CMM INTERNATIONAL
COMMERCIAL MICRO MANUFACTURING

Beyond Moore: Design and Manufacturing Challenges in System-on-a-Wafer

JUNE 4, 2024 10:18 AM

Chetan Arvind Patil. Senior Product Engineer. NXP USA Inc

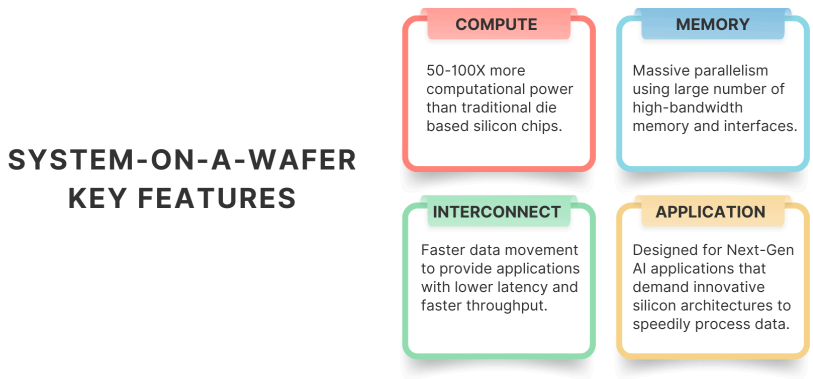
For decades, Moore's Law has been the guiding principle of the semiconductor industry, driving exponential growth in computing power by doubling the number of transistors on a chip approximately every two years. However, as the semiconductor chip design and manufacturing approach the physical and economic limits of traditional scaling, the industry must look beyond Moore's Law for innovative solutions.

There have already been many solutions on this front, such as multi-core chips, System-On-A-Chip (SoC), 2.5D to 3D-based SoCs, and most recently, the advent of heterogeneous integration using chipset-based solutions, an extension of System-In-A-Package (SiP).

Lately, however, one more promising solution beyond Moore's is the System-On-A-Wafer (SoW), which integrates multiple dies onto a single wafer to create an extensive, contiguous system.

SoW not only goes beyond the limitation of silicon area but is also seen as an alternative to providing silicon chips geared towards never-before-seen memory and compute-intensive applications, i.e., Artificial Intelligence.

Understanding System-On-A-Wafer (SoW)



June-2024 - 1

System-On-A-Wafer (SoW) is an advanced technology that integrates various functional dies, such as processors, memory, and specialized accelerators, onto a single wafer. Unlike traditional System-On-Chip (SoC) or System-In-Package (SiP) designs, SoW aims to create a unified, large-scale system that offers higher performance and efficiency.

One key distinction of SoW is that it typically produces a single finished silicon product per wafer, unlike SoC or other techniques that often yield multiple finished silicon products. This unique characteristic of SoW can be seen as a trade-off, as it can be more risky and costly. However, it also brings several benefits tailored to specific applications and use cases, making it a compelling choice in particular scenarios.

There are several ways to visualize how SoWs get designed and then fabricated is by comparing them with traditional die-based chips, where there are numerous (depending on die size/area) die per wafer. In contrast, SoW takes one wafer to create one single die, i.e., one single chip. Also, the focus of SoW is on providing an end solution 50-100x more potent than traditional die-based chips.

Aspect	Traditional Die-Based Chip Fabrication	System-On-A-Wafer (SoW) Fabrication
Integration And Architecture	- Discrete chips are fabricated, packaged, and assembled on a PCB. - Modular approach allows easy upgrades or replacements.	- Multiple dies (for the same end product) are integrated on a single wafer, functioning as a unified system. - Complex design is required for the seamless integration of diverse components.
Performance And Latency	- Data transfer between chips occurs through external interconnects, introducing latency. - Performance is limited by the speed and efficiency of PCB traces and interconnects.	- Direct interconnects between dies on the wafer minimize latency and maximize throughput. - Improved signal integrity and reduced interconnect lengths enhance performance.
Manufacturing Complexity	- Utilizes established semiconductor manufacturing processes. - Each die is packaged separately, simplifying some aspects of manufacturing.	- Requires cutting-edge lithography, bonding, and etching techniques. - The entire wafer may be packaged as one system, adding complexity to packaging.
Yield And Cost	- Defects in one die do not typically affect others, allowing for better yield management. - Lower initial costs due to established processes.	- Defects in the wafer can affect multiple dies, leading to higher yield challenges. - Higher initial costs due to advanced techniques and lower yields.
Power Efficiency And Thermal Management	- Separate chips generate heat independently, often requiring individual cooling solutions. - Power distribution is managed at the PCB level.	- Integrated design allows for more efficient heat dissipation across the wafer. - Power can be distributed more efficiently within the wafer, reducing losses.
Scalability And Flexibility	- Adding or upgrading components involves adding more chips to the PCB. - High flexibility in changing or upgrading individual components.	- Scaling requires integrating more dies onto the wafer, which is limited by wafer size and yield. - While heterogeneous integration is possible, making changes or upgrades is more complex
Wafer Size	- Can make use of 6, 8, and 12-inch wafer - Thus, the cost can be determined based on the business case and demand.	- By default, has to make use of a 12-inch wafer to reach the break-even point - Even then, the negative yield impact can derail the ROI

SoW Versus GPU For Artificial Intelligence Application

Artificial Intelligence (AI) applications have already revolutionized industries by enabling advanced data processing, machine learning, and deep learning capabilities. Both System-On-A-Wafer (SoW) and Graphics Processing Units (GPUs) target this market by offering robust computational solutions. However, SoW can potentially significantly

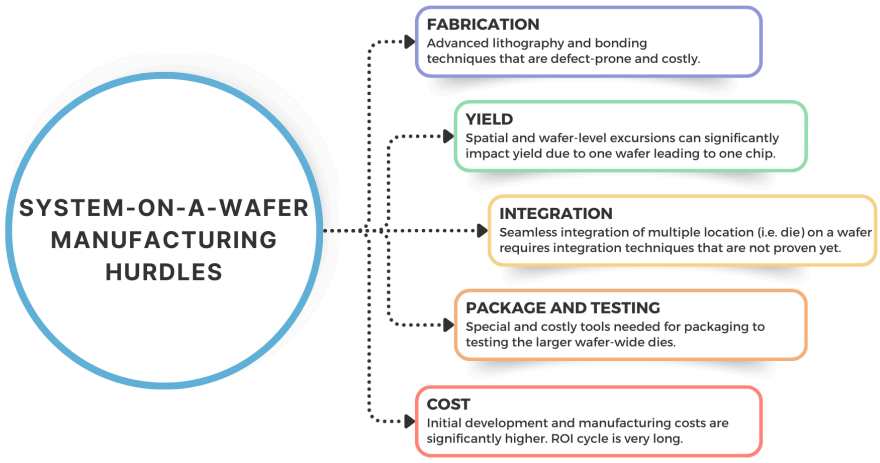
disrupt and enhance the landscape of AI silicon and the applications that use it.

GPUs In AI: GPUs have long been the backbone of AI and deep learning applications. Their architecture, designed for parallel processing, equips them to handle the massive computational requirements of AI workloads efficiently. GPUs shine in training deep neural networks, executing complex mathematical operations, and processing large datasets in real time. Industries from autonomous driving and healthcare to finance and entertainment heavily depend on GPUs to power their AI systems, underscoring their broad application.

SoW In AI: System-On-A-Wafer (SoW) technology is emerging as a compelling alternative to GPUs, especially for applications that demand unprecedented levels of integration and performance. SoW integrates multiple functional dies, including CPUs, GPUs, and specialized accelerators, directly onto a single wafer. This innovative approach offers significant benefits in terms of performance, power efficiency, and latency, making it a highly viable option for the same AI applications that are currently dominated by GPUs.

