of unintended covert channels. Prospective solutions include
extensions of what is used in the industry today for HBM s, for power delivery at the edge, the use of local power regulation within chiplets etc. Appropriate soft-IP development and standards are needed in this respect to specify physical parameters of the power connections.

10.2.9. SiP Security and Privacy Issues

Goals/Needs: Security/Privacy Concerns at ALL levels of the system

Increasingly, sensitive computations—finance, health, even computations and data related to national security—are being deployed on shared pay-per-use infrastructures owned and operated by Cloud services that are able to leverage economies of scale and comparative advantage to drive down costs. While traditional isolation and virtualization technologies (like hypervisors and OSes) have served us well, the prevalence of the Cloud has exposed software systems to new security vulnerabilities and heightened the risk of old ones. There are security issues specific to heterogeneously-integrated systems. Overall, the security challenges introduced are as follows:

- Chiplet IP compromise through reverse engineering of the integrated SiP or during the integration process itself.
- Security issues stemming from the use of chiplets from different vendors that may be tampered deliberately and or be flawed in ways that unitedly introduce security threats when deployed with other chiplets inside the package.
- Covert channels that are established either via the interposer or via power converters/regulators shared by chiplets.
- Run-time compromises that can lead to DoS attacks and/or information leakage via side channels and covert channels exacerbated by the close proximity of chiplets in both 2.5D and 3D configurations.
- Run-time compromises that are enabled by malware or by design flaws within the firmware of chiplets.
- System-scale challenges of a general nature need to be addressed as well and include minimizing overhead of isolation techniques (virtualization, containerization, etc.), providing proof of confidentiality of remote resources (edge, cloud, HPC data center), incorporating security/privacy concerns efficiently across distributed systems/chiplets.

Potential solutions to address the general threats are emerging and will not be emphasized here. But solutions for threats specific to heterogeneously integrated systems include:

- Prevention of reverse engineering using techniques such as logic locking, design obfuscation, partitioning of sensitive functions across independently sourced chiplets, packaging solutions for obfuscation etc. Research and development are still needed here to generalize the techniques and mitigate the performance and energy impact of the existing solutions.
- The development of general-purpose tools for detecting chiplet tampering, both in the EDA tool suite and other solutions that have been proposed such as optical/X-ray scanning and others. New material developments are needed to expose tampering during or after integration.
- The use of appropriate EDA tools for pre-integration multiphysics analysis that identifies covert channels. CAD tool vendors already have some products, but a more general suite of EDA tools is needed in this regard.
- The use of solutions at the chiplet boundaries for avoiding leakage. This requires the development of efficient solutions for use at run-time that do not affect run-time performance and energy needs.
adversely. Simple isolation techniques at the inter-chiplet interfaces are a start, but these can be tampered with.

- For higher-end SiPs in general, the use of a security monitor engine, serving as a root-of-trust will detect anomalous behavior, authenticating dynamically loaded firmware, for deploying adaptive solutions against run-time attacks. Machine learning techniques for anomaly detection will play a key role here. Any anomaly detection technique will have to rely on appropriate sensors (such as those for physical parameter sensing for temperature, power etc.) as well as existing instrumentation counters within chiplets as well as added sensors implanted at chiplet boundaries (for error, traffic monitoring etc.). These sensors have to also be authenticated or protected themselves. The security monitor engine can also be extended to handle system management function related to key management/distribution, power management, reliability monitoring etc. In general, the security monitor engine should also be resilient to inherent software vulnerabilities with the firmware it runs. Further R&D activities are warranted in the coming years with respect to the entire monitoring/system management infrastructure to make it customizable, lightweight and be verifiable to serve as a root-of-trust.

10.3. Materials, Processes and Devices

10.3.1 Logic and Memory devices

10.3.1.1 Logic devices

Goals/Needs: The number of connected devices and the amount of digital data transferred during computation continue to grow exponentially. Simultaneously, the total fraction of available power consumed by electronics is increasing at an exponential rate. To maintain pace, energy consumed per computation will demand that device area and volume must continue to decrease, and that new materials, device designs, processes and equipment be realized for novel energy-efficient CMOS logic that is more than 1,000,000 x more energy efficient than current systems [??]. CMOS scaling is aggressively moving to 3D stacking to reduce power consumption and enable added functionality, so new strategies must be pursued to deliver power and dissipate unwanted heat. Chemical processing and new integration schemes are needed to provide new pathways to efficient and effective 3D scaling.
Fig. 10.3: Projection of progress in CMOS transistor density and corresponding switching energy [??].

