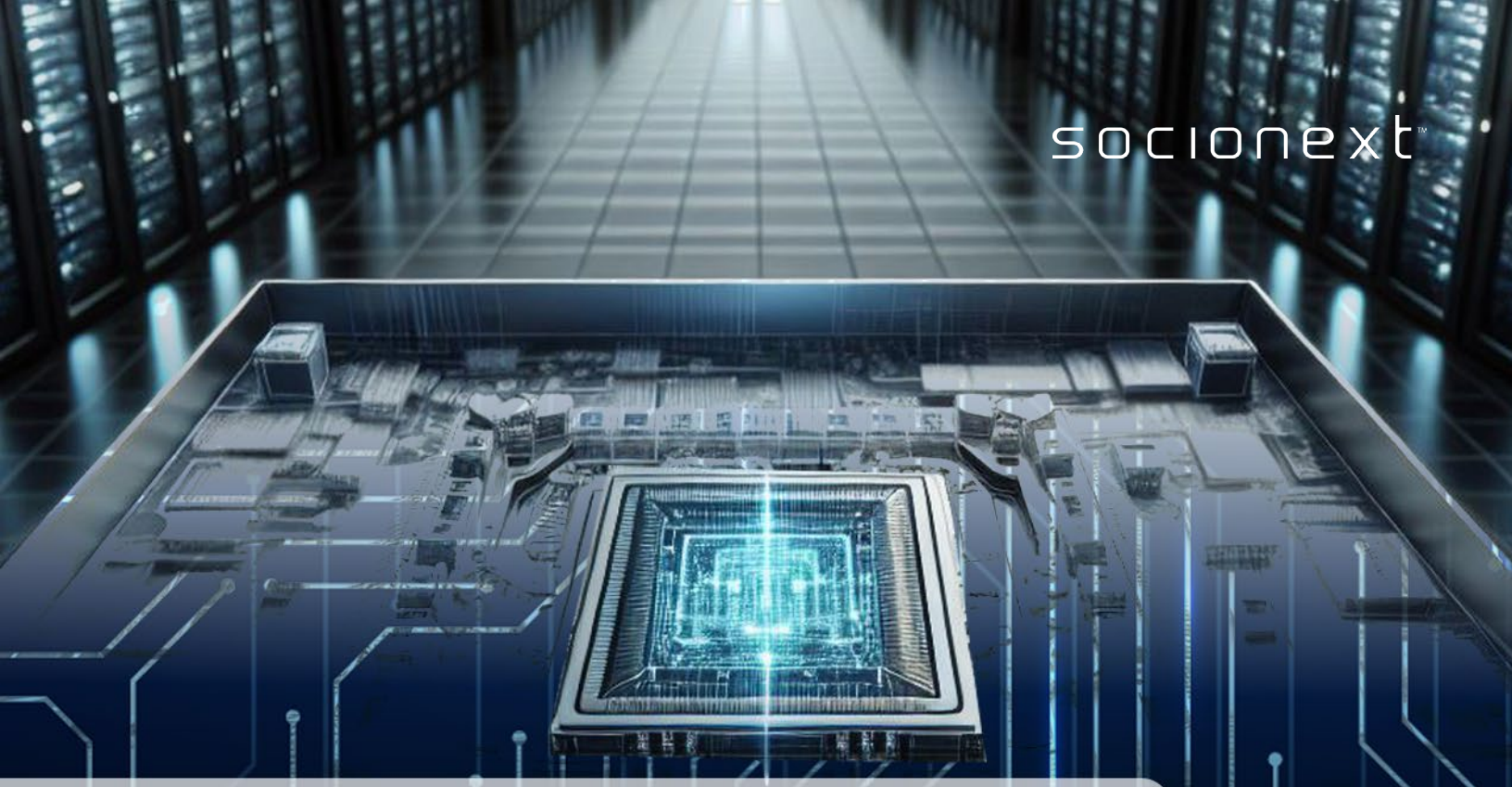
socionext



WHITE PAPER

Accelerating AI with Chiplets

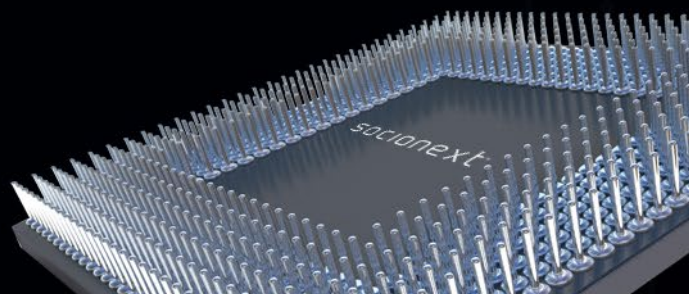


Table of Contents

Introduction

The demands of AI and machine learning workloads are driving rapid growth in compute capacity through advancements in multi-core processors, GPUs, and specialized AI accelerators. However, individual nodes can no longer handle large-scale tasks, leading to the shift toward parallel processing architectures where multiple processing elements (PEs) are linked via high-speed interconnects in a cluster. While compute power is expected to grow 10x in the next 3-4 years, I/O bandwidth is only projected to double every eight years, causing bottlenecks. The key challenge is scaling these interconnects.

ASIC designs are approaching physical limits with factors leading to reduced return of investments; those include reticle sizes, cost, complexity, and increased power consumption. Evolving system requirements are outpacing hardware development, making continuous redesigns less viable and inefficient.

Chiplets offer a promising solution by partitioning key functions such as AI acceleration and I/O into separate modules. They allow designers to focus on optimizing compute without being hindered by I/O bottlenecks.

By addressing the expansion of compute capabilities and scalable interconnects, chiplets enable easier design evolution through modular upgrades, eliminating the need for full ASIC redesigns. However, their effectiveness hinges on optimizing the interconnects between chiplets to ensure they keep pace with the performance scaling of compute elements.

Growth Drivers and System Design Challenges

The exponential growth of AI has been driven by algorithms, data, and compute.

Algorithms are becoming increasingly complex thanks to advances in machine learning techniques and the collaborative nature of open-source development. Breakthroughs in model architectures, such as transformers, have rapidly advanced AI capabilities, enabling larger and more complex models.

