



Micron DDR5 96GB monolithic RDIMM

Power and performance

This technical brief compares the power and performance of Micron's DDR5 96GB registered dual in-line memory module (RDIMM) with a commercially available competitive high-capacity DIMM.

The comparison is based on the workload testing framework in the figure below, which shows how each workload is bound (or sensitive) to bandwidth, latency, or capacity. For example, high performance computing applications tend to be more bandwidth-bound and data analytics applications are more capacity-bound.

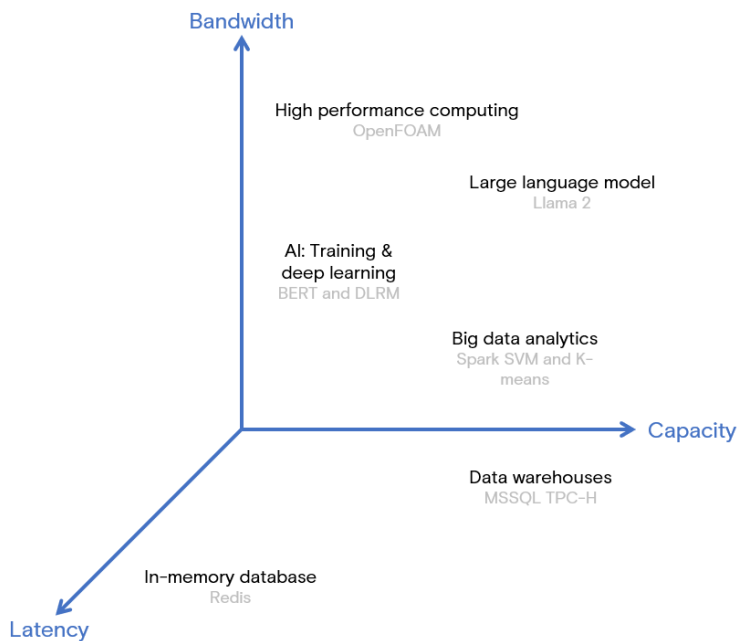


Figure 1: Workload testing framework

Fast facts

Micron's DDR5 96GB monolithic RDIMM provides high capacity using a single-die package (SDP) manufacturing process, enabling data center infrastructures to operate with lower latencies and reduced power consumption.¹

Reduced power

24%

Similar or better performance with up to 24% lower power for capacity-bound workloads. Data center workloads are computationally intensive, leading to increased power consumption, especially when training large models. Micron's high-capacity DIMMs help to keep data centers energy efficient by offering lower power than commercially available competitive modules.

Lower latency

17%

Up to 17% lower latency for bandwidth-bound workloads. For AI inference, low latencies are required to provide real-time results.²

Higher bandwidth

High performance computing, AI, machine learning, and big data analytics involve large-scale data processing, complex computations, and parallel execution. These kinds of applications require high memory bandwidth to support high throughput and low latency.

1. As compared with commercially available competitive DDR5 3DS modules with 128GB capacity.
2. 17% is based on JEDEC specifications. 9% lower latency is based on the MSSQL TPC-H test.

Overview of workload results

We tested a range of micro-benchmarks and real-world workloads comprised of BERT (Bidirectional Encoder Representations from Transformers) and DLRM (Deep Learning Recommendation Model) recommendation models, Llama 2 large language model, OpenFOAM, Spark Support Vector Machine (SVM) and K-means, Microsoft SQL with TPC-H, and Redis as an in-memory database.

As compared with commercially available competitive DDR5 128GB 3D stacking modules, Micron's DDR5 96GB monolithic RDIMM offers up to 17% lower latency, higher memory bandwidth, and better performance for bandwidth-bound workloads, and similar performance for capacity-bound workloads, all while consuming up to 24% less power.

Micron DDR5 96GB monolithic RDIMM		
AI training and deep learning	Similar performance	Up to 24% lower power
Large language model		
High-performance computing		
Big data analytics		
Data warehouses		
In-memory database		

Table 1: Summary of performance and power consumption for Micron DDR5 96GB monolithic RDIMM³

Advantages of single-die package (monolithic)

Single-die package (SDP) and 3D stacking (3DS) are two different ways of integrating multiple components into a single system. Single-die package (monolithic) uses a single large die that contains all the components, while 3D stacking uses multiple smaller dies that are stacked vertically and connected through-silicon vias (TSVs). The table below highlights the advantages of a single-die package manufacturing process.

Single-die package (monolithic)		
		
Power	Scalability	Latency
A monolithic solution uses a single die, resulting in a less complex circuit that generally consumes less power and dissipates less heat within the package.	A monolithic DRAM design provides a more effective way to scale bit density by enhancing the storage capacity within a single die.	Lower latencies ⁴ due to shorter signal pathways within a single die, allowing for quicker data retrieval between the die and memory controller.

Table 2: Advantages of monolithic technology (single-die package)¹

3. The entire benchmark testing is based on Intel x86 systems, using memory reference codes from Intel's Maintenance Release 1 (MR1) for Dell Power Edge R760 from Bios release 1.2. The choice between SDP and 3DS modules should be based on the capacity requirements of an application and the trade-offs should align with those requirements.

4. As defined by the JEDEC specification.

High-capacity DIMMs are needed by many data center workloads in AI training and inference, large-scale data analytics, and in-memory key-value databases for processing and quickly deriving insights from vast datasets. Often these workloads must retain as much data as possible in the system's main memory. In several cases, such as in-memory databases and generative AI, either all the data or the model in its entirety (in the case of large language models) are required to stay resident in memory. Moreover, as deep learning models increase in size and complexity, more memory is needed to process and store large datasets efficiently.

These rigorous demands from current and emerging workloads have driven the need for data center systems to architect solutions with significant amounts of memory. The table below shows details about the workloads tested and their respective domain.

