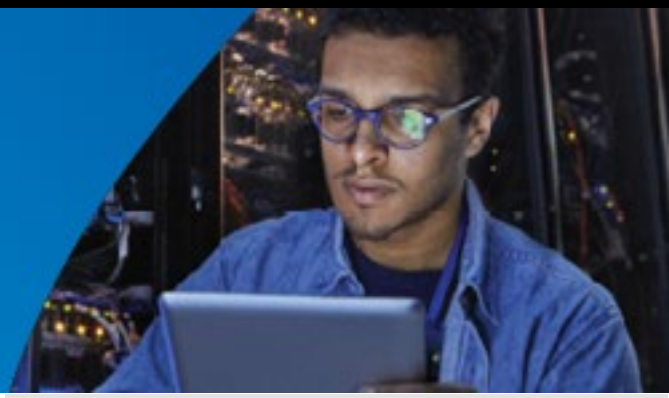
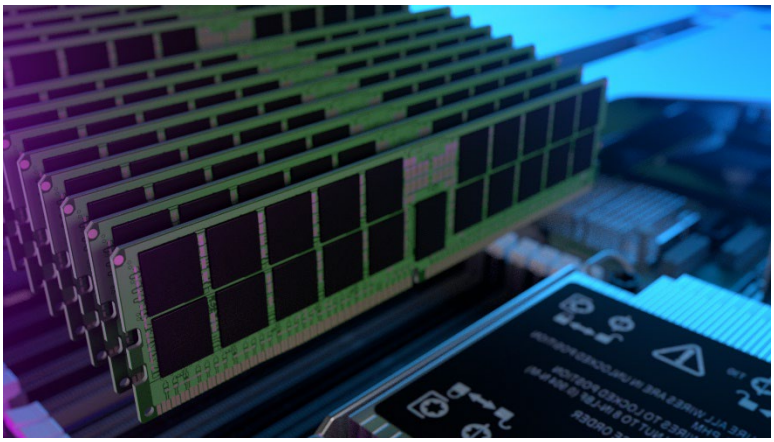
Micron DDR5: Artificial Intelligence Inference Workload Performance



Artificial intelligence (AI) inference workloads are rapidly growing in data centers, due in large part to the recent proliferation of large language models, which further augment recommendation, vision and natural language applications. Execution of these workloads requires significant compute and memory subsystem capabilities.

Our testing has shown gains as high as seven times for machine learning (ML) models when Micron DDR5 along with 4th Gen Intel® Xeon® Scalable processors are deployed in systems, which improve memory bandwidth and more instructions per cycle with silicon-based accelerators in modern CPU systems. These systems offer a potent solution for AI workloads.



Fast facts

Optimized server configurations for AI inference with Micron DDR5 and 4th Gen Intel® Xeon® Scalable processors using Advanced Matrix Extensions (AMX) provide a platform for AI inference.

Computer vision gain of 7.3x

Image classification used in computer vision is capable of annotating and labeling over 14,000 samples processed per second.

Natural Language Processing (NLP) gain of 4.9x

NLP of questions and answers on blocks of text occurs with over 1,200 samples processed per second.

Recommender system gain of 4.3x

Recommendation engines discover and predict data trends faster for more effective cross-selling strategies, with over 99,000 samples inferenced per second.

Overview

AI inference workloads are rapidly growing in the data center, requiring significant computational and memory subsystem resources. Various applications of AI inference in the form of computer vision and object detection, natural language processing (NLP), and recommendation models continue to push the boundaries of memory capabilities. AI inference uses optimized, trained models to predict and estimate outcomes and requires performance-based CPUs and memory resources with high throughput. Data centers today must grapple with the AI revolution and how best to deploy practical solutions for their applications based on performance, scale and costs.

Our tests have shown that using Micron DDR5 and 4th Gen Intel Xeon processors using Intel® Advanced Matrix Extensions (AMX), a new built-in accelerator for deep-learning, training, and inference on the CPU, provided the necessary computing power, memory bandwidth, and capacity for AI applications. Micron DDR5-4800 delivered a 2x improvement in memory bandwidth compared to DDR4-3200. 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 the 3rd Gen Intel Xeon 8380 CPU and improves the cache architecture (i.e., speed and capacity) to boost performance for AI inference. To fuel the CPU core counts, Micron DDR5 increased the burst length, enabling two independent channels per DIMM, doubling the server platform's available memory channels for more concurrent operations.

