McKinsey & Company

Industrials & Electronics

Domain-specific architectures and the future of compute

July 18, 2023 | Article

Domain-specific architectures take center stage for compute innovation as transistor scaling falls behind Moore's law and compute domains proliferate.

Long-standing scaling trends in semiconductor process-technology innovation are slowing down. After several decades of remarkable compliance with Moore's law—the observation that transistor density on a semiconductor wafer roughly doubles every two

years—transistor scaling has meaningfully slowed in past years and is behind where Moore's law would have predicted by a factor of about ten. Dennard scaling, the projection that power consumption per unit chip area remains constant as transistor density increases, is also failing, leading to an increasing need for complex cooling solutions in large data centers and other high-performance compute environments.

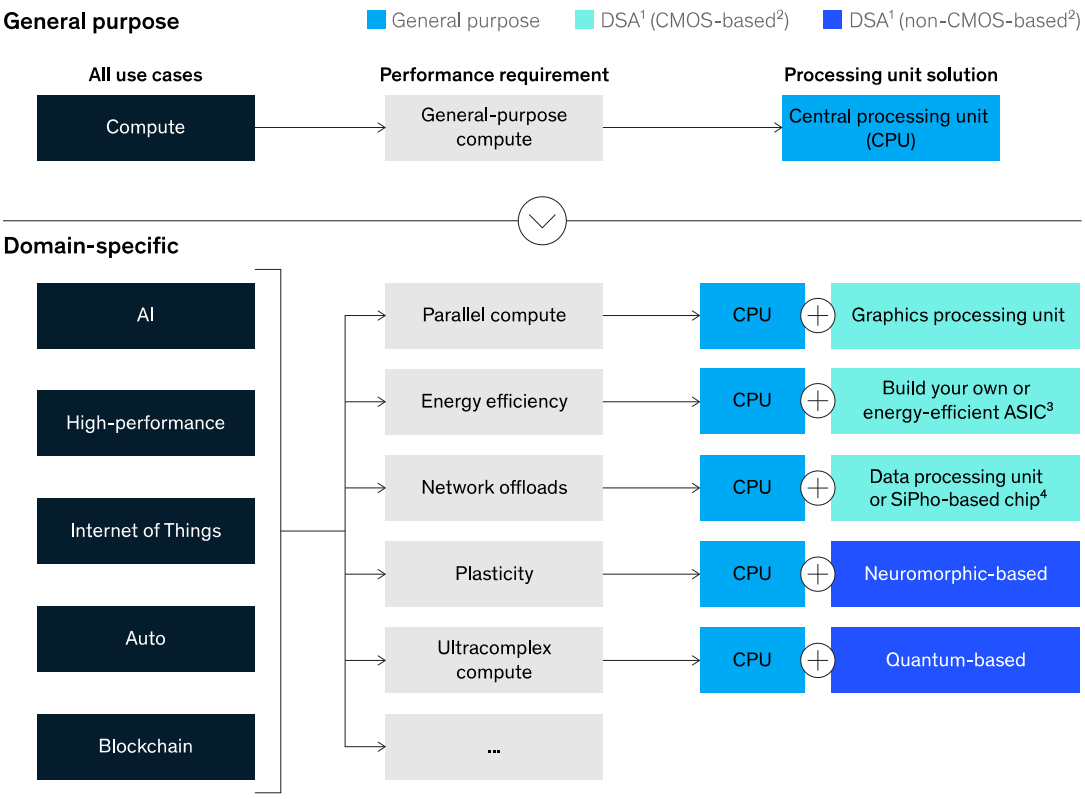
The architectural response to slowing semiconductor process innovation

In their Turing lecture in 2018, famed computer architects John Hennessy and David Patterson observed that slowdowns in process technology innovation would steadily increase the incentive for architectural innovations—that is, the way in which integrated circuits are designed to perform

computational tasks.[1] They argued that the inherent inefficiencies of general-purpose compute architectures (for example, CPUs) will start to yield to (or be complemented by) the compute power and cost efficiency of architectures optimized for specific computational tasks, also called domain-specific architectures (DSAs) (exhibit).

Exhibit

Domain-specific architectures complement or substitute general-purpose compute by offering workload- and application-specific features.



[1]Domain-specific architecture.
[2]CMOS stands for complementary metal–oxide–semiconductor.
[3]Application-specific integrated circuit.
[4]Silicone photonics–based.

McKinsey & Company

At the same time, as compute and digitization are proliferating into many application domains, such as cloud (AI and high-performance computing), networking, edge, the Internet of Things (IoT), and autonomous driving, highly domain-specific computational workloads are expanding the opportunities for DSAs to offer meaningful

performance advantages. Large language models, a core engine for generative AI, through applications such as ChatGPT provide further specialization within AI workloads at very high volume and may lead to further hardware specialization.

