Agentic AI in Data Centers: The Critical Role of High-Bandwidth Memory

Authors: Kevin Assogba, Moiz Arif, and Sudharshan S. Vazhkudai

Introduction

As agentic AI systems grow more autonomous and capable, their performance is increasingly shaped not just by computation power, but also by how efficiently they leverage available memory resources. This blog explores the evolving role of the memory subsystem in enabling high-throughput, low-latency agentic workflows, focusing on performance of HBM3e and projections for HBM4. We delve into HBM bandwidth and utilization patterns to offer a data-driven perspective on memory systems for agentic AI and uncover how memory architecture influences responsiveness and scalability of AI agents.

Key Takeaways

- Generational **leaps in high-bandwidth memory (HBM) are enabling faster inference and scaling of large agentic AI workflows**, pushing the boundaries of autonomy and responsiveness.
- Newer GPU generations with higher FLOPs **increasingly leverage higher available memory bandwidth** to deliver higher throughput gen-over-gen.
- **HBM4 (288GB) is projected to deliver up to 1.4x–2.6x higher throughput** for the agentic AI workflow, fueled by a 2.82x increase in bandwidth and 1.5x more memory capacity compared to HBM3e (192GB).

From LLMs to agentic AI: A shift in design and demands

Traditional large language models (LLMs) perform zero-shot inference, generating responses without revisiting their outputs. Retrieval-augmented generation (RAG) enhances these LLMs by querying external knowledge sources, e.g., vector databases, to provide additional context to the user's query. However, RAG systems lack autonomy, perception, and planning. Agentic AI introduces a paradigm shift. These systems are designed to perceive their environment, plan multi-step actions, evaluate intermediate outputs, and collaborate with other agents. They also leverage short-term and long-term memory (short-term memory for contextual awareness, and long-term memory for persistent knowledge) to improve response quality.

The speechwriter: A task-oriented agentic workflow

To evaluate agentic AI in action, we developed a workflow called “The Speechwriter” with the goal of generating speeches, given a user-defined prompt, according to the following format:

*“Write a “5-minute” (**D0**) speech to report to the “Executive Leadership Team” (**D1**) on the trend of “DRAM Market Penetration” (**D2**) in “AI Data Centers” (**D3**) industries from “2022 to 2024” (**D4** to **D5**), focusing on causes, challenges, and provisions to improve the state of “DRAM Market Penetration” (**D2**) for the upcoming years.”*

The submission of the request involves providing decision points: (**D0**) for speech duration; (**D1**) for audience type; (**D2**) for topic; (**D3**) for target industry; and (**D4-D5**) for a reference timeframe.

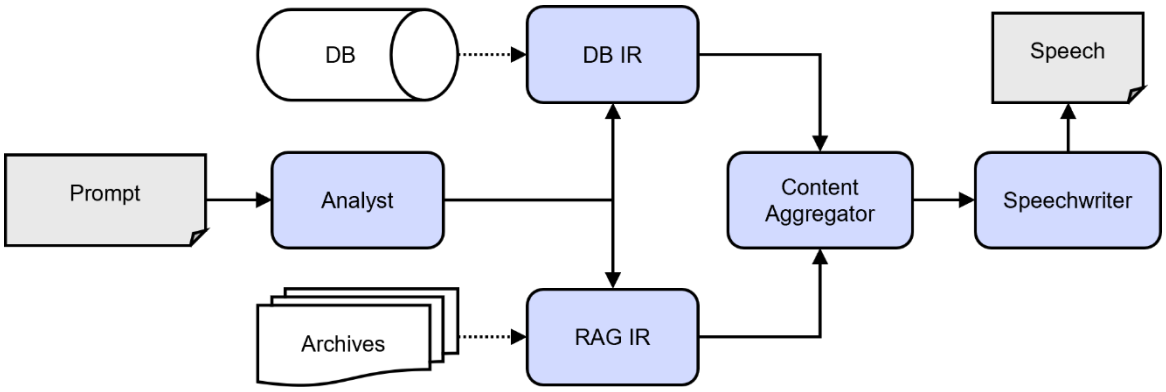

Figure 1: Illustration of the proposed speechwriter agentic workflow

Figure 1 illustrates the design of the speechwriter workflow, enabling parallel execution and dynamic tool use. The workflow involves five specialized agents:

(1) The Analyst extracts key parameters from the prompt;
(2) The Internal Researcher (DB IR) queries structured databases, e.g., relational databases containing quarterly manufacturing, supply chain, and sales data;
(3) The Internal Researcher (RAG IR) augments context through a similarity search against a vector database, e.g., a database containing embeddings of technical reports, blogs, product data sheets, and internal and outbound documentation;
(4) The Content Aggregator summarizes collected data;
(5) The Speechwriter crafts the final speech using information from previous agents.

