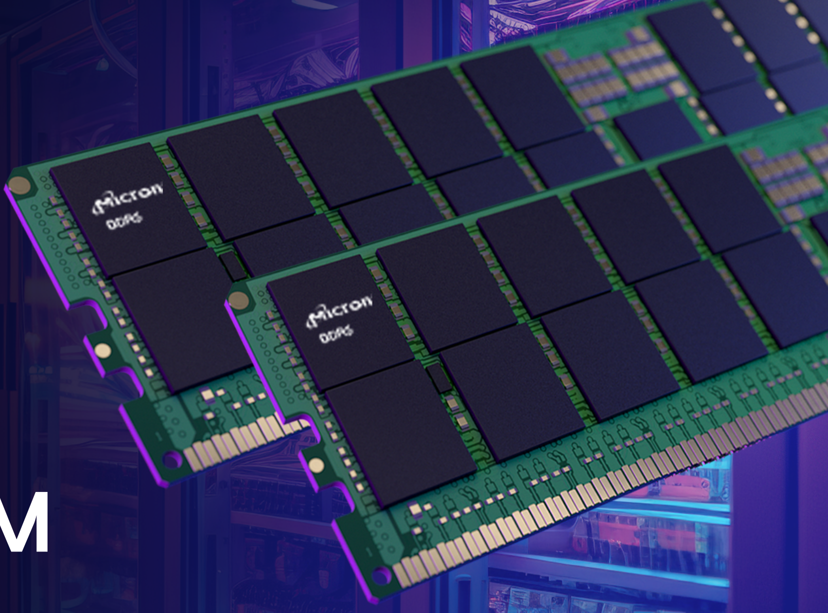


Micron® DDR4 and DDR5 Server DRAM

Choose the best DRAM for your data center workload

Data center managers often face the challenge of deciding between the performance benefits of DDR5 and the cost-effectiveness of DDR4.

DDR5 represents a significant leap forward compared to the performance of DDR4. However, transitioning from DDR4 to DDR5 can be expensive, since it often necessitates an entirely new server system with a motherboard designed to support DDR5.

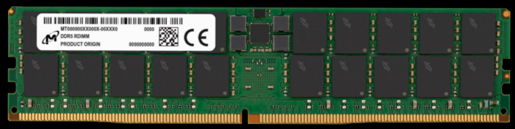
In this tech brief, we explore several common scenarios and examine how selecting the appropriate memory solution can improve server performance, translating to better real-world results.

DDR5

- Classifying images
- Recommendation throughput
- Natural language processing (NLP)

DDR4

- MS SQL
- Apache Spark



Micron DDR5 Server Memory

Micron DDR5 Server Memory delivers higher bandwidths along with improved reliability, availability, and scaling, when compared[1] to DDR4. It's ideal for resource-intensive tasks like AI and HPC.

- Speed (MT/s): 4800, 5600
- Densities: 16GB, 24GB, 32GB, 48GB, 64GB, 96GB, 128GB
- Form factor: RDIMM, ECC UDIMM, ECC SODIMM

Micron DDR4 Server Memory

With module densities up to 128GB, Micron DDR4 Server DRAM lets you maximize system performance by increasing the installed memory capacity of each server. This can provide a significant performance boost for mainstream applications and legacy servers.

- Speed (MT/s): 3200
- Densities: 8GB, 16GB, 32GB, 64GB, 128GB
- Form factor: RDIMM, LRDIMM, ECC UDIMM, ECC SODIMM

Improved image classification

Image classification is a critical task for many modern technologies. For example, it allows autonomous vehicles to recognize objects around them and helps generative AI recognize the photos it uses as a reference.

With DDR5, it's possible for image classification systems to annotate and label samples more than seven times faster[2] than DDR4.

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

Figure 1: ResNet interferencing throughput comparison – SUT1 vs. SUT2 exhibits 7x gain throughput for ResNet

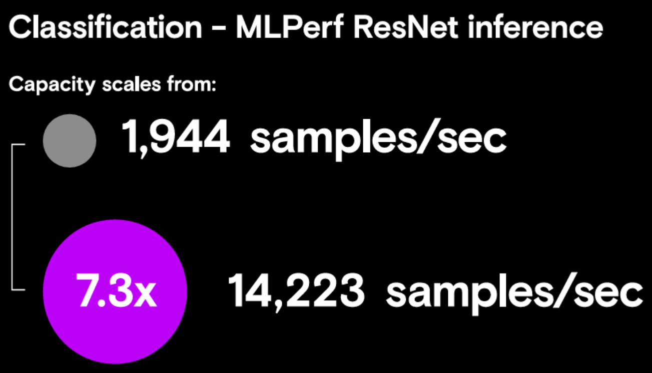
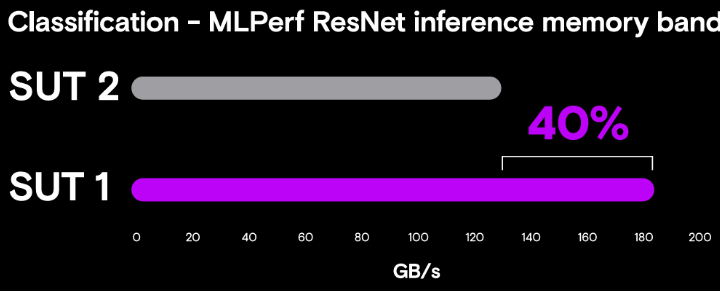


Figure 2: ResNet interferencing bandwidth comparison – SUT1 vs. SUT2 provides 40% gain in memory bandwidth for computer vision



Use case: Content moderation for a social media platform

Social media platforms receive an enormous volume of user-generated content daily, including images and videos. DDR5 can help improve the processes used to analyze uploaded images and categorize them based on their content.

- Faster processing to handle image data more swiftly
- Improved accuracy by training on larger datasets
- Scalability for the requirements of a fast-growing platform

Better recommendation throughput

The internet has become highly personalized to provide content relevant to the individual user. Personalized recommendations drive search results, streaming recommendations, product recommendations, and more.

Using DDR5, businesses can improve personalization by speeding predictions and forecasting to discover more effective cross-selling strategies. Tests show[2] that it can more than quadruple the recommendation throughput of DDR4.

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

Figure 3: DLRM interferencing throughput comparison – SUT1 vs. SUT2 delivers 4.3x gain in throughput for DLRM

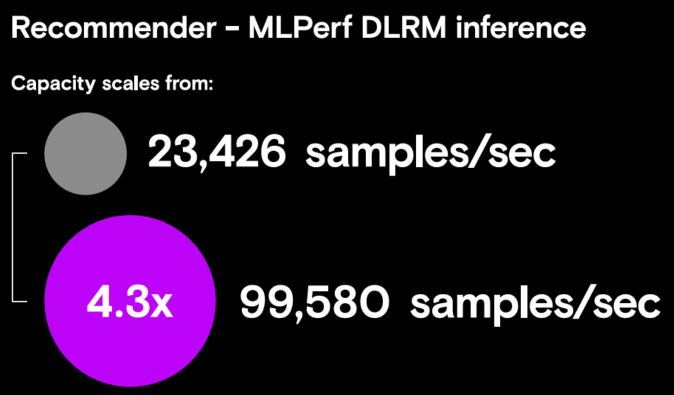
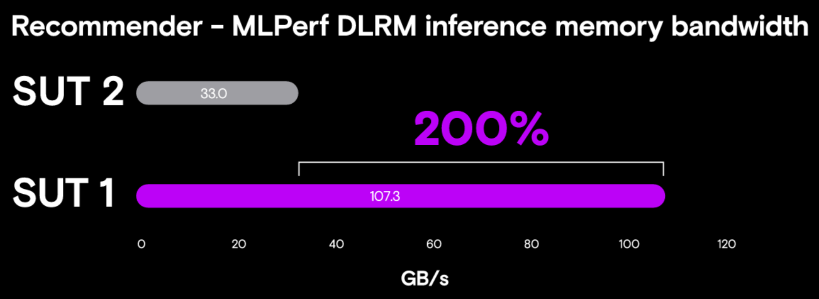


