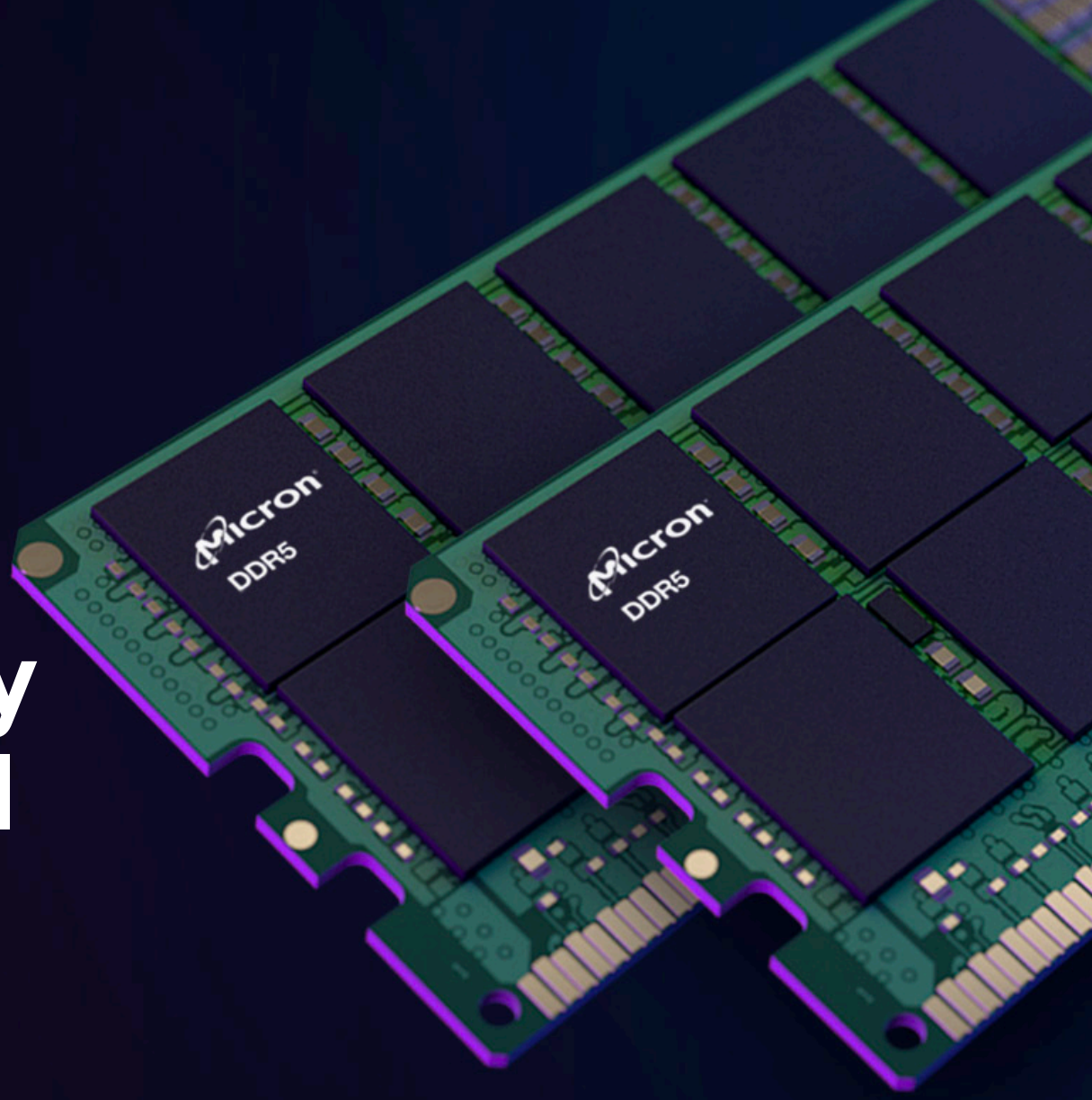
Advanced server memory is foundational for AI