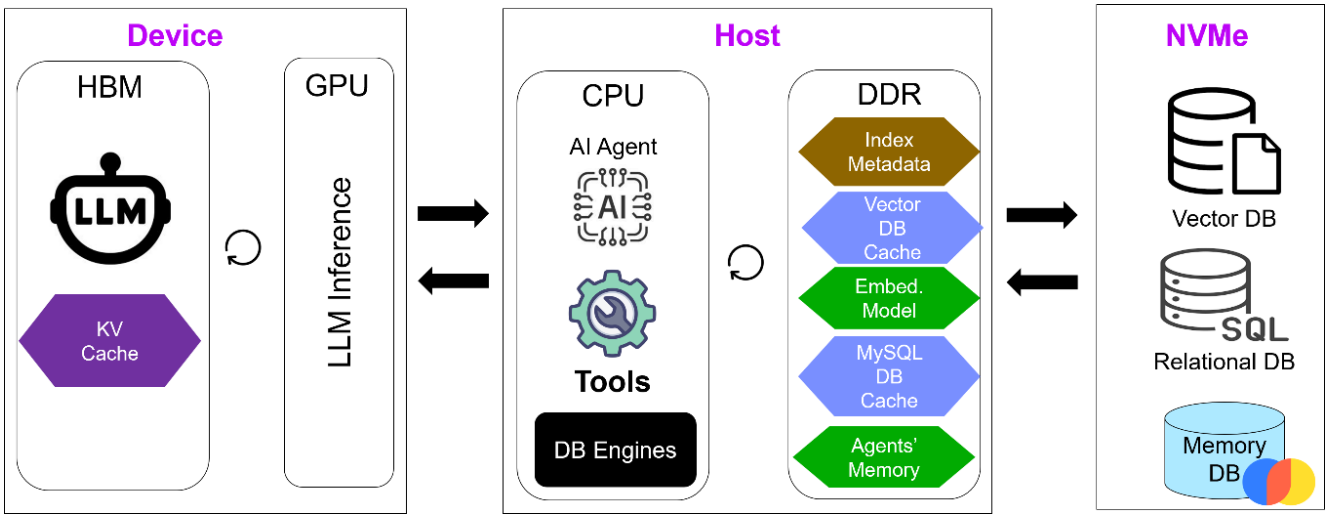

Figure 2: Architectural deployment of the proposed agentic workflow

Performance insights

We deployed the CrewAI framework to orchestrate the agentic workflow and use the Meta-Llama-3.1-8B-Instruct-FP8 model to support all agents' inference requests. Figure 2 presents the system architecture, highlighting that the GPU handles inference, whereas the host CPU manages tool use and agent execution.

HBM Bandwidth and GPU Utilization for a Single User Request: We deployed the workflow using a system with 192GB HBM3e and 2TB DDR5 system memory.[1] A breakdown of resource utilization from each phase in the agentic workflow, as shown in Figure 3, depicts up to **36%** HBM bandwidth and up to **83%** of GPU utilization spikes during LLM interactions.

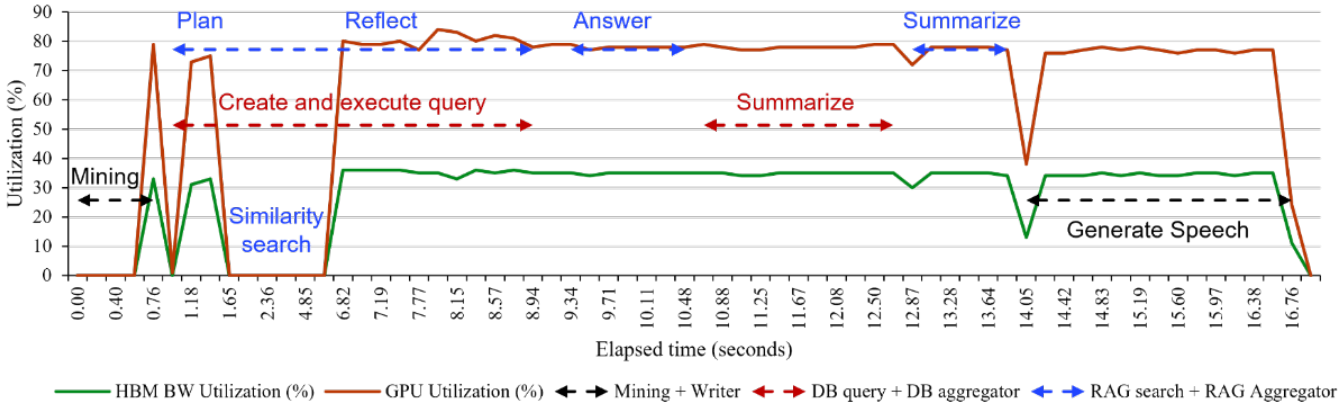

Figure 3: HBM3e bandwidth and GPU utilization of the Speechwriter workflow for a single request

Scaling Agentic AI System: The performance of agentic AI increases at scale with the processing of large concurrent requests in a batch. We evaluate the performance of **8 TB/s HBM3e** (192GB from eight HBM cube placements), using a batch size of 100, 1,000, and 10,000 queries from the Google Research Natural Questions benchmark. We deployed an agentic AI workflow with one RAG agent supported by an LLM and a vector DB. Figures 4a-4c report the normalized time spent during inference on the GPU, the normalized KV cache usage and the maximum HBM bandwidth.

