White Paper, sponsored by Micron Technology | January 2025

The Importance of Memory in High-Performance Computing and AI

Jeff Janukowicz

Research Vice President
Solid State Drives and Enabling Technologies

Table of Contents

Use the links below to navigate this document.

IN THIS WHITE PAPER

IDC explores the rise of AI in datacenters and memory's critical role in powering the next generation of AI and computing applications to deliver efficiency, scalability, and sustainability.

SITUATION OVERVIEW

Artificial intelligence is rapidly transforming industries worldwide, with datacenters playing a pivotal role in this evolution. AI is reshaping the business landscape by automating processes, enhancing decision-making, and driving innovation. Through AI, companies can process and analyze vast amounts of data, revealing insights and trends that were previously inaccessible. This enables organizations to optimize operations, personalize customer experiences, and develop new products and services that meet evolving market demands.

For organizations, AI-powered automation not only cuts operational costs but also boosts efficiency by taking over repetitive tasks, allowing employees to focus on strategic initiatives. Furthermore, AI enhances decision-making by providing real-time data analysis and predictive capabilities, enabling businesses to respond swiftly to market changes and maintain a competitive edge. As AI continues to evolve, its impact on businesses is expected to grow, ushering in a new era of innovation and growth. In fact, IDC predicts that AI will generate a cumulative global impact of $19.9 trillion by 2030.

Data, algorithms, and models form the foundation of AI, creating a powerful force that revolutionizes operations and drives the need for datacenters to support AI workloads. As AI models grow more complex, the demand for real-time data processing is increasing, driving the need for advanced memory and storage technologies. Traditional datacenter infrastructure often struggles to meet the high-bandwidth demands of AI workloads, such as ML, deep learning, and big data analytics. These workloads require powerful memory solutions to process and analyze data seamlessly in real time, underscoring the importance of modern datacenters in supporting AI growth.

The Impact of AI on Datacenter Infrastructure

Artificial intelligence is revolutionizing industries, transforming datacenters, and driving the need for advanced infrastructure to manage the rapid growth of data, support massive computational demands, and ensure energy efficiency and sustainability. The surge in AI workloads, including machine learning and deep learning models, and other data-intensive applications presents unique challenges to traditional datacenter infrastructure, forcing significant changes in the processing, storing, and powering of data.

Data Growth and Management Complexity

AI's ability to analyze vast amounts of data means that datacenters must now handle exponentially larger data sets than ever before. This explosion in data generation — driven by digital devices, IoT, social media, and online services — presents a significant challenge to datacenters in terms of storage capacity, quick access to data sets, and efficient data management. Innovations in high-performance memory solutions, such as High Bandwidth Memory (HBM) and DDR5, are addressing the need for high bandwidth and low latency for efficient data processing, which are crucial for AI workloads that process large data sets in real time. In addition to memory improvements, storage solutions have evolved to keep up with AI demands. Nonvolatile memory express and solid state drives are seeing wide adoption for their high speed and capacity, enabling faster data retrieval and storage. These advancements ensure that the massive amounts that AI applications generate can be efficiently managed and processed, reducing bottlenecks in dataflow between processors and memory.

AI-driven data management software is also playing a crucial role in improving data organization, retrieval, and security. The complexity of handling large data sets requires intelligent systems capable of automating these tasks, further enhancing the operational efficiency of datacenters as they evolve to accommodate AI workloads.

Increasing Computational Demands

AI workloads demand significantly more computational power than traditional applications. These models are made up of trillions of parameters that must be processed simultaneously, requiring immense processing capabilities. To meet these demands, datacenters are leveraging heterogeneous architectures and increasingly adopting specialized hardware, such as graphics processing units (GPUs) and tensor processing units. GPUs, originally developed for graphics rendering, have become crucial for AI tasks because of their superior ability to manage parallel processing compared with traditional CPUs. As AI models increase in complexity, there is growing demand for custom AI chips tailored to specific tasks, intensifying competition in the hardware market. This evolving need is driving datacenters to constantly upgrade their infrastructure to accommodate the rising computational requirements of AI applications.