Radically increase AI performance with Micron® DDR5 Server DRAM



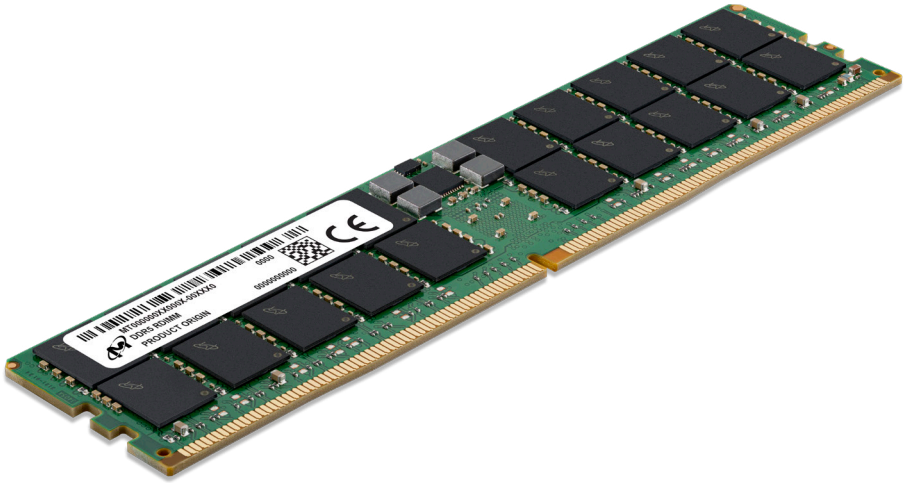
You can't have an AI revolution without advanced memory. Micron® DDR5 Server Memory is built for such intensive workloads. And we've proven it with recent testing[1] completed on three fundamental AI use cases:

- **Computer vision**
- **Predictions/forecasting** (entity extraction, topic detection, text summarization, document classification, etc)
- **Natural language processing**

Stop and think for a moment about how ubiquitous these use cases are in everyday AI.

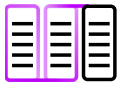
- **Classifying images** makes it possible for autonomous vehicles to recognize objects around them and allows generative AI to recognize photos it uses as a reference.
- **Personalized recommendations** drive search results, streaming recommendations, product recommendations, etc.
- **Natural language processing** makes digital assistants possible in the first place and also makes them more accurate. Speeding up natural language processing increases ease of use on a daily basis.

We know DDR5 is a game changer in these regards. Speeding up AI means speeding up its uses.



Best for

Artificial intelligence



Data mining

Predictive analytics

Intensive simulations

In tests detailed in this ebook, we'll show how much faster the DDR5 server is when employed with the 4th Generation Intel® Xeon® 8490 H Processor vs. the 3rd Generation Intel Xeon 8380 Processor[1]. It's natural to see an increase in this situation, but the numbers gathered are exceptional and can make a big difference in performance.

Keep reading to find out how to make your AI function as successfully as possible.

7x faster classifying images
Image classification used in computer vision is capable of annotating and labeling more than 14,000 samples inferenced per second.

4.3x faster personalized recommendations
Predictions and forecasting discover and predict data trends faster for more effective cross-selling strategies with more than 99,000 samples inferenced per second.

4.9x faster natural language processing
NLP of questions and answers on blocks of text is greatly improved with more than 1,200 samples inferenced per second.

AI inference workloads are rapidly growing in the data center, requiring significant computational and memory subsystem resources. The advent of Micron® DDR5-capable systems and 4th Gen Intel® Xeon® processors, such as the Supermicro Hyper SuperServer™, provides the necessary compute power, memory bandwidth and density for AI applications.

Micron DDR5-4800 offers a 50% increase in data rates compared to DDR4-3200 at the component level. In addition to the increased data rates, Micron DDR5 adds two times the bank groups, burst length (BL16), and improved refresh schemes to deliver much higher effective bandwidth than DDR4-3200, beyond what is enabled by the higher data rate alone. The latest 4th Gen Intel Xeon 8490H CPU increases the core count by 50% compared to 3rd Gen Intel Xeon 8380 CPU and improves the cache architecture (i.e., speed and density) to boost performance for AI inference. To fuel the CPU core counts, Micron DDR5 increased the burst length that enables two independent channels per DIMM, effectively doubling the available memory channels on the server platform for more concurrent operations.

