Micron® MRDIMM Technology Delivers Performance and Efficiency

Authors: Henrique Pötter, Sravani Gomatam, Moiz Arif, Sujit Somandepalli, Sarthak Thakkar, and Sudharshan S. Vazhkudai

Abstract

This report comprehensively evaluates Micron® Multiplexed Rank Dual In-line Memory Module (MRDIMM) technology, emphasizing its impact on memory subsystem latency, bandwidth, and system-level power efficiency across a range of server workloads.

MRDIMM modules operating at 8800 MT/s, with enhanced bus efficiency, delivering up to 41% higher sustained bandwidth and up to 40% lower loaded latency compared to conventional RDIMM solutions at 6400 MT/s. These architectural improvements contribute to greater CPU power efficiency, as evidenced by reductions in memory-induced pipeline stalls and increases in instruction retirement rates in bandwidth- and latency-sensitive workloads.

Empirical results across high-performance computing, data analytics, and AI inference workloads demonstrate that MRDIMM at 8800 MT/s accelerates task completion while delivering significant energy savings, with up to 20% lower task energy at 64GB, 128GB, and 256GB capacities. This is achieved by minimizing execution time and optimizing memory access patterns. Furthermore, PCIe Gen6 will double per-lane throughput, resulting in a surge in data movement, intensifying DRAM bandwidth requirements, and solidifying the demand for future iterations of MRDIMM technology.

Looking ahead, MRDIMM Gen. 2, operating at 12800 MT/s, is projected to offer even greater improvements in task-level energy efficiency for bandwidth-bound applications. Additionally, it is expected to provide lower latencies under high memory load conditions, further benefiting latency-sensitive use cases.

MRDIMM's scalable architecture is positioned as a critical enabler for balanced, energy-efficient data center architecture, ensuring that memory subsystems can keep pace with escalating I/O and computational demands in next-generation platforms.

MRDIMM benefit
We show that MRDIMM not only improves performance on HPC, data analytics, and vector database applications, but it can also increase server power efficiency by up to 20% on bandwidth-bound applications.
We assess how MRDIMM can decrease the average and tail latency under low and heavily memory-loaded scenarios.
Also, we assess how tall form factors enable solutions with 128GB and 256GB capacity per MRDIMM module at 8800 MT/s, offering both high performance and large capacity.

Table 1: Findings summary

This paper is organized into discrete sections outlining the MRDIMM architecture and operational principles, detailing the benchmark methodologies and workload characteristics used in testing, system configurations, and performance analysis.

Introduction

In the rapidly evolving landscape of data centers, performance, power efficiency, and capacity demand on the memory subsystem are ever-increasing. Strategic investment in advanced memory technologies with enhanced capacity and bandwidth enables data centers to achieve and sustain high operational efficiency, helping meet the escalating performance demands of a rapidly evolving computational landscape.

The demand to manage complex and data-heavy applications, such as high-performance computing (HPC), artificial intelligence (AI), large-scale data analytics, and cloud services, is driving this trend. As we add more CPU cores and faster PCIe devices, we need more memory to keep up with the computations. If we lack memory capacity and bandwidth, the CPU cores and PCIe devices can slow (stall), leading to poor performance and wasted processing power.

For instance, HPC applications such as OpenFOAM can be bandwidth-bound and directly benefit from extra memory bandwidth. Apache Spark applications may be sensitive to both memory capacity and bandwidth, while also relying heavily on high-throughput NVMe™ storage for data ingestion and checkpointing. Performance often scales with increased memory availability and speed, enabling more effective caching and reuse of intermediate computational results across distributed worker nodes.

Micron's latest innovation, the Multiplexed Rank Dual In-line Memory Module (MRDIMM), helps address these needs by offering greater memory bandwidth and higher capacity than prior solutions, all while increasing the system's power efficiency on memory-bound applications. MRDIMM enables higher bandwidth to help ensure that multi-core processors and PCIe devices can operate at their full potential, avoiding bottlenecks and maximizing performance. MRDIMM enables the host CPU to run twice as fast as the memory backplane clock speed, allowing the internal components to experience only half of the host memory clock speed. This enables the use of lower-speed specification components (as found in modern RDIMM solutions) to build a faster DRAM solution. A summary of the benefits is shown in Table 2.

Micron's MRDIMM portfolio includes a range of capacities to suit a broad range of server solution requirements: 64GB, 96GB, 128GB, and 256GB. These modules are designed to meet the rigorous demands of modern data centers, providing a scalable solution that enhances performance while maintaining cost-effectiveness.

MRDIMM benefits for data centers

MRDIMM benefit	How it works
Enhanced performance	First-generation MRDIMM modules operating at 8800 MT/s provide up to 41% more effective memory bandwidth than traditional RDIMMs at 6400 MT/s. This improvement is crucial for high-performance computing (HPC), artificial intelligence (AI) inference, data analytics applications, and PCIe I/O devices relying on Direct Memory Access. The second-generation MRDIMM is expected to improve performance further.
Improved scalability	With capacities ranging from 64GB to 256GB, MRDIMM modules offer flexible options for scaling memory in data centers. Tall Form Factor (TFF) modules offer higher capacity and improved bandwidth efficiency by having more memory ranks, enhancing mixed read/write access patterns.
Decreased latency	MRDIMM technology reduces latency by up to 40%, ensuring faster data access and processing. This particularly benefits latency-sensitive mission-critical workloads, such as financial trading systems, industrial automation, and virtual multi-tenancy in cloud environments. Moreover, MRDIMM exhibits significantly lower tail latencies and tighter latency distributions, which are essential for ensuring real-time, deterministic system behavior with predictable response times, especially for workloads with stringent low-latency and high availability requirements.
Greater energy efficiency	By enhancing CPU performance, MRDIMM boosts task-energy efficiency, enabling more work to be completed in a shorter time. This, in turn, reduces the total energy consumption per completed task.

Table 2: MRDIMM data center benefits

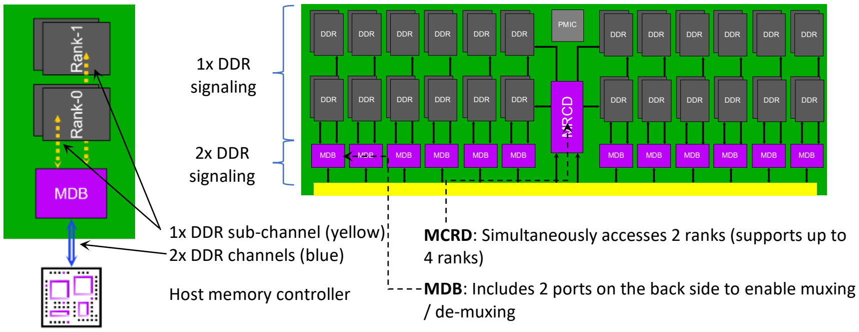
The Multiplexed Rank Dual In-line Memory Module (MRDIMM)

The DDR5-based Registered Dual In-line Memory Module (RDIMM) introduced several enhancements that improved memory bank-level parallelism. One of these features is the presence of two independent 32-bit subchannels per DIMM (40 bits with full ECC), each capable of delivering cache lines with a burst length of 16. This translates to 512 bits or 64 bytes (the typical cache line size) per burst.

Because the subchannels operate independently, banks within each can concurrently serve 64B cache lines, boosting overall throughput. Increasing the number of ranks on RDIMMs (e.g., dual-rank [2R] or quad-rank [4R]) further enhances bank parallelism by distributing banks across additional ranks with relaxed timing constraints. However, this comes with trade-offs, like higher rank-to-rank switching latency, increased electrical loading, and limited concurrent data access (as only one rank can actively perform a read or write data burst at a time).

MRDIMMs help address these limitations by incorporating a high-speed data buffer aggregating data from multiple ranks within a DDR5 subchannel into a single, high-bandwidth stream. This buffer effectively isolates the CPU and memory controller from the electrical load of additional ranks, enabling greater parallelism and bandwidth efficiency without the penalties associated with traditional multi-rank RDIMMs.

Figