Moving artificial intelligence workloads to edge AI processors

Fortify edge server infrastructure with Micron memory and storage

The specialized requirements of data-intensive workloads like artificial intelligence (AI), machine learning (ML), and deep learning (DL) are already taxing traditional data center infrastructures. AI processing is shifting from the cloud/data center to the network edge (see Figure 1). In fact, edge devices and edge infrastructure have become the largest market for AI acceleration technology[1]. These edge AI/ML workloads add value to supply chain management, private 5G networks, industrial processes, and smart manufacturing.

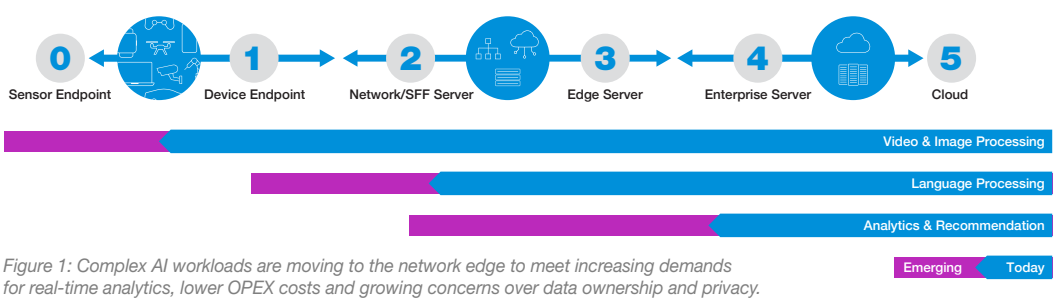

Figure 1: Complex AI workloads are moving to the network edge to meet increasing demands for real-time analytics, lower OPEX costs and growing concerns over data ownership and privacy. This increases the need for highly efficient memory and storage solutions on edge platforms.

Accelerating competitive advantage with edge AI

What's the main motivator for the shift to the intelligent edge? Competitive advantage!

By 2025, 50% of enterprises will have devised AI orchestration platforms to operationalize AI, says Gartner[2]. But for best results, AI and ML depend on speed. To quickly ingest, analyze and transform torrents of data into strategic insight, compute and storage must be located as close as possible to the place where data is being created. And so AI moved to the edge.

Edge AI can help deliver higher bandwidth, lower latency, and improved security. In addition, improvements in flash technology can help shrink the physical storage footprint and reduce power and cooling costs — all while improving the overall speed, flexibility, and failure rates. Data egress fees between the edge and the cloud (or any two different networks) can be expensive. Performing AI training and inference at the edge — only sending back certain high-value device data to the cloud for ad hoc trend analysis and predictions — reduces data transfer costs.

Challenges of edge AI will demand more memory and storage

As edge AI/ML deployments move away from simple gateway infrastructures that simply aggregate, process and forward, new complexities will require more memory and storage (see Figure 2). This evolution will also unleash improved edge analysis, more microservices, aggregation with time functions (important in predictive analysis) and a host of exciting opportunities. Infrastructure technology is already addressing some of AI/ML's key challenges (see Table A).

#	Challenge	How memory and storage infrastructure helps
1	Massive amounts of data to aggregate, store, and transfer	Machine learning and AI model training move huge amounts of data through the network. Placing servers at the network edge saves on data transfer. High-performance PCIe connectivity in NVMe™ SSDs streamlines and accelerates the process.
2	Specialized workloads use complex parallel processes	Moving specialized workload data can cause bottlenecks and hamper efficiency. The new generation of server DRAM has doubled several functions to enable 2x the concurrent operations, transfer more data per cycle, and respond quicker to each data queue[1].
3	AI training uses huge amounts of energy	Edge servers that are housed away from centralized data centers need low power. The latest NAND includes NV-LPDDR4, a low-voltage interface for per-bit transfer savings of more than 30%*. The newest high-performance SSDs reduce power usage by 77%[3].

Table A: Next-generation memory and storage help meet edge AI challenges

The business value of data is unleashed at the edge

Register the for e-book

Businesses are increasingly making data decisions at the network edge. This free e-book, written by the experts at Omdia and commissioned by Micron, covers the innovation that has enabled data capture and AI/ML training and inference to expand out of server farms. New business processes are springing up at remote locations where computing devices could never have existed before. The critical element: responsive, flexible, next-gen IT infrastructure.

Push past traditional IT limitations with Micron

Micron's broad memory and storage portfolios are key components of an edge AI infrastructure strategy that can exceed performance, bandwidth, latency and capacity challenges in the new data economy. Our products will continue to be invaluable as edge AI inevitably grows more sophisticated. In addition to edge servers, our LPDDR4 and LPDDR5 technology provides intelligent endpoints with low power and small footprint memory.