Figure 4: DLRM interferencing bandwidth comparison – SUT1 vs. SUT2 provides 200% gain in memory bandwidth for recommendation



Use case: Personalized recommendations for e-commerce

E-commerce platforms thrive on personalized recommendations. Whether it's suggesting the perfect pair of sneakers or a binge-worthy TV series, these tailored suggestions drive user engagement and sales.

- Higher throughput for faster responses times and a smoother user experience
- Increased memory bandwidth improves the efficiency of data retrieval
- More samples inferenced per second to accommodate a larger user base

Faster natural language processing

Natural language processing (NLP) is the foundation that enables digital assistants to exist. It not only brings them to life but also enhances their accuracy. Accelerating NLP has a direct impact on daily usability, making interactions smoother and more efficient.

DDR5 greatly improves[2] NLP of questions and answers on blocks of text. It can handle nearly five times as much NLP as DDR4.

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

Figure 5: BERT interferencing throughput comparison – SUT1 vs. SUT2 delivers 4.9x gain in throughput for NLP

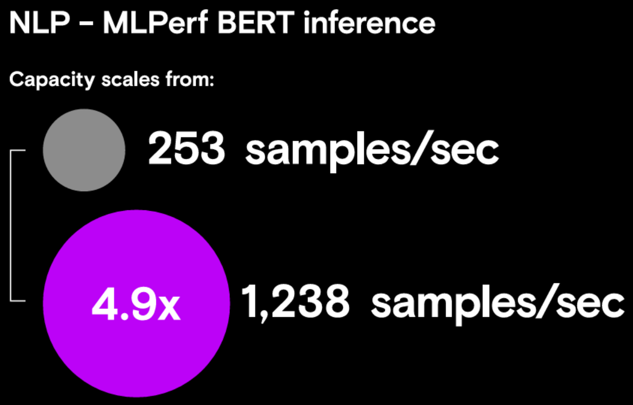
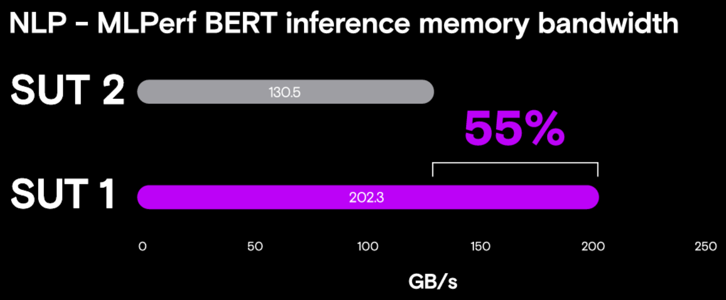


Figure 6: BERT interferencing bandwidth comparison – SUT1 vs. SUT2 provides 55% gain in memory bandwidth for NLP



Use case: Customer support chatbots

Data centers deploy chatbots powered by NLP to handle routine inquiries, troubleshoot issues, and provide instant responses. These chatbots can assist IT support teams and improve user experience.