The commercial potential of DSAs—hardware and software developed for a specific application domain—is sizable. Already, graphics processing units (GPUs) and tensor processing units (TPUs) have gained significant market share in data centers, where they outperform CPUs for workloads that benefit from a high degree of parallelization, such as AI workloads (learning and inferencing). The performance improvements can be dramatic, with workload-specific accelerations of 15 to 50 times being common. In automotive, customized solutions from leading providers deliver the low-latency, high-performance inference needed to safely support increasing levels of autonomous driving.

As we see DSAs expand to other application domains, we estimate that DSAs will account for roughly $90 billion in revenue (or about 10 to 15 percent of the global semiconductor market) by 2026—up from approximately $40 billion in 2022. Therefore, it is not a surprise that we have seen a marked increase in venture capital inflows into domain-specific design start-ups, with $18 billion of funding supporting roughly 150 start-ups cumulatively in the past ten years, a stark difference with the decade prior, when hardware investments were shunned in favor of software.

Firms in the semiconductor value chain, manufacturers of compute systems, and end users of compute solutions should prepare to take advantage of this trend rather than be caught off guard.

Key enablers for commercial viability of

DSAs are increasingly in place

Historically, in addition to benefiting from the massive tailwind of Moore's law, CPUs have benefited from large economies of scale to offset theoretical benefits from competing domain-specific chips, which logically face smaller volume demand due to their specificity and which may require specialty software to deploy efficiently. The more expensive the chip (driven by die size, complexity, and process technology node), the larger the scale required from the application domain to justify a DSA. Fueling the disruptive potential of DSAs are the following five important enablers that are coming together to narrow the economic gap between general-purpose and domain-specific designs:

1. Access to mature and leading-edge

semiconductor technology manufacturing through foundries

Foundries—companies that focus on semiconductor manufacturing services—have taken increasing share of global semiconductor manufacturing because they can aggregate demand and achieve the efficiencies of scale needed to offset the escalating cost of producing modern semiconductors. (The cost of a leading-edge semiconductor fabrication plant, or “fab,” is more than $10 billion.) Not only have foundries steadily gained manufacturing market share across technology nodes, but they also offer access to the most advanced technology nodes, an edge held until recently by integrated device manufacturers. As a result, any start-up with a clever idea for an outperforming DSA design can rapidly access the most advanced manufacturing

to have it built without having to invest a single dollar in manufacturing capacity.

2. Fast go-to-market scaling through mature cloud platforms

Would-be providers of superior DSAs, specifically those targeting enterprise, AI, or high-performance computing (HPC) workloads, do not necessarily need to develop their own go-to-market infrastructure. They can rely on a mature ecosystem of cloud service providers (CSPs) that offer compute as a service. If they can demonstrate to CSPs and their customer base that their DSA offers superior compute performance (per dollar and per watt) for specific workloads, their hardware solutions can be integrated into the CSP data center infrastructure and be offered as a hardware instance to end customers of the compute cycles.

3. Rich libraries of open-source and licensed IP to jump-start DSA design

Even though DSAs are, by definition, designed for domain-specific workloads, this does not mean DSA designers have to start from scratch when designing the circuitry. Licensable instruction set architectures (ISAs), such as Arm and x86, and open-source ISAs such as RISC-V democratize chip design and provide a rich set of building blocks and ready-to-go design components. They also permit access to the respective ecosystems of compilers and application-level software solutions. Choosing between these different ecosystems as a foundation will be a trade-off between software stack maturity, cost, and domain-specific hardware performance.

4. Advances in 2-D and 3-D chip packaging

supporting heterogeneous integration of DSA chiplets

Increasingly, leading-edge compute devices are no longer made up of a single chip. As high-performance chips have grown bigger and process technology has become more expensive and more difficult to deliver at high process yields, leading players have moved to a disaggregation strategy, building chiplets rather than a single large, monolithic die. These chiplets, potentially optimized on their own process technology and for their own functions, are subsequently integrated in an advanced package. Where chip packages used to contain only one chip, advanced packaging allows for the heterogeneous integration of tens of chips in a single package, arranged in 2-D and even in 3-D. This technology trend is favorable for firms focusing on DSA chiplets, since these chiplets can now be integrated in advanced packages, allowing for

connectivity with other compute, communication, memory, and analog components with extremely high bandwidth and low latency.

5. Physical-layer innovations enabling new types of DSAs

Alternatives to complementary metal–oxide–semiconductors (CMOS) for the physical compute layer, such as photonics and neuromorphic and quantum architectures, promise to offer specific advantages for domain-specific compute needs, such as energy efficiency, plasticity, task-specific speed, and linear scaling with specific NP-hard problems (for more on DSA use cases, see sidebar, “Domain-specific architecture use cases, ready to go”). As these physical-layer solutions mature, they will open up new classes of DSAs.

Further innovations across the technology stack are needed to extract full value from DSAs

Across the tech stack, from physical-layer to application-level workload management, further innovations are needed to propel the feasibility and commercial success of DSAs.