Data creation is accelerating at an unprecedented rate, with massive amounts of labeled and unlabeled data generated daily from various sources. This data serves as the fuel for AI, enabling deeper learning and more accurate models. However, efficiently utilizing this data requires substantial processing power and memory bandwidth.

Compute makes this AI revolution possible. Initially, many hyperscalers and data center service providers have leveraged standard ASSPs (Application-Specific Standard Products) for AI training and inference. However, AI models and workloads have grown in scale and complexity, causing a shift towards custom SoC (System-on-Chip) based hardware that can efficiently handle massive datasets and complex algorithms. Custom SoCs are designed with processing elements (PEs) that are tightly coupled with high-bandwidth memory (HBM) and high-speed interconnects so they can handle AI workloads. These systems must be capable of scaling up (increasing resources within a single node) and scaling out (distributing workloads across multiple nodes) seamlessly.

As data and models are scaled across different processing elements, they are either divided into independent slices for parallel processing or processed in a pipelined fashion for tasks with dependencies between stages. As we move outward from the core of an ASIC, interconnect bandwidth decreases at each level, from high-speed connections between processing elements within the ASIC to inter-die links in multi-chip modules (MCMs) through the substrate and inter ASIC connections at the PCB level.

Scaling efficiently hinges on the speed and bandwidth of the interconnects between processing elements. Still, because PCIe, Ethernet, and other chip-to-chip links are not evolving as rapidly as compute capacity, several challenges arise, including rising costs due to the need for more complex hardware design and lack of scalability if hardware cannot be easily customized for system requirements. The result is extended time-to- market for new AI systems as more complex designs take longer to implement and validate.

