SEMICONDUCTOR MEMORY AT A TURNING POINT

— THE TECHNOLOGY AND VALUE CHAIN SUPPORTING THE GROWTH OF GENERATIVE AI —

Reina Ogawa
Technology & Innovation Studies Div., Industry Innovation Dept.
Mitsui & Co. Global Strategic Studies Institute

SUMMARY

- The functions of semiconductors can be compared to human office work: logic corresponds to the person, memory to a desk, and storage to a bookshelf. Within DRAM, the most common type of memory, market growth is being driven by high bandwidth memory (HBM) for generative AI, with the HBM market projected to reach USD 44 billion by 2029.
- Memory is considered a bottleneck in AI development. In the next generation of HBM, part of the manufacturing process is expected to be outsourced to logic foundries.
- Changes in the value chain starting with next-generation HBM are expected to expand the range of suppliers for elemental technologies and create new business models.

1. ESSENTIAL FACTS ABOUT MEMORY

Within the semiconductor market, which is projected to reach USD 1 trillion by 2030, memory is the second-largest segment after logic devices (such as CPUs and GPUs) and is expected to experience strong growth. The market size for dynamic random-access memory (DRAM), a representative type of memory, is forecast to grow from USD 96 billion in 2023 to USD 136 billion in 2029[1]. The main driver of this growth will be high bandwidth memory (HBM) for generative AI applications, which is expected to expand from USD 5.4 billion in 2023 to USD 44 billion[2] in 2029—more than an eightfold increase over six years. This section introduces the essential facts about memory.

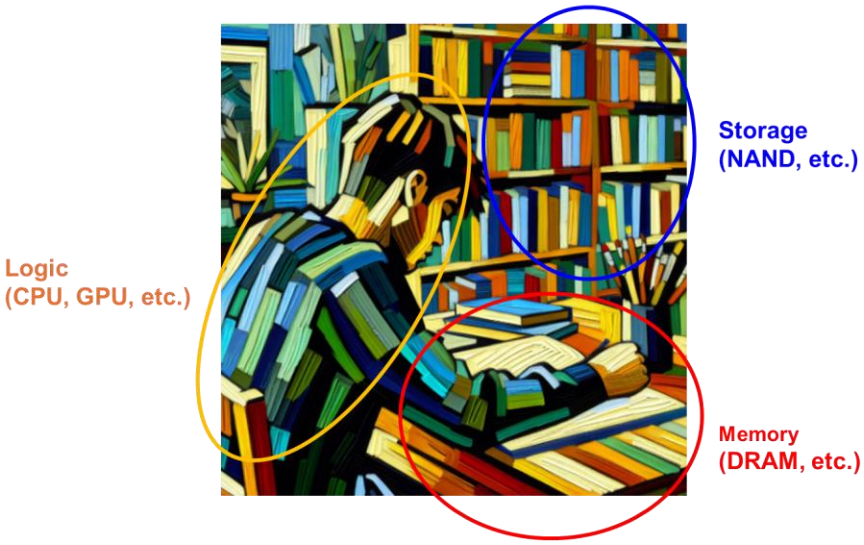
1-1. WHAT IS MEMORY?

Memory is a semiconductor storage device that temporarily stores data in order to supply it to logic components that perform computations in PCs, data centers, and the like. It is also called random-access memory (RAM). If we compare the functions of semiconductors to human work, memory corresponds to a desk where data is arranged for processing, logic to the person performing the work, and storage to a bookshelf for keeping data over the long term (Figure 1).

---

[1] Actual figures for 2023 are taken from the World Semiconductor Trade Statistics (WSTS), and the market forecasts for DRAM and HBM are based on projections by Yole Intelligence.
[2] Source：EE Times Japan, Global Memory Market to Exceed Record USD 220 Billion in 2025, Driven by Generative AI

Figure 1: Conceptual Illustration Comparing the Roles of Semiconductor Devices to Human Work