Workloads tested

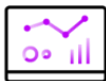
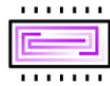
Bandwidth bound			
			
AI Training & deep learning	High performance computing	Large language model	
BERT	DLRM	OpenFOAM	Llama 2
Excels at understanding the context and relationships between words in sentences.	Used in recommendation systems employed by online stores or streaming services.	Specialized software used to simulate complex physical processes (airflow over an aircraft or motorcycle).	An open-source large language model (LLM) based on transformers.
Capacity bound			
			
Big data analytics	Database warehouse	In-memory database	
SVM	K-means	MSSQL TPC-H	Redis
A supervised machine learning method used for classification and regression tasks.	A clustering algorithm for knowledge discovery and data mining.	TPC-H evaluates databases for OLAP (online analytical batch processing) use cases.	An in-memory key-value store that serves a multitude of functions, including its role as a database.

Table 3: Capacity-bound and bandwidth-bound data center workloads

Performance

Below we showcase performance and power results from the workloads tested. We first evaluated the memory modules with a batch of MLC micro-benchmarks to set the performance baseline for each memory module and identify bandwidth and latency idiosyncrasies.

AI inference

Deep learning models

We see similar performance for both Bidirectional Encoder Representations from Transformers (BERT) and Deep Learning Recommendation Model (DLRM) because these workloads are compute-intensive and bandwidth-bound as opposed to capacity-bound.

BERT

We used a pre-trained BERT model with 340 million parameters to perform a series of experiments across four different batch sizes (number of samples processed): 64, 128, 256, and 512 (with the highest throughput). Throughput quantifies the rate at which data can be transferred or processed in memory, measured here as samples inferred per second.

Micron result: Similar throughput across all batch sizes tested (two DIMMs per channel)

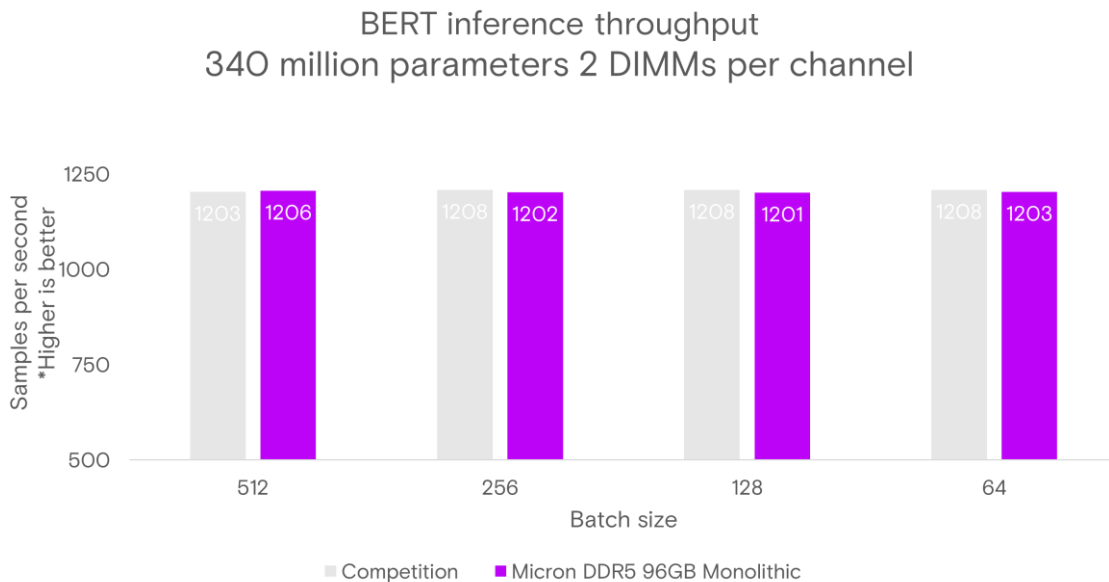


Figure 2: BERT inference throughput

DLRM

We used a pre-trained DLRM model with a batch size of 128,000. Performance is based on throughput (number of samples per second) to understand the rate of inference.

Micron result: Similar throughput for batch size tested (one and two DIMMs per channel)