Let us take a more detailed look at different technical characteristics of how SoW will challenge GPUs:

Aspect	System-On-A-Wafer (SoW)	Graphics Processing Unit (GPU)
Architecture	Integrating multiple dies on a single wafer creates an extensive, contiguous system.	Comprises multiple cores optimized for parallel processing on a single die or a small number of interconnected dies.
Performance	Higher performance due to improved interconnect and thermal management.	High performance through massive parallelism, with thousands of cores designed for concurrent processing.
Scalability	Adding more dies to the wafer achieves high scalability, but this is limited by wafer size and yield management.	Scalable through multi-GPU configurations but limited by inter-GPU communication overhead and power constraints.
Power Efficiency	Potential for better power efficiency due to optimized interconnects and thermal management.	Generally, there is high power consumption, though recent advancements focus on improving efficiency for AI workloads.
Latency	Lower latency due to shorter interconnects between dies.	It is designed to minimize latency in parallel processing tasks but can introduce delays in multi-GPU setups.
Cost	High initial cost due to advanced manufacturing techniques and lower yield rates.	Generally, there is a lower cost per unit due to mature manufacturing processes, but costs can add up in large-scale deployments.
Manufacturing Complexity	Highly complex manufacturing process requiring advanced lithography, etching, and bonding techniques.	Less complex compared to SoW, benefiting from established fabrication processes for single-die GPUs.
Flexibility	High flexibility in integrating heterogeneous components (CPUs, GPUs, specialized accelerators).	Limited to the architecture of the GPU, though modern GPUs can include tensor cores and other specialized units.
Heat Dissipation	Improved heat dissipation across the wafer surface.	Effective heat dissipation mechanisms, but can be challenged in high-density or multi-GPU configurations.
Applications	Ideal for large-scale AI applications requiring vast computational resources and high efficiency.	Widely used in AI and deep learning due to optimized parallel processing capabilities and software support.



June-2024 - 2

Adoption And Way Forward For SoW

The adoption of System-On-A-Wafer (SoW) technology in the semiconductor industry is poised to bring about a paradigm shift in how high-performance computing systems are designed and manufactured. While promising significant advantages, this transition will involve overcoming several challenges and strategically navigating the existing landscape.

One of the primary driving factors for adopting SoW is the increasing demand for enhanced computational power. AI and machine learning applications and data-intensive tasks such as big real-time data interpretation and decision-making requires unprecedented computational power. SoW technology, by integrating multiple dies on a single wafer, can more effectively meet these performance requirements than traditional chip architectures. Additionally, the proximity of very large integrated components in SoW reduces latency and increases data transfer rates, providing a significant advantage for applications requiring real-time processing and high-bandwidth communication.

Despite its advantages, the adoption of SoW faces several challenges. Manufacturing complexity is a significant hurdle. The fabrication of SoW requires cutting-edge lithography, bonding, and etching technologies, which are more complex and costly than traditional semiconductor manufacturing processes. Yield management is another critical issue, as defects in the wafer can affect multiple integrated dies, posing significant yield challenges. Advanced defect detection and yield improvement techniques are essential to make SoW commercially viable.

Cost considerations are also significant. The initial costs of developing and manufacturing SoW are higher due to the advanced techniques required. However, economies of scale may help reduce these costs as the technology matures and scales. Significant investments in new infrastructure and equipment will also be necessary to transition from traditional manufacturing processes to SoW technology.

In the end, adopting System-On-A-Wafer (SoW) technology represents a significant opportunity for the semiconductor industry to meet the growing demands for high-performance, energy-efficient, and scalable computing solutions. While there are considerable challenges to overcome, the potential benefits of SoW in terms of performance, efficiency, and integration make it a promising technology for the

future. With continued innovation and collaboration across the industry, SoW has the potential to transform high-performance computing and drive the next wave of technological advancements.

JUNE 4, 2024 10:18 AM