The AI Infrastructure Market from a Custom SOC Perspective

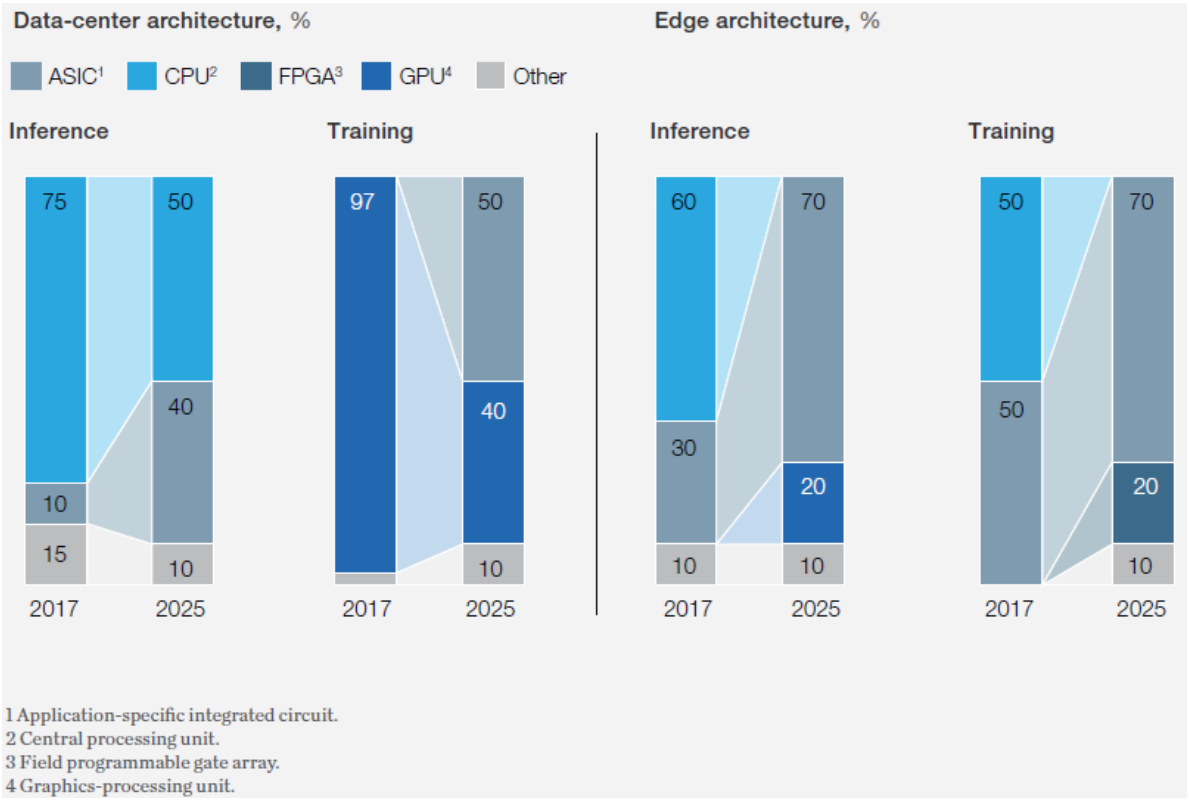


Fig. 1. The preferred architectures for compute are shifting in data centers and the edge[1]

The move toward custom SoCs, particularly ASICs, across data centers and edge environments reflects the trend to address workload challenges with specialized compute architectures. As AI inference and training workloads grow more demanding, GPUs and general-purpose CPUs are losing ground to custom ASICs, which can handle specific tasks more efficiently.

In 2017, GPU-based architectures dominated AI training workloads, with 97% of compute relying on GPUs, which excel at handling the highly parallelizable nature of AI model training. However, by 2025, ASICs are expected to surpass GPUs, taking a 50% share. Optimized for deep learning algorithms, these ASICs will offer superior cost efficiency and performance-per-watt, making them ideal for large-scale training in data centers.

The projected dominance of ASICs for both training and inference workloads at the edge by 2025 represents a dramatic shift in edge computing architectures. This shift highlights the need for architectures to meet growing AI demands while maintaining energy efficiency at the edge.