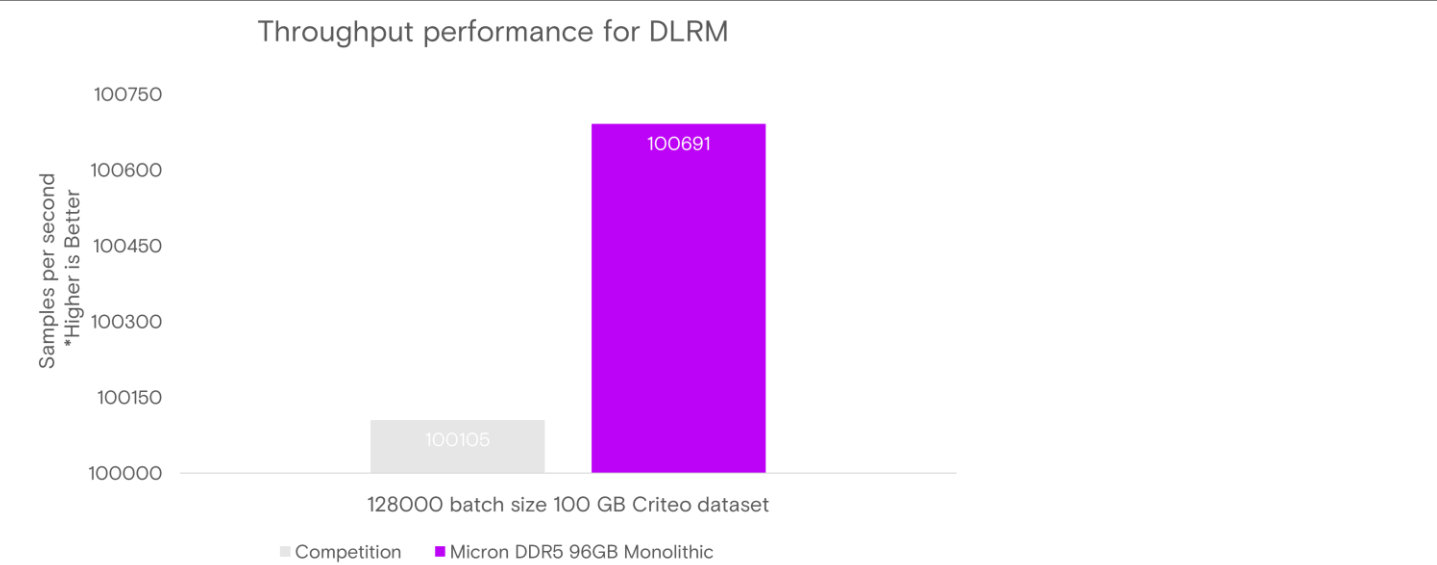


Figure 3: DLRM inference throughput

Large language model

Llama 2

Llama 2 was chosen because it best reflects modern large language models such as ChatGPT. We executed 1024 token inputs, meaning 1024 words or symbols are allowed as input per stream. Tokens are individual units of data or records. We then used 112 streams, where streams are parallel processes.

Performance is measured by tokens per second, a throughput metric that indicates how efficiently a system can read, write, and process data. The performance improvement is due to the higher bandwidth offered by Micron's 96GB module (see MLC bandwidth test result on page 8). The lower power is due to fabrication and architectural differences between the modules.

Micron result: 3% better performance and consumes 22% less power (one DIMM per channel)

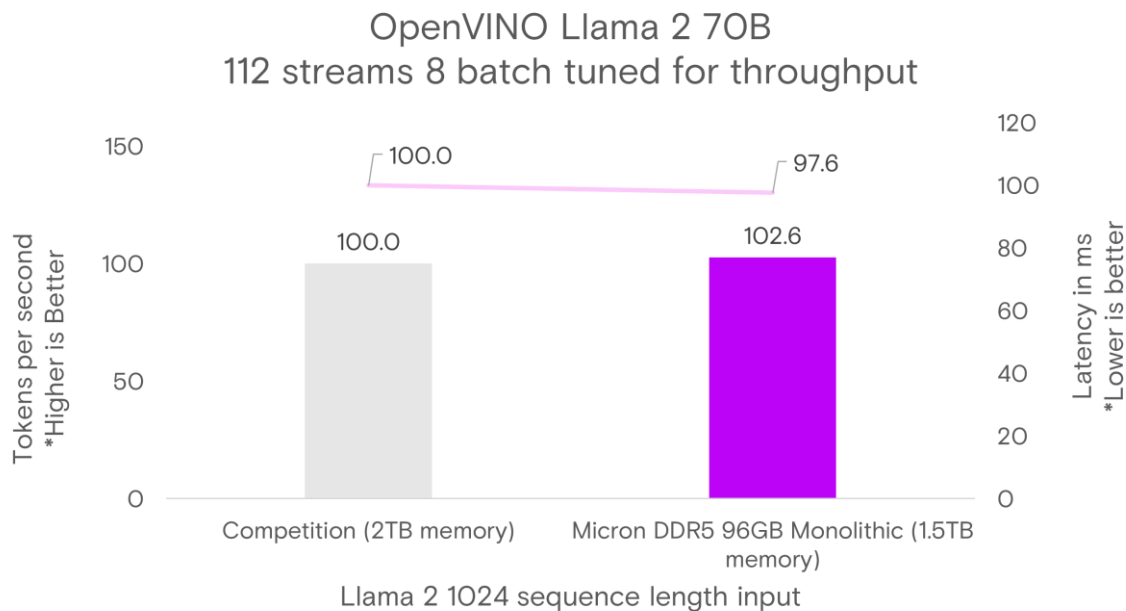


Figure 4a: Normalized throughput for Llama 2 with secondary latency axis

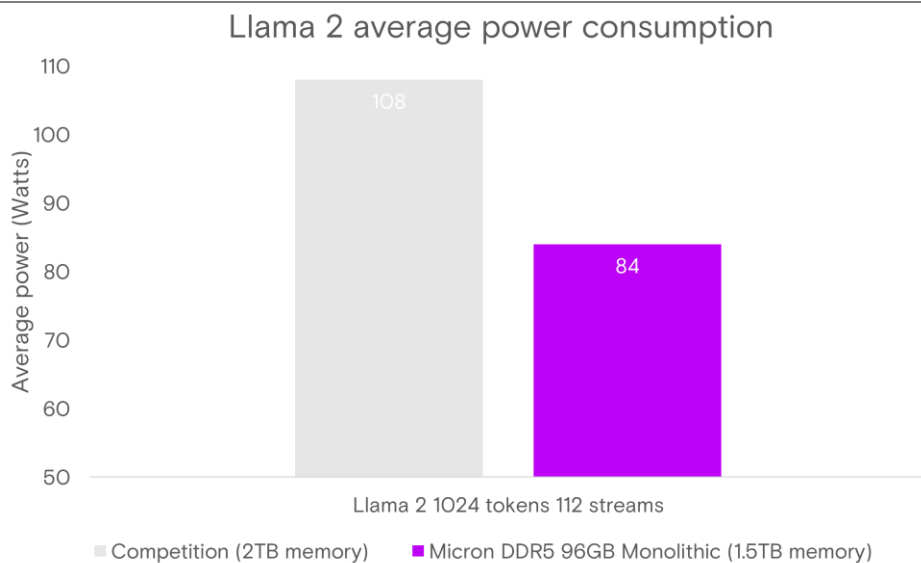


Figure 4b: Average power consumption for Llama 2

Bandwidth

OpenFOAM

The computational fluid dynamics OpenFOAM application is the most bandwidth-sensitive workload we evaluated in our benchmark testing.⁵ We ran the OpenFOAM default motorbike example, which renders a 600x240x240 motorbike mesh to calculate the steady flow around a motorcycle and rider.

High performance computing (HPC) workloads are typically bound by DRAM bandwidth. Micron’s 96GB monolithic RDIMM shows 4% lower runtime because the memory is faster than the competition. The lower power is due to fabrication and architectural differences between the modules.

Micron result: 4% lower runtime and consumes 20% less power (one DIMM per channel)

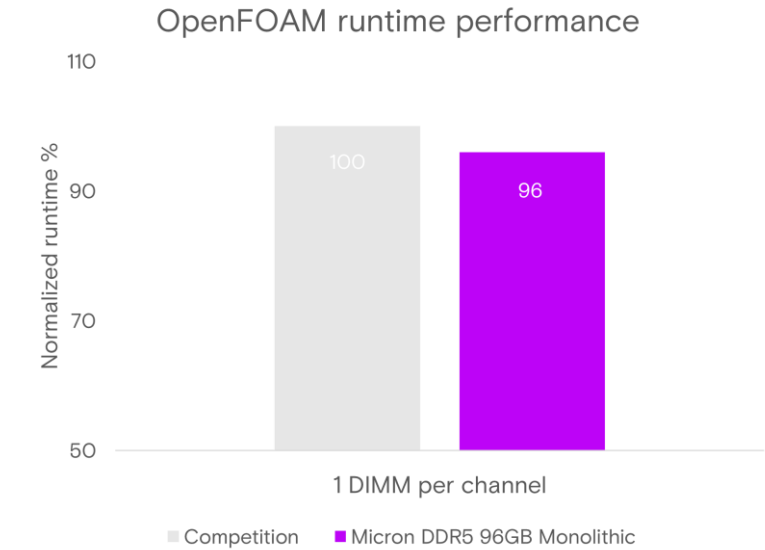


Figure 5a: HPC OpenFOAM runtime

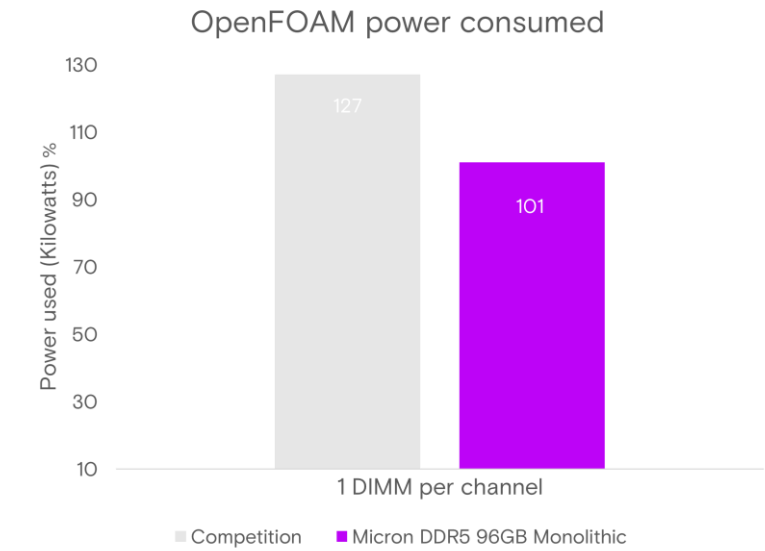


Figure 5b: HPC OpenFOAM power consumed

5. **System A:** Micron DDR5 96GB Monolithic system is a Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 1.5 TB memory fully populated, and Micron 9300 NVMe SSD. **System B:** Competition DDR5 128GB TSV-based system is a Dell PowerEdge R760 with Intel Xeon 8480+ Scalable processors, 2TB memory fully populated, and Micron 9300 NVMe SSD. The benchmark is Open Foam MPI 4.1 112 cores Motorbike size 600 * 24 * 240 (Large). We used OpenMPI 4.1 to parallelize the OpenFOAM processing. System A completed the benchmark in 713 seconds and consumed 66278 joules. System B completed in 743 seconds and consumed 83176 joules.

Intel Memory Latency Checker (Intel MLC)

The Intel MLC tool measures the latency and bandwidth of the memory subsystem in a computer system. Using the MLC tool, we performed a core sweep, which is the incremental use of cores to push bandwidth.

Micron result: 6% higher bandwidth

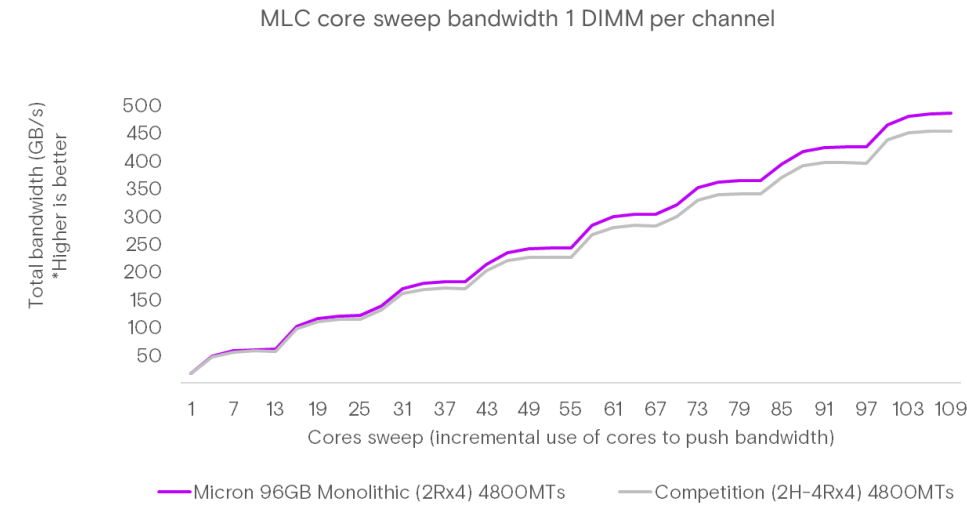


Figure 6: MLC core sweep bandwidth

STREAM triad access pattern

To measure the peak memory bandwidth, we ran MLC with a specific memory access pattern that mimics the STREAM triad access pattern (from the actual STREAM benchmark). Triad is the most complex scenario and is also relevant to high performance computing.