Combined, Micron DDR5 and 4th Gen Intel Xeon 8490H CPU offer a theoretical maximum memory bandwidth of 614GB/s (at a DDR5 speed of 4,800MT/s) compared to a DDR4-based system offering 410GB/s (at the speed of 3,200MT/s), which is 50% higher for dual-socket systems.

Testing parameters

Understanding MLPerf Workload Benchmarking for AI Inference

Standard application workload benchmarks are used across the industry to evaluate a collection of programs with a goal of comparing both hardware and software system performance. MLPerf — Machine Learning Performance, developed by MLCommons, a consortium of AI leaders from academia, research labs, and industry — is an independent objective performance benchmark that evaluates software frameworks, hardware platforms and cloud platforms for machine learning models. The MLPerf benchmark suite allows developers to evaluate architectures for AI training and inference for ideal deployments. Micron performed workload tests focused on MLPerf inference benchmarking, which measures how fast the system runs models in a deployment scenario that includes:

- **NLP using BERT** (Bidirectional Encoder Representations from Transformers) that allows for language-related use cases for text relationships, questions and answering, sentence paraphrasing and others.
- **Recommendation using DLRM** (Deep Learning Recommendation Model) creates personalized results for user-facing services such as social media, online shopping and content streaming.
- **Image classification using ResNet** that can categorize and label images from computer vision or fixed image use cases.

To choose the best CPU-based AI inference platform, it is important to understand how platform resources are exercised heaviest when running the workloads (benchmarks). This helps the system scale and maintain performance for given application use case scenarios.

System testing configuration and analysis

Micron's Data Center Workload Engineering (DCWE) team performed the testing and validation in collaboration with Supermicro and Intel to determine an ideal CPU-powered platform optimized for AI inference workloads. The following is the configuration for the systems under test (SUT) used throughout the extended testing.

System under test 1 (SUT 1):

- **Supermicro Hyper SuperServer**
- **Dual CPU** – Intel Xeon 8490H-60C with Intel AMX
- **Memory** – 8x DDR5-4800 64GB RDIMM per CPU socket
  1TB total memory
- **OS** – Alma Linux 9 (kernel 5.14)

System under test 2 (SUT 2):

- **DDR4 server platform**
- **Dual CPU** – Intel Xeon 8380-40C
- **Memory** – 8x DDR4-3200 64GB RDIMM per CPU socket
  1TB total memory
- **OS** –Alma Linux 9 (kernel 5.14)

Both DDR4 and DDR5 capable systems were configured by setting the same software stack to Alma Linux 9 (kernel 5.14). The systems were populated with 64GB RDIMMs in a one-DIMM-per-channel (1DPC) configuration for both DDR4 and DDR5 systems for a total of 1TB.

7.3x gain in classifying images[1]

MLPerf uses a ResNet® model with the ImageNet dataset that is a representation of millions of annotated images with labels and bounding boxes for image classification and computer vision. Observed results for image classification benchmark shows 7x throughput gain for the number of samples inferenced per second on SUT 1 over SUT 2 (Figure 1). In addition to higher throughput gain, SUT 1 also achieves 40% higher sustained memory bandwidth over SUT 2 as shown in Figure 2.

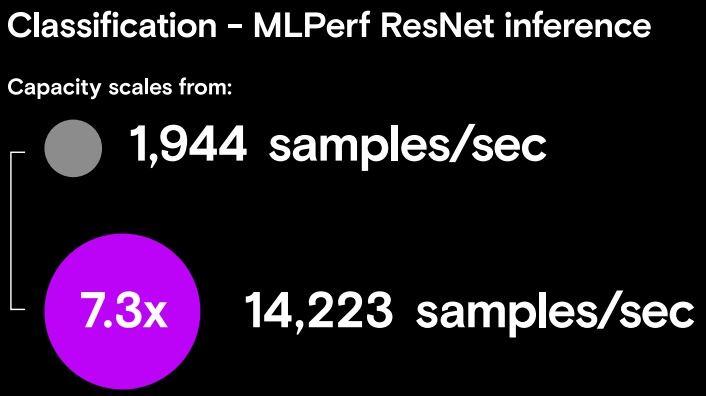


Figure 1: ResNet inferencing throughput comparison — SUT 1 vs SUT 2 exhibits 7x gain throughput for ResNet.