Roadblocks/Challenges: Computational power consumption in traditional CMOS logic is determined by the supply voltage, the extent of gate current leakage, device capacitance including gate, interconnect and parasitic capacitance, and clock frequency. Increasing frequency puts immense demands on reducing voltage, capacitance, and leakage, which are determined by: i) device materials (i.e. gate dielectric permittivity and thickness, N-P abruptness and isolation, interconnect metal conductivity); ii) physical device size and design (i.e. gate-to-contact capacitance, cell height, interconnect capacitance, source/drain overlap, etc.); and iii) device integration and packaging (i.e. 2D planar vs 3D stacking). New techniques will be needed to realize high aspect ratio and asymmetric architectures, particularly when feature alignment requirements are very small relative to element thickness. Backside power delivery needs attention to be compatible with 3D integration while minimizing deleterious capacitances.

Possible Solutions:
Device structure and design: The evolution of device structure from FinFET to Ribbon FET and Stacked Ribbon FET will be helpful, but may not be sufficient to achieve target efficiency needs. Vertical FET, Tunnel FET, and Negative Capacitance FET structures must also be aggressively pursued. Cryo-FETs operating at very low temperatures can also be considered, but feasibility is limited for most applications. These structures can promote low power by reducing short channel effects and minimizing deleterious contact-to-gate capacitance and via and contact resistance. Negative Capacitance FETs reduce power by improving gate capacitance through effective higher dielectric constant achieved by stabilizing negative capacitance in ferroelectric materials, and Tunnel FETs can overcome the 60mV/decade subthreshold swing limit in conventional CMOS transistors. The overall reliability of many of these device schemes has yet to be verified, and for each approach, consistent long-term performance is expected to depend on materials choice and fabrication methodology.

Feature size scaling: Very aggressive area scaling can be achieved in 3D architectures by making use of advances in Design Technology Co-Optimization (DTCO) to evaluate and optimize power, performance and area (PPA) at the block level. Layout dependent effects are also important to consider, and may be addressed, for example by the transition from double diffusion break to single diffusion break, and
developing contact over active gate FinFET designs. Innovation is needed for low thermal budget 3D conformal metal and gap fill processes, budget stacking methods that are low thermal-budget and low cost, and new thinner channel structures will help reduce device volume and thermal mass. Advanced 2D semiconductors and insulators that are conformally deposited and etched in desired geometries may provide better electrostatic control over the channel and reduce power consumption with smaller channel resistance and low parasitic capacitance. Chiplet design and integration will also benefit from emerging System Technology Co-Optimization (STCO) methodologies.

Channel materials: Channel materials will continue to evolve from strained Si to SiGe, and further to Ge, 2D Transition Metal Dichalcogenides (TMDs), and nanotube materials. Low-dimensional materials promote electrostatic control to achieve high carrier density, high mobility, reduced parasitic capacitance, and reduced tunneling due to high carrier effective mass. Channels must be robust with few defects, highly reliable, enable low resistance ohmic contact, and be compatible with low temperature gate dielectric and gate metal with tunable work function. Oxide semiconductors with reasonable mobility (~100 cm²/Vsec) and low leakage can be deposited at low temperatures, but integration requires improved thermal stability and means to avoid unwanted H₂ doping effects.

Interconnect materials: Interconnect materials also need to be improved. Currently used copper requires a thin TaN diffusion barrier, and in small vias, the barrier layer (with a relatively large resistance) can occupy a large fraction of the via volume. Ruthenium, cobalt and molybdenum are potential candidates, but attention to other materials could have substantial impact.

Fabrication and processing: Realizing these devices in manufacturing will require improved accuracy of feature size and alignment (i.e., smaller Edge Placement Error, and uniform gate length control) as well as sub-nm uniformity in vertical structures where aspect ratio exceeds 50-100:1. For vertical device schemes, the power consumed per device is reduced but power per unit volume becomes very large, demanding new schemes for heat dissipation. Vertical device fabrication also requires advances in materials synthesis at low temperatures. High-quality low-defect materials are generally achieved at high temperatures where thermodynamic driving forces promote crystallization and defect reduction, but high thermal budgets can damage underlying materials and material junctions. Better understanding of chemical surface processes is needed, to enable low-temperature kinetic control of material assembly at the atomic scale.