Micron result: Maximum memory bandwidth⁶ of 487GB/s

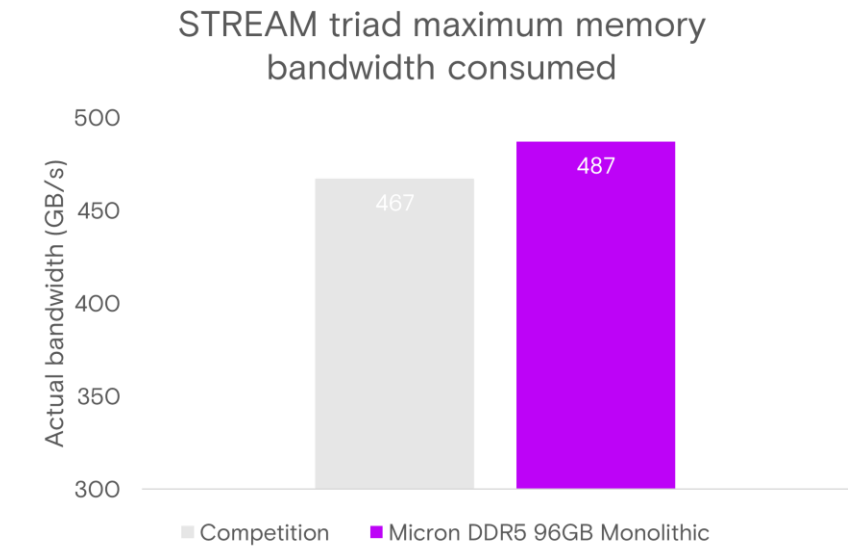


Figure 7: Maximum memory bandwidth STREAM triad

6. For a 2-CPU system

Capacity

Big data analytics

Spark Support Vector Machines (Spark SVM)

We ran Spark SVM with the largest HiBench dataset of 360GB in disk space and up to 1.3TB of memory usage during execution. Even the largest dataset was insufficient to force Spark to use enough memory to justify the competition module's 2TB of total capacity. Therefore, we created a larger dataset of 560GB, which uses up to 2.3TB of memory during execution. Performance is based on normalized runtime, the time it takes to complete a task divided by a reference time. It shows how much faster or slower a task is compared to the reference.

Micron result: Similar performance (two DIMMs per channel) and consumes **23% less power** (one DIMM per channel)⁷

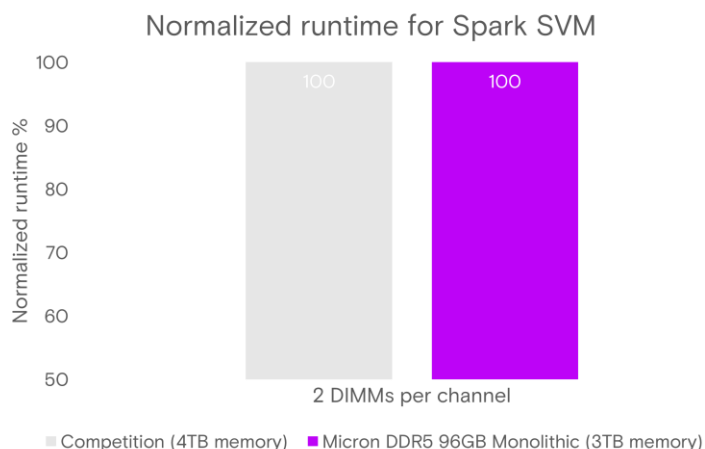


Figure 8a: Normalized runtime SVM using Spark

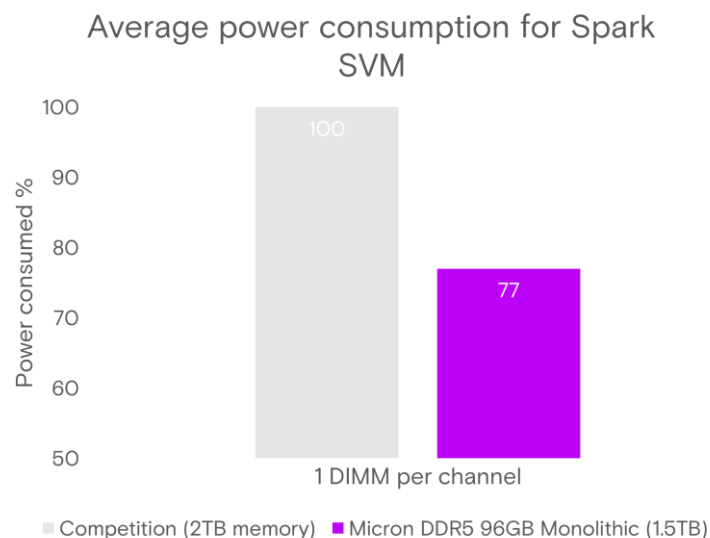


Figure 8b: SVM average power consumed using Spark

7. **System A:** Micron DDR5 96GB Monolithic system is a Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 3TB memory fully populated, and Micron 9300 NVMe SSD. **System B:** Competition DDR5 128GB TSV-based system is Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 4TB memory fully populated, and Micron 9300 NVMe SSD. Spark SVM 360GB GitHub - Intel-bigdata/HiBench: HiBench is a big data benchmark suite. The dataset was expanded to 560GB and run on both systems. Benchmark completed in 468 seconds in System A and 466 seconds in System B.

Spark K-means

We ran Apache's Spark K-means with a HiBench dataset, "Bigdata" of 240GB, in disk space. Spark K-means can use up to 2.9TB of memory during execution. Performance is measured by normalized runtime.

For both Spark SVM and K-means, performance is dependent on the size of the input (here, the size of the dataset). For these workloads, the amount of available memory with Micron's 96GB monolithic module is enough to process the largest dataset size (560GB), resulting in similar performance. However, if the input size exceeds the available memory, we would see a delta.

Micron result: Similar performance (two DIMMs per channel) and consumes **23% less power** (one DIMM per channel)

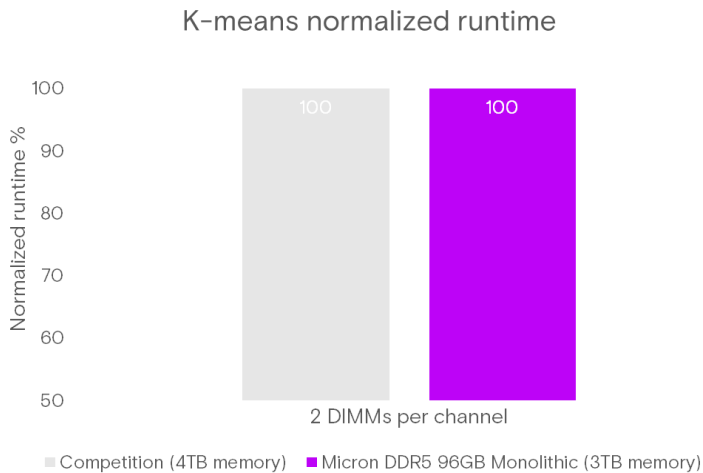


Fig 9a: K-means normalized runtime performance

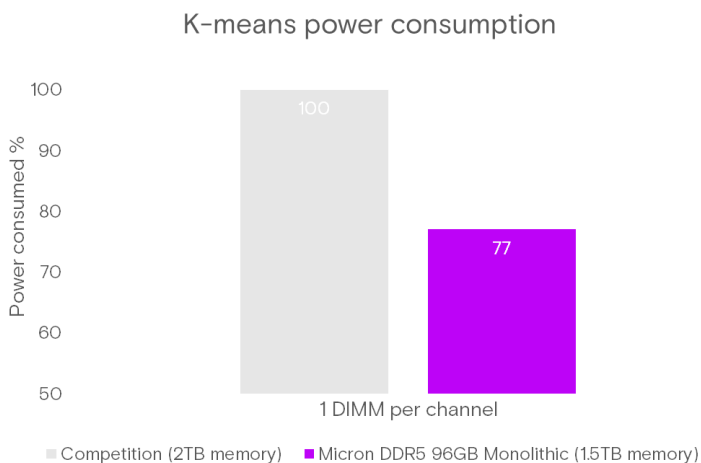


Figure 9b: K-means power consumption

Latency

JEDEC standard

Single-die package (SDP) has lower read/write latency timings compared to 3D Stacking (3DS) as defined by JEDEC. As shown in the figures below, Micron 96GB DDR5-4800 MT/s outperforms in read/write latency, where it is 17% lower in latency, resulting in faster data transfer and processing speed.⁸

Micron result: 17% lower in read/write latency

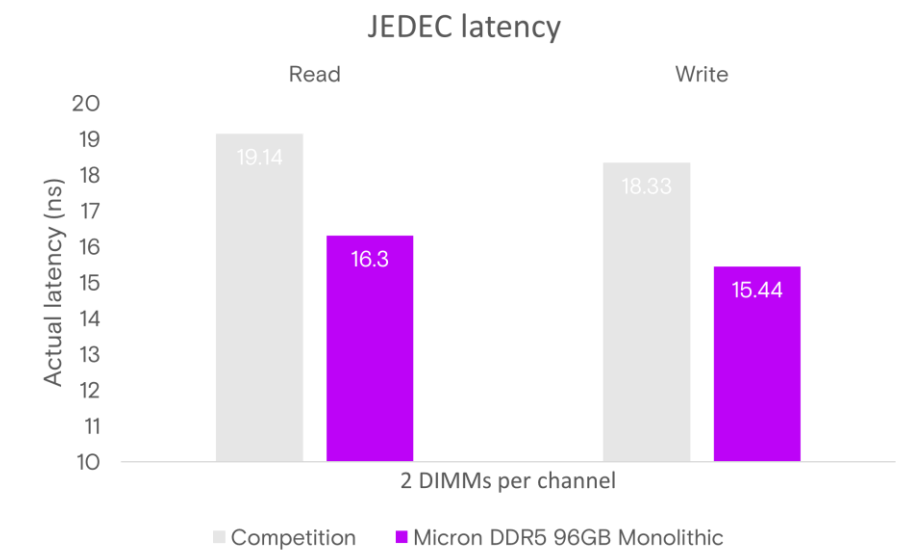


Figure 10: Latency based on JEDEC standard

Note: Micron’s production DIMMs meet or exceed the JEDEC standard.

8. Latency Read command to First Data, CL-nRCD-nRP would represent the sequence of delays from when a read command is issued to when the first set of data is available. Micron 96GB DDR5 4800 has a latency of 16 ns for Read Command to First Data and competition 3DS DDR5 4800 128GB has a latency of 18.75 ns per DDS and JEDEC specification.