At the physical and circuit layer, open-source ecosystems such as Arm and RISC-V need to mature further to support a full software stack on top of DSAs based on these building blocks. Without efficient software stacks, many hardware-level performance advantages will not translate into real-life workload accelerations.

At the system-in-package (SiP) level, standardization of chiplet interfaces will be required to allow economic and ubiquitous

integration of DSAs. Industry consortiums such as the Universal Chiplet Interconnect Express (UCIe) have started to form to define these standards. Furthermore, in the United States, the CHIPS Act and DARPA (Defense Advanced Research Projects Agency) recognize that the realization of collaborative development platforms for advanced packaging is an important investment area and are directing incentives to stimulate their development.

At the operating-system and compiler level, higher-level compilers will need to account efficiently for the potential coexistence of multiple ISAs in a single package.

At the data center level, advanced hypervisors and orchestrators will be needed to coordinate workload containers optimally across different DSA compute instances and balance utilization across the data center to deliver the DSA-level benefits to the end customer's applications. Furthermore, CSPs will develop tools to support their end customers to understand the optimal

configuration of hardware instances for their specific compute needs to avoid inefficient deployment of compute resources.

Firms in the semiconductor value chain and end users should prepare for DSA-driven disruptions

DSAs will likely activate disruptions throughout the semiconductor value chain. Here's how to prepare for them.

Semiconductor firms: Prepare for value chain disruptions

Materials providers should understand the impact of advanced packaging (for example, the need for new substrate materials that form the foundation of 2-D and 3-D integration and the need to be thermally and mechanically more stable than current substrate materials), as well as the impact of new physical-layer paradigms on front-end and back-end material flows.

Front-end tool manufacturers will want to participate in the advanced-packaging and heterogenous-integration boom supporting DSA integration, which will require precision definition and alignment similar to front-end manufacturing.

Foundries will need to gear up to meet demand for smaller batches and a higher mix of domain-specific chips and chiplets and find ways to support smaller players efficiently and economically. In addition, support for non–leading edge and new physical-layer solutions such as photonics will become ever more important as functionality is disaggregated into task-optimized chiplets.

Chip design firms will require talent that can think through end-to-end domain-specific workload performance, from gate layout and hardware architectural design choices to software stacks and workload management, to take full advantage of DSAs' architectural optimization.

Electronic design automation (EDA) and hardware IP firms should respond to two challenges. The first is how to adapt their business models to support smaller DSA disrupters that may not have the financial firepower to buy expensive up-front licenses. The second is how to expand their IP, design, and simulation suites from chip level to SiP level to support system-level, multiphysics (logical, electrical, thermal, optical, and mechanical) EDA across multiple chiplets and ISAs working together at bandwidths and latencies previously seen only at chip level.

Compute consumers: Use optimized DSAs, and

invest in expertise to leverage them well

CSPs have already recognized the value of workload-specific chip architectures, as demonstrated through their increasing adoption of GPUs, as well as by moving to in-house chip designs for data center hardware instances. As design start-ups continue to target workload specificity for AI and HPC use cases, CSPs will want to keep close tabs on emerging winners that they can potentially support and propel to scale.

Enterprise customers need to educate themselves on the benefits of using DSAs for their specific workloads. The benefits of moving their compute infrastructure into public clouds will only increase because access to the DSAs can be offered at scale by CSPs that can aggregate demand for specialized hardware instances and efficiently manage the workload deployment to them. However, enterprises should gain or retain the expertise on how to best leverage these hardware

instances, based on a deep understanding of how their specific workloads and cloud hardware instance configurations can be optimized to deliver maximum total cost of ownership benefits.

Domain-specific OEMs—such as IoT and edge device makers, networking-equipment vendors, car manufacturers, and blockchain platforms—will want to deeply understand the evolution of their domain-specific compute needs and workloads rather than relying on general-purpose compute progress. They will also want to familiarize themselves with the architectural options that exist to meet those needs in chip design that are optimized for their energy, cost, footprint, and performance needs. This may require them to reach beyond traditional supply chain structures: for example, automotive OEMs may need to directly scout for and engage with DSA start-ups rather than relying on tier-one and tier-two suppliers to have all the latest insights.

Moore's law has propelled the compute industry with incredible longevity, driving decades of performance improvements in general-purpose computing that largely negated the need for investments in workload specialization. With transistor scaling slowing down, DSAs will increasingly gain a use-case-specific performance edge and drive significant disruptions for value chain participants and their customers.

1. "John Hennessy and David Patterson deliver Turing lecture at ISCA 2018," Association for Computing Machinery, accessed June 27, 2023.

ABOUT THE AUTHOR(S)

Mena Issler is an associate partner in McKinsey's Bay Area office, where **Mark Patel** is a senior partner and **Rutger Vrijen** a partner. **Sheila Zingg** is a consultant in the Zurich office, and **Wendy Zhu** is an associate partner in the Denver office.

EXPLORE A CAREER WITH US