The Emerging Memory-Centric Data Center

And the new role memory providers play in the industry.

Karl Freund
Cambrian-AI Research, LLC
October 14, 2024

Introduction

During the last few years, we have witnessed explosive growth in the data center, driven by the pervasive use of public cloud services, big data analytics, and most recently by the development of Artificial Intelligence (AI) enabled by power and data-hungry GPUs. This growth and changes in the type of workloads being run have prompted a re-thinking of the data center infrastructure, with new demands for high-density power distribution and increasingly water-cooled servers.

The data center landscape is profoundly transforming, driven by the relentless demand for higher performance, improved efficiency, and increased scalability. At the heart of this evolution lies a revolution in memory and computing technologies, especially around deploying AI-specific infrastructures and reshaping the architecture and capabilities of modern data centers. All data centers will become AI data centers to some extent.

This white paper explores the impact of new technologies on memory providers and asserts that advanced technologies such as high bandwidth memory (HBM) are fundamentally altering the industry landscape. Instead of being relegated to lower-margin commodity status, memory companies are now enjoying an elevated role in the critical supply chain for AI.

Changing the Industry Control Points and Leadership

AI system performance depends on memory capacity, bandwidth, and latency. Consequently, memory technology is a critical design criterion for selecting infrastructure components like GPUs and other accelerators. Data center revenue has already shifted to GPUs, which need faster memory across the entire memory hierarchy. Consequently, memory, storage, and their suppliers deliver more value and enjoy a more critical position in the data center value chain. Data center infrastructure companies are now positioning their products as having a superior value proposition based partly on the HBM they selected and integrated into their GPUs and accelerators. Memory providers are now in a position of significantly differentiating the final solution, unlike in the past where memory and storage were commodities bought primarily on price.

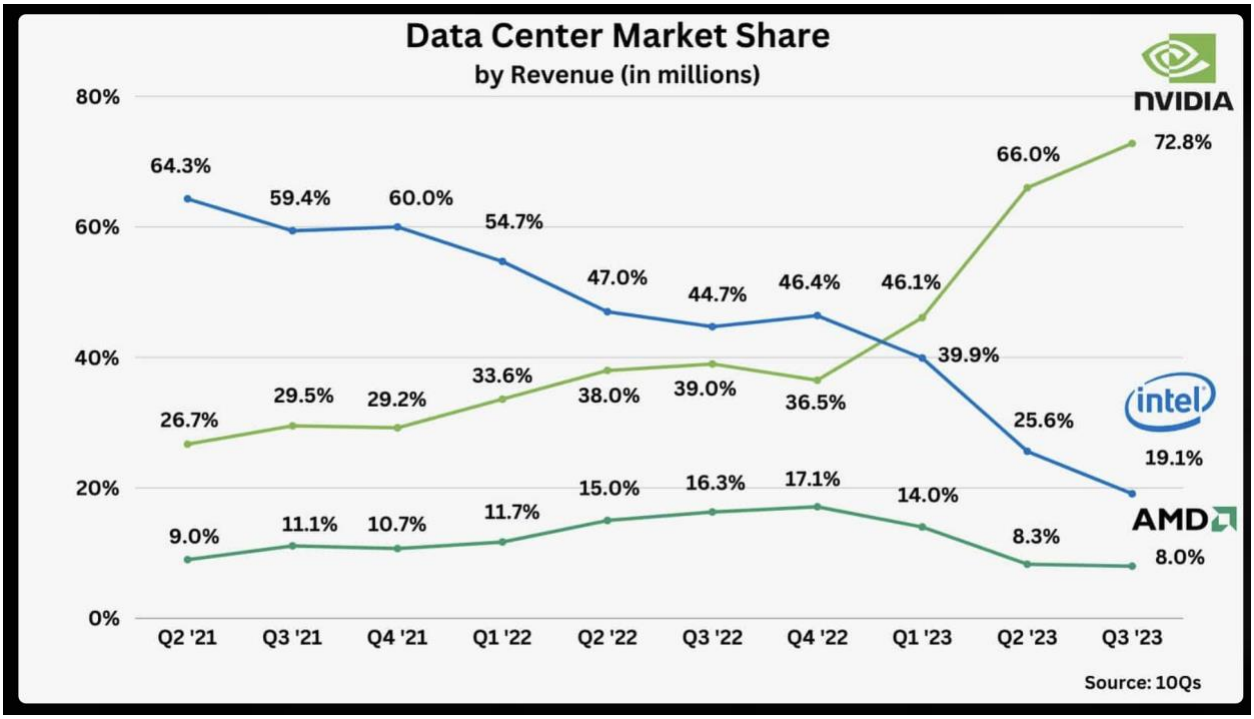


Figure 1: A rough estimate of market share from Eric Flaningam, of Generative Value.