MSSQL and OLAP TPC-H

We benchmarked Microsoft SQL (MS SQL) server database with the OLAP TPC-H workload. We ran 16 streams in parallel (with each stream containing 22 sequential queries). The throughput test is based on a 3000-scale factor dataset, which amounts to a total of 3TB. MSSQL loads the necessary data on a per-query basis.

Micron's 96GB monolithic module shows 9% lower p99 latency (a statistical measure that refers to 99% of the queries being satisfied under the quality-of-service latency metric).⁹ A lower p99 latency is due to the 24Gb components used in the 96GB module.

Micron's module also took 2935 seconds to complete the benchmark, resulting in a 3% higher runtime performance.¹⁰ Slightly higher runtime is due to the lower capacity of the 96GB module and is considered within the standard deviation range (that is, performance is similar).

Micron result: 9% lower latency (two DIMMs per channel)

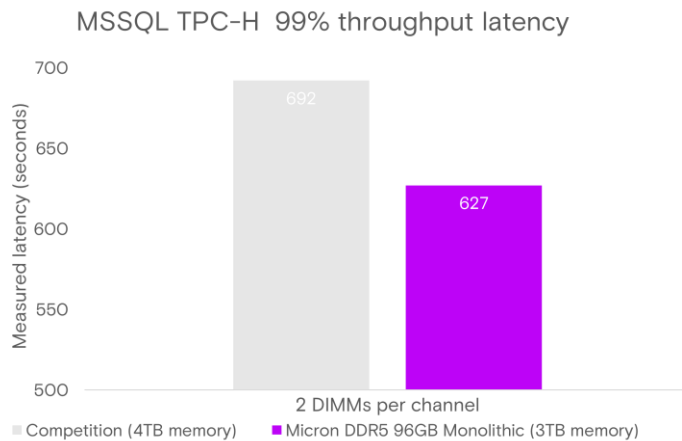


Figure 11a: p99 latency for MS SQL server with TPC-H

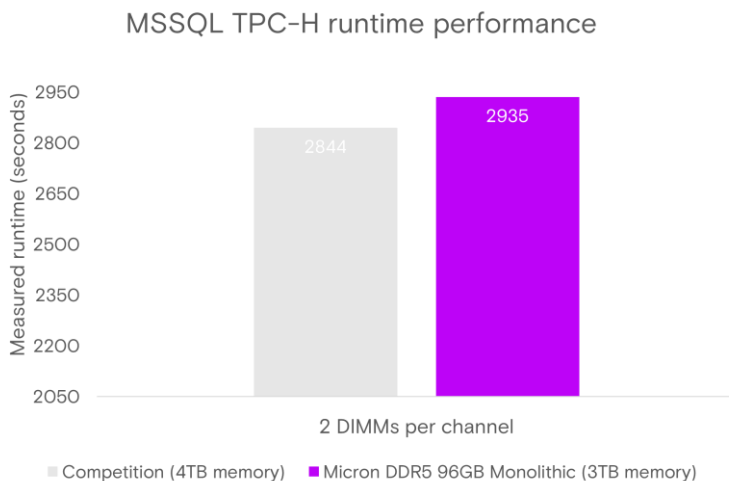


Figure 11b: Runtime performance for MS SQL server with TPC-H

9. TPC-H workload configuration: **System A:** Micron DDR5 96GB Monolithic system is a dual-socket Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 3TB memory fully populated, and Micron 9300 NVMe SSD. **System B:** Competition DDR5 128GB TSV-based system is a dual-socket Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 4TB memory fully populated, and Micron 9300 NVMe SSD. MSSQL TPC-H 3.0.9 scale factor 3000 was used and executed in throughput mode. **System A:** Finished in 2935 seconds and consumed 15 TB memory. **System B:** Executed in 2844 seconds with peak memory consumption of 2TB.

10. On a two-CPU system with 3TB of memory and using two DIMMs per channel. A 3% delta is considered within the standard deviation, meaning performance is similar between the memory modules.

Power

Redis

In our performance analysis of Redis, using the YCSB benchmark (Yahoo Cloud Serving Benchmark) across five different read/write ratios for a 2.9TB database, we observed a slight performance drop of 2% overall with Micron DDR5 96GB monolithic module as compared to competition 128GB 3D stacking DIMMs.¹¹

The slightly lower performance result is primarily because in-memory databases, such as Redis, scale well with increased capacity. However, what is noteworthy is that the additional memory capacity offered by the competition module did not result in a substantial performance gain for the 2.9TB in-memory database. Micron DDR5 96GB monolithic module also has 21% lower power, where the lower power is due to fabrication and architectural differences between the modules.