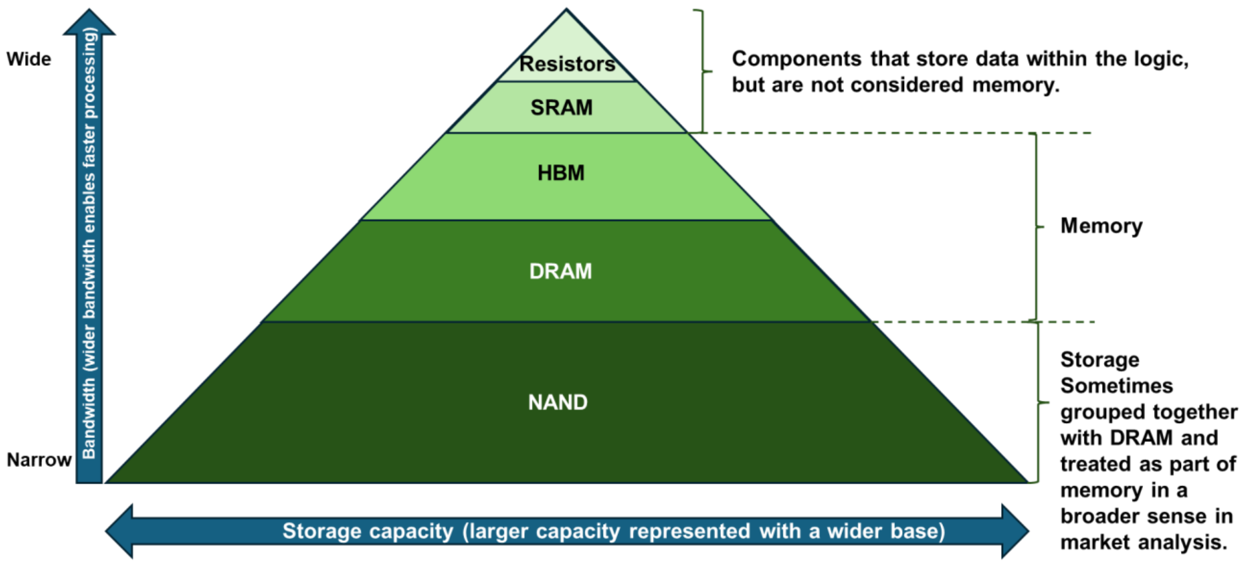


Source: Created by MGSSI with Copilot

1-2. TYPES AND PERFORMANCE OF MEMORY DEVICES

There are various types of memory devices, classified by storage capacity and processing speed (Figure 2). Processing speed is expressed by bandwidth—the wider the bandwidth, the faster the processing speed. Since there is a trade-off between processing speed and storage capacity, memory devices with the optimal combination of storage capacity and bandwidth are selected according to the application. Dynamic random-access memory (DRAM) is commonly used as RAM in personal computers, data centers, and smartphones. In market analysis, NAND (Not AND) is sometimes treated as memory in the broadest sense, but since its processing speed is slow, it is generally used as storage. There are also memory devices such as registers and static random-access memory (SRAM), which temporarily store data for computation by logic. However, these are typically located inside the logic itself and are generally considered separately from memory.

Figure 2: Types and Performance of Memory Devices



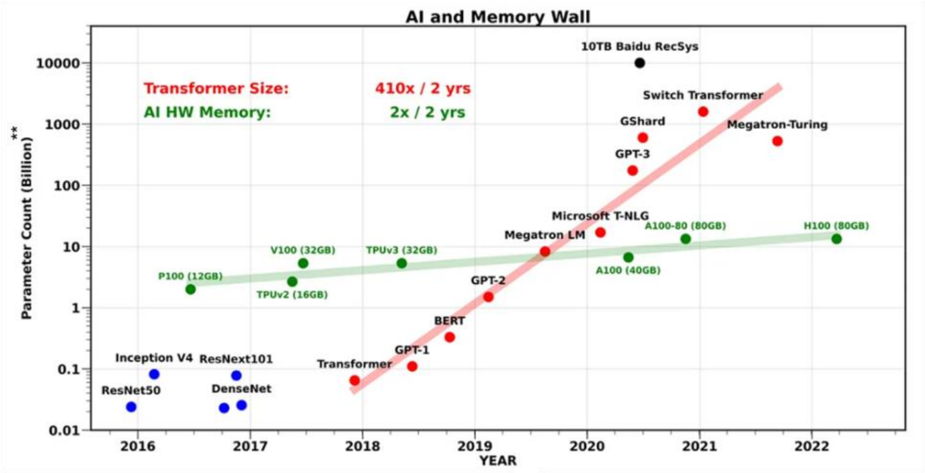
Source: Compiled by MGSSI based on various sources