As AI models grow in complexity, these specialized architectures must evolve while balancing factors like cost, performance, and time-to-market. With Moore's Law slowing down, scalable and efficient architectures, such as chiplets, can help keep pace with the future of AI computing.

[1] Expert interviews; McKinsey analysis

The Cluster is the New Node

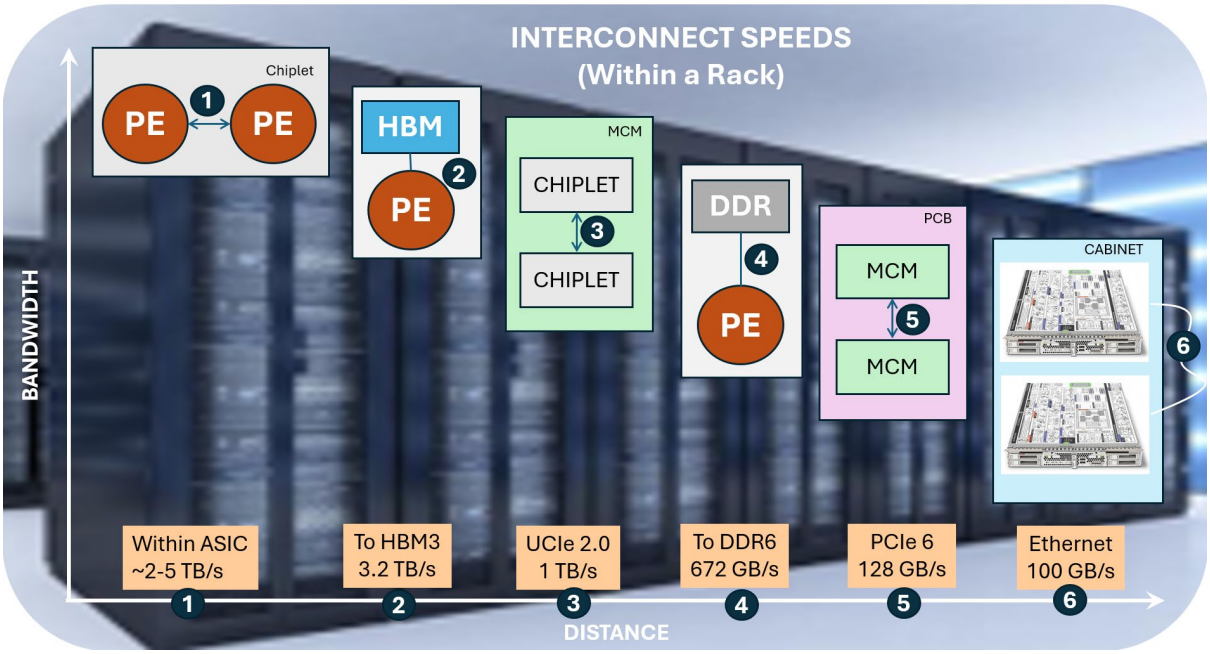


Fig. 2 - Cluster is the new node

Moving outward from the processing elements within an ASIC to external memory, chiplets, and eventually across server racks, available bandwidth steadily decreases while the distance between components increases, limiting how much data can be efficiently transferred between different parts of a system. This is especially true, as on-die connections give way to inter-board and inter-rack communications.

AI model's growth in size and complexity means a single node can no longer handle the size of modern training workloads. The volume of data and the number of parameters involved in training state-of-the-art models exceed the capacity of any individual chip or system, even with the most advanced compute architectures. This trend is driven by the demand for more accurate, capable models that require massive parallel processing and memory resources to handle larger datasets and more sophisticated algorithms.

So, the focus has shifted from maximizing the power of a single node to distributing workloads efficiently across multiple nodes in a cluster. Architectures that allow clusters of interconnected nodes to work seamlessly together are better than trying to fit everything into a monolithic system. This shift requires new strategies for scaling out, where workloads are partitioned and executed across many nodes so the entire cluster can handle the demands of fast-growing models. The key is to ensure efficient communication between nodes to maintain high performance as model sizes expand.