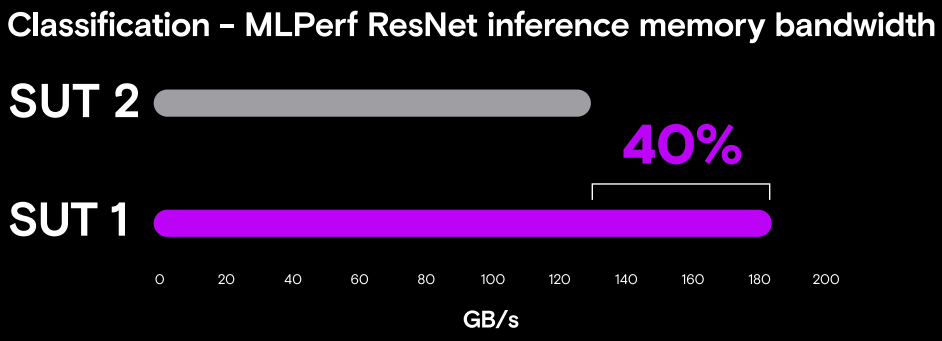


Figure 2: ResNet inferencing bandwidth comparison — SUT 1 vs SUT 2 provides 40% gain in memory bandwidth for Computer Vision.

4.3x gain in recommendation throughput[1]

MLPerf Recommendation using DLRM with Criteo 1TB click log dataset is used as a benchmark for click-through-rate (CTR) prediction for online shopping, content ranking and social media platforms. The dataset contains click logs of 4 billion user and item interactions over a 24-hour period. The performance results observed show a 4.3x gain in throughput in samples inferenced per second on SUT 1 (8490H/DDR5) over SUT 2 (8380/DDR4). The results in Figure 4 also show the average memory bandwidth was increased by 200% on SUT 1 with 107GB/s when compared to SUT 2 at 33GB/s.

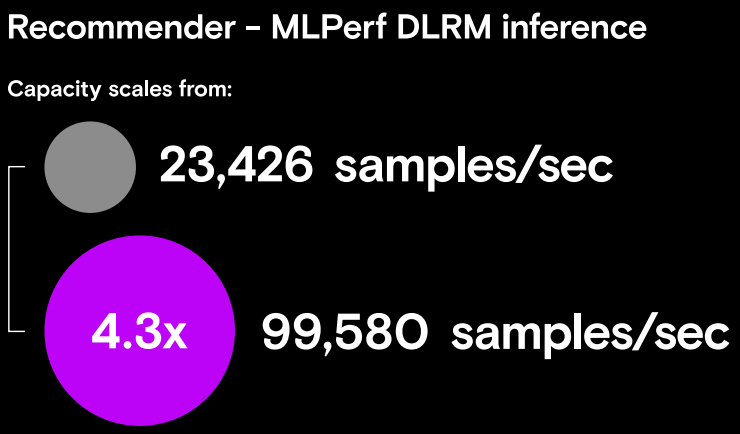


Figure 3: DLRM inferencing throughput comparison — SUT 1 vs SUT 2 delivers 4.3x gain in throughput for DLRM.

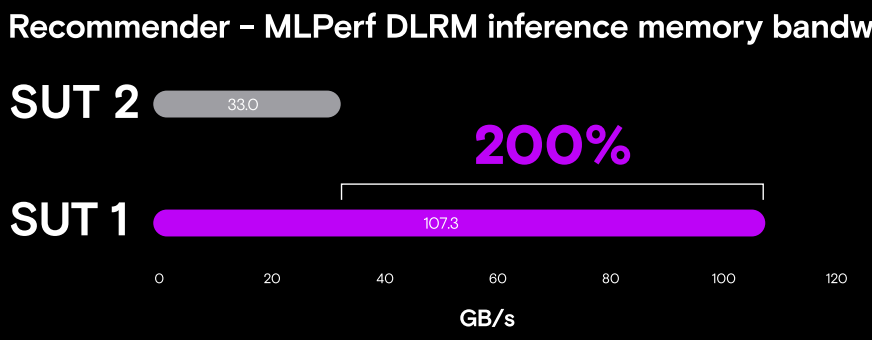


Figure 4: DLRM inferencing bandwidth comparison — SUT 1 vs SUT 2 provides 200% gain in memory bandwidth for recommendation.