1-3. HBM DRIVING MEMORY GROWTH FOR GENERATIVE AI

In the field of data processing for generative AI, that has been attracting increasing attention in recent years, engineers are criticizing the processing speed of memory and the communication speed between logic and memory as performance bottlenecks, known as the "memory wall[3]." The main reason for the bottleneck is that improvements in memory bandwidth have lagged significantly behind the increase in "parameters," which correspond to the data processing speed required for generative AI development. By mid-2019, the processing capability of memory installed in a single AI semiconductor was overtaken by the processing capability required by AI models. Since then, the gap has continued to widen (Figure 3).

[3] Medium, "AI and Memory Wall" (Written by Amir Gholami, https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8)

Figure 3: Comparison of Memory Data Processing Capability and the Data Volume Required for Transformers*



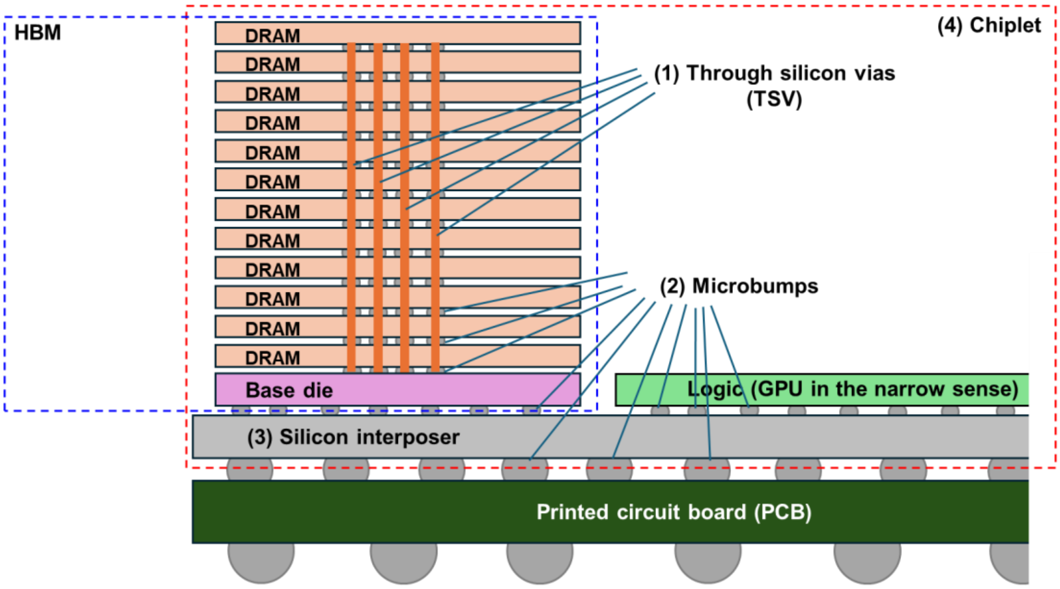
*Transformers are models that learn the context of input information and generate responses in generative AI and other applications.
Parameter Count on the vertical axis represents the amount of information required to implement a transformer. The graph plots it by overlaying **the amount of data that can be processed by the memory installed in a single AI semiconductor (green) with the amount of data required by transformers (red).

Source: Medium, "AI and Memory Wall"
(Written by Amir Gholami, https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8)

High bandwidth memory (HBM) is a type of memory in which DRAM is stacked on a substrate called a base die, which handles control and communication functions, in order to increase memory capacity per unit area. In GPUs[4] from NVIDIA (US), the logic and HBM are further connected via a substrate called a silicon interposer (Figure 4), which reduces the communication time between the logic and memory. The realization and evolution of HBM require advances in elemental technologies spanning a wide range of fields, such as manufacturing processes, materials, and design, as well as their implementation in manufacturing sites. Specifically, technological improvements in (1) through silicon vias (TSV), (2) microbumps, (3) silicon interposers, (4) chiplets, (5) power supply voltage quality stabilization, and (6) thermal management are directly linked to HBM performance (Figures 4 and 5).