- Faster NLP results in near-instantaneous answers
- More memory bandwidth allows chatbots to retrieve relevant information faster, improving their accuracy
- Improved performance on large chunks of text to efficiently handle lengthy user inputs

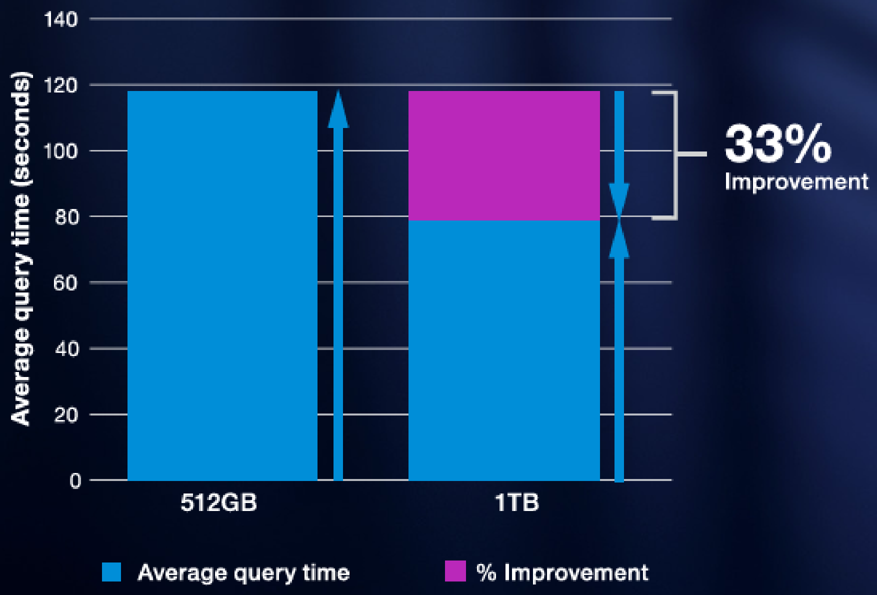
Improve MS SQL performance by adding more capacity

While DDR5 can provide a sizeable performance upgrade, it's worth exploring how servers can still use DDR4 to enhance existing server infrastructure. By simply increasing DDR4 capacity (while maintaining the same memory channels) or adding more capacity with additional channels, you can significantly improve workload performance for data analytics applications like Microsoft® SQL/TPC-H.

For instance, consider the popular TPC-H data analytics workload using the MS SQL database. When the available memory is doubled from 512GB to 1TB (distributed across two sockets) using DDR4, servers experience substantial benefits[3] to the workload.

- 2.2x performance increase
- 33% improvement in average query time

Figure 7: Apache Spark™ TPC-DS
DDR4 capacity expansion benchmark (lower is better)



Use case: Data warehousing

Data centers can aggregate data from various sources (sales, marketing, customer interactions) into a single MS SQL data warehouse for business intelligence. By expanding their DDR4 capacity, they can more efficiently store, process and analyze large volumes of data.

- Increased memory capacity for scalable data warehouses
- Faster query times reduce latency and leads to quicker data retrieval
- Cost-efficient solution that balances performance and affordability