Energy Efficiency and Sustainability

AI workloads are not only computationally demanding but also power intensive, making energy efficiency a critical factor in modern datacenter operations. As AI continues to grow, datacenters are consuming increasingly large amounts of a nation's electric grid capacity. In 2024, IDC conducted a survey showing that electricity accounted for 46.3% of operating costs in enterprise datacenters, a figure expected to rise as AI workloads expand. In fact, worldwide datacenter energy consumption is projected to grow from 352TWh in 2023 to 857TWh by 2028, representing a compound annual growth rate (CAGR) of 19.5%.

With AI IT (servers, storage systems, and networking) accounting for 12% of the global datacenter energy consumption in 2024, energy efficiency and sustainability have become a major focus for datacenter operators as they face growing pressure to monitor operating costs and minimize their environmental impact. AI-driven energy management systems are being deployed to monitor and optimize power usage in real time, reducing waste and improving overall efficiency. In addition, lower-power memory technologies, advanced cooling systems for better heat dissipation, and renewable energy sources are being incorporated into datacenter designs to support sustainability initiatives. These innovations help reduce the carbon footprint of AI datacenters, aligning them with corporate environmental, social, and governance (ESG) goals and broader environmental regulations.

The rapid adoption of AI is reshaping datacenter infrastructure in profound ways. From meeting the high computational demands of AI workloads to managing the growing complexity of ever-increasing data sets, datacenters must continuously evolve to keep pace with the advancements in AI. At the same time, the focus on energy efficiency and sustainability is pushing datacenters to adopt more power-efficient technologies and AI-driven energy management systems. As AI continues to drive technological innovation, datacenters will remain at the heart of this transformation, providing the necessary foundation for AI's future growth while balancing performance, energy efficiency, and environmental responsibility.

FUTURE OUTLOOK

The rapid growth of AI workloads and other data-intensive applications has placed significant demand on datacenter infrastructure. According to *IDC's Worldwide AI and Generative AI Spending Guide*, AI spending on hardware, software, and services is expected to grow at a CAGR of 29% from 2024 to 2028. Memory has emerged as a critical component in enabling the performance and efficiency of these systems. From managing large data sets to ensuring real-time processing, memory technology plays a central role in the execution of AI models. For the foreseeable future, the role of memory in AI infrastructure — and the technological innovations driving its advancement — and its impact on datacenter efficiency and sustainability will be paramount.

Memory as a Performance and Efficiency Enabler

Memory plays a crucial role in the performance and efficiency of AI and data-intensive workloads, acting as a key enabler for advanced applications. AI models require large-scale data processing, and the performance of these workloads is directly linked to memory attributes such as bandwidth, capacity, and latency. High memory bandwidth allows for faster data movement between memory and processing units, boosting computational speed, while higher memory capacity ensures that vast data sets can be stored and processed without CPU offloading and delays associated with CPU-to-GPU or GPU-to-GPU data transfers, which is critical for AI tasks. Low latency further enhances performance by reducing the time it takes to access data, speeding up real-time computations and decision-making, and improving the user experience in the case of inferencing.

In real-time applications, memory technologies are essential for handling and processing large data sets. These workloads demand high memory capacity and high bandwidth to ensure that vast amounts of data are efficiently processed and readily available. Efficient memory usage enables faster model training and inference, unlocking the full potential of AI models and allowing businesses to analyze data and make real-time decisions. Ultimately, memory technologies are crucial in driving AI innovation and maximizing infrastructure efficiency.

Innovations in Memory Technology: High-Bandwidth Memory

HBM is a pivotal innovation in AI infrastructure. It offers several benefits, making it essential for handling the massive data and high-performance requirements of AI workloads. Today's latest HBM technology is HBM3E, with all major players rapidly developing HBM4 technology.