[4] GPU stands for Graphic Processing Unit. Originally designed as logic for high-speed image processing, GPUs have come to be widely used as AI semiconductors because they are also capable of accelerating machine learning computations, including generative AI. As semiconductor products, HBM and logic (specifically, GPUs in the narrow sense) are connected via a silicon interposer and sold mounted on a printed circuit board. For this reason, the entire assembly shown in Figure 4 is often referred to collectively as a GPU.

Figure 4: Overview of HBM and NVIDIA's GPU



(5) Power supply voltage quality stabilization and (6) thermal management are elemental technologies that apply to the entire HBM and GPU, so they are not shown in the figure.
Source: Compiled by MGSSI based on various sources

Figure 5: Overview of Elemental Technologies Required for HBM

Elemental Technology	Overview of the Technology	Technical Challenges and Development Trends
(1) Through silicon via (TSV)	Electrode formed by drilling through-holes in silicon chips and filling them with conductive material. Enables stacked chips to be connected over a shorter distance than with wires.	Miniaturization increases the difficulty of through-hole and electrode formation (resist, etching, plating, planarization, etc.).
(2) Microbump	Solder protrusion formed to electrically connect electrodes between semiconductor devices and circuit boards.	The connection distance caused by bumps limits integration density and reduction of power loss, so hybrid bonding that enables connection without bumps is being considered.
(3) Silicon interposer	A silicon substrate with wiring formed on it, onto which multiple chips are mounted.	Alternatives being explored include silicon bridges that use interposers only for connection points, and glass interposers to address warpage issues.
(4) Chiplet*	A semiconductor integrated circuit block designed to exchange information with other chips (chiplets) to form large, complex circuits.	For AI semiconductors, standardization of inter-chip connections for the development of large-scale integrated circuits is advancing, and design tools are maturing.
(5) Power supply quality stabilization	High-speed, large-capacity data transmission by HBM requires large currents. To ensure stable operation, a power supply capable of providing stable voltage even at high currents is necessary.	Power semiconductors made from materials such as gallium nitride (GaN), which has higher breakdown voltage and lower loss than silicon, are being considered for power supply units.
(6) Thermal management	Control of heat inside and outside the device through structural design, power circuits, and monitoring systems to prevent HBM from exceeding its operating temperature.	HBM generates significant heat due to high-current operation and its stacked structure makes heat dissipation more difficult than with single-layer semiconductors.

*Details on chiplet are discussed in the report "Chiplets" in Biz Tech Focus 2025.
https://www.mitsui.com/mgssi/ja/report/detail/__icsFiles/afieldfile/2025/02/07/2501btf_ogawa_ishiguro.pdf

Source: Compiled by MGSSI based on various sources

2. STRUCTURE OF DRAM AND HBM, AND CHANGES IN THE VALUE CHAIN

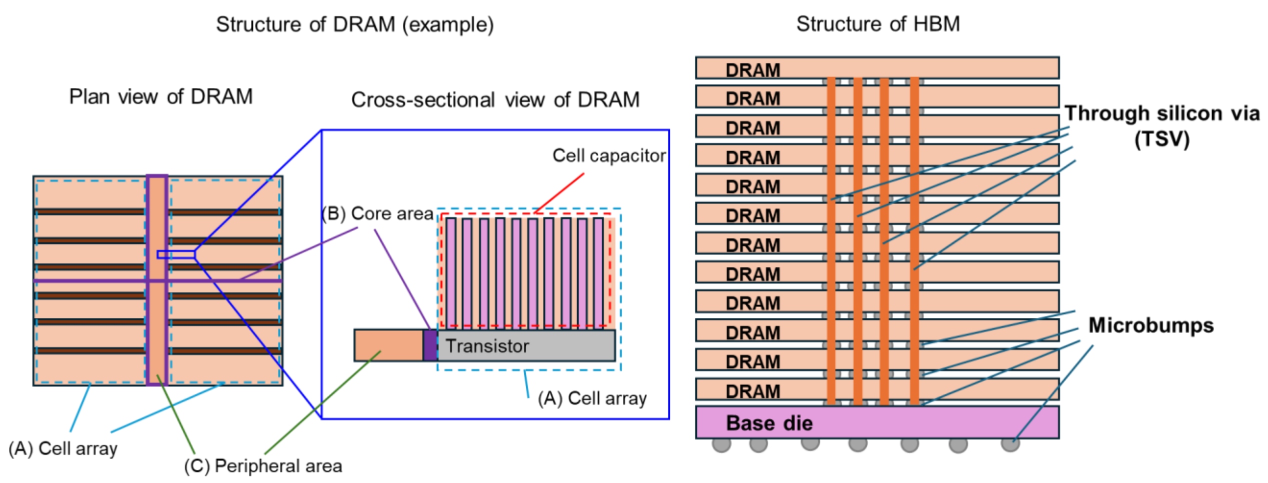
This section explains the structure of DRAM and HBM, as well as changes in the value chain as seen in the HBM development roadmap.