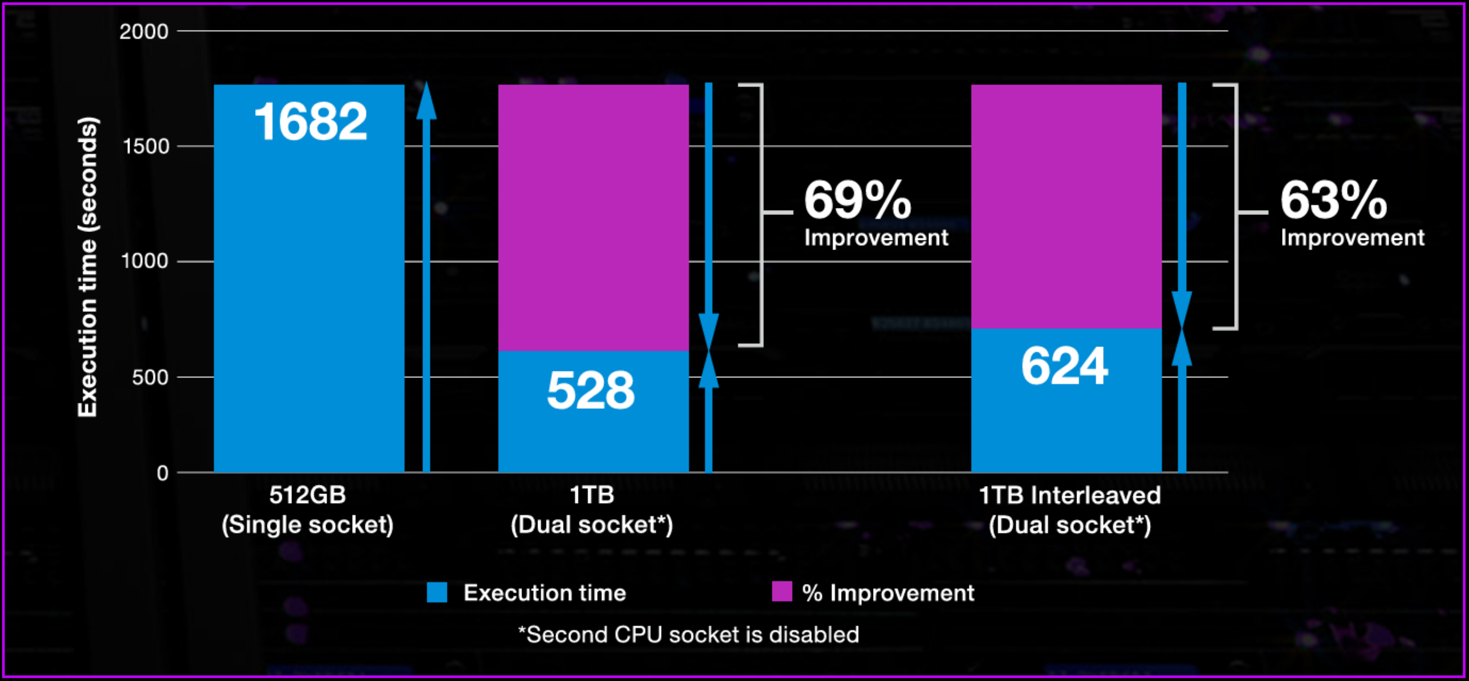
Accelerate Apache Spark™ SVM execution

Support Vector Machine (SVM) is a powerful supervised algorithm used for data classification, pre-processing, and anomaly detection. It identifies a hyperplane that optimally separates input data into distinct classes.

Benchmarks have shown[3] that by increasing the server's memory capacity from 512GB (8x64GB DIMMs) to 1TB (16x64GB DIMMs) with Micron DDR4 server Server DRAM, Spark's SVM execution improved dramatically.

- 3.2x performance increase
- Up to 69% improvement in execution time

Figure 8: Apache Spark™ SVM (360GB input)
DDR4 capacity expansion benchmark (lower is better)



*Second CPU socket is disabled

Use case: Real-time fraud detection

Financial institutions, such as banks and credit card companies, process millions of transactions daily. They can use Apache Spark with SVM to flag transaction anomalies and help with real-time fraud detection. Increasing the amount of DDR4 can improve fraud detection systems to help financial institutions — and their customers — stay safe.

- Increased performance allows Apache Spark with SVM to process transactions swiftly, minimizing the window for fraudulent activities
- Improved execution times lead to quicker fraud identification
- Higher capacity enables more concurrent SVM model executions, improving overall system responsiveness

Micron: Your memory solution

For over 45 years, Micron has led the way in memory and storage solutions. Whether you're upgrading existing systems or building new servers, our experts can help you find the right memory solution and capacity.

We rigorously test configurations across diverse platforms and workloads to innovative solutions to for complex challenges. Our expert insights and test data can give you the tools you need to upgrade and optimize your system, no matter your workload.

Discover your perfect memory solution at microncpg.com/memory

Footnotes:

1. DDR5 launch data rate of 4800MT/s transfers 1.5x (50%) 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.

2. 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)

3. Based on Micron internal benchmarks. Actual results may vary.
   Learn more: Micron DDR4 Server DRAM Campaign Maximize IT Infrastructure eBook (EN) (microncpg.com)