HBM has become a pivotal factor in enhancing the performance of AI and high-performance computing systems. The transition from NVIDIA's H100 to H200 GPUs exemplifies this impact. The H200, equipped with 6xHBM3E, delivers twice the performance for large language models (LLMs) compared to the H100, which uses 5xHBM3. Specifically, the H200 boasts 141GB of memory and 4.8 TB/s bandwidth, significantly surpassing the H100's 80GB and 3.35 TB/s. This substantial increase in memory capacity and bandwidth underscores that HBM is no longer a mere commodity but a critical performance driver in cutting-edge computing applications.

Consequently, memory companies have transitioned from commodity suppliers to critical partners in AI advancements, with increased influence on product development, pricing, competitive positioning, and supply chain dynamics. As the amount of HBM capacity continues to grow, with the new 12-high stacks now available, companies such as Nvidia and AMD will, in part, position their products based on HBM features.

Major memory manufacturers prioritize HBM production due to its growing demand and critical role in AI acceleration. Micron, for instance, is building new fabs in Boise, Idaho, and Central New York to ensure a reliable domestic supply of leading-edge memory. These facilities will support the increasing need for advanced memory in AI and other high-performance applications.

Memory providers have become crucial partners in the AI hardware supply chain. The burgeoning demand for HBM in AI applications is delivering memory providers a more

stable and growing market, potentially reducing the memory industry's historically cyclical nature.

Workloads Driving Change

Data center operators are trying to keep up with business demands for AI-capable infrastructure, which entirely changes power densities and cooling technologies. Platforms like the upcoming Nvidia GB200 NVL72, with 720 PetaFlops of FP8 performance that consumes ~120 KW per rack. Compare that to the “typical” 13 KW/rack, and you can see the challenges ahead.

The key drivers of change in the modern data center we will discuss below include:

- A shift from data retrieval to analytic service
- The exponential growth in data itself
- The rise of AI and ML workloads

As we delve into these trends, we examine how memory hierarchies are addressing the challenges presented by modern workloads. From HBM near memory storage to expansive data lakes, new technologies can reshape the data center landscape, enabling organizations to extract maximum value from their data assets. By incorporating these new technologies, organizations can better prepare for the future and make informed decisions about their infrastructure investments.

Advancing from Data Retrieval to Analytic Services

Historically, data centers have provided simple access and analysis to stored information in the cloud or an on-premises facility. That information was stored in files or databases on spinning disks and SSDs.

While analytics have always played a part, processing has not been particularly compute-intensive, nor have analytics required interconnects between processors beyond that which Ethernet has been able to provide. Only a single core on a CPU typically fetches a file and extracts the requested data. While much of the internet remains based on this simple usage model, the fast-growing demand for complex analytics is rewriting that model into one in which multi-processor collaboration is needed, typically demanding acceleration from scores or even thousands of fast GPUs.

As AI-based analytic services have emerged, the latencies and bandwidth provided by DRAM DIMMs have proved inadequate and need to be augmented by HBM. Similarly, the existing interconnects could not deliver the point-to-point connectivity required by neural network processing and are being extended with Nvidia NVLink within a node and Nvidia InfiniBand between nodes. AMD's response to Nvidia's NVLink is the development of UALink (Ultra Accelerator Link), an open standard to provide an alternative to Nvidia's proprietary interconnect technology. Similarly, the Ultra Ethernet Consortium is developing an open alternative to InfiniBand called the Ultra Ethernet

Transport. The UEC 1.0 specifications are slated to be released publicly in later in 2024[1].

From Petabytes to Zettabytes

All this innovation is being spurred by two developments: fast accelerators for parallel computation and massive amounts of data. The growth of data stored in data centers is experiencing a rapid and significant increase. As of 2024, approximately 402 million terabytes of data are created daily, equating to around 147 zettabytes annually. This figure is expected to rise to 175 zettabytes by 2025, a substantial increase from the 33 zettabytes generated in 2018. The growth trajectory has been consistent, with a 74-fold increase in data generation from 2010 to the present.

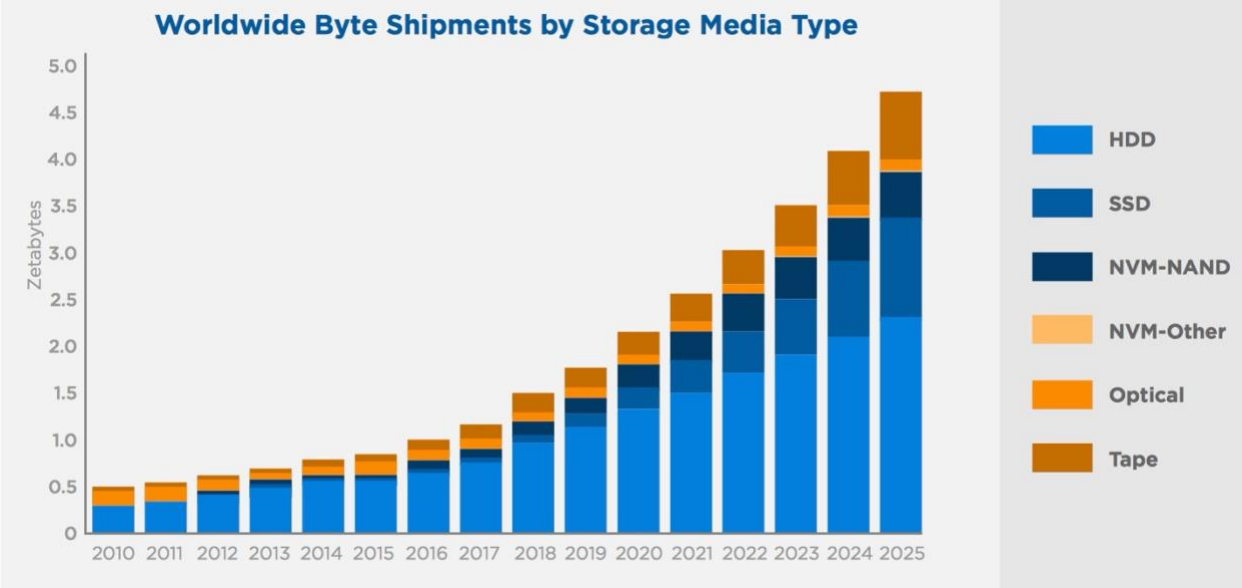


Figure 2: The growth in the Datasphere. https://www.storagenewsletter.com/2018/11/28/global-datasphere-from-33zb-in-2018-to-175zb-by-2025/