Efficient Interconnects Drive the Cluster

Efficient Interconnects ensure that data can move swiftly and seamlessly from one processing node to another and across racks in a data center.

Architecting these interconnects demands collaboration with partners with the expertise needed to optimize cost, performance, and power for complex SOC designs.

The **Universal Chiplet Interconnect Express (UCIe)** standard is vital for enabling high-speed, low-latency data exchange between chiplets on the same package. The UCIe interface handles massive bandwidth requirements but requires finely tuned controllers to handle protocol translation, error correction, and resource management.

Table 1 – UCIe Bandwidth Summary

Generation	Year Introduced	Data Rate per Lane	Bandwidth per Lane	Total Bandwidth (x16)
UCIe 1.0	2022	16 GT/s	32 GB/s	512 GB/s
UCIe 2.0	~2024/2025 (Expected)	32 GT/s	64 GB/s	1.024 TB/s (1024 GB/s)
UCIeS	~2025/2026 (Expected)	64 GT/s	128 GB/s	2.048 TB/s (2048 GB/s)
UCIeA	~2027/2028 (Expected)	128 GT/s	256 GB/s	4.096 TB/s (4096 GB/s)

Similarly, HBM (High Bandwidth Memory) and GDDR (Graphics Double Data Rate) controllers manage the high-speed access to memory critical for handling large datasets in AI workloads. This effort requires engineering expertise to minimize latency, balance power consumption, and meet the bandwidth demands of next-generation models. Well-structured collaboration is essential.

Table 2 – HBM Bandwidth Summary

Generation	Year Introduced	Bandwidth per Stack	Total Bandwidth (4 stacks)	Frequency	Stacks (DRAM Dies)	Interface Width
HBM	2015	~128 GB/s	Up to 512 GB/s	500 MHz	2 to 4	1024-bit per stack
HBM2	2016	Up to 256 GB/s	Up to 1 TB/s	1.0 GHz	4 to 8	1024-bit per stack
HBM2E	2019	Up to 410 GB/s	~1.64 TB/s	1.2 – 1.6 GHz	Up to 8	1024-bit per stack
HBM3	2022	Up to 819 GB/s	Up to 3.276 TB/s	2.0 – 2.5 GHz	Up to 16	1024-bit per stack
HBM4	~2026 (expected)	~1.2 – 1.4 TB/s	~5-6 TB/s (estimated)	~3.0 GHz (est.)	TBD	1024-bit per stack

Table 3 – DDR Bandwidth Summary

Generation	Year Introduced	Effective Frequency	Bus Width per Chip	Bandwidth per Chip	Total Bandwidth (256-bit interface)
GDDR5	2008	Up to 7 GHz	32 bits	56 GB/s	224 GB/s
GDDR5X	2016	Up to 12 GHz	32 bits	96 GB/s	384 GB/s
GDDR6	2018	Up to 16 GHz	32 bits	128 GB/s	512 GB/s
GDDR6X	2020	Up to 21 GHz	32 bits	168 GB/s	672 GB/s
GDDR7	~2025 (expected)	Up to 32 GHz	32 bits	256 GB/s	1024 GB/s

Advanced Packaging: A Key Factor for Success

Advanced packaging techniques, like 2.5D integration and 3D stacking, enable chiplets to be interconnected with high-speed links like UCIe. IO chiplets can feature UCIe on one side and the latest PCIe and Ethernet ports on the other, facilitating flexible connectivity and expanding interconnect scalability.

2.5D packaging uses an interposer to connect multiple dies, but optimizing the interposer to handle the necessary power and signal integrity requires years of refinement. 3D stacking, which vertically integrates memory and logic, pushes the limits of chip design but also introduces significant challenges in thermal management and die-to-die communication. Similarly, EMIB offers an alternative to full interposers, connecting dies with lower complexity, but requires careful planning to manage signal integrity and bandwidth.