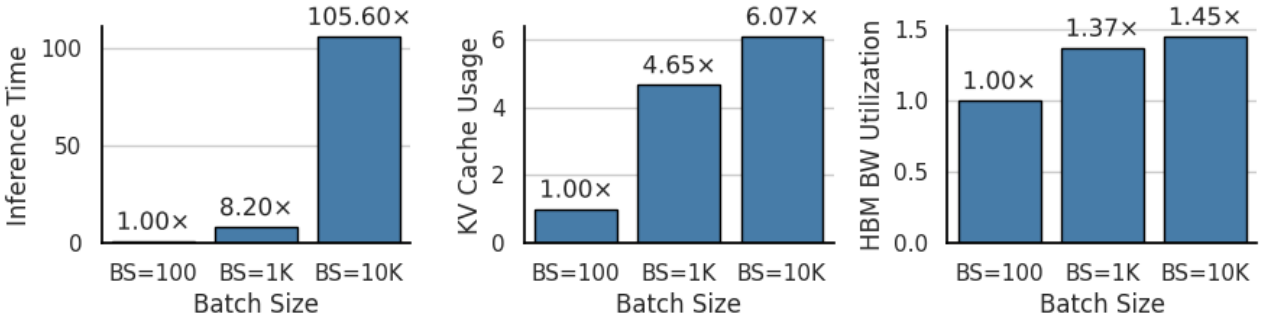

Figure 4: (a) Normalized Inference Time (s) (b) Normalized HBM Capacity (GB) (c) Normalized HBM Bandwidth (GB/s)

The observed performance of scaling concurrent agentic AI requests highlights three main observations on the importance of HBM.

- **Inference time** increases sharply with concurrency, rising from baseline at 100 requests to ~8.2x at 1K and ~105.6x at 10K. The steep increase at 10K is due to the 100% KV-cache utilization not just the RAG overheads, indicating that once KV capacity is exhausted, additional concurrency leads to severe memory pressure and rapidly degraded inference, despite available compute and bandwidth.

[1] *1x NVIDIA Blackwell GPU*

- **HBM bandwidth** utilization increases at higher concurrency. As batch size increases, HBM bandwidth utilization rises to ~1.37x (1K) and ~1.45x (10K) compared to the baseline of 100 requests, indicating improved bandwidth utilization at scale.
- **KV-cache demand** scales super-linearly with batch size, increasing from baseline at 100 requests to ~4.65x at 1K and ~6.07x at 10K requests with a fully utilized available KV cache budget at 10K batch size. Larger HBM capacity directly supports higher concurrency by accommodating larger KV-cache footprints, enabling agents to operate with richer context. As concurrency increases, effective HBM capacity utilization becomes increasingly dependent on incoming request rate controlled by the RAG execution and the overall batch composition.

HBM4 projections for agentic AI

As the AI ecosystem continues to evolve at a fast pace, newer GPUs with HBM4 are projected to efficiently serve large-context reasoning and retrieval, aligning with the requirement of agentic workflows to combine intermediate agents' output with the user query in multi-agent deployments. Building on the observed performance of agentic AI workflows with earlier HBM and GPU generations, we estimate throughput projections for HBM4, specifically focusing on total serving throughput (tokens per second). To illustrate the impact of next-gen memory on agentic AI workflows, we utilize performance data from an HBM3e (192GB) system. Key specifications of HBM4 include:

- **HBM4 memory configuration:** 36GB per stack in a 12-high arrangement.
- **Interface:** 2048-bit, delivering up to ~22 TB/s of memory bandwidth.
- **Total capacity:** 288GB, enabled by eight HBM cube placements.

