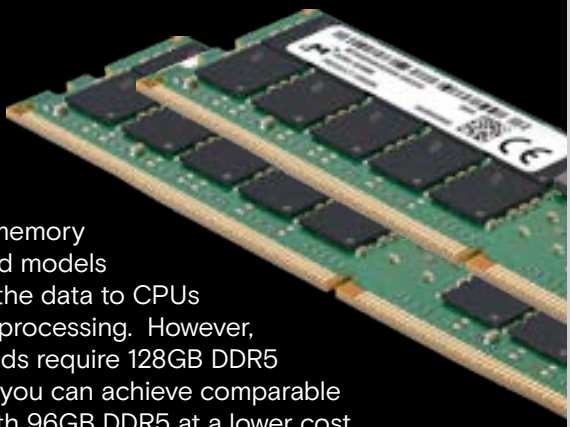
Fast capacity for AI inferencing

96GB DDR5 server DRAM has the density to accelerate AI while minimizing costs

AI inference — or the ability of a trained model to make predictions, solve problems and complete tasks in the real world — is tied to the availability of high-capacity, high-speed and reliable memory, which is needed to perform the millions of calculations necessary to successfully run an AI model.



Similar to AI training memory requirements, inference demands fast, high-density memory to store trained models while feeding the data to CPUs and GPUs for processing. However, not all workloads require 128GB DDR5 densities, and you can achieve comparable throughput with 96GB DDR5 at a lower cost.

We tested Micron 96GB DDR5 Server DRAM to demonstrate how it fares with these challenges when compared to 128GB DDR5. This solutions brief shows test results using BERT NLP.

AI Inference – BERT NLP

BERT (Bidirectional Encoder Representations from Transformers) is based on Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. The input is a sequence of tokens, which are first embedded into vectors and then processed in the neural network, and the output is a sequence of vectors in which each vector corresponds to an input token with the same index. As an unsupervised model, BERT learns from the unlabeled text and can improve even as it's being used in practical applications (i.e., Google search).

Hardware and software setup used in the BERT evaluation:

- BERT model version 3
- 1x MU 9300 NVMe of 15TB

Test methodology

We used a BERT-large-uncased model, which has 24 layers, 1024 hidden, 16 attention heads and 340 Million parameters.

BERT was pre-trained using only an unlabeled, plain text corpus (namely the entirety of the English Wikipedia and the Brown Corpus). We used the SQuAD dataset consisting of questions posed by crowd workers on a set of Wikipedia articles for BERT inferencing (100,000+ questions on 500+ articles).

We tested a model's ability to read a text selection and answer questions. The answer to every question is a text segment from the corresponding reading passage. Our series of experiments used varying batch sizes (number of samples processed) of 64, 128, 256, and 512, with the highest throughput (samples per second).

Results

Figure 1 presents the results. For 2DPC MU 96GB vs. 2DPC 128GB 3DS and 1DPC MU 96GB vs. 1DPC SS 128GB 3DS we observed similar throughput (which is the number of samples inferenced per second). We also observed that the average memory bandwidth was also almost the same on these five different memory configurations.

Key takeaways from these results are that:

- 2DPC MU 96GB is on par with 2DPC SS 128GB 3DS
- 1DPC MU 96GB is on par with 1DPC SS 128GB 3DS

A microarchitectural analysis of BERT did not suggest any major differences in bottlenecks (in terms of memory backend boundedness for bandwidth or latency) across the various memory configurations.

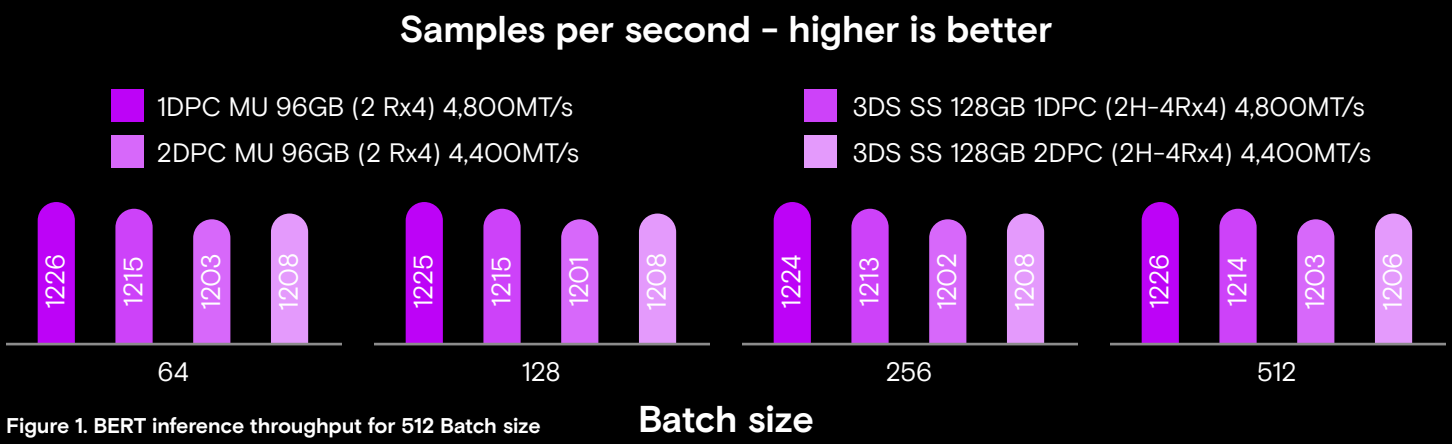


Figure 1. BERT inference throughput for 512 Batch size

Why these test results matter

These tests show that 96GB throughput is comparable to 128GB and can provide substantial cost savings in the long run. 96GB also saves on energy demands to lower energy use and related costs. The high throughput delivered by 96GB DDR5 can accelerate AI inference for applications like home security systems, predictive text features and large language models like ChatGPT. That means faster identification of patterns in data, quicker image generation and real-time NLP use cases.

96GB server DDR5 is ideal for budget-constrained projects that require higher capacities to meet workload demands.

With 96GB, you can fit larger model sizes with more parameters; the higher bandwidth and lower latency compared to 3DS will give you more throughput and response time on AI inference so that you can process more queries per second.

DDR5 96GB offers a lower-cost solution to inferencing needs.

Pattern recognition

When training complex AI models (like for generative AI), large datasets are processed through training models to teach them to recognize patterns and generate new data. Inference happens when prompting or making a request to generative AI; the model uses the training patterns to generate new data similar to the input.

The challenge, of course, is that training requires large amounts of fast memory and storage to hold the data and feed it to the compute as it is processed.