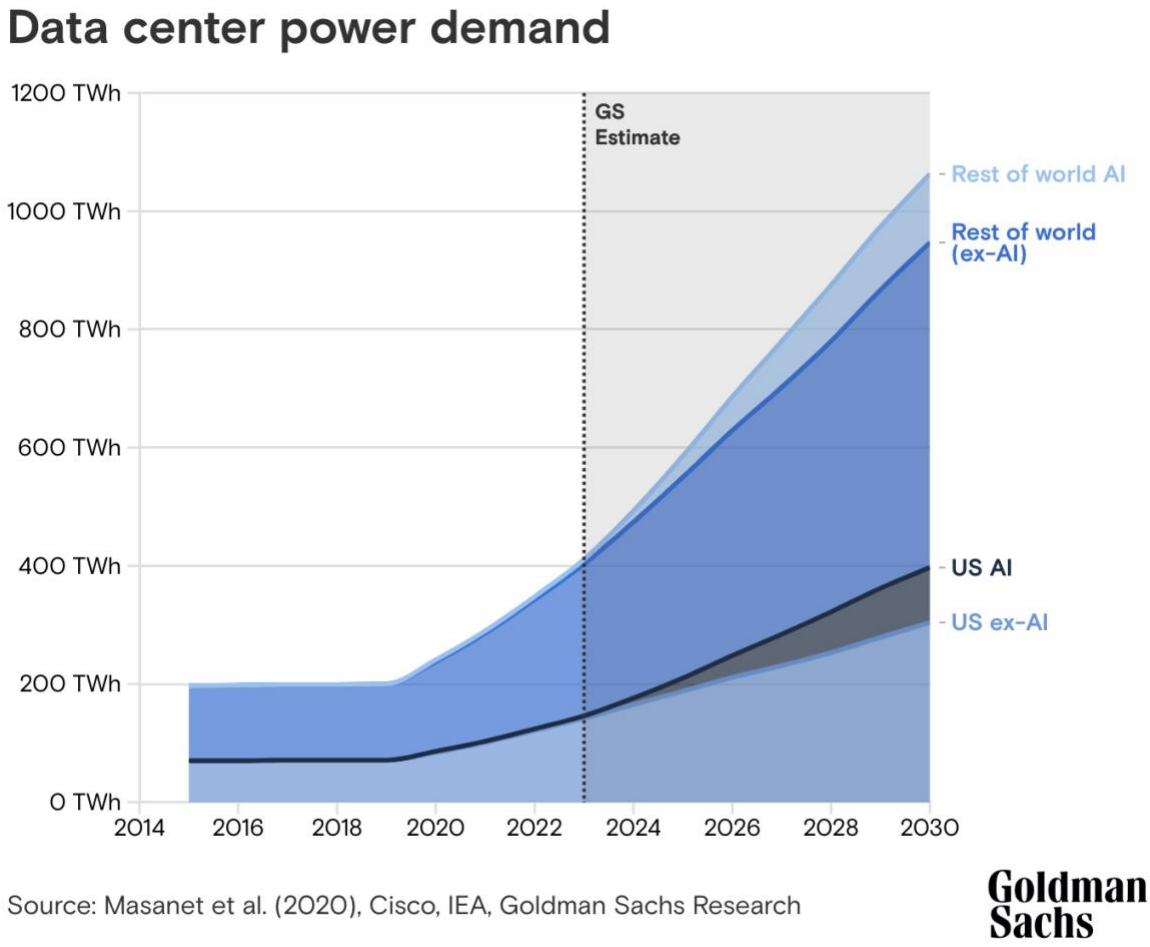
From Big Data to AI

During the last two decades, companies began to combine disparate data stores and databases into data lakes, simplifying access to big data analytics. While many companies benefitted from this investment and data consolidation, the analytic services remained rudimentary. Now, AI promises to accelerate and expand the utility these data repositories can provide. Data lakes can be curated to train or fine-tune AI models for enterprise use cases such as customer support and HR. Thanks to historical investments in building big data, enterprises can be in a solid position to take on AI projects that can impact their business.

And All This Takes Power - a Lot of Power

[1] https://www.nextplatform.com/2024/06/26/what-if-omni-path-morphs-into-the-best-ultra-ethernet/

Data center power consumption is increasing rapidly, spurred primarily by the adoption of GPUs and AI. Globally, data center power is projected to increase by 160% to 3-4% of total power production, with estimates for US power consumption ranging up to 9.1%, a significant increase over the 4% we consume today[2]. And with high-power GPUs, the power density is increasing dramatically; an Nvidia NVL-72 consumes approximately 125 KW in a single rack, roughly ten-fold the historical power per rack that data centers were designed to deliver. This level of power density is leading to a lot of wasted space in data centers, and the use of liquid cooling is expected to grow significantly.



Moreover, the latest bottleneck facing data centers is the amount of energy they can supply. Many data centers are increasingly energy-limited, not compute-bound, especially for those hosting AI servers. AI is now poised to drive a 160% increase in data center power demand.[3]

Micron's HBM3E, which consumes 30% less power than competing products, effectively addresses this challenge. Additionally, Micron's 9550 NVMe SSD, designed for AI and data-intensive workloads, offers up to 43% less average SSD power consumption while

[2] https://www.epri.com/about/media-resources/press-release/q5vU86fr8TKxATfX8IHf1U48Vw4r1DZF
[3] https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

delivering superior performance. Similarly, the Micron 6500 ION NVMe SSD promotes sustainability with 20% lower active power consumption and 56% better power efficiency than competing QLC SSDs. Anything that can be done to lower power consumption can equate to increased performance and the quantity of computations that can be done.

Memory Hierarchies Enable Higher Performance and Accessibility

With so much data analytics and AI being deployed, memory has evolved to keep up with latency and capacity demands. We have seen the emergence and rapid growth of HBM for GPU local memory, and DDR5 memory has largely replaced DDR4 for CPU main memory. CXL, the protocol for sharing memory disaggregated from the server CPUs, continues to provide an elegant, high-performance memory-sharing solution in the data center.

Let's examine each of these memory technologies, starting with the Near Memory solutions enabled by HBM.

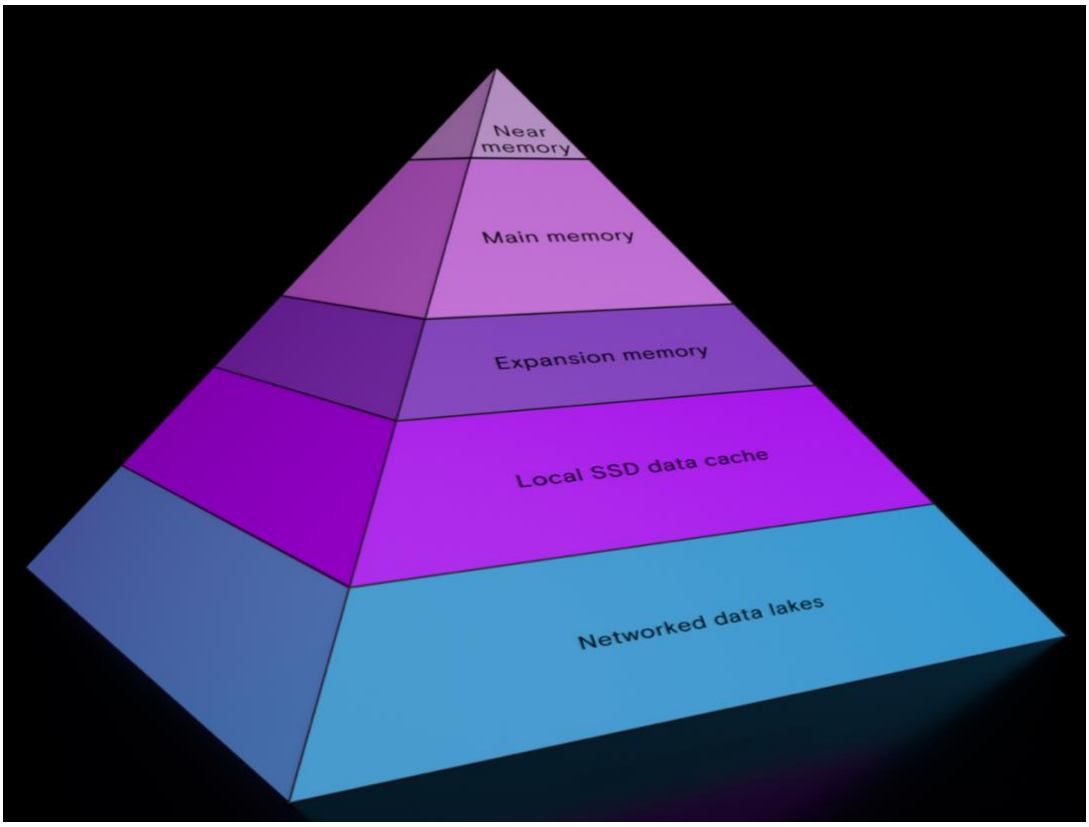


Figure 3: Hierarchy of memory and storage Micron

Near Memory

One of the most significant changes in data centers over the last few years has been adopting high bandwidth memory or HBM. AI requires high-speed memory, especially for training, and HBM provides the best current solution. Even with HBM, training an AI model with hundreds of billions or trillions of parameters can take days, weeks, or even months. Each parameter is updated thousands or millions of times as the data steers the model to yield adequate precision. Accessing these parameters in DDR5 memory would be a non-starter, increasing training times to many years. Hence, high bandwidth memory has quickly become the standard for AI processing.

While training a model is highly compute-bound, inferencing processing of generative AI tends to be more memory-bound. The cost of running a single inference is much smaller than training. Still, the aggregate cost of inference over the model's lifetime often exceeds that of training due to the sheer volume of inferences. Consequently, there is tremendous focus on higher HBM capacity per accelerator, and the memory providers have now developed a 12-high HBM stack to meet this need.

HBM is built from specialized DRAM chips stacked vertically to increase density, with Through-Silicon Vias (TSVs) interconnecting the chips. The current 8-high HBM3E generation provides up to 24GB of memory on the same package as the processor and can transfer data at 1.2 TB per second. In the case of Micron, their HBM3E solution consumes 30% less energy than competitors' offerings. In addition to AI, HBM is used in High-Performance Computing, enabling more efficient data processing in memory-intensive applications. To help access larger near memory, vendors like Micron have recently extended their offerings to 12-high DRAM dies, upping capacity to 36GB.