4.9x gain in natural language processing[1]

MLPerf inference using BERT with SQuAD dataset tests a model's ability to read a passage of text and then answer questions about it. SQuAD has 100,000+ questions on 500+ articles. There is 4.9x gain in throughput which is the number of samples inferenced per second on SUT 1 (8490H/DDR5) when compared to SUT 2 (8380/DDR4) shown in Figure 5. Other observations indicate that there is an approximate 6-7% decrease in throughput on SUT 2 as the batch size increases.

The model tasks are run in accuracy mode over the entire dataset and the results show the average memory bandwidth recorded on SUT 1 is 202GB/s, shown in Figure 6. This equates to SUT 1 reaching 39% of the theoretical memory bandwidth of 614.4GB/s. SUT 1 achieves 55% higher memory bandwidth compared to SUT 2, resulting in a higher throughput.

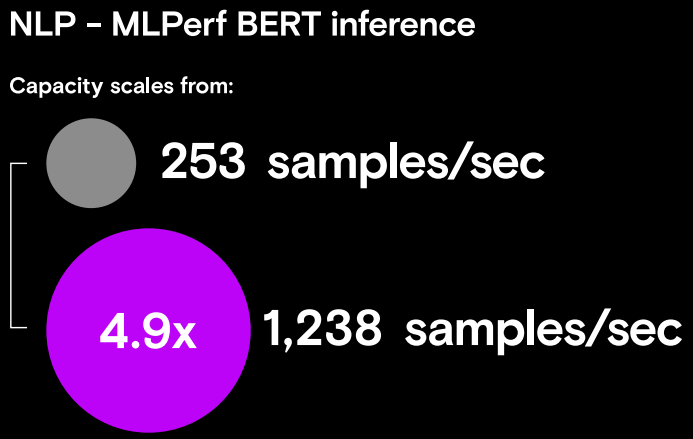


Figure 5: BERT inferencing throughput comparison — SUT 1 vs SUT 2 delivers 4.9x gain in throughput for NLP.

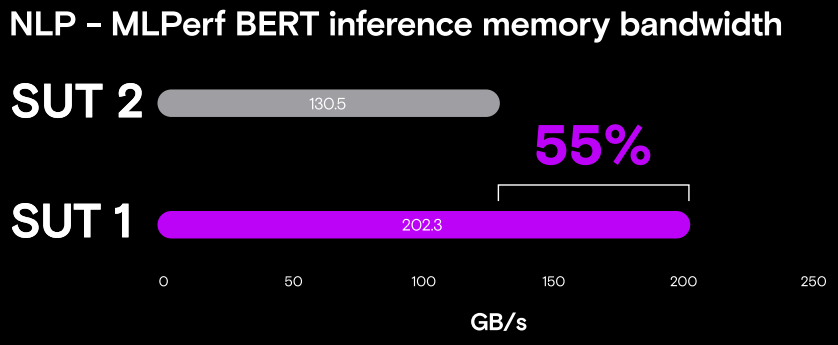


Figure 6: BERT inferencing bandwidth comparison — SUT 1 vs SUT 2 provides 55% memory bandwidth gain for NLP.

How these systems will scale

It is also important to understand how these systems will scale by running more model instances in parallel[1]. BERT inference can run with multiple instances that take advantage of multi-socket, multi-core systems to improve overall efficiency. In this test case scenario, the number of BERT instances is increased to run on each system concurrently while maintaining performance. The total number of cores is equally distributed among all BERT instances executed in the specific system. Despite both the systems SUT 1 and SUT 2 having same capacity of memory, SUT 2 with Micron DDR4 could not support more than four instances, whereas SUT 1 with Micron DDR5 can provide consistent throughput and can scale to more than 24 instances (Figure 7). Figure 8 shows the 30% memory bandwidth gain by SUT 1 over SUT 2. The L3 cache miss rate is also higher on SUT 2 compared to SUT 1, which has 4x the cache size. So, L3 cache capacity and DDR5 speed both support better performance in SUT 1. Effectively, Micron DDR5 memory can sustain multiple instances of BERT inference workload, without compromising on throughput.