2-1. STRUCTURE OF DRAM AND HBM

A DRAM chip consists of (A) a cell array that stores data, (B) a core area that controls data access, and (C) a peripheral area for communication with external components (Figure 6, left). The cell array is made up of cell capacitors and transistors, and occupies most of the chip area. The miniaturization of DRAM focuses on reducing the size of both the cell capacitors and transistors, which differs significantly from logic, which can solely focus on the miniaturization of transistors.

HBM has a structure in which DRAM is stacked on top of a base die and connected using TSVs and microbumps (Figure 6, right). This structure increases memory capacity per unit area without compromising DRAM processing speed. As a result, HBM has been adopted as memory capable of achieving both the large capacity and high-speed processing required for generative AI applications.

Figure 6: Structures of DRAM and HBM



Source: Compiled by MGSSI based on various sources

2-2. CHANGES IN THE VALUE CHAIN SEEN IN THE HBM DEVELOPMENT ROADMAP

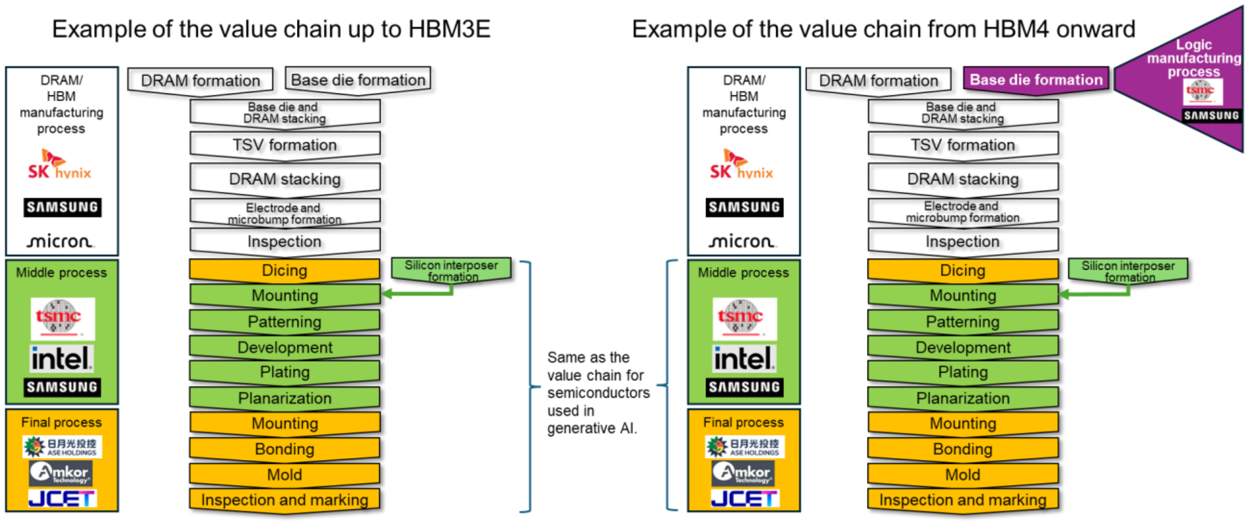
The HBM industry has established a roadmap and is developing products accordingly (Figure 7). Since the release of the first generation in 2016, three generations have been developed as of June 2025, with the latest being HBM3E. The three companies that develop and market HBM are SK Hynix (South Korea), Samsung Electronics (South Korea), and Micron Technology (US), all of which are major DRAM manufacturers. The release of HBM4 is scheduled for late 2025 to 2026. Until now, HBM has been manufactured entirely using DRAM processes, but starting with HBM4, the base die will be produced using logic processes (Figure 8). This process change will double the bandwidth compared to HBM3E.

Figure 7: HBM Development Roadmap

Period	2016-2018	2018-2020	2020-2022	2022-2024	2024-2026	2026-2028	2028-2030
Product	HBM	HBM2	HBM2E	HBM3	**HBM3E**	**HBM4**	HBM4E
Bandwidth	256GB/s	307GB/s	461GB/s	819GB/s	1.0TB/s	2.0TB/s	2.5TB/s
Memory capacity per die	8GB		16GB		24GB		32GB
No. of layers	4 or 8			8 or 12		12 or 16	
Total HBM capacity	4 or 8 GB		8 or 16 GB	16 or 24 GB	24 or 36 GB	36 or 48 GB	48 or 64 GB
Base die manufacturing process	DRAM process					**Logic process**	

Source: Compiled by MGSSI based on ISSCC 2025 Forum 3.4: Memory Solutions for AI Era: High-Bandwidth Memories and Dense-Data Stage

Figure 8: Changes in the HBM Value Chain



Source: Compiled by MGSSI based on various sources

The manufacturing of the HBM4 base die is expected to be handled primarily by TSMC (Taiwan), the world's largest foundry for advanced logic production. It is reported that not only SK Hynix and Micron Technology, which do not have their own logic manufacturing facilities, but also Samsung Electronics, which operates its own foundry, are exploring partnerships with TSMC.

3. DEVELOPMENT TRENDS FOR EVEN FASTER AND MORE POWER-EFFICIENT MEMORY

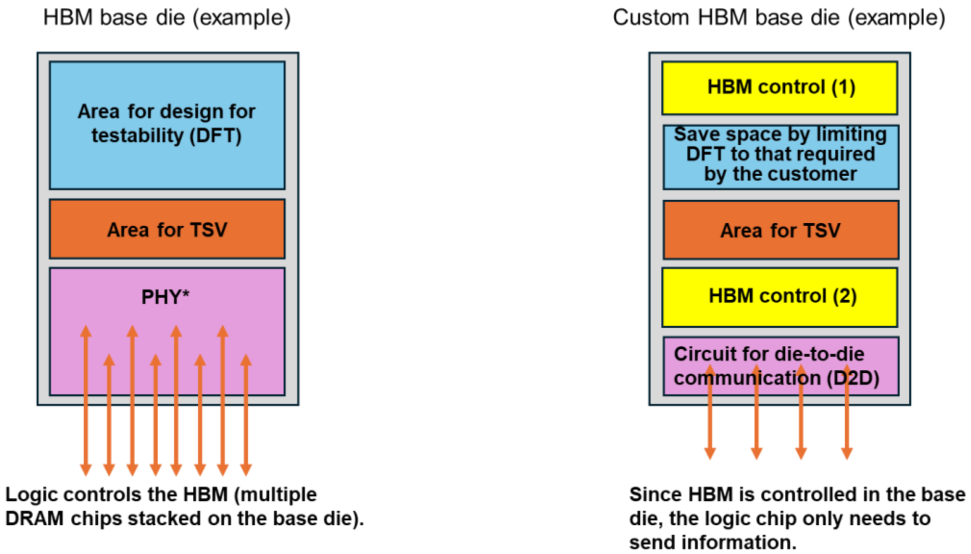
This section explains custom HBM and compute-in-memory, two technologies expected to be implemented in society ahead of the release of HBM4.

3-1. CUSTOM HBM

The outsourcing of HBM4 manufacturing to foundries marks the beginning of horizontal specialization in memory production. Unlike memory manufacturers that mass-produce identical products, foundries such as TSMC offer services that allow different chip designs to be produced for each wafer or even within the same wafer. It is believed that applying this service to base die production will enable the customization of HBM.