Looking forward, the JEDEC standard for HBM4 has been ratified and released. The new HBM4 standard enhances capacity and bandwidth compared to its predecessors. HBM4 will support 24Gb and 32Gb layers, with the capability to stack up to 16 memory chips in a single package (16-Hi stacking), supporting up to 512Gb (64GB) per stack. HBM4 features a 2048-bit memory interface, doubling the current 1024-bit interface used in HBM3. This will enable a theoretical peak memory bandwidth of over 1.5 TB/s per stack.

Main Memory

While HBM gets a great deal of attention, main memory remains the workhorse of the data center. According to Micron Technologies, the server's main memory is shifting to DDR5 DRAM, which gives servers a boost of 45% in memory density, 17% lower latency for memory-intensive workloads, and 24% better power efficiency.

Low-power DDR5 is beginning to make inroads into the data center, especially for power-hungry AI servers. LPDDR5 runs up to 6400 Mbps with many low-power and RAS features. The latest generation, LPDDR5X, reaches speeds as high as 9600 Mbps. DDR5 DRAMs, with a data rate of up to 6400 Mbps, support higher density, including a dual-channel DIMM topology for higher channel efficiency and performance.

Monolithic high-capacity dies, such as Micron's 32Gb monolithic die-based 128GB DDR5 RDIMM, offer significant advantages over traditional multi-die systems, including improved performance, reduced latency, and enhanced energy efficiency. These advancements are crucial for meeting the growing demands of AI and high-performance computing applications.

Expansion Memory

An emerging trend in data centers is shared memory pools that are disaggregated from servers, a memory tier that fits between faster main memory and slower SSDs. Shared disaggregated memory represents a significant departure from traditional server-centric memory architectures. Decoupling memory resources from compute nodes creates a memory pool that can be dynamically allocated and shared across the data center, increasing memory capacity and utilization, improving rack-level performance, and lowering costs.

CXL (Compute Express Link) is an industry standard that allows disaggregated memory, providing cache-coherent access by multiple applications and servers. The adoption of CXL is gaining momentum, particularly in hyperscale data centers, and is expected to grow significantly in the coming years. The introduction of CXL-enabled products and active industry engagement are driving this growth. CXL 2.0 can provide over 2TB of expansion memory at 36 GB/s latency.

Local SSD Data Cache

NVMe Solid-State Drives (SSDs) have replaced most spinning disk drives in the modern data center, offering high-performance and large-capacity drives for memory-intensive workloads. Advantages include dramatically improved I/O performance, lower power consumption compared to HDDs, and increased reliability and durability.

While HBM has become the primary store for LLM parameters, one still must load the initial weights into the HBM, typically from SSDs. Since that load time can be lengthy, faster SSDs are required to lower the time to the first token, which is burdened by the data load more than the compute time.

Networked Data Lakes

Massive data lakes are becoming the source of training data for company AI models. Consequently, performance is rising to the top of many buyers' buying criteria, shifting demand from spinning HDDs to SSDs. The Micron 6500 ION, for example, can improve performance by 48% over capacity-centric alternatives and can ingest 100TB of data four days faster than HDD alternatives with five times more density.

Recommendations

The industry landscape is changing dramatically, and the rising importance of memory technologies like HBM, CXL, and DDR5 disrupts the technology suppliers and value chains. The industry has never seen such dramatic evolution in data center

technologies, design, cooling, and power, mostly brought to bear to support AI applications and the data-centric world of Artificial Intelligence. To navigate the evolving data center landscape, organizations should plan their evolution to shift to AI accelerators, NVMe SSDs, and CXL disaggregated memory pools to increase performance and lower costs. We also recommend collaborating closely with memory companies to identify and implement energy and power efficiency solutions.

Innovations in memory technology can significantly reduce energy consumption and operational costs. They provide insights into the latest advancements and help tailor solutions to meet specific energy efficiency goals. By staying ahead of these trends and challenges, organizations can build resilient, high-performance data centers that meet the demands of modern workloads and support future growth.

IMPORTANT INFORMATION ABOUT THIS PAPER

Author and Publisher

Karl Freund, Founder and Principal Analyst, Cambrian-AI Research LLC

Inquiries

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

Citations

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Cambrian-AI Research". Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

Licensing

Disclosures

This document was developed with Micron Inc. funding and support. Although the document may utilize publicly available material from various vendors, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

Disclaimer

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.