A typical enterprise and data center server architecture includes dual-socket systems with 16 memory channels. Micron DDR5 offers 50% higher theoretical maximum memory bandwidth of 614 GB/s (at a DDR5 speed of 4800 MT/s) compared to DDR4 based systems, offering 410GB/s (at the speed of 3200 MT/s) based on analysis by Micron Data Center Workload Engineering teams. The additional memory bandwidth serves to alleviate the performance bottleneck faced by AI problems and also unlock the parallelism offered by emerging server system configurations. For example, it is common for AI inferencing to run multiple instances of the machine learning model on high core count systems to effectively leverage the concurrency and serve numerous clients. In such a workload-optimized server system configuration, it is essential for the memory subsystem to be able to sustain high memory bandwidth. Micron DDR5 provides an excellent solution for such settings.

Below is a summary of the results from our Micron Data Center Workload Engineering team for throughput for the MLPerf Inference benchmark on a Micron DDR5-based system compared to a DDR4-based system:

- ResNet computer vision AI/ML inference achieves a seven times improvement and a 40% higher memory bandwidth
- BERT natural language processing (NLP) AI/ML inference achieves a 4.9 times improvement and a 55% higher memory bandwidth
- Deep learning recommendation model (DLRM) inference achieves a 4.3 times improvement and a 200% higher memory bandwidth compared to a DDR4-based system
- In addition, the AI inference models are able to sustain high memory bandwidth at increased levels of concurrency on the Micron DDR5 subsystem, serving several clients in the process

MLPerf workload benchmarking for AI inference explained

Standard application workload benchmarks are used across the industry to compare both hardware and software system performance. MLPerf 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. The workload tests performed by Micron focus on MLPerf inference benchmarking, which measures how fast systems run models in a deployment scenario that includes:

- Image classification using ResNet that can categorize and label images from computer vision or fixed image use cases,
- NLP using bidirectional encoder representations from transformers (BERT) that allow for language-related use cases for text relationships, questions and answers, sentence paraphrasing and others,
- Recommendations using deep learning recommendation model (DLRM) creates personalized results for user-facing services such as social media, online shopping, content streaming, etc.

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

System testing configuration and analysis

The testing and validation were performed by Micron to determine an ideal CPU-powered platform optimized for AI inference workloads. The following is the configuration for the systems under test.

- DDR5 System: Dual CPU with Intel Xeon 8490H-60C and 16 DDR5 Micron 64 GB RDIMMs per CPU.
- DDR4 System: Dual CPU with Intel Xeon 8380-40C and 16 DDR4 Micron 64 GB RDIMMs per CPU.
- Both the DDR4 and DDR5 systems were running the same Alma Linux 9 (Kernel 5.14) operating system.

Computer vision results

We tested MLPerf using the 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 7.3 times throughput gain for the number of samples inferenced per second on DDR5 System over DDR4 System (Figure 1). In addition to higher throughput, DDR5 system also achieves 40% higher sustained memory bandwidth over DDR4 System, as shown in Figure 2.

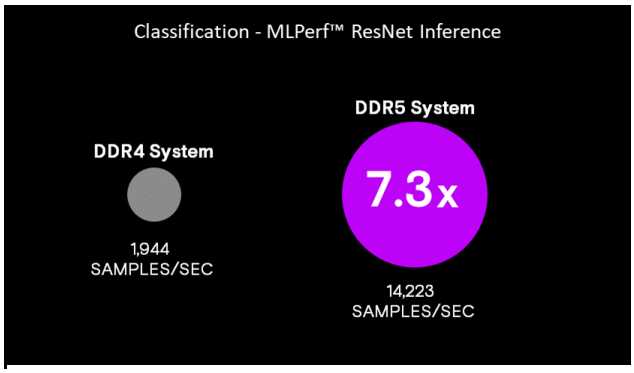


Figure 1: ResNet inferencing throughput comparison – DDR5 SYSTEM versus DDR4 SYSTEM exhibits seven times gain in throughput for ResNet

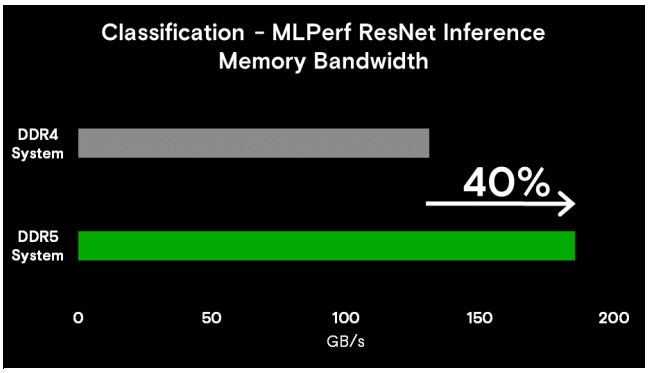


Figure 2: ResNet inferencing bandwidth comparison – DDR5 SYSTEM versus DDR4 SYSTEM provides 40% gain in memory bandwidth for computer vision