In conventional HBM, the same chip is used for all customers, so a large portion of the base die is occupied by circuits for verifying the proper operation of the stacked DRAM, known as Design for Testability (DFT). Due to the limited area of the base die, HBM control has been handled by the logic side. In contrast, with custom HBM, the DFT area can be minimized to only what is necessary for each customer, creating space on the base die to integrate HBM control circuits. This eliminates the need for customers to design HBM control circuits on the logic side. Furthermore, the data transmission previously required to control HBM from the logic side becomes unnecessary, and communication between HBM and the logic die is limited to sending only the data required for computation (Figure 9). In this way, custom HBM reduces the design burden on the logic side and offers the additional benefit of lowering power consumption for data transmission.

Figure 9: Comparison of Base Dies for HBM and Custom HBM



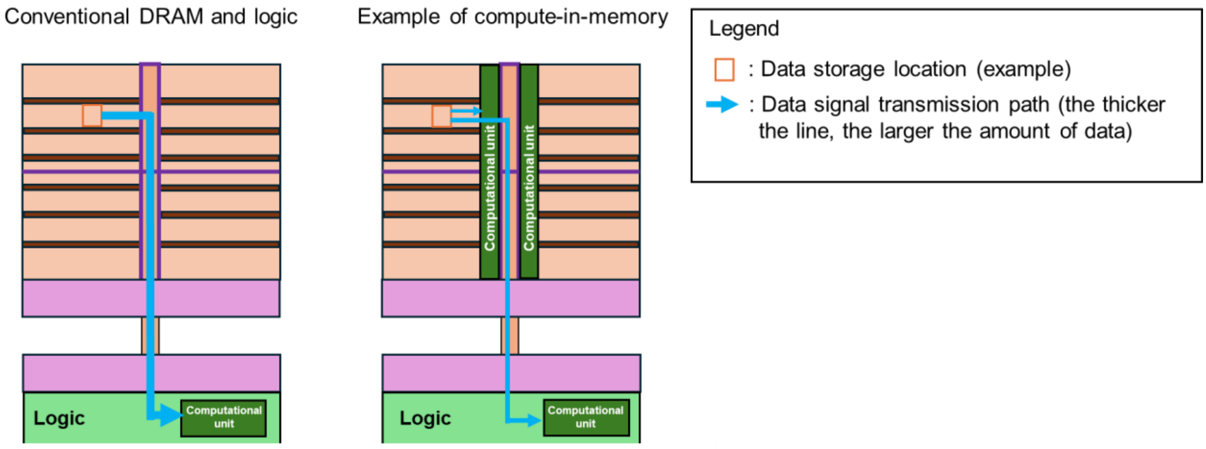
*PHY: A circuit for transmitting information. Also referred to as the physical layer.

Source: Compiled by MGSSI based on various sources

3-2. COMPUTE-IN-MEMORY

Compute-in-memory is also gaining attention as a technology capable of further reducing power consumption in memory for generative AI. The power consumption of semiconductors is proportional to the distance signals travel. Conventional DRAMs transmit data to logic to perform operations, resulting in long signal travel distances and higher power consumption. In contrast, compute-in-memory integrates part of the logic function (computational unit) directly into the DRAM, reducing the amount of data transmission from DRAM to logic and enabling lower power consumption (Figure 10). DRAM-based compute-in-memory technology is being developed and promoted by SK Hynix and Samsung Electronics under the name "processing-in-memory (PIM)." Meanwhile, TSMC and other semiconductor manufacturers than DRAM manufacturers are exploring ways to integrate computational units into memory devices such as magnetoresistive random-access memory (MRAM), which consumes less power than DRAM.

Figure 10: Differences Between Conventional DRAM and Compute-in-Memory



Source: Compiled by MGSSI based on various sources

4. CONCLUSION: NEW BUSINESS OPPORTUNITIES

This report explained the fundamentals of memory, highlighting that technological innovation is needed to overcome memory-related bottlenecks limiting generative AI performance. It also outlined how HBM, which is currently adopted for generative AI, will use logic processes for manufacturing the base die starting with HBM4 scheduled for release between 2025 and 2026, bringing significant changes to the HBM value chain.

Manufacturers with new ideas, as well as functions that connect people and technologies from different industries and technical fields are likely to become increasingly important.