Keep hungry GPUs fed with lightning-fast performance

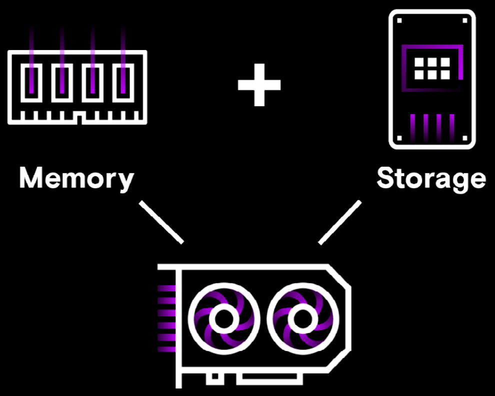
Tips to get the most from your GPUs

Solve memory and storage challenges to avoid under- or over-utilizing costly GPUs

You're investing a lot of resources into GPUs, and for good reason. GPUs are key for AI, ML, GNN, and other innovations that are forever changing the ways we use data. As with any large investment, you want to make sure you're getting the best returns you possibly can. You need to look at your system and ask, "Am I getting the most from my GPUs?"

- Under-utilizing GPUs means you are wasting resources. GPUs are sitting idle when they should be hard at work. In this case, you aren't getting the returns you deserve from your investment, and you're wasting power, space, and potential performance.
- Over-utilizing GPUs occurs when you push your GPUs too hard. Your power consumption increases, the GPUs are at risk of damage from overheating, and your performance suffers as you encounter bottlenecks, slow processing times, and reduced efficiency.

Memory and storage play a critical role in achieving optimal GPU usage, especially for resource-intensive tasks like AI, ML, and GNN. In these situations, high-performance memory and storage can make all the difference.



Right-size your storage for AI workloads

AI is dependent on fast, scalable GPU architectures, but right-sizing the storage infrastructure can be equally important to achieving your goals. By choosing the right-size storage, you can ensure there is less GPU idle time while significantly reducing power consumption, allowing for larger AI deployment within a given power budget.

The Micron 9550 SSD is a breakthrough, high-performance storage device that offers strong performance, latency, and power efficiency to keep GPUs running at optimal levels during demanding data center workloads.

- Optimized for AI and other high-performance applications
- Significantly reduces power consumption1
- Micron-designed controller ASIC, 8th-generation NAND, and DRAM

Use NVIDIA Magnum IO GPUDirect Storage to create a direct route between GPUs and SSDs

Creating a direct path between GPUs and SSDs[2] is a method to reduce latency and improve data transfer speeds between storage and GPUs. One popular way to accomplish this task is by using NVIDIA Magnum IO GPUDirect Storage.

With GPUDirect Storage, the performance speed of SSDs has a major impact. Since the GPU can pull data directly from the SSD, it's important to have high IOPS and throughput to ensure the GPU doesn't become idle while waiting for data.

The Micron 9550 NVMe SSD is an ideal fit for this role[3], since its powerful performance translates to better GPU utilization, especially for workloads like AI, ML, and GNN.

Compared to other high-performance SSDs, the Micron 9550 provides a dramatic surge in system bandwidth and a significant reduction in power used across different training workloads using GPUDirect Storage[4].

34%
faster throughput

76%
better power efficiency

81%
less energy used

33%
faster training workload completion

60%
higher SSD performance

29%
less power used by training system

Make NVMe storage a third tier of "slow" memory

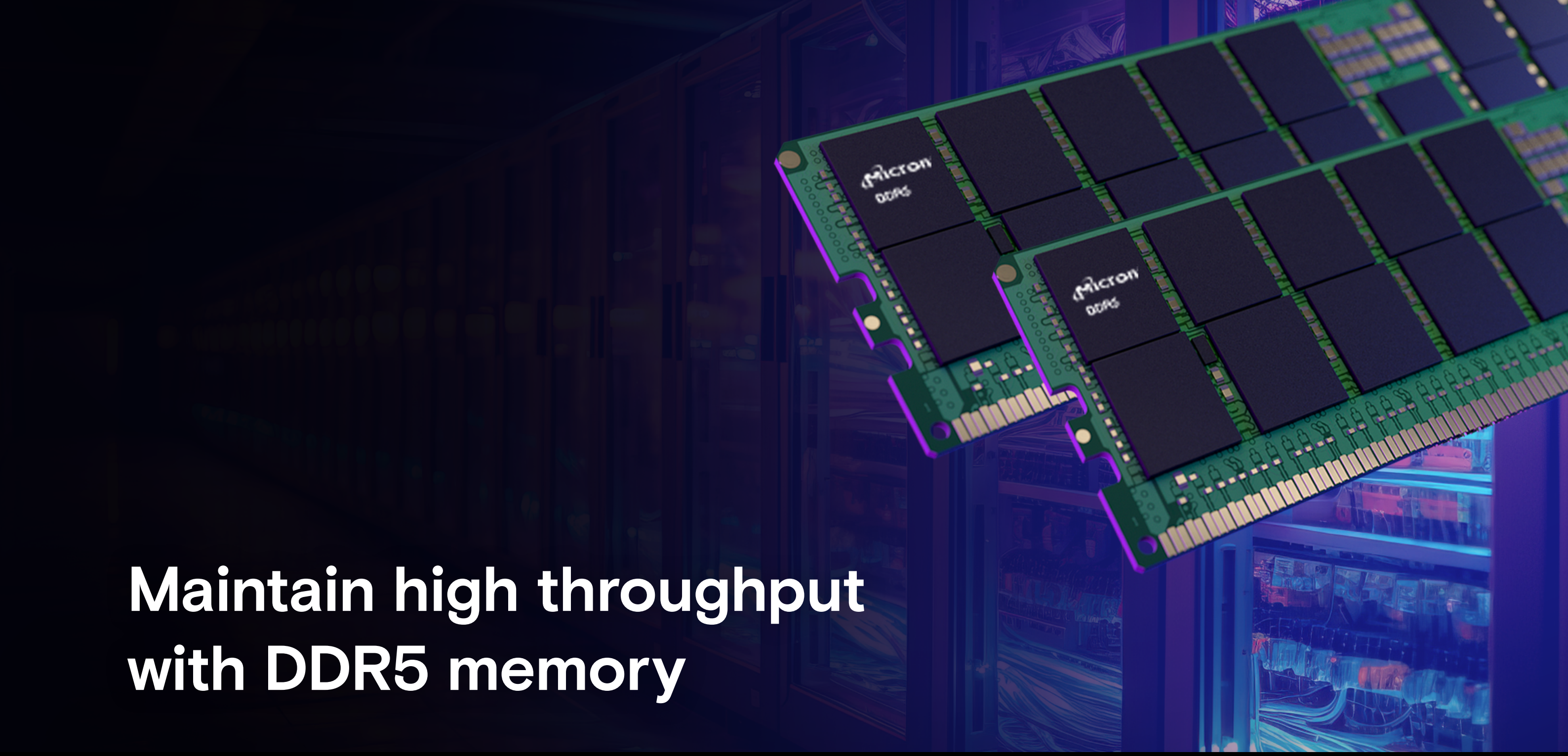
One technique for training large models is to have as much high-bandwidth memory as possible on the GPU, along with as much system DRAM as possible. If a model doesn't fit in this HBM + DRAM, it can be parallelized over multiple GPU systems.