10.3.1.2 Memory Devices

Goals/Needs: New memory technologies are in high demand, particularly for applications such as artificial intelligence. The desire for electronic memory is growing at such a fast pace that silicon needed for memory will soon exceed globally available. The energy efficiency of memory technology has not kept pace with the advance of computing power, further pushing the need for radical new technologies. Significant improvements in power of memory that can be manufactured at scale with > 100 x the storage density of current systems is required to bring data closer to the computing units, to enable large bandwidth while eliminating significant power consumption when data is transferred between CPU and DRAM.

Roadblocks/Challenges: To be viable for cache-level memory that mitigates the power issue due to data transfer between CPU and DRAM, strict device-level requirements include read/write time close to 10 ns for last-level cache or embedded DRAM (eDRAM), and even faster read/write times (about 2 to 3 ns) for L2/L3-level cache. Endurance needs to approach 10¹⁵ to 10¹⁸ read/write cycles, and the operation voltage needs to be low enough so that the memory devices can be embedded with advanced logic transistors to
compete with state-of-art SRAM in terms of density. Hence, significant improvements in the leading memory options are required to enable next-generation energy-efficiency high-speed dense embedded memory.

As current memory design geometries get squeezed in size, feature scaling becomes non-viable, such as what has happened with NAND Flash where vertical strings of cells have replaced older lateral designs—the first true 3D homogenous integration. DRAM scaling will also need a similar transition to 3D, but it is more difficult because the memory cells are larger than NAND Flash. Vertical stacking will need new BEOL-compatible select devices with appropriate drive characteristics and very low leakage. For all new integration methods, there is need for new materials and new processes, especially deposition and etch control through ALD and ALE, respectively. Relief for this continued scaling approach may be possible through new architectural schemes that bring memory closer to compute, so-called in- and near-memory processing. Additionally, new architectural schemes will drive new hybrid and heterogeneous 3D integration technologies that require innovations in chip stacking beyond current High Bandwidth Memory (HBM) implementations.

Possible Solutions:

**Emerging Memory:** “Emerging” memory, including Resistive RAM (ReRAM), Phase Change RAM (PCRAM), Magnetic RAM (MRAM), Ferroelectric RAM (FERAM), Compressed RAM (Z-RAM), Thyristor RAM (T-RAM) is still “emerging”, with nothing recognized as having the combined set of properties that can supplant incumbent SRAM, DRAM and NAND Flash technologies, firmly rooted in storing electrons. Emerging memories must overcome their known shortcomings in cost, scaling, device variability, repeatability, cyclability, reliability, and other metrics. Early applications may be niche uses. For instance, MRAM as a non-volatile SRAM replacement, and FERAM as a relatively fast, non-volatile memory that works well for low power, low cycling applications such as smart cards. Unlike devices that utilize electron storage, atomistic motion in ReRAM, and to a lesser extent in PCRAM, is inherently stochastic (i.e., uncontrollable and potentially non-deterministic) leading to substantial variation that limits implementation. New architectures that make memory more central to processing could point to memory solutions with properties inherent to one or more of the emerging memories.

**Ferroelectric Memory:** Several forms of ferroelectric memory can be considered including: i) Ferroelectric random-access memory; ii) Ferroelectric transistors; and iii) Ferroelectric tunnel junctions or diodes. Ferroelectrics for memory systems can be fluorite-based materials, e.g., doped hafnia, perovskites, e.g. BaTiO$_3$, and wurtzite, e.g. doped AlN, ZnTe, BeS. There is substantial need for deeper understanding of materials, including how defects impact switching dynamics, wake-up, fatigue, and dielectric breakdown, particularly to allow scaling to sub-1V operation to be compatible with advanced logic technology nodes. Improved analytical techniques are also needed to characterize defect type and density, and identify phase composition. Devices require advances in interfacial metal engineering to enable abrupt polarization switching with applied field for disturb immunity, and endurance to $10^{15}$ to $10^{18}$ read/write cycles for cache-level memory replacement. Processes are needed to achieve vertically stacked RAM device configurations for high-density integration. For Ferroelectric transistors, a goal is to reduce trap-charging/discharging effects through modulation of the ferroelectric polarization and carrier density, as well as gate stack and source/drain engineering to reduce defects and enable low voltage operation. Tunnel junctions in particular will benefit from the ability to form an ultra-thin (< 3 nm) ferroelectric layer (sub-3nm) with low defects to enable high $I_{on}$ with high on/off and high tunneling electroresistance (TER).

