The Role of Foundry for High-Performance AI Semiconductors

Young-hwa Kwun ,Bo-young Kim

ABSTRACT

Recently, the AI semiconductor market has been experiencing rapid advancements across multiple dimensions, including computational performance, energy efficiency, and integration density. As a result, not only the design capabilities of fabless companies but also the manufacturing capabilities of foundry companies that implement these designs have become more critical than ever. In particular, achieving high yield, possessing advanced process technologies, and securing advanced packaging capabilities have emerged as key factors for realizing high-performance AI semiconductors.

This study selects TSMC, the world's largest foundry company, as a case and analyzes how the company secures manufacturing competitiveness in AI semiconductors through yield management, process control, and advanced packaging technologies. The findings indicate that TSMC enhances AI semiconductor performance through integrated management across the entire value chain—from design and manufacturing to packaging and mass production—while maintaining stable production capabilities and making proactive R&D investments in advanced packaging. These results demonstrate that the integrated management of yield, process, and packaging constitutes a core mechanism for improving AI semiconductor performance.

☞ keyword : AI semiconductor, Fabless company, Foundry company, Manufacturing capability, TSMC

1. Introduction

Artificial Intelligence(AI) refers to a technology that imitates and extends human intelligence [1], and the AI market is currently growing at an annual rate exceeding 20%. Along with advances in science and technology, AI is rapidly transforming nearly every aspect of daily life and is increasingly approaching human-level capabilities [2]. The changes driven by AI are already profound and are expected to accelerate further in the future. At the center of this transformation lies AI semiconductors [3].

All chips used in AI applications can be broadly categorized as AI semiconductors [4]. As AI is being adopted across diverse industrial sectors, the demand for AI semiconductors has correspondingly increased [5]. In particular, significant advances in deep learning technologies have led to the development of a wide range of AI semiconductors [6]. AI semiconductors are specifically designed to accelerate applications based on artificial neural networks (ANNs) [7] and play a critical role across AI models and devices. Without high-performance AI semiconductors, most AI functionalities would be severely limited.

Moreover, the role of foundries is essential in realizing high-performance AI semiconductors. Achieving superior chip performance requires not only advanced design capabilities but also excellence in manufacturing processes. The continuous emergence of new semiconductor technologies has further increased the complexity of manufacturing facilities [8]. Consequently, close collaboration between fabless companies and foundries has become increasingly critical to the development of high-performance AI semiconductors.

This study therefore examines how advanced foundry technologies determine the core performance of high-performance AI semiconductors, using TSMC as a representative case. While prior studies have primarily focused on AI semiconductors from a design perspective [9][10], manufacturing capability is equally, if not more, decisive. Even the most sophisticated chip designs cannot achieve their intended performance without advanced manufacturing execution. Accordingly, the foundry's role is fundamental to the realization of high-performance AI semiconductors.

This study contributes to the literature by approaching AI semiconductors from a manufacturing perspective. Furthermore, the findings derived from the analysis of TSMC's technological capabilities are expected to provide meaningful academic and practical implications.

2 Main Body

2.1. AI Semiconductor Companies

AI semiconductor firms can be broadly classified into three categories. The first category consists of Integrated Device Manufacturers (IDMs). The second category comprises fabless firms. The final category includes Big Tech companies. This study examines AI semiconductor firms according to each of these categories.

2.1.1 Integrated Device Manufacturer (IDM) Firms

IDM firms are vertically integrated companies that conduct design, manufacturing, and packaging in-house. Representative examples include Samsung Electronics and Intel [11]. These firms also develop their own AI semiconductors.

Samsung Electronics operates across memory semiconductors, fabless, and foundry businesses, and conducts AI semiconductor development within each division. In the memory semiconductor division, Samsung is developing Processing-in-Memory(PIM) technologies and is actively researching neuromorphic semiconductors. Under the conventional von Neumann architecture, where the processor and memory are physically separated [12], large-scale data transfers inevitably result in bottlenecks and substantial power consumption [13]. Consequently, in the era of AI, neuromorphic semiconductors that integrate processing and memory are increasingly recognized as a critical technological direction.

In addition, Samsung previously collaborated with Naver through its System LSI division to develop an NPU chip called Mach-1 for data center inference applications. NPUs represent the forefront of AI semiconductor development [14]. Currently, Samsung continues the development of its NPU technology independently. Furthermore, within its foundry division, Samsung has begun mass production of 2nm process chips based on Gate-All-Around(GAA) technology from the end of 2025. Recently, Samsung Foundry reportedly secured a large-scale contract to manufacture Tesla's A16 chips using its 2nm process [15]. These chips are expected to be used primarily in Tesla's autonomous driving systems and robotics platforms. While the previous A15 chip was manufactured by TSMC, production has shifted to Samsung Foundry for the A16 generation.

Intel also launched the chiplet-based Gaudi 3 accelerator in 2024. Gaudi was developed based on the technology of Habana Labs, an Israeli AI semiconductor company acquired by Intel in 2019. This chip is positioned primarily for inference workloads and is reported to offer competitive performance at a lower price compared to NVIDIA's H100, along with the advantage of an open software ecosystem. Intel is also reportedly preparing to release Gaudi 4 in the near future.

Moreover, Intel's recently released Core Ultra processors integrate internally developed GPUs and NPUs on-chip. These processors are designed for AI PCs and support on-device AI functionalities. In parallel, Intel Foundry is preparing advanced manufacturing capabilities, including a GAA-based 1.4nm process, to enable the production of high-performance AI semiconductors.

2. 1. 2 Fabless Firms

Fabless firms are design-specialized companies that do not operate their own fabrication facilities. Nevertheless, they have the advantage of being able to commercialize products under their own brand names. This section examines major fabless companies engaged in the design of AI semiconductors.

First, among fabless firms, NVIDIA is the most prominent company in the AI semiconductor sector. NVIDIA is estimated to account for over 90% of the GPU market for data centers. The company currently releases new product generations on an annual basis. The GPU integrated into NVIDIA's Blackwell accelerator is the B200, while the next-generation product, Blackwell Ultra, incorporates the B300 GPU. After designing its GPUs, NVIDIA outsources fabrication to TSMC. High Bandwidth Memory(HBM) is supplied primarily by SK hynix, and advanced packaging is subsequently performed at TSMC to complete the accelerator products. NVIDIA provides not only hardware such as GPUs but also a comprehensive full-stack solution, including equipments, software platforms and services. As a result, customers face high switching costs, and NVIDIA's AI ecosystem has become highly entrenched.

Second, AMD is rapidly emerging as NVIDIA's primary competitor. In an effort to capture market share in the GPU market dominated by NVIDIA, AMD has developed ROCm(Radeon Open Compute), a software platform comparable to NVIDIA's CUDA(Compute Unified Device Architecture). However, due to the strong lock-in effect of NVIDIA's ecosystem, AMD GPUs have been adopted primarily by customers who experience supply constraints in

accessing NVIDIA products. Nevertheless, AMD's market share has shown a continuous upward trend. AMD's recently introduced accelerator, MI350X, is based on the CDNA4 architecture and is manufactured using a 3nm process. It has been reported to offer advantages over NVIDIA's B200 in terms of both performance and price competitiveness.

Third, Qualcomm, widely known for its Snapdragon system-on-chip(SoC) platforms, is another representative fabless company. Qualcomm integrates its proprietary Hexagon NPU within its SoCs to enable AI functionalities. Snapdragon platforms are deployed across a wide range of applications, including smartphones, AI-enabled laptops, smart glasses, and autonomous vehicles. In particular, Qualcomm has established a strong competitive position in the on-device AI segment.

Fourth, Broadcom is a fabless firm with particular strength in communication networks and customized semiconductor solutions. Broadcom primarily provides ASIC design services tailored to customer-specific requirements. Typically, new circuits are first implemented and validated using Field-Programmable Gate Arrays(FPGAs) to verify functionality and performance, and are subsequently converted into ASICs for final deployment. FPGAs offer high flexibility and reconfigurability [16], and while they provide strong performance, their power consumption is relatively moderate [17]. However, ASICs significantly outperform FPGAs in terms of both performance and power efficiency.

Many Big Tech companies outsource the design and development of customized ASICs to Broadcom in order to optimize chips for their own end products. Recently, a notable trend has emerged in data center inference workloads, where there is increasing interest in replacing NVIDIA GPUs with NPUs or ASICs.

Finally, MediaTek, the largest fabless semiconductor company in Asia, represents another important player. MediaTek develops application processors(SoCs) for edge devices such as smartphones and Internet of Things(IoT) devices and integrates NPUs within these chips to support AI functionalities.

2. 1. 3 Big Tech Firms

The development of AI semiconductors has also become increasingly active among Big Tech firms. By developing in-house chips, these companies can reduce their dependence on external supply chains while simultaneously optimizing semiconductor performance for their own products and services. Moreover, in-house semiconductor development is widely recognized as offering significant cost advantages. For these reasons, the trend of semiconductor development by Big Tech firms is expected to continue.

This section reviews the current status of AI semiconductor development among major Big Tech companies.

First, Apple has pursued an aggressive strategy of internalizing semiconductor development to design nearly all chips required for its products. Apple's AI semiconductor technology is widely recognized through its Neural Engine, which represents a type of neural processing unit(NPU). The Neural Engine is embedded in both the A-series and M-series processors and supports on-device AI functionalities.

Second, Google has developed its own AI accelerator known as the Tensor Processing Unit(TPU), which is also a form of NPU. TPUs are utilized across both training and inference workloads within Google's data center infrastructure.

Third, Microsoft has introduced an AI chip known as Maia, which is designed to support both training and inference tasks. This chip is deployed within Microsoft's cloud infrastructure, particularly in services such as Azure and Copilot.

Fourth, Amazon, the world's leading cloud service provider, has developed proprietary AI chips for use within its cloud ecosystem. Amazon has introduced Trainium for training workloads and Inferentia for inference tasks, and these chips are actively utilized across its cloud services.

Fifth, Tesla has developed custom AI semiconductors for autonomous driving applications. The company's Full Self-Driving(FSD) chip, known as the A15, was manufactured by TSMC and was deployed in Tesla's autonomous driving systems. In addition, Tesla previously developed a dedicated AI supercomputer called Dojo to efficiently train large-scale data collected from its vehicles. The Dojo system utilized the D1 chip and was originally intended to replace NVIDIA's A100 accelerators. However, Tesla has recently decided to discontinue the Dojo project, and development of the D2 chip has also been reportedly halted.

Finally, Meta has developed its own AI accelerators while also operating the open-source large language model (LLM) known as LLaMA. Although Meta continues to procure a substantial number of GPUs from NVIDIA, it has introduced the Meta Training and Inference Accelerator (MTIA), which, as its name suggests, is designed to support both training and inference workloads.

In addition to U.S.-based Big Tech firms, several Chinese companies are also actively developing AI semiconductors. Huawei has developed the Ascend series for data center applications and integrates NPUs into its Kirin chips for smartphones. Similarly, companies such as Alibaba and Baidu have introduced their own AI semiconductors and are deploying them within their respective product and service ecosystems.

(Table 1) Classification of AI Semiconductor Firms and their Characteristics

Category	IDM Firms	Fabless Firms	Big tech Firms
AI Semiconductor Firms	Samsung Electronics, Intel	NVIDIA, AMD, Qualcomm, Broadcom, MediaTek	Apple, Google, MS, Amazon, Tesla, Meta, and other Chinese companies
Types of AI semiconductor	Samsung Electronics: PIM, Mach 1 Intel: Gaudi, NPU	NVIDIA: GPU AMD: GPU Qualcomm: NPU Broadcom: ASIC MediaTek: NPU	Apple: Neural Engine Google: TPU MS: Maia Amazon: Trainium, Inferentia Tesla: A15, D1 Meta: MTIA Other Chinese companies: Various AI semiconductors
AI semiconductor manufacturing	Capable of in-house manufacturing through proprietary foundries	Mainly in TSMC	Mostly manufactured by TSMC, and Chinese firms mainly rely on SMIC

2. 2 Foundry Firm

2. 2. 1 Definition and Types of Foundries

In the semiconductor industry, a foundry refers to a firm that manufactures semiconductor chips designed by customers. Foundry companies do not commercialize chips under their own brand, whereas customers retain ownership of their branded products.

Historically, semiconductor businesses were typically vertically integrated, with design, manufacturing, and packaging conducted within a single firm. However, as manufacturing costs in advanced process nodes have increased substantially, it has become increasingly difficult for a single firm to maintain full integration across the entire value chain. Moreover, foundry operations require extremely large capital investments, resulting in very high entry barriers. A single fabrication facility typically contains approximately 300–400 pieces of equipment [18]. In contrast, fabless firms can operate with relatively lower capital requirements, which leads to comparatively lower entry barriers.

Within the semiconductor ecosystem, foundries function as central actors in technological innovation and maintain a strong interdependent relationship with fabless firms. Modern foundries are no longer limited to passively manufacturing customer designs; rather, they are increasingly expected to provide more proactive support, including suggesting improved design methodologies from the perspective of process optimization. Likewise, customers are heavily dependent on foundries for achieving high-performance products and must therefore consider the strategic position of foundry partners. In practice, large-scale foundries are often unable to accommodate all requests from smaller customers due to capacity and prioritization constraints.

Although foundries specialize in manufacturing, they cannot operate independently and must collaborate closely with a wide range of ecosystem partners. These include Electronic Design Automation(EDA) providers, Intellectual Property(IP) vendors, design houses, and Outsourced Semiconductor Assembly and Test(OSAT) firms. Strong partnerships with these stakeholders are essential for foundries to deliver optimized semiconductor solutions that meet customer requirements.

Foundries can generally be classified into two types: pure-play foundries and IDM-based foundries. Pure-play foundries focus exclusively on foundry services and do not engage in semiconductor design or product development. Representative examples include TSMC, UMC, and GlobalFoundries. In contrast, IDM-based foundries simultaneously operate foundry services while also engaging in semiconductor design activities. Intel and Samsung Electronics are prominent examples of this model.

The primary advantage of pure-play foundries is the absence of conflicts of interest with customers, allowing them to focus entirely on customer needs. By contrast, IDM-based foundries may raise concerns among customers regarding potential risks related to intellectual property protection, as customers may perceive the foundry as a potential competitor. Nevertheless, IDM-based foundries also possess distinct advantages, such as deeper integration between

design and manufacturing capabilities. Table 2 presents a comparison between pure-play foundries and IDM-based foundries.

(Table 2) Comparison between Pure-Play Foundries and IDM-Based Foundries

Category	Pure-play foundries	IDM-based foundries
Definition	A company that engages exclusively in manufacturing operations	a company that covers the entire process from design and manufacturing to packaging
Representative companies	TSMC, UMC, GlobalFoundries	Samsung Electronics, Intel
Advantages	-No conflict of interest with customers -Ability to collaborate with a wide range of customers -Customized services tailored to customer requirements	-Integrated management from design to manufacturing -Comprehensive understanding of technology and products -Ability to leverage internal design assets
Disadvantages	-No proprietary branded products -Limited internal design capabilities -High dependence on customer demand	-Potential competition with customers (Overlapping design domains) -Trust issues may arise when working with certain customers
Customer perspective	High reliability	-Concerns regarding leakage of proprietary design assets -Customers may hesitate to engage due to competitive relationships

2. 2. 2 The Current Status of Foundry Firms

Over the past several decades, semiconductor manufacturing has driven an increasingly significant portion of the global economy [19][20]. Semiconductor manufacturers have continuously pursued improvements in process performance [21]. Foundry companies are primarily located in the United States, China, Taiwan, Japan, and South Korea. However, most foundries currently operate at process nodes of 10nm and above, and only a limited number of firms possess the capability to manufacture advanced AI semiconductors.

Foundry firms can generally be classified into two broad categories. The first group consists of mature-node foundries that primarily operate at process nodes of 10nm and above. The second group includes advanced-node foundries capable of manufacturing at process nodes of 7nm and below. In particular, extreme ultraviolet(EUV) lithography equipment is considered almost essential for processes below 7nm, making it difficult for firms with limited financial resources to enter the advanced foundry business.

Representative firms in the former category include UMC, GlobalFoundries, PSMC, Hua Hong Semiconductor, and DB HiTek. The latter category includes TSMC, Samsung Foundry, Intel Foundry Services, and Rapidus. Notably, China's Semiconductor Manufacturing International Corporation (SMIC) has been unable to import EUV equipment due to U.S. export restrictions. As a result, SMIC has reportedly achieved 7nm production using deep ultraviolet(DUV) lithography through multi-patterning techniques rather than EUV-based processes.

As described above, foundry firms can be broadly divided into mature-node and advanced-node categories. While some mature-node foundries are attempting to enter advanced process technologies, the entry barriers remain extremely high due to the substantial capital investment and technological capabilities required.

In the current AI-driven era, advanced-node foundries have become increasingly prominent. AI semiconductors are predominantly manufactured using 4nm and 5nm process technologies, which has drawn significant attention to TSMC and Samsung Foundry. Other foundries currently lack sufficient capability to manufacture AI semiconductors at advanced nodes. Although Intel Foundry Services and Rapidus possess the technological potential to enter advanced AI semiconductor manufacturing, their capabilities have not yet been fully validated.

Under these circumstances, demand for advanced-node capacity currently exceeds supply. TSMC has secured the majority of AI semiconductor orders, resulting in near-full utilization of its fabrication facilities. In contrast, Samsung Foundry has thus far obtained relatively fewer customer orders. However, as demand for AI semiconductors is expected to increase sharply, Samsung Foundry may also benefit from expanding opportunities in the future. A comparative analysis of TSMC and Samsung Foundry in AI semiconductor manufacturing is presented in Table 3.

(Table 3) Comparison between TSMC and Samsung Foundry in AI Semiconductor Manufacturing

Category	TSMC	Samsung Foundry
Business Model	Pure-play foundry	IDM-based foundry
Market share (2025)	-Over 70% of the global foundry market -Over 90% share in AI semiconductor manufacturing	-Approximately 10% of the global foundry market -Relatively low share in AI semiconductor orders
Main customers	NVIDIA, AMD, Apple, Google, Amazon, Meta etc.	Tesla(FSD chip), Qualcomm(Snapdragon), Samsung System LSI
AI Semiconductor Manufacturing Process	-Based on 5nm, 4nm, and 3nm processes -2nm GAA mass production in progress -High yield stability	-4nm and 3nm processes in mass production -First to adopt 3nm GAA (Yield instability reported) -2nm GAA mass production in progress
Advanced Packaging Capabilities	-World-leading technologies including CoWoS(2.5D), InFO, and SoIC(3D) -Operates six dedicated packaging fabs, with plans to expand to twelve by 2030 -Optimized for HBM–GPU integration	I-Cube and X-Cube(2.5D/3D) packaging technologies -Increasing internalization of advanced packaging capabilities -Relatively lower market credibility compared to TSMC
Yield performance	Very high (Industry-leading) →ensures stable supply	Relatively lower yields →considered a limiting factor for customer acquisition
Strength	-High customer trust -Stable yield and process management -Central position in the global AI chip ecosystem	Ability to offer HBM and foundry services simultaneously (One-stop AI semiconductor solution) →strategic partnerships with selected customers such as Tesla
Weakness	-No proprietary end products(high dependence on customers) -Geopolitical risk(Taiwan)	-Insufficient yield stability and customer trust -The company's image was hit by issues related to 3nm GAA

2. 3 Requirements for Foundries to Enable High-Performance AI Semiconductors

Foundry firms must fundamentally understand customer requirements in order to secure manufacturing orders. In addition, activities such as marketing and customer service are necessary to attract and retain customers. However, from a technological perspective, particularly in the production of high-performance chips such as AI semiconductors, the most critical factors are yield performance, process capability, and advanced packaging technologies.

High yield performance is essential to ensure timely delivery of products to customers. Advanced process capabilities are necessary to manufacture high-performance AI semiconductors, while advanced packaging technologies play a crucial role in achieving superior PPA(Performance, Power, and Area) characteristics.

Accordingly, this study examines foundry capabilities from these three dimensions: yield performance, process technology, and advanced packaging capability.

2. 3. 1 Yield Performance

Yield is a critical factor in foundry operations. As the design complexity of chips increase particularly in the case of advanced AI semiconductors, yield typically decreases. Consequently, customers inevitably prefer foundries with high yield performance.

This section examines the importance of yield in AI semiconductor manufacturing from three perspectives.

First, insufficient yield directly undermines economic viability. Yield refers to the proportion of functional(good) chips produced during the manufacturing process. If yield performance is low, the defect rate increases while the proportion of usable chips decreases, which can lead to significant financial losses. Under such conditions, a foundry cannot sustain its business operations.

Second, low yield makes it difficult to meet customers' time-to-market requirements. When yield is poor, products cannot be delivered on schedule, which in turn prevents customers from launching their products in a timely manner. This issue has been identified as one of the factors that previously limited Samsung Foundry's ability to secure large-scale orders.

Third, yield performance is closely associated with product reliability. Poor yield suggests the presence of potential problems in the manufacturing process. From the customer's perspective, even chips classified as functional may be perceived as carrying a higher risk of latent defects, thereby reducing trust in the foundry.

Given the critical importance of yield, foundries must prioritize continuous efforts to improve yield performance. Factors that negatively affect yield can generally be categorized into three sources.

First, yield issues may arise during the chip design stage. In the complex process of designing high-performance chips, design errors can occur.

Second, yield degradation may occur during manufacturing processes. For example, challenges in etching, lithography, or deposition processes can significantly affect yield when producing advanced AI semiconductors.

Finally, environmental factors within the fabrication facility can also impact yield. Inadequate control of temperature, humidity, or airborne particles inside the fab can lead to defects in semiconductor devices.

Overall, strong yield performance is a prerequisite for customer trust and is a fundamental requirement for foundries seeking to participate in advanced AI semiconductor manufacturing.

2. 3. 2 Process Capability

Process capability refers to a foundry's ability to deliver chips that meet customer-required performance specifications at competitive cost and within a short time frame. To satisfy these conditions, process capability constitutes a critical component of a foundry's overall competitiveness. Superior process capability not only enhances customer satisfaction but also enables substantial reductions in manufacturing costs, thereby strengthening a firm's competitive position.

Most AI semiconductors are manufactured using advanced process nodes at 7nm and below. Accordingly, a foundry's ability to execute advanced scaling technologies has become a key determinant of competitiveness. As foundries are specialized manufacturing enterprises, the effectiveness with which process technologies are managed and maintained plays a decisive role in operational success. Through appropriate process control and management, foundries must be able to prevent potential issues that may arise during manufacturing in advance.

At present, TSMC reportedly secures more than 90% of AI semiconductor orders, and one of the primary reasons for strong customer preference is its superior process capability. In particular, foundries must be able to respond proactively to customer requirements and provide optimized process solutions in order to remain competitive. Effective customer-oriented process optimization contributes directly to higher levels of customer satisfaction. For example, customers differ in their requirements regarding power consumption and performance, and foundries must tailor their process solutions accordingly.

Furthermore, the stability of process control significantly affects key product attributes such as performance, power efficiency, and thermal characteristics. The ability to introduce new process technologies into production at an early and stable stage is also a crucial factor in attracting and retaining customers. In addition, through process optimization, foundries can minimize material and component usage and improve equipment utilization efficiency, thereby substantially reducing manufacturing costs.

Ultimately, customers select foundry partners based on their ability to supply high-quality chips at competitive cost and within required delivery timelines.

2. 3. 3 Packaging Capability

Historically, packaging processes were not regarded as a critical component in the semiconductor industry. Improvements in semiconductor performance were achieved primarily through front-end processes, such as lithography and device scaling, rather than through back-end packaging processes. Packaging was mainly responsible for protecting chips, dissipating heat, and providing electrical connections between the chip and the substrate.

In recent years, however, packaging has become a strategically important area for foundry firms. As further process scaling has become increasingly difficult and manufacturing costs have continued to rise, advanced packaging has emerged as an important alternative for improving chip performance while controlling costs. In some cases, appropriate packaging solutions can enhance chip performance more effectively than advancing to the next process node. Accordingly, leading foundries such as Samsung Electronics and TSMC have shown a clear trend toward internalizing advanced packaging capabilities. In particular, the ability to provide one-stop services that integrate both process technology and packaging has become a key advantage in securing large-scale customer orders.

With the full-scale transition into the AI era, growing demand for high-performance chips has accelerated the development of diverse packaging technologies. Representative examples include 2.5D and 3D packaging, which involve stacking and integrating heterogeneous chips in close proximity. One of the most effective approaches to improving performance at the packaging level is to reduce the physical distance between chips. Given that data transfer between processors and memory remains essential under the von Neumann architecture, shortening interconnect distances through 2.5D and 3D integration provides a significant performance advantage. Currently, most AI accelerators primarily adopt 2.5D packaging; however, a gradual transition toward 3D packaging is widely anticipated.

Other advanced packaging approaches that have gained increasing attention include chiplet architectures, which improve performance by modularly stacking and interconnecting multiple dies, and hybrid bonding technologies, which enable direct vertical interconnection between chips without the use of bumps. Furthermore, advanced packaging allows for optimized interconnect structures, which can reduce latency and enhance overall efficiency by optimizing power delivery paths.

Recently, major foundries such as Samsung Electronics, TSMC, and

Intel have expanded their packaging fabrication facilities. This expansion is largely driven by the rapid increase in demand for AI chips used in AI servers and high-performance computing systems. As a result, advanced packaging capability has become one of the most critical determinants of competitiveness in the AI semiconductor era.

Looking ahead, foundry firms are expected to continue investing in research and development of novel packaging technologies to further enhance chip performance.

2. 4. **Case Analysis of TSMC**

TSMC continues to lead the global foundry market in AI semiconductor manufacturing. The company manufactures not only NVIDIA's AI accelerators but also the vast majority of AI semiconductors, effectively holding a dominant position in this segment. In contrast, the market shares of Samsung Foundry and Intel Foundry Services in AI semiconductor manufacturing remain relatively limited. Therefore, selecting TSMC as the sole subject of analysis is considered appropriate for this study.

In particular, revenue from the high-performance computing segment, where AI semiconductors are extensively utilized, reportedly accounts for more than 60% of TSMC's total revenue. In response to rapidly increasing demand, TSMC has been transitioning portions of its legacy 6-inch wafer capacity toward advanced 12-inch wafer production lines. These strategic shifts further reflect TSMC's strong focus on AI semiconductor manufacturing.

Given that TSMC represents the most prominent foundry in the production of AI semiconductors, this study adopts TSMC as the primary case for analysis. Specifically, the study examines how TSMC enables high-performance AI semiconductor manufacturing through its capabilities in yield performance, process technology, and advanced packaging.

2. 4. 1 Yield Performance Perspective

Yield is a critical factor in the production of high-performance AI semiconductors. High yield not only ensures timely delivery of products to customers but also enables cost reduction through increased production volume. Compared with competing foundries, TSMC consistently achieves superior yield performance. One of the primary reasons is that TSMC has accumulated extensive process know-how by focusing almost exclusively on the foundry business for nearly four decades.

In addition, TSMC hired more than 10,000 employees in 2024 and reportedly plans to maintain a similar level of recruitment in 2026. The company provides an organizational environment that supports engineers in enhancing yield through continuous technological innovation.

Beyond these structural advantages, TSMC has implemented a range of concrete strategies to sustain high yield performance in AI semiconductor manufacturing. This study examines how TSMC has maintained superior yield performance through systematic technological and operational practices.

First, TSMC adopts rigorous end-to-end integration across the design, manufacturing, packaging, and mass production stages. To this end, it operates integrated frameworks such as Design for Manufacturing(DFM) and Design-Technology Co-Optimization(DTCO).

DFM supports customers in ensuring that their chip designs achieve optimal performance and yield on TSMC's process platforms. For this purpose, TSMC provides pre-verification and simulation services.

DTCO focuses on improving yield by jointly optimizing design and process technologies. Through DTCO, TSMC adjusts transistor layout, structure, and pattern configurations according to customer design characteristics.

TSMC also maintains one of the lowest levels of defect density(D_0) in the industry prior to wafer processing, recognizing that poor initial wafer quality inevitably degrades yield. Its efforts to reduce defect density can be categorized into three areas: process equipment, materials & components, and process control & design methodologies.

With respect to process equipment, TSMC has reduced mask-related defects through the adoption of high-NA EUV lithography systems and strengthened uniformity control using advanced etching and deposition equipment. Equipment upgrades have also enabled automated defect inspection at each process step, while big-data-based analytics are used to trace root causes of defects.

Regarding materials and components, TSMC has secured epitaxial growth technologies that enable the uniform formation of multilayer nanosheet channels for GAA-based devices. In addition, the company utilizes high-purity materials and applies low-K dielectrics to reduce electrical defect rates.

In terms of process control and design methodologies, major customers such as Apple and NVIDIA incorporate process-friendly patterns through DFM at the design stage and conduct pre-simulations to identify potential hotspot regions with high defect probability. Furthermore, TSMC detects yield degradation factors at an early stage and tunes equipment using AI-based defect prediction models. The company also enhances yield by connecting distributed fabs into clusters and sharing operational data across facilities. Through these efforts, TSMC reportedly achieved an SRAM yield of approximately 90% at the pilot stage of its 2nm process [22], while overall logic chip yield has exceeded 60%.

Once wafer processing begins, TSMC employs an advanced yield management system that collects and analyzes massive volumes of production data in real time. This enables the prediction of potential defects and identification of root causes. During wafer fabrication, various process deviation data are continuously monitored, and AI-based models detect patterns and immediately correct process errors. Moreover, millions of process parameters, including equipment temperature and pressure, are statistically controlled to minimize variability.

TSMC also applies repair processes that correct defects in cells and interconnects to convert defective chips into functional products. In addition, the company conducts pilot production runs on small-scale lines prior to mass production in order to identify and resolve potential issues in advance. When identical processes are operated across multiple fabs, TSMC synchronizes process conditions and equipment parameters to prevent yield deviations between facilities.

2. 4. 2 Process Capability Perspective

TSMC's process capability is widely regarded as among the highest in the global semiconductor industry. The company is known for achieving relatively high initial mass-production yields and for transitioning from process development to volume manufacturing more rapidly than its competitors. TSMC commercializes next-generation process nodes approximately every two years, representing the shortest cycle in the industry. In practice, TSMC began mass production of its 7nm process in 2018, 5nm in 2020, and 3nm in 2022. Moreover, TSMC's process technologies are generally operated with a high level of stability. This is largely because the company prioritizes yield improvement and process maturity over the early adoption of highly uncertain next-generation technologies.

For example, in 2022, TSMC continued to adopt the FinFET architecture for its 3nm process, citing uncertainties associated with Gate-All-Around(GAA) technology, thereby achieving higher yield and greater process stability. In contrast, Samsung Foundry adopted GAA earlier, but reportedly experienced challenges related to yield and process stability, which limited its ability to secure customer orders. As a result, TSMC continues to maintain a competitive advantage in terms of process stability. From the customer's perspective, this stability allows for more predictable delivery schedules.

In addition, TSMC is widely perceived as a highly reliable company, supported by its corporate philosophy of integrity and its strong reputation for technological maturity. TSMC's wafer defect density in its 5nm process has been reported to be below 0.1 defects/cm² [23], which is up to 30% lower than that of Samsung Foundry during the same period. This level of process maturity has enabled TSMC to consistently meet the mass-production schedules of major customers such as Apple and NVIDIA.

Furthermore, TSMC operates a diverse range of process platforms across its fabs, allowing it to optimize manufacturing processes according to customer-specific requirements. The company also maintains advanced control systems capable of rapidly detecting and addressing process defects.

TSMC frequently proposes design improvements to customers from an early stage and incorporates these insights into process optimization. This strong customer-oriented approach is widely recognized as one of the primary reasons why customers continue to select TSMC as their foundry partner.

At the same time, TSMC has deployed more EUV lithography systems from ASML than any other foundry and has achieved earlier stabilization of EUV-based manufacturing. For example, TSMC reportedly applies EUV patterning to approximately 14 layers in its 5nm process and more than 20 layers in its 3nm process, representing up to twice the level of EUV utilization compared with competing firms. This extensive use of EUV has contributed not only to improved pattern precision but also to enhanced power efficiency.

TSMC has also entered mass production of its 2nm process. At this node, transistor performance is expected to improve by up to 15% compared with the 3nm process, while power consumption is projected to decrease by up to 30%. These improvements illustrate TSMC's strong capability in transistor architecture optimization.

Moreover, TSMC collaborates closely with customers from the early

stages of chip design, jointly developing process technologies and maximizing process maturity by the time of tape-out. The company maintains strong partnerships with key equipment suppliers such as ASML, Applied Materials, and Tokyo Electron, as well as with materials and component suppliers across the ecosystem. Supported by high profitability, reportedly exceeding 50% gross margin, TSMC is able to make aggressive early investments in new process lines, enabling faster mass production of next-generation technologies than its competitors. In addition, even within the same process node, TSMC offers customer-specific customization to meet diverse requirements related to power, performance, and other design constraints.

2. 4. 3 Packaging Capability Perspective

As further scaling toward sub-1nm process nodes is increasingly expected to face substantial technical and economic limitations, many foundry firms are placing strategic emphasis on advanced packaging technologies. Among them, TSMC is widely regarded as the most advanced company in the field of semiconductor packaging. Its advanced packaging technologies demonstrate superior performance compared to those of other firms, enabling the simultaneous achievement of low power consumption, high performance, and high integration density. Consequently, TSMC is reported to hold the leading position in the global advanced packaging market.

In response to the rapid growth in demand for AI semiconductors, TSMC anticipated that advanced packaging would become a critical growth area. Accordingly, the company has significantly increased its investment in research and development and has expanded its advanced packaging capacity through the construction of new fabrication facilities. These strategic investments have enabled TSMC to strengthen its competitiveness in advanced packaging and to secure a large number of AI semiconductor customers.

TSMC is reportedly operating more than six dedicated packaging fabrication facilities, of which three are advanced packaging fabs, as summarized in Table 4. In addition, TSMC is currently constructing a total of six additional packaging facilities across Taiwan and the United States. As a result, the number of TSMC's packaging fabs is expected to exceed twelve by 2030.

(Table 4) Major Advanced Packaging Fabs of TSMC(2025)

Category	Location	Key Technologies	Remarks
Advanced Packaging	Hsinchu (新竹)	CoWoS, InFO, SoIC	Focused on HPC and AI applications
Advanced Packaging	Taichung (台中)	CoWoS	Primarily for AI chip production
Advanced Packaging	Tainan (台南)	InFO	Mobile and HPC applications

TSMC has integrated advanced packaging technologies for AI semiconductors such as CoWoS(Chip-on-Wafer-on-Substrate), InFO(Integrated Fan-Out), and SoIC(System on Integrated Chips) into its manufacturing ecosystem. Through these technologies, TSMC optimizes performance and power efficiency while offering services at relatively competitive cost. This integrated capability has enabled TSMC to secure a dominant position in the AI semiconductor market.

The current status of TSMC's major advanced packaging technologies can be summarized as follows.

First, with respect to CoWoS packaging, TSMC's production capacity is projected to exceed 130,000 wafers per month by 2026 [24], which represents a substantially faster expansion pace than that of competing firms. TSMC currently performs CoWoS packaging by integrating NVIDIA's GPUs(e.g., B100 and B200) with SK hynix's HBM3E. TSMC's market share in this segment is reported to exceed 90%.

Second, InFO packaging is widely applied to mobile application processors(APs) in smartphones. Apple is currently the primary customer for InFO packaging, which is used in its A-series and M-series chips. InFO adopts a fan-out architecture that increases I/O density and enables thinner substrates, thereby supporting ultra-lightweight and low-power packaging solutions.

Third, SoIC is a three-dimensional(3D) packaging technology that vertically stacks logic dies and utilizes through-silicon vias(TSVs). Major companies such as AMD, NVIDIA, and Apple are reportedly evaluating SoIC for next-generation high-performance computing applications. Furthermore, the adoption of hybrid bonding allows for direct chip-to-chip interconnection without micro-bumps, which contributes to improvements in both power efficiency and processing speed.

In addition, TSMC has internalized not only packaging processes but also testing processes, thereby enhancing overall customer

satisfaction. The company considers packaging strategies from the early stages of chip design and provides customers with optimized packaging solutions tailored to their requirements. Moreover, because TSMC integrates both front-end(process) and back-end(packaging and testing) operations within its ecosystem, it is able to provide rapid feedback and issue resolution when problems arise during development or production.

3. Conclusion

This study examined a wide range of AI semiconductor firms and analyzed the current landscape of foundry companies. Based on this analysis, the study identified yield performance, process capability, and advanced packaging capability as the most critical technological factors in advanced foundry operations. Furthermore, through an in-depth case study of TSMC, this research investigated how a leading foundry enables the manufacturing of high-performance AI semiconductors from the perspectives of yield, process technology, and packaging.

The findings indicate that TSMC secures a competitive advantage in AI semiconductor manufacturing by implementing rigorous end-to-end integration from design and fabrication to packaging and mass production, maintaining highly stable process capability, and making proactive and substantial R&D investments in advanced packaging technologies. These combined capabilities enable TSMC to sustain superior competitiveness in the production of high-performance AI semiconductors.

This study is meaningful in that it approaches AI semiconductor research from the perspective of foundries, which has been relatively underexplored compared to design-focused studies. One key implication of the findings is that AI semiconductor firms should place greater strategic emphasis on foundry selection, as the performance and competitiveness of AI chips can be significantly influenced by the technological capabilities of foundry partners.

This study adopts a single-case research design and focuses on the role of foundries in terms of yield, process capability, and packaging capability. Therefore, its findings are subject to limitations in terms of generalizability. Future research may extend this work by examining additional critical factors in AI semiconductor manufacturing and conducting comparative analyses across multiple foundry firms to provide broader empirical insights.

Reference

[1] Er-Ping Li et al., “An Electromagnetic Perspective of Artificial Intelligence Neuromorphic Chips,” Electromagnetic Science, Vol. 1, No. 3, Sep. 2023

[2] Heyang Xu, “FPGA: The super chip in the age of artificial intelligence,” Journal of Physics: Conference Series, doi:10.1088/1742-6596/2649/1/012018, 2023

[3] Dhruvitkumar V. Talati, “Silicon minds: The rise of AI-powered chips,” International Journal of Science and Research Archive, Vol. 01, No. 02, pp. 097-108, 2021.

[4] Sorna Mugi Viswanathan, “AI Chips: New Semiconductor Era,” International Journal of Advanced Research in Science, Engineering, Technology, Vol. 7, No. 8, pp. 14687-14694, Aug. 2020

[5] Hyunji Kim, Seyoung Yoon, Hwajeong Seo, “Research Trends in Domestic and International AI Chips,” Smart Media Journal, Vol. 13, No. 3, pp. 36–44, Mar. 2024.

[6] Hiroshi Momose, Tatsuya Kaneko, and Tetsuya Asai, “Systems and circuits for AI chips and their trends,” Japanese Journal of Applied Physics 59, 2020

[7] P. Ebby Darney, “A Review on Artificial Intelligence Chip,” Recent Research Reviews Journal, Vol. 1, No. 1, pp. 99-109, Dec. 2022

[8] Prashis Raghuwanshi, “Revolutionizing Semiconductor Design and Manufacturing with AI,” Journal of Knowledge Learning and Science Technology, Vol. 3, No. 3, Sep. 2024

[9] Shaojun Wei et al., “Reconfigurability, Why It Matters in AI Tasks Processing: A Survey of Reconfigurable AI Chips” IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I, Vol. 70, No. 3, pp. 1228- 1241, Mar 2023

[10] Siwoo Eum, et al., “Development Trends in Artificial Intelligence Semiconductors,” Journal of the Korea Institute of Information Security and Cryptology, Vol. 35, No. 3, pp. 63–71, Jun. 2025.

[11] Robert Huggins et al., “Competition, open innovation, and growth challenges in the semiconductor industry: the case of Europe’s clusters,” Science and Public Policy, 50, pp. 531–547, Mar. 2023.

[12] Wenqiang Zhang et al., “Neuro-inspired computing chips,” Nature Electronics, Vol. 3, pp. 371–382, Jul. 2020

[13] S. Ambrogio et al., “An analog-AI chip for energy-efficient speech recognition and transcription,” Nature, Vol. 620, 24 Aug. 2023

[14] Q. Zhang, H. Deng, and K. Song, “Latest VLSI Techniques for

3nm Technology for Building Efficient AI Chips," "Fusion of Multidisciplinary Research, An International Journal (FMR)," Vol. 5, No. 2, 2024

[15] https://economist.co.kr/article/view/ecn202507310045 (accessed Aug. 24, 2025)

[16] Wei Gao, and Pingqiang Zhou, "Customized High Performance and Energy Efficient Communication Networks for AI Chips," IEEE, Vol. 7, 2019

[17] Lennart P. L. Landsmeer et al., "Tricking AI chips into simulating the human brain: A detailed performance analysis," Neurocomputing, Vol. 598 Sep. 2024

[18] Yon-Chun Chou and I.-Hsuan Hong, "A Methodology for Product Mix Planning in Semiconductor Foundry Manufacturing," IEEE Transactions on Semiconductor Manufacturing, Vol. 13, No. 3, pp. 278-285, Aug. 2000

[19] Wojciech T. Osowiecki et al., "Achieving Sustainability in the Semiconductor Industry: The Impact of Simulation and AI," IEEE Transactions on Semiconductor Manufacturing, Vol. 37, No. 4, pp. 464-474, Nov. 2024

[20] S. Fore and C.T. Mbohwa, "Cleaner production for environmental conscious manufacturing in the foundry industry,

Author Biography

Young-hwa Kwun

B.A. in English Literature, Kwangwoon University, 1994

M.S. (Business Administration), Sungkyunkwan University, 1998

M.A. in Japanese Studies (International Area Studies), Hanyang University, 1999

Ph.D. in Business Administration, Seoul School of Integrated Sciences & Technologies (aSSIST), 2013

Ph.D. Candidate in AI Convergence Engineering, Seoul School of Integrated Sciences & Technologies (aSSIST), 2025–Present

Research Interests: AI semiconductors, edge devices, etc.

E-mail:690730@daum.net

Bo-young Kim

B.A. in Information Design, Ewha Womans University, 1996

M.A. in Information Design, Ewha Womans University, 2000

Ph.D. in Engineering, School of Engineering and Design, Brunel University, 2006

Professor, Graduate School of Business, Seoul School of Integrated Sciences & Technologies (aSSIST), 2008–Present

Research Interests: Media processing, context awareness, etc.

E-mail:bykim2@assist.ac.kr