NLP results

MLPerf inference using BERT with SQuAD 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 4.9 times gain in throughput, which is the number of samples inferenced per second on DDR5 System (8490H/DDR5) compared to DDR4 system (8380/DDR4), as shown in Figure 3.

The model tasks run in accuracy mode over the entire dataset, and the results show that the average memory bandwidth recorded on DDR5 System is 202 GB/s, shown in Figure 4. DDR5 system achieves 55% higher memory bandwidth compared to DDR4 SYSTEM on 3rd Gen Intel Xeon 8380 and DDR4, resulting in a higher throughput for the 4th Gen Intel Xeon 8490H and Micron DDR5 system.

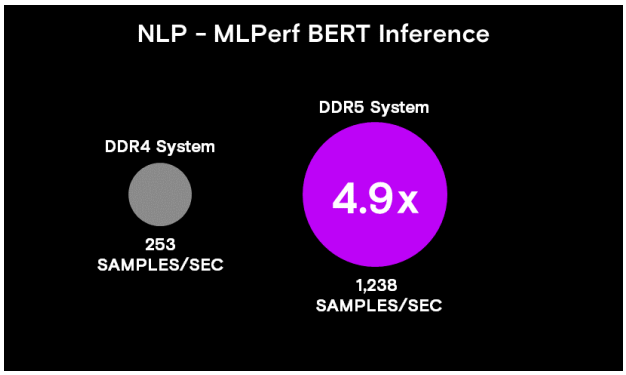


Figure 3: BERT inferencing throughput comparison – DDR5 SYSTEM versus DDR4 SYSTEM delivers 4.9 times gain in throughput for NLP

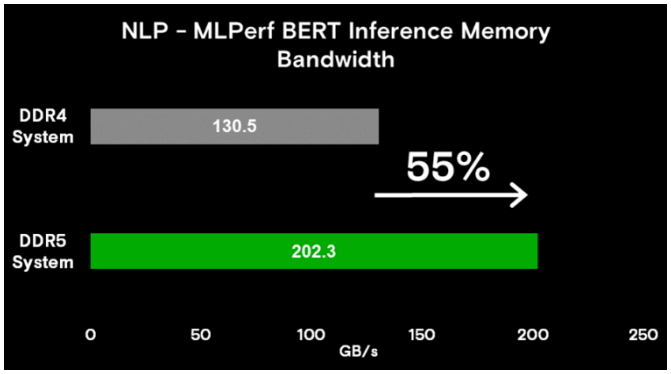


Figure 4: BERT inferencing bandwidth comparison — DDR5 SYSTEM versus DDR4 SYSTEM provides 55% memory bandwidth gain for NLP

NLP multi-instance results

It is also important to understand how these systems will scale by running more model instances in parallel. 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. As described in Figure 5, DDR5 system provides consistent throughput and scales to more than 24 instances whereas DDR4 system could not support more than four instances. The sustained scaling is, exhibited by DDR5 system, due to the superior memory subsystem and CPU architecture. Figure 6 shows the 30% memory bandwidth gain by DDR5 system over DDR4 system. The L3 cache miss rate is also higher on DDR4 system compared to DDR5 system, which has a four times larger cache size. Thus, L3 cache capacity and DDR5 speed both support higher performance in DDR5 system. Effectively, 4th Gen Intel Xeon 8490H together with Micron DDR5 memory can sustain multiple instances of BERT inference workload without compromising on throughput.