However, there is a heavy cost to parallelized training over multiple servers: lower GPU utilization and decreased efficiency due to data flow over network and system links. These can easily become bottlenecks.

To overcome these issues, NVMe storage can be used as a third tier of "slow" memory. This can be achieved using Big accelerator Memory (BAM) with GPU Initiated Direct Storage (GIDS) to relace and streamline the NVMe driver, handling data and control paths to the GPU.

The BaM software stack relies on the low latency, high throughput, large density, and high endurance of NVMe SSDs as a memory extension. These requirements make the Micron 9550 NVMe SSD a solid choice.

BaM and GIDS testing shows the Micron NVMe SSD enables faster GNN training, demonstrated higher SSD performance, used less system power, and provided strong scaling results[5].



Maintain high throughput with DDR5 memory

DDR5 delivers higher bandwidths along with improved reliability, availability, and scaling when compared to DDR4[6]. It keeps data flowing smoothly so GPUs can stay fed and run at peak efficiency.

For AI/ML workloads, DDR5 is essential to reach full server potential. Test data[7] shows DDR5 offers massive performance improvements over DDR4 in these types of workloads.

Image classification

- 7.3x gain in classifying images
- 40% higher sustained memory bandwidth

Recommendation throughput

- 4.3x gain in recommendation throughput
- 200% gain in memory bandwidth for recommendations

Natural language

- 4.9x gain in natural language processing
- 55% higher memory bandwidth, resulting in higher throughput

Get expert advice about how to optimize your GPU usage

We work closely with customers at engineering sites across the world to streamline processes and reduce the load on your engineering teams. Micron's experts rigorously test server architecture for purpose-built solutions that keep GPUs running at optimal levels while reducing power consumption and improving overall efficiency.

Learn more at micronc pg.com/datacenter

1. Based on Micron engineering test results in AI training offload, measured SSD-to-GPU direct data transfer rate with a 1TB dataset, and standard AI performance benchmarks.
2. See https://developer.nvidia.com/gpudirect-storage for additional details on the IO path differences.
3. See the NVIDIA GPUDirect Storage Overview Guide for additional information on GDS.
4. Micron internal engineering analysis of AI training workloads shows that different IO sizes are seen depending on model and data formats. Therefore, this document focuses on small (4KB), medium (128KB), and large (1MB) transfer sizes in two test scenarios.
5. Values are maximums observed during testing. Competitive PCIe Gen5 SSDs chosen from the top 10 PCIe SSD suppliers shown in the Forward Insights analyst report “SSD Supplier Status Q1/24 May 2024.”
6. DDR5 launch data rate of 6400MT/s transfers 2x (100%) more data than the maximum standard DDR4 data rate of 3200MT/s. JEDEC projected speeds of 8800MT/s are 2.75x faster than DDR4’s maximum standard data rate of 3200MT/s.
7. Micron’s Data Center Workload Engineering (DCWE) team performed testing and validation in collaboration with Supermicro and Intel to determine an ideal CPU-powered platform optimized for AI inference workloads. Workload tests performed by Micron focused on MLPerf (Machine Learning Performance) inference benchmarking, which measures how fast systems run models in a deployment scenario that includes NLP using BERT (Bidirectional Encoder Representations from Transformers); DLRM (Deep Learning Recommendation Model); and Image classification using ResNet. Actual results may vary. Learn more: Micron Server DDR5 AI Use Case Test Results eBook (EN) (microncpg.com)