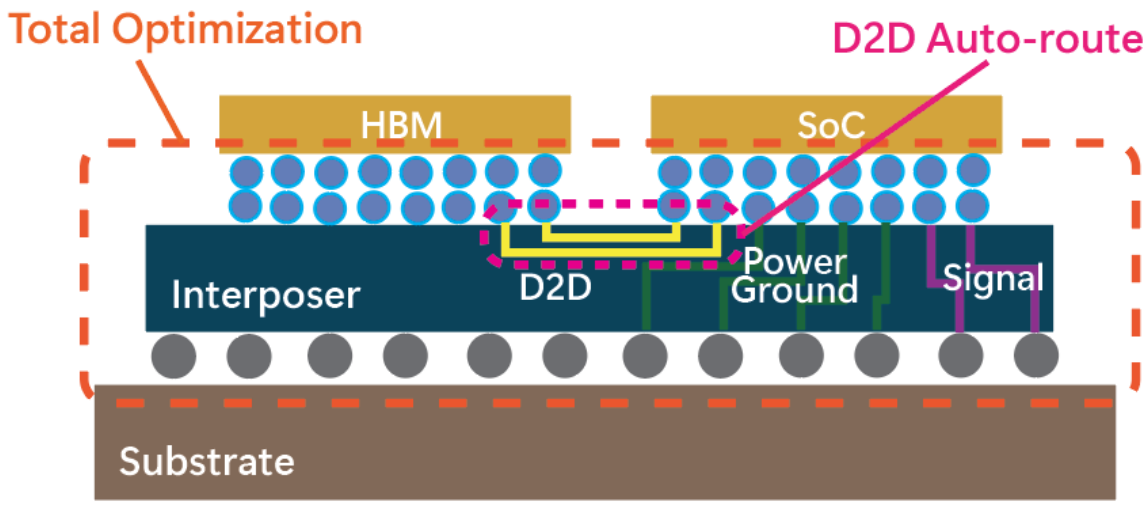


Fig. 3 – Advanced 2.5D packaging

Working with partners who have mastered these advanced packaging technologies and developed reliable IPs for UCIe, HBM, SerDes, PCIe, and optical interconnects can make it easier to integrate chiplets into a cohesive, high-performance system. The results: reduced risk and time-to-market delays.

Transitioning from Copper to Optical: Co-Packaged Optics

Copper, while still prevalent in short-range connections, is reaching its limits when scaling bandwidth over longer distances. The shift from copper to optical interconnects further illustrates the complexity of designing efficient interconnects. Co-packaged optics (CPO), where optical transceivers are integrated directly into the chip package, can maintain high-speed data transfer at the rack or system level. But this isn't just a simple upgrade - it requires a deep understanding of photonics, packaging, and thermal management.

Developing co-packaged optics requires expertise in optical transmission and integrating these components with silicon designs. Working in tandem with a partner company that has the right IP and expertise can reduce risk and accelerate the development of scalable, high-bandwidth systems.