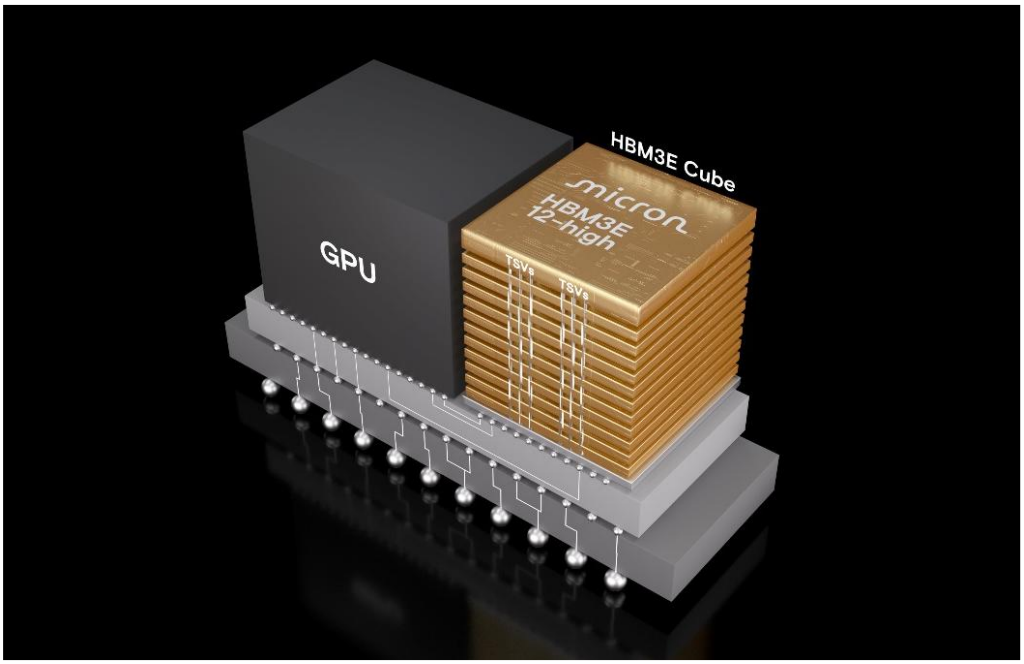
Key benefits of HBM include:

- **High data access performance:** HBM's high-speed data transfer capabilities and higher number of I/Os enable processing units to handle vast amounts of information quickly.
- **Reduced latency:** The technology's design ensures that data is accessed with minimal delay, improving overall processing speed.
- **Scalability:** HBM's architecture is scalable by stacking multiple DRAM dies within the package, supporting larger data sets and more complex AI models. Future developments in packaging, modular designs, and interconnect technologies might allow for more flexible upgrades.
- **Energy efficiency:** The technology's power-efficient design significantly reduces energy consumption to move data, making it ideal for AI workloads that demand both high performance and sustainability.

One of the critical innovations in HBM is its 3D stacked design (see Figure 1), which reduces the space required for memory and cooling. Multiple layers of memory are stacked vertically and connected using through-silicon vias (TSVs). A TSV is required to connect 8 or 12 memory dies to one another, forming a memory cube. The memory cube created with this stacking structure is connected to an interposer (the light gray box in Figure 1). The processor sits on the same interposer right next to the memory cube. This design reduces the energy required for data transfer, while the smaller physical footprint allows for more effective cooling, contributing to overall energy efficiency.