3D crossbar array architectures using ferroelectric memories for neuromorphic computing are of importance to be explored due to multi-level states enabled by analog conductance of devices, originated from partial polarization switching depending on voltage strength and pulse duration in ferroelectric
materials. For FeRAM, array scalability with proper sensing circuit schemes needs to be studied for non-destructive read operations. Comprehensive understanding in variations of multi-current levels and endurance in scaled FeFETs is required for high-density configuration, and FTJs with diode-like current-voltage characteristics potentially provide selector-free two-terminal multi-level memory cells in a stacked crossbar array.

**Spintronic memory:** Options for spintronic memory include Spin-Transfer Torque MRAM (STT MRAM) and Spin-orbit torque MRAM (SOT MRAM). STT MRAM requires ferromagnetic material engineering that enables lower switching current for magnetization reversals, and device designs that enable lower switching current due to larger effective spin-transfer torque at the same time without impacting tunneling magnetoresistance (TMR). SOT MRAM needs new materials with larger spin-orbit effect to enable lower switching current for magnetization reversals, and device designs that combine with other effects such as voltage-controlled magnetic anisotropy or magnetoelectric effects for high-density configuration.

### 10.3.2 Device Architecture *(This section still under development)*

### 10.3.3 Interconnects

**Goals/Needs:** Interconnect scaling enables circuits to improve the metric of power, performance and area (PPA) by optimizing the RC time constant, reducing line resistance, and minimizing capacitance, while simultaneously maximizing reliability and minimizing cost. In addition to new low-κ materials, this will require new interconnect conductors and liner materials. Moreover, the size of vias and lines must be reduced by introducing new lithography along with improved deposition and etching techniques.

**Roadblocks/Challenges:** Copper requires a thin TaN diffusion barrier, and in small vias, the barrier layer (with a relatively large resistance) can occupy a large fraction of the via volume. Due to line wall scattering and other phenomena, the resistivity of copper increases as line size decreases. As trenches become narrower and deeper, metal voids become more dominant, and the electromigration (EM) reliability degrades. A key challenge for advanced low-κ is that CMP of soft low-κ oxides can introduce leakage paths that degrade reliability. Additional challenges include enabling high aspect ratio contacts and vias, achieving a low thermal budget for the back-of-die interconnect stack, deep trench etching with acceptable line edge roughness (LER) and side-wall roughness, advanced metrology to monitoring quality and variability in real-time, improved high numerical aperture (NA) lithography, and approaches to integrate complex elements for back-end interconnect beyond power delivery.

**Possible Solutions:** New approaches are needed including hybrid metallization and semi-damascene subtractive etch methodology to reduce resistance. To replace copper, candidates include ruthenium, cobalt and molybdenum but other compounds could also have substantial impact. Liner materials beyond TaN, MnN, and Co need to be explored, and precursor and reaction chemistry needs to be developed for these materials to enable liner deposition by thermal Chemical Vapor Deposition (CVD) Plasma Enhanced CVD (PECVD) and Atomic Layer Deposition (ALD). Organic materials, including self-assembled monolayers (SAMs), small molecule layers, and polymeric thin films, including aliphatic and conjugated polymers, formed for example by Molecular Layer Deposition (MLD) may also provide options. New techniques to develop and realize air-gap insulators with high mechanical stability within tighter pitch sizes are also needed.
10.3.4 Two-Dimensional (2D) Materials

**Goals:** To solve device, processing, and system integration challenges, two-dimensional materials offer new options to replace or complement existing conductors, semiconductors, and insulators currently in use. 2D materials have started to be integrated in some commercial products that do not require a high integration density, such as sensors [1] and specialty cameras [12]. However, ultra-scaled devices exploiting the properties of 2D materials are more challenging because native defects in the 2D materials play a more important role and degrade yield and variability. Better understanding of 2D materials and methods to deposit, etch, and pattern these materials while maintaining high quality and low defect density is an important goal to make use of these materials in future devices.