Breaking Up a Monolithic ASIC

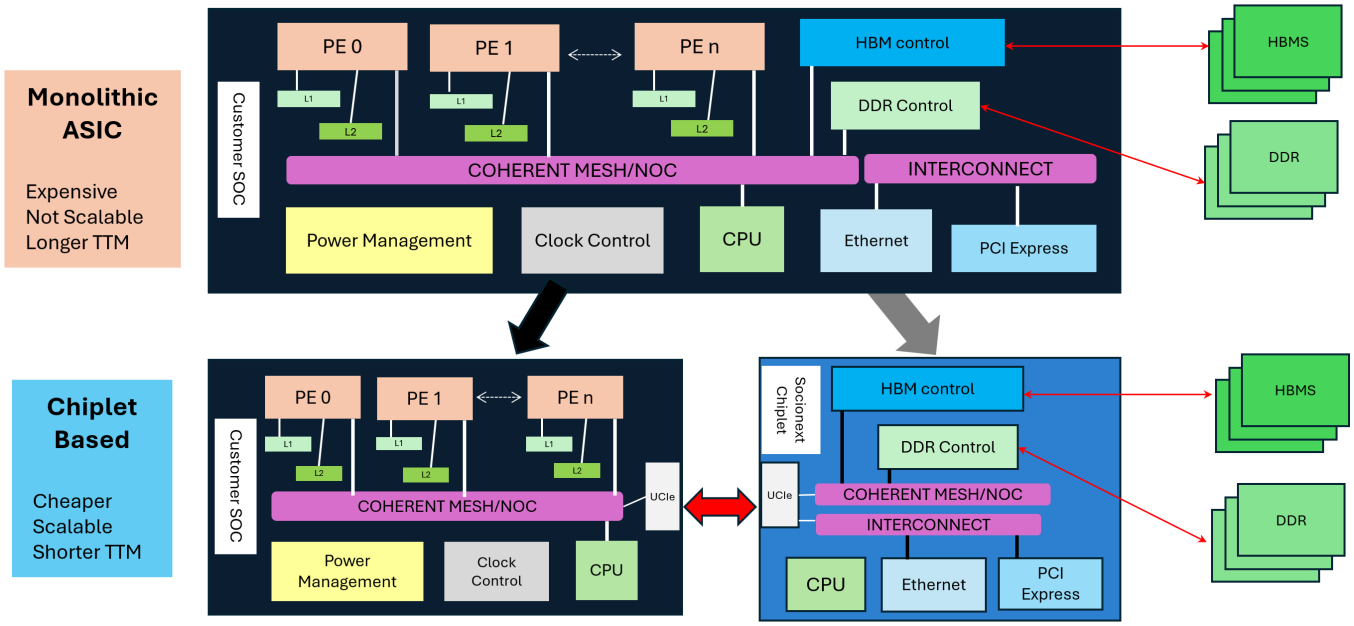


Fig. 4 - Breaking up an ASIC

The example above demonstrates a potential strategy for decomposing a monolithic ASIC into a chiplet-based architecture. In a traditional monolithic design, the processing elements are interconnected through a coherent mesh or Network-on-Chip (NOC), each with its dedicated L1/L2 cache and access to a shared L3 cache (not depicted). The design also includes functional blocks for power management and clocking. All these elements are tightly integrated and communicate through the coherent mesh/NOC, with memory and IO controllers interfaced via a central switch.

There are several approaches to breaking this monolithic ASIC into chiplets. One option is to move both the IO and memory controllers into a separate chiplet, connecting it to the primary compute chiplet via UCIe. This design can help reduce the complexity and size of the main chip while simplifying advanced IO and memory integration.

Alternatively, in latency-sensitive designs, a more selective approach could be adopted using only the non-latency-sensitive functional blocks, such as certain IOs in the second chiplet. The primary chiplet would retain the HBM memory controllers and other critical functions in order to minimize latency.

In either scenario, cache coherence must be carefully managed, as communication between the chiplets introduces new complexities in maintaining a consistent and coherent memory hierarchy. The design of the interconnect, especially with respect to handling latency, ensures optimal performance in a chiplet-based architecture.

End Note

The growing size and complexity of AI models necessitate a parallel processing scale-out architecture with optimized interconnects and distributed functionalities using chiplets, which enable a modular approach. This allows designers to choose the best process technology for different parts of the system, offering power, cost, scalability, and time-to-market advantages.

About Socionext

Socionext was formed in 2015 by the fusion of Fujitsu semiconductor and Panasonic LSI, two well-established semiconductor companies with over 40 years of experience as technology leaders delivering solution SoCs for high-performance computing, networking, and consumer applications. The challenge of developing and bringing novel AI architectures to the market is substantial, but it can be managed by partnering with Socionext.

Socionext is developing chiplet-based Solution SoCs to address the most demanding applications in the Automotive, Data Center, and Industrial segments, and recently announced a collaboration with Arm and TSMC to develop 2nm chiplet Solution SoCs aimed at hyperscale data center server, 5/6G infrastructure, DPU, and edge-of-network markets.

socionext™

Accelerating AI with Chiplets