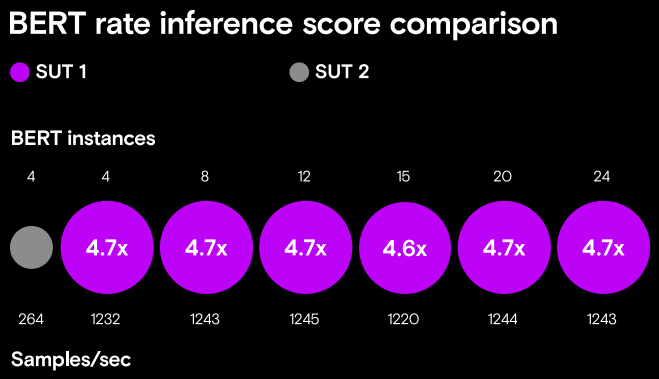


Figure 7: BERT inferencing throughput comparison — SUT 1 vs SUT 2 provides 5x gain in throughput for NLP multi-instance.

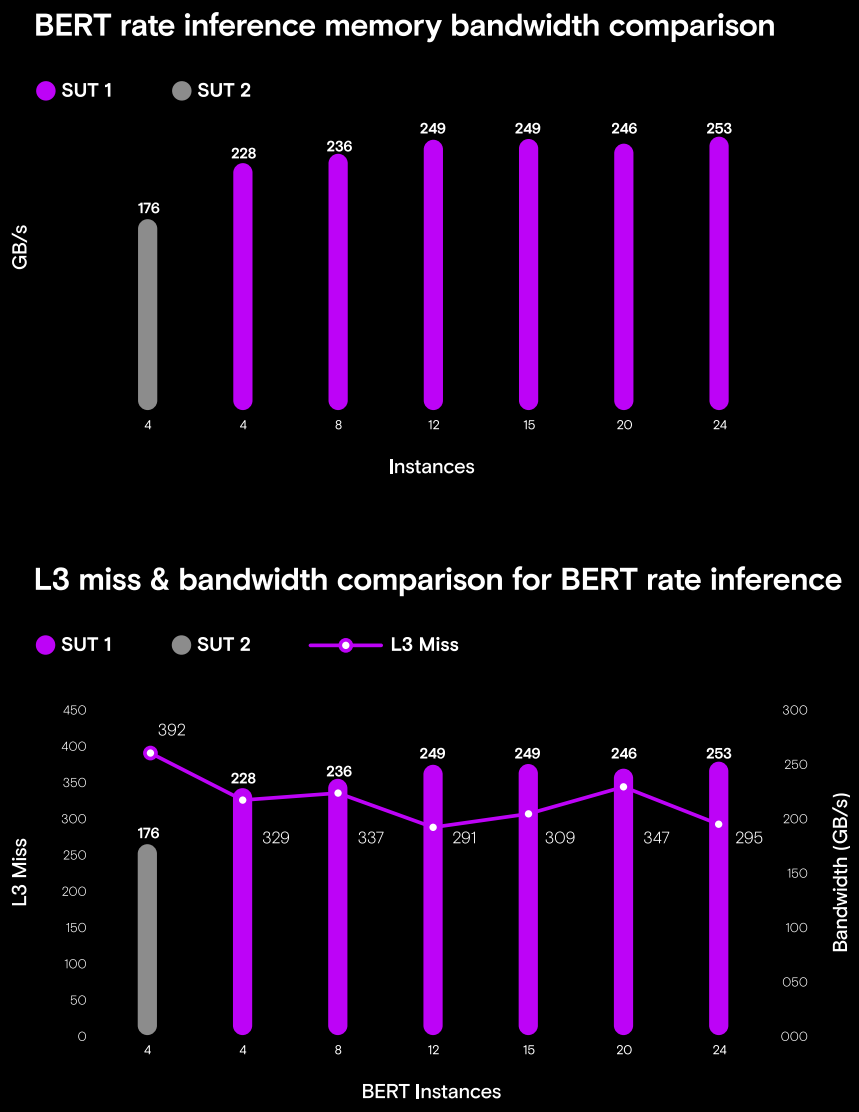


Figure 8: BERT inferencing L3 miss rate and memory bandwidth comparison SUT 1 vs SUT 2 for NLP multi-instance.