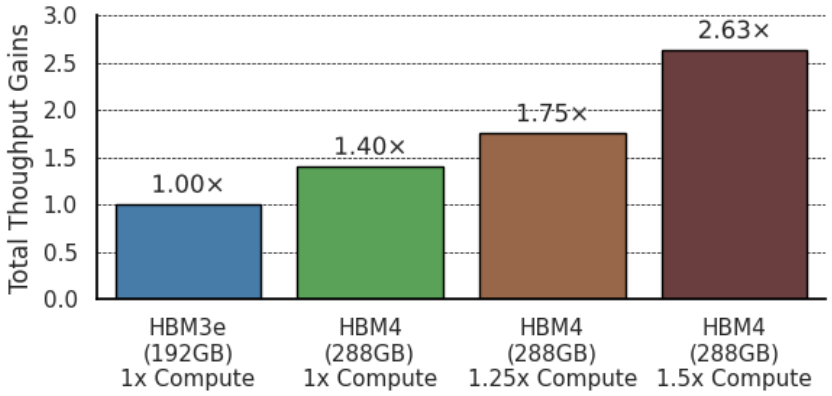

Figure 5: Projections for HBM4 compared to HBM3e (192GB) with varying GPU compute

Using a roofline model that we developed, we project performance gains for the HBM4 compared to HBM3e in Figure 5 with varying levels of GPU compute, with key takeaways summarized below:

- Under a fixed compute budget, HBM4 (288GB) delivers **~1.4x** throughput gains over HBM3e (192GB), enabling faster execution of agentic AI workflows poised to demand long-horizon planning and dynamic memory retrieval.
- As the GPU compute scales, HBM4 (288GB) scales more effectively, reaching **~1.75x** throughput with 25% higher compute and **~2.63x** with 50% higher compute, reflecting its superior bandwidth and capacity utilization.

These projections, illustrated in Figure 5, align with the **2.82**x higher peak theoretical bandwidth of HBM4 compared to HBM3e (192GB). The projections account for increase in HBM capacity and bandwidth, while the bulk of the gain comes from HBM4 bandwidth improvements. As the number of concurrently deployed agents grows, each backed by a dedicated LLM, and context windows continue to expand, the KV cache is expected to become a potential bottleneck. To ensure high-throughput and low-latency agentic workflows, future systems will require not only greater HBM bandwidth but also significantly higher capacity to accommodate the increasing memory demands.

Conclusion

Agentic AI presents new opportunities to advance HBM technologies by emphasizing the growing demand for memory bandwidth and capacity across HBM generations. At scale, higher HBM capacity and bandwidth utilization on HBM3e (192GB) systems drives higher performance. This results in reduced LLM inference time and improved serving throughput. As we move toward HBM4, with **2.82x** peak HBM bandwidth and **1.5x** HBM capacity increase compared to HBM3e (192GB), the memory subsystem steadily emerges as a central enabler of intelligent system performance and scalability. Our projections indicate that HBM4 could deliver up to **2.63x** higher throughput compared to HBM3e (192GB).

Author Bios

Kevin Assogba is a Systems Software Engineering Intern in the Data Center Workload Engineering team, working at the intersection of heterogeneous memory solutions and AI workloads. Kevin is a Ph.D. candidate in Computer Science at the Rochester Institute of Technology (Rochester, NY), focusing on high-performance computing (HPC) and data movement optimization for HPC and AI applications.

Moiz Arif is a Principal Engineer in the Data Center Workload Engineering team, specializing in performance engineering for AI workloads including agentic AI, Vector DBs, RAG, reasoning models, and physical AI along with developing proof-of-concepts, articulating the value of Micron products, and designing system architectures to optimize workload performance. Dr. Arif specializes in tiered memory management, disaggregation, and tiering for a wide range of AI and HPC workloads. Dr. Arif holds a Ph.D. in computer science from the Rochester Institute of Technology (Rochester, NY).

Dr. Sudharshan S. Vazhkudai is a Fellow of systems design engineering at Micron Technology. He established the Data Center Workload Engineering team, which brings an end-to-end systems perspective in understanding how deep-memory hierarchy is used to create modern system architectures optimized for workloads. Prior to this, for over two decades, he worked at the US Department of Energy's Oak Ridge National Lab as a Distinguished Scientist and a Program Director, building HW/SW codesign solutions for generations of world's fastest production supercomputers, serving a large user base in AI and HPC. Dr. Vazhkudai holds a Ph.D. in computer science from the University of Mississippi and has also served as a joint faculty at the University of Tennessee. He has published over a hundred peer-reviewed papers on the fundamental underpinnings of data center systems design, system architecture, deep-memory/storage hierarchy, heterogeneous CPU/GPU architectures, data management, and distributed systems.

micron.com

©2026 Micron Technology, Inc. All rights reserved. All information herein is provided on as "AS IS" basis without warranties of any kind, including any implied warranties, warranties of merchantability or warranties of fitness for a particular purpose. Micron, the Micron logo, and all other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners. No hardware, software or system can provide absolute security and protection of data under all conditions. Micron assumes no liability for lost, stolen or corrupted data arising from the use of any Micron product, including those products that incorporate any of the mentioned security features. Products are warranted only to meet Micron's production data sheet specifications. Products, programs and specifications are subject to change without notice. Rev. A 04/2026 CCM004-1681249710-11878