FIGURE 1

High-Bandwidth Memory



Source: Micron Technology, 2025

Energy Consumption and Operational Efficiencies of HBM

HBM improves energy efficiency through several design innovations, making it a preferred memory solution in AI infrastructure and other data-intensive applications. The following aspects highlight how HBM optimizes energy consumption:

- **High bandwidth at lower frequencies is achieved through:**
  - **Efficient bandwidth usage:** HBM achieves high data throughput at lower operating frequencies because of its wide memory bus (1,024 bits for HBM3E; HBM4 will have a bus width of 2,048 bits), which contrasts with traditional memory such as GDDR (e.g., GDDR7's bus width is 256 bits) that requires higher clock speeds to achieve high data transfer rates. This reduces the power consumed during data processing.
  - **Lower clock rates:** By operating at lower clock rates, HBM lowers the overall power requirement for processing data, leading to energy savings and reduced heat generated.
  - **Energy-efficient data transfer:** HBM consumes significantly less power per bit of data transferred compared with traditional memory technologies such as DDR5 and GDDR. Its wider data bus allows for large data transfers at lower clock speeds, reducing power consumption while maintaining high performance.

- **Scaling AI workloads with memory:** As AI models continue to grow in complexity, the role of memory in supporting this growth is critical. HBM and other advanced memory technologies enable the scaling of AI workloads by offering high capacity and efficient data transfer. This allows datacenters to support larger data sets and more complex computations without dramatically increasing power usage.
- **Advanced process nodes and power management:** Leveraging advanced process nodes, such as the 1-beta process node, which lowers power consumption, further optimizes modern technologies such as HBM. The 3D stacking design in HBM3E also allows for more effective power management by minimizing the energy required for data transfer and integrating power management systems directly into the memory module, further enhancing operational efficiency.

The Impact of Lower-Power Memory on Datacenter Efficiency

While memory is not the most power-intensive component in a datacenter, the reduced power consumption of memory can become significant in memory-bound workloads. The thermal properties of the memory and the mechanisms by which heat is dissipated are crucial. Lower-power memory has a significant positive impact on datacenter efficiency in several key ways:

- **Energy efficiency and cost savings:** Energy-efficient memory technologies such as HBM are transforming AI datacenters by reducing power consumption. Memory is a significant component of the total power usage in AI datacenters, especially in models that require massive data processing. By reducing the energy needed for memory operations, lower-power memory helps datacenters achieve energy savings. These savings directly translate into lower operational costs, particularly in facilities running large AI workloads, where electricity bills form a significant portion of ongoing expenses. Surveys that IDC has conducted show that electricity accounts for 46.3% of operating costs in enterprise datacenters.
- **Enhanced performance for AI workloads:** Lower-power memory also enhances thermal performance, which in turn improves the efficiency of AI workloads. Fast data access and high memory throughput are essential for large language models and models such as deep learning. Power-efficient memory delivers the necessary speed and bandwidth for processing while minimizing energy consumption. This enhanced capacity to handle large volumes of data reduces bottlenecks, enabling faster and more efficient AI computations. In addition, efficient thermal management helps maintain optimal operating temperatures for memory and processing units, preventing thermal throttling — a condition in which components reduce performance to avoid overheating. This ensures the system operates at peak efficiency.
- **Increased scalability of AI models:** As AI models grow larger and more complex, they require more memory for data storage and processing. Lower-power memory allows datacenters to scale their AI infrastructure without a proportional increase in power consumption. Technologies such as HBM and DDR5, with their high memory density and capacity, support the growing demand for AI workloads while maintaining energy efficiency.

- **Improved cooling efficiency:** Memory components contribute significantly to the heat generated in datacenters, requiring extensive cooling and heat dissipation systems. With better thermal management, lower-power memory reduces heat output, minimizing the energy required for cooling. This leads to more efficient cooling systems, such as liquid cooling or air-cooled solutions, which further reduce the energy demand of datacenter operations.
- **Sustainability and environmental impact:** The adoption of lower-power memory technologies in AI datacenters contributes to sustainability goals. As energy consumption decreases, the carbon footprint of these centers is reduced. This is particularly relevant as organizations face increasing pressure to adopt energy-efficient and environmentally friendly practices. Implementing power-efficient memory is a crucial step in supporting corporate ESG initiatives and achieving carbon neutrality.

For organizations, lower-power memory enhances datacenter efficiency by reducing energy consumption and cooling needs, enabling scalability, and contributing to hardware longevity — all of which result in lower operational costs and more sustainable performance.

Micron's HBM Solutions

Micron Technology is a global leader in advanced memory and storage solutions that offers a broad portfolio of products, including DRAM, NAND, and NOR flash memory, as well as leading-edge technologies such as HBM. With a focus on high performance and energy efficiency, Micron's memory products power critical applications across AI, datacenters, and the automotive and mobile industries. Designed for data-intensive workloads, Micron's memory technologies, such as GDDR, DDR, low-power DDR, and HBM, provide speed, bandwidth, and efficiency to meet the evolving demands of modern computing. Micron's latest HBM3E is currently shipping in AI systems today, and the company has shipped mechanical samples of HBM4 to customers to enable the next generation of future AI systems.

Micron's HBM solutions are specifically tailored for AI, machine learning, high-performance computing (HPC), and other data-intensive applications. These solutions offer:

- **High bandwidth and low latency:** Utilizing 3D stacking and through-silicon vias, Micron's HBM delivers extremely high bandwidth at low power, significantly boosting performance in AI accelerators, supercomputers, and datacenters. The HBM3E 12-high 36GB offers more than 1.2TBps of memory bandwidth, facilitating rapid data access for compute-heavy tasks.
- **Energy efficiency:** Micron's HBM solutions minimize power consumption by reducing the distance data travels within the memory, making them ideal for power-hungry AI workloads. Innovations in Micron's 1-beta process node and advanced packaging ensure lower power usage and improved thermal management, enabling sustainable datacenter operations. According to Micron, this enables its HBM3E solution to be up to 30% lower power than existing solutions, helping customers achieve operating expense savings.
- **Scalability for AI and HPC:** Optimized for handling large AI models and deep learning tasks, Micron's HBM provides increased memory density and capacity. The HBM3E 12-high 36GB, a

50% capacity increase over current offerings, supports the growing need for real-time data processing and analytics, enabling seamless scalability for AI and HPC workloads.

CHALLENGES

Micron and the High-Bandwidth Memory market face several challenges despite its critical role in high-performance computing, artificial intelligence, and data-intensive applications. Some of these challenges are:

- **High manufacturing costs:** HBM is more expensive to produce than traditional memory technologies because of its complex 3D stacking architecture and the need for advanced packaging techniques such as through-silicon vias. This raises the overall cost of systems that use HBM, limiting its adoption to specialized high-end applications where high-bandwidth memory is most critical.
- **Integration complexity:** Incorporating HBM into systems requires precise integration with processors, such as GPUs or AI accelerators. The specialized packaging techniques and close coordination between memory and processing units add design and manufacturing complexity, increasing the difficulty and cost of implementation.
- **Supply chain constraints:** The advanced technology required to produce HBM means that only a few manufacturers can produce it at scale. This limited production capacity can lead to supply chain bottlenecks, particularly as demand grows with the expansion of AI and HPC applications.

Overcoming these challenges is crucial for HBM to continue evolving and gaining broader adoption across industries.

CONCLUSION

The datacenter is at the heart of significant IT market trends, including the proliferation of data and data types, diversified workloads, heterogeneous computing, distributed computing, and AI. Memory technology is crucial for the performance and efficiency of AI and data-intensive infrastructure. As AI models grow in complexity, innovations in memory technology, such as HBM, are essential for managing large data sets, ensuring real-time processing, and reducing energy consumption. Lower-power memory not only enhances AI workload performance but also contributes to significant energy savings, improved cooling efficiency, and greater sustainability in datacenter operations leading to improvements in the total cost of ownership. As AI continues to drive datacenter evolution, advancements in memory technology will be vital for scaling infrastructure while maintaining energy efficiency and environmental responsibility.

IDC Custom Solutions

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. This IDC material is licensed for external use and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

IDC Research, Inc.
140 Kendrick Street, Building B, Needham, MA 02494, USA
T +1 508 872 8200

X (Twitter) @IDC | LinkedIn @IDC | www.idc.com

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.