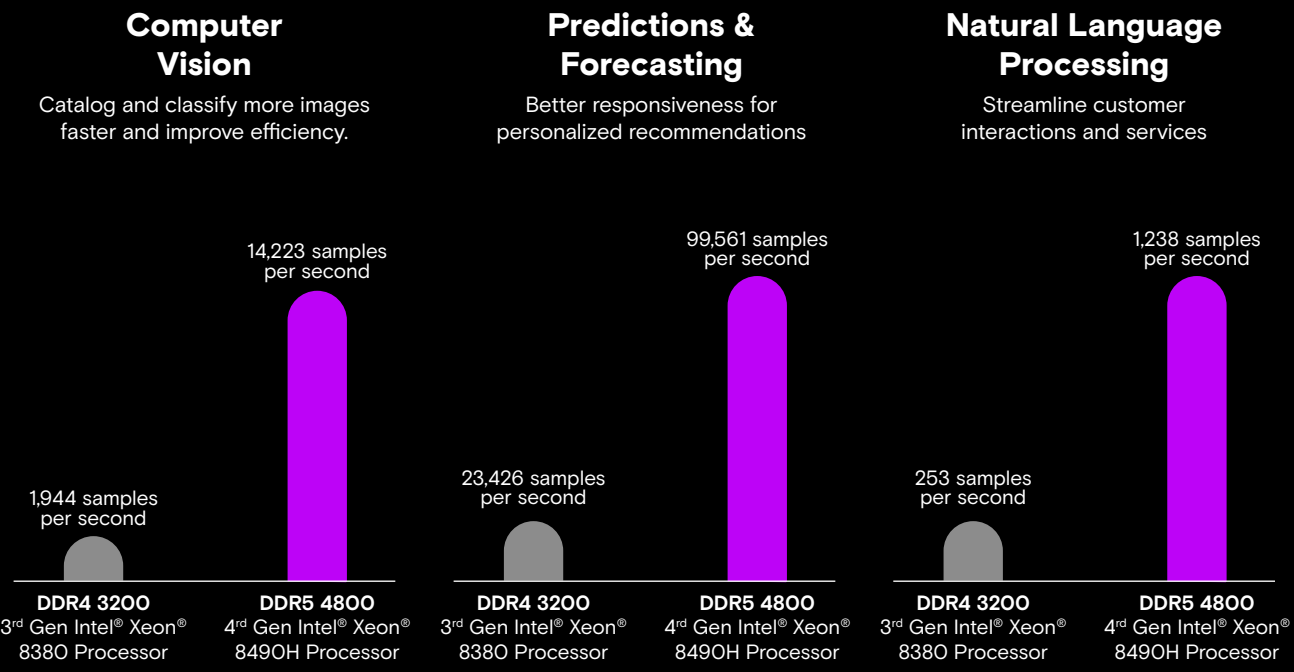
Conclusion

In-house testing reveals[1]

The results from the MLPerf benchmark testing[1] are a strong indicator that properly configured CPU-powered platforms have the horsepower to deliver the performance on optimized trained models to predict and estimate outcomes on AI inference workloads. The combination of Micron DDR5 and 4th Gen Intel Xeon 8490H CPU enabled with Intel AMX on the Supermicro Hyper SuperServer SYS-121H-TNR provides higher throughput for AI inference models, better memory bandwidth utilization and increased scalability compared to the prior generation[1].

When it comes to deploying practical AI solutions, IT administrators have the information to weigh their options for an optimized CPU-powered server to run their AI applications based on performance, scalability and costs. Micron DDR5 Server DRAM has proven it can improve the speed and usefulness of AI[1]. Developers close to this revolution ask for more out of memory — and Micron is already ahead of these growing needs. Employ Micron DDR5 Server DRAM to support your AI and gain the competitive advantage.

Learn more at **micronpg.com/serverDDR5**



MLPerf ResNet benchmarks performed on the Supermicro SYS-121H-TNR server with 4th Gen Intel® Xeon® 8490H Scalable processors with Intel® Advanced Matrix Extensions (Intel® AMX) and Micron DDR5

Sources

1. 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.

Rev. 08/2023