**Roadblocks/Challenges:** 2D materials have significant opportunities but there are many fundamental challenges that need attention. 2D materials can contain a high density of native defects that reduce yield and performance while increasing variability. Most synthesis requires high temperatures (>800ºC) incompatible with CMOS processes, and scaling of transfer processes is particularly difficult for non-planar substrates. Coating 2D materials with metals or insulators, or etching patterns or vias, can produce damage. Improved materials and processes are needed for ohmic contact and for gate insulator integration.

**Possible Solutions:** Most important is the need for improved 2D material deposition and etch, as well as deposition and etch of other materials, such as contact metals and gate dielectrics, that would be integrated with 2D materials. Exfoliation and transfer could be envisioned, but wafer-scale implementation has inherent challenges. Discovering new 2D materials that overcome the limits of high-temperature processing possibly including CaF_{2}, PTCDA, BiSO_{5}, SrTiO_{3}, and others, could provide new options.

For conducting applications, 2D conductors such as graphene and MXenes could decrease sheet resistance, dissipate heat and avoid electromigration[4]. Semiconducting 2D molybdenum disulfide (MoS_{2}) and tungsten diselenide (WSe_{2}), could be used as FET channels, where a well-defined ultra-thin (sub 1 nm) channel depth can improve electrostatic control, reduce vertical leakage current, minimize horizontal phonon scattering, and avoid short channel effects[5], while maintaining device-to-device consistency[6]. Insulating 2D hexagonal boron nitride (h-BN) could reduce leakage current and scattering at interfaces and slow dielectric breakdown in FETs[7]. In addition, conducting defects in 2D insulators can be highly localized, enabling confinement that could be exploited for memristor systems.

**Figure 10.4:** Schematic visualization of the technology roadmap for the introduction of 2DMs in CMOS-compatible technology. Adapted from [8] (*This item still under development*).

10.3.5 Processes and Manufacturing

10.3.5.1 High NA Lithography and Directed Self Assembly

**Goals:** Lithography has enabled chipmakers to develop smaller and faster devices at advanced nodes. Utilizing the 13.5 nm wavelength, ASML’s 0.33 NA EUV scanners are being used by major companies for advanced chip production. However, beyond the 3 nm node it will become difficult to pattern future chips using existing EUV. This problem can be addressed using 0.55 high-NA EUV currently developed by ASML. Used to pattern the most critical back-end-of-line patterns, with one exposure it can print metal line/space...
patterns with 16 nm full-pitch. However, the transition high-NA presents challenges for resist performance, optimizing underlayer, etch photomask, pellicles and other aspects of the pattern transfer process. Hyper NA is a novel area that will enable hyperscaling transistor needs, this will require new resists, power sources, and metrology techniques.

Roadblocks/Challenges: Resist materials are analyzed by quantifying resolution [R], line-width roughness [L], and sensitivity [S], but optimizing one generally degrades at least one of the others. This “RLS trade-off” is a dominant problem receiving significant attention. Most current EUV resists in production are chemically amplified resists (CARs) or metal oxides. However, industry continues to search for resists that meet the RLS requirements for high NA EUV at an acceptable dose. Small features can be attained by EUV double patterning, but the process is expensive and complex, and double patterning cannot overcome the resolution limits of the current tools compared to that provided by high NA.

Possible Solutions: The High-NA platform is designed to enable multiple future nodes, starting at the 3 nm Logic node and followed by Memory nodes at similar density. New resist materials and resist processing, including dry deposition and dry development, i.e. by Molecular Layer Deposition and chemically selective dry etching, respectively, are important directions for enhanced study. Modifying deposited and/or developed resists with further chemical processing, such as polymer addition via Directed Self Assembly, inorganic modification via vapor phase infiltration, and resist shaping by atomic layer deposition and atomic layer etching are possible directions for study.

10.3.5.2 Atomic Scale Processing, including Atomic Layer Deposition and Atomic Layer Etching

Goals: New device structures and materials will require new understanding and improved control of materials synthesis and processing at the atomic scale. Atomic Layer Deposition (ALD), using a controlled sequence of self-limiting surface reactions, is a current work-horse method, and quasi-self-limiting Atomic Layer Etching (ALE) is quickly gaining speed toward full-scale manufacturing. Area-selective Deposition (ASD), often achieved using combined deposition and etching, is also of substantial interest. Other Atomic Scale Processes, such as Atomic Layer Annealing, are also of considerable interest. New advances in molecular engineering are needed, including precursor molecule design and synthesis, atomistic modeling and directed chemical synthesis based on self-limiting principles of ALD or using new control strategies yet to be realized.

Roadblocks/Challenges: The extent of film growth during an individual self-limiting ALD cycle depends on the structure of the reactive precursors and the nature of the surface reactive sites. This means that different mechanisms typically dominate, and different deposition occurs during nucleation and growth initiation. These non-conventional mechanisms become more dominant when ultra-thin films are being formed. Similarly, many mechanisms in ALE are described, but how mechanisms evolve as etching proceeds, and how they apply to ultrathin film, are still not well known.

New devices need to be constructed using many different materials exposed on the surface during fabrication. This poses a challenge for Atomic Scale Processing where desired reactions on one exposed material may cause detrimental results or unwanted damage to adjacent materials. The advent of 2D materials, for example, provides additional challenges since the 2D structure is defined by built-in chemical anisotropy, with reactive edges and relatively passive exposed surfaces.

The reactants used during ALD and ALE are molecules with built-in atomic scale precision, and a key challenge is to identify viable and scalable low temperature processes that can translate this precision to
solid and thin film materials and material junctions. The constraint of low temperatures is critical to avoid underlying dopant and metal diffusion in underlying layers, and therefore, unlike high temperature processes such as epitaxial growth, low temperature imposes limits on the extent to which equilibrium thermodynamics can be employed to drive desired outcomes. Moreover, ALD and ALE are inherently random, so that steric hindrance and other molecular effects can lead to intrinsic non-uniformities at the atomic scale.

Possible Solutions:

**Chemical Selectivity:** Means to identify, promote, analyze and quantify chemical selectivity in both ALD and ALE are a high priority for future processing. Advances in selectivity, including Area Selective Deposition, and selective etching, will require a balance of underlying thermodynamic driving forces and chemical reaction rates. Surface passivation may block an undesired energetically favorable reaction, but surface control to attain intrinsic selectivity, where molecular passivation is not needed, may be more favorable if the rate of a desired process can be increased to limit the extent of ancillary unwanted nucleation. While ALD allows the thickness of uniform films to be controlled on high-aspect ratio surfaces with sub-monolayer precision, the chemical selectivity achieved in ASD may also provide a pathway to control lateral growth. Researchers are beginning to understand how precursor interactions on non-reactive surfaces are influencing the shape and extent of lateral overgrowth of resulting ASD patterns. Better means to quantify pattern shape, such as analysis of vertical selectivity, could lead to new understanding to address the fundamental limits of stochastic surface reactions during ALD and ALE.

**Precursor, Process Co-Design:** Synergistic design of precursors and processes is an important direction for expanded research. New understanding of precursor stability, liability and reactivity, including for example the development of novel Hf compounds, has led to substantial advances in ALD processes and ALD integration into semiconductor manufacturing. Precursors specifically designed, for example, for selective reaction on a desired surface would be of high value. Precursor/process co-design could also lead to reactants that follow desired predetermined reaction pathways when specific co-reactants are used under predesigned reaction conditions. This could allow, for example, low temperature deposition of stable, low-defect crystalline materials for advanced device systems.

**Process Intensification:** Atomic scale processing may also benefit from new synergistic approaches where deposition and etching are integrated, either as a repeated cyclic sequence, or as simultaneous co-localized or adjacent reactions. The importance of coupling multiple synthesis steps into a single intensified process is well recognized in the chemical industry, and analogous advances in process intensification are important to pursue in electronics manufacturing. Deposition and etching can be balanced by adjusting temperature to control equilibrium product distributions, but this balance typically requires high temperatures where unwanted dopant and metal diffusion often sets in. New techniques combining process reactions at low temperature may hold promise to guide reaction pathways needed for future device structures.

References

*Decadal Plan for Semiconductors - SRC*