Micron result: Similar performance and consumes 21% lower power

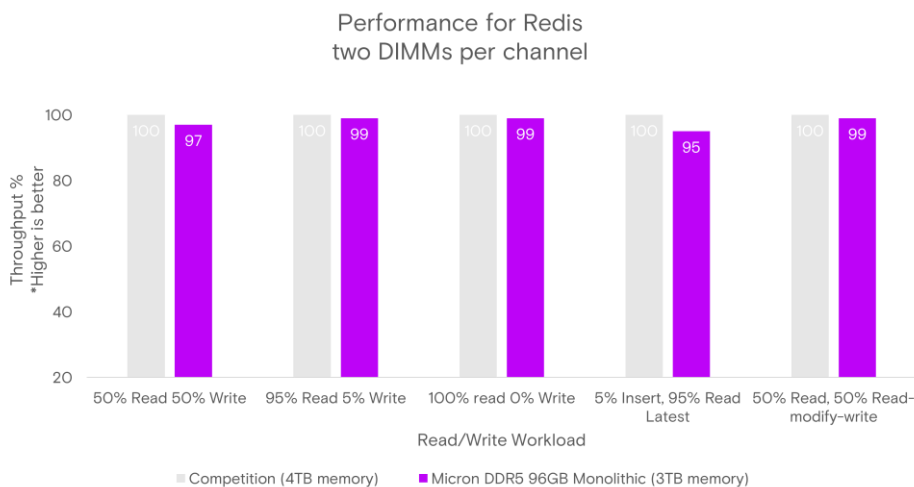


Figure 12a: Throughput performance for Redis with YCSB clients

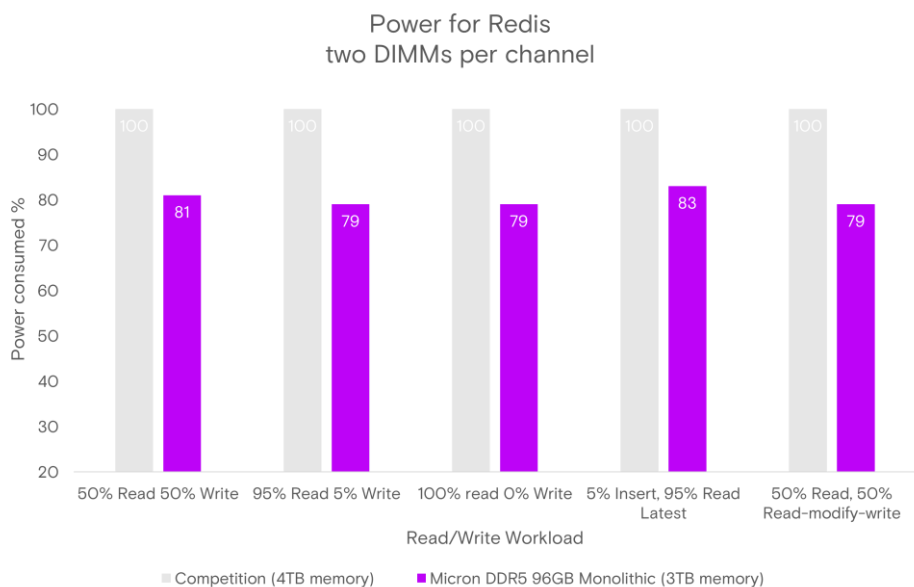


Figure 12b: Average power consumption for Redis with YCSB clients

11. **System A:** Micron DDR5 96GB system is a Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 3TB memory fully populated, and Micron 9300 NVMe SSD. **System B:** Competition DDR5 128GB TSV-based system is a Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 4TB memory fully populated, and Micron 9300 NVMe SSD. Redis YCSB Workload A 200M records loaded with 4KB per record size. **System A:** 1110286 ops/seconds by consuming 72 Watts. **System B:** 1195304 ops/second consuming 84 Watts. Keeping power (Watts) as the baseline for 84W, System A can process 1277589 ops/second or an 8% improvement over System B. A 2% delta is considered within the standard deviation, meaning performance is similar between the memory modules.

MLC

For the results in the figure below, power was collected through Intel's PMC tool while under a heavy memory access load.¹²

For total DRAM power on a single socket, Micron 96GB DDR5 has 22% lower power (idle, underload) for one DIMM per channel and 24% lower power for two DIMMs per channel. The lower power is due to fabrication and architectural differences between the modules.

Micron result: 24% lower power (two DIMMs per channel)

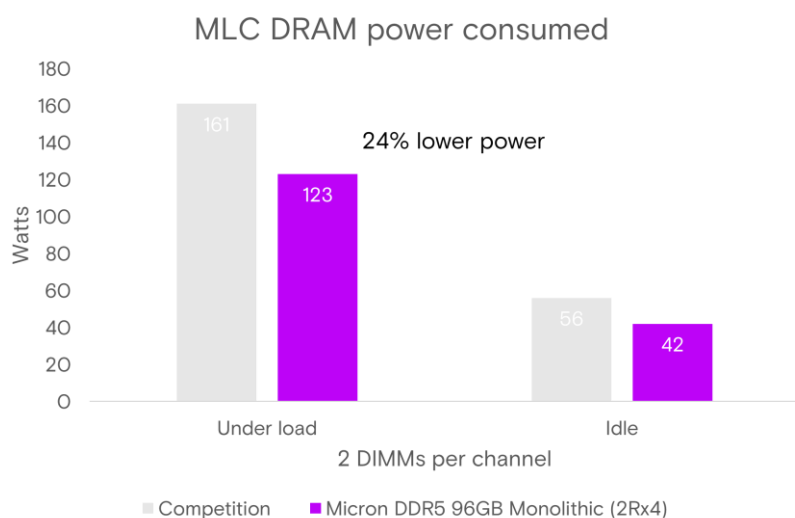


Figure 13: MLC DRAM power (idle, under load) consumption

12. Power test via Intel's PCM-Power tool while under Heavy Memory Access load (Intel MLC with W6 flag streaming writes, Loaded Latency with no injection delay, 2GB buffer per thread)
System A: DDR5 Micron 96GB system is a Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 3TB memory fully populated, and Micron 9300 NVMe SSD. **System B:** Competition DDR5 128GB system is a Dell PowerEdge R760 with Intel Xeon 8480+ Scalable Processors, 4TB memory fully populated, and Micron 9300 NVMe SSD. **System A:** Consumes 161W under load and 56W idle latency. **System B:** Consumes 128W under load and 42W idle latency.

Conclusion

High-capacity memory along with high memory bandwidth and low latency is essential for data center infrastructures to handle the computational complexity, large model sizes, and enormous datasets characteristic of AI, database, and data analytics applications. Micron's DDR5 96GB monolithic RDIMM offers a powerful solution for these environments as shown by our workload results.

Upgrading your enterprise or AI infrastructure or HPC environment and want to learn how to locate the right DDR5 configuration, contact [Micron sales network](#).

micron.com/ddr5