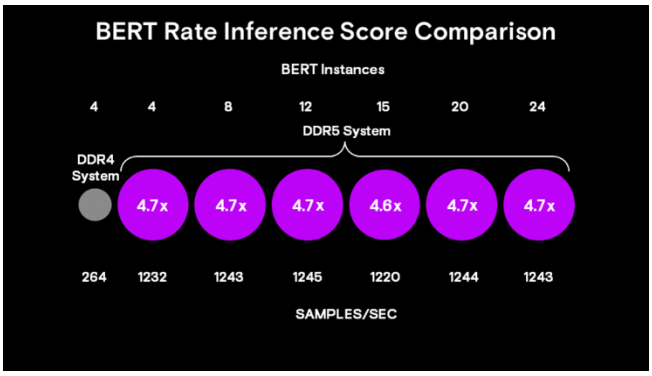


Figure 5: BERT inferencing throughput comparison DDR5 SYSTEM versus DDR4 SYSTEM provides five times gain in throughput for NLP multi-instance

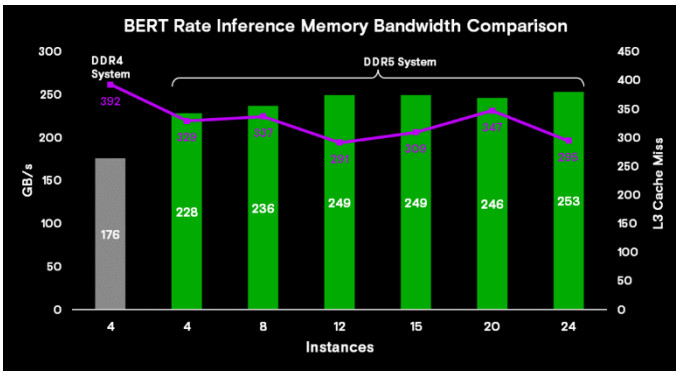


Figure 6: BERT inferencing L3 miss rate and memory bandwidth comparison DDR5 SYSTEM versus DDR4 SYSTEM for NLP multi-instance

Recommendation results

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 24 hours. The performance results show a 4.3 times gain in throughput in samples inferenced per second on DDR5 SYSTEM over DDR4 SYSTEM. The results in Figure 8 also show that the average memory bandwidth was increased by 200% on DDR5 SYSTEM with 107 GB/s compared to DDR4 SYSTEM at 33 GB/s.

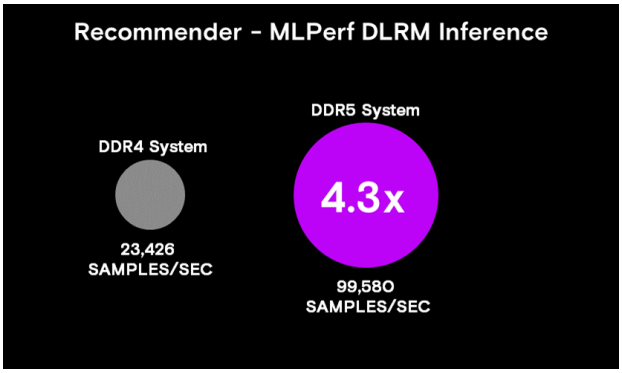


Figure 7: DLRM inferencing throughput comparison — DDR5 SYSTEM versus DDR4 SYSTEM delivers 4.3 times gain in throughput for recommendation

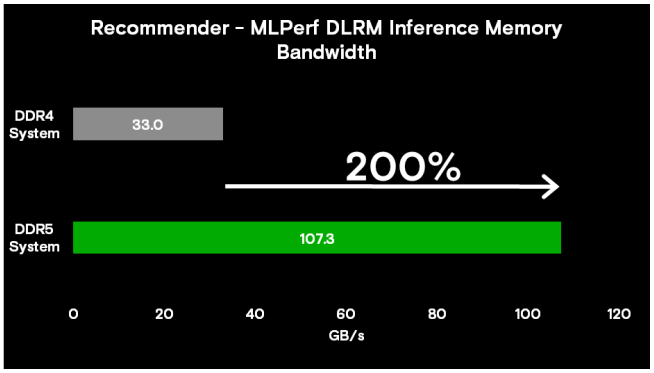


Figure 8: DLRM inferencing bandwidth comparison -DDR5 SYSTEM versus DDR4 SYSTEM provides 200% gain in memory bandwidth for recommendation

Conclusion

DDR5-based server systems offer the much-needed memory bandwidth for AI applications and serve to unlock the potential of high core-count systems. The results from our workload testing demonstrate that the Micron DDR5 memory subsystem delivers higher memory bandwidth at a sustained rate for memory intensive AI applications. We found several AI inferencing problems achieved 40%-200% higher memory bandwidth on Micron DDR5-based platforms over DDR4-based platforms. Upgrading your Enterprise or HPC environment or AI infrastructure and want to learn locate the right DDR5 configuration, contact Micron Sales Network.

micron.com/intel