Micron memory and storage solutions already:

- Set new standards for AI/ML storage and memory performance
- Accelerate business-critical applications while maximizing IT budgets
- Make AI training and inference deployments faster and more efficient

Our proven portfolio of innovation for the edge and AI

Micron innovations help our memory and storage technology to optimize mission-critical AI workloads at the edge, especially those that require high speed, security and power-efficient execution of massive capacities. Our technology enables edge servers to run the same sophisticated workloads as centralized facilities — so organizations can go where the data takes them. Some key innovations include:

Micron performance NVMe™ storage

- Massive capacity and high-performance storage are vital to tackling AI workloads at the edge
- The Micron® 9400 NVMe SSD offers up to an industry leading capacity of 30.72TB[4], outperforming industry competitors up to 2.3x in mixed workloads[5]. The Micron 9400 was a class leader in tests performed with Aerospike Database, known for its massive parallelism that enables billions of transactions in real time[6] (see Figure 3).

Micron performance NVMe™ storage

- Low power and high performance are a perfect fit for power- and cooling-constrained edge environments
- Small form factors, led by M.2, meet the need for high-performing boot drives at the edge
- The Micron® 7450 NVMe SSD consistently delivers 2ms and lower latency for 99.9999% QoS[7] and has a broad range of PCIe® Gen4 SSD form factors

Micron® memory

- Multiple server memory configuration options support both training and inference
- Fully validated DRAM multi-access edge computing modules thrive in challenging remote edge architectures
- Server and client DRAM flexible memory support V-RAN, O-RAN networks, and other virtualized workloads
- Micron® DDR5 Server DRAM feeds rapidly growing processor core counts with memory bandwidth and capacity, all while enabling 2x the data rates of DDR4[8]
- Micron® DDR4 Server DRAM is a mainstream DRAM leader, delivering power savings, performance enhancement and density

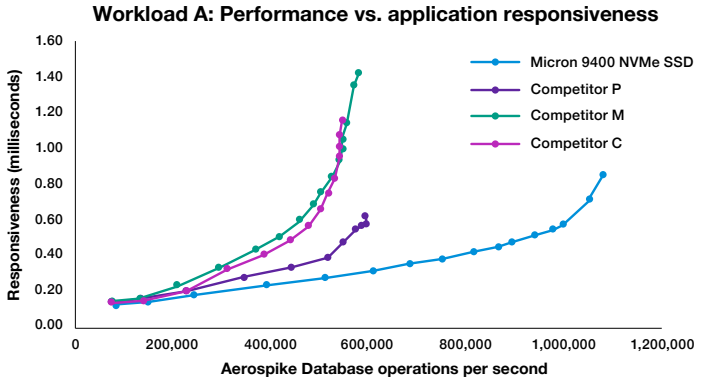

Figure 3: The Micron 9400 vs. competitors tested on Aerospike in an update heavy workload; the flatter curve for the 9400 shows the SSD application responsiveness remains consistent as demand increases.

Partner with a true memory manufacturer.

Talk with our experts about strategic, tactical, and logistic edge AI business objectives. For example, what specific problems can be solved? What will the data chain of custody be across the network? What specific endpoint compute needs can be met across the network? Extract maximum business value from edge AI deployments with our proven guidance.

Micron memory and storage innovation have been foundational for the successful integration of AI and ML. We are one of three global DRAM manufacturers and a leader in NAND storage. Our nearly 45 years of memory and storage experience and vertical integration give partners like you access to unparalleled innovation that spans from the cloud to the intelligent edge.

For more information on edge AI, visit Micron cpg.com/datacenteredge

Sources:

1. From the "Edge AI Processor Market Report," May 2022 by Research DIVE. More at https://www.researchdive.com/8514/edge-ai-processor-market.
2. Gartner, "The 2022 Market Guide for AIOps Platforms," May 2022, more at https://www.gartner.com/en/documents/4015085.
3. Micron 9400 SSD: 94,118 4K random read IOPS/watt vs 53,100 IOPS/watt for prior generation Micron 9300 NVMe SSD.
4. 30.72TB capacity is the highest U.2/U.3, SSD capacity available on the open market at the time of this document's initial publication. Unformatted capacity. 1GB = 1 billion bytes, formatted capacity is less.
5. Comparisons are made based on other leading PCIe Gen4 data center U.2/U.3 NVMe SSDs based on data center market share as noted in the Forward Insights SSD Supplier Status Q2/22 report and available on the open market at the time of this document's initial publication. 1GB = 1 billion bytes, formatted capacity is less.
6. Source: https://aerospike.com/resources/solution-brief/aerospike-real-time-data-platform/#:~:text=The%20Aerospike%20Real%2Dtime%20Data%20Platform%20enables%20organizations%20to%20act,the%20smallest%20possible%20server%20footprint.
7. Up to queue depth = 64 for 4KB, 100% random, 90% read workload and up to queue depth = 32 for 4KB, 100% random, 70% read workload.
8. Under memory-intensive workloads, DDR5 is designed to deliver 1.87x the bandwidth as a result of double burst length, double the banks and bank groups, and significantly higher speed than DDR4, as established by JEDEC, an independent organization that develops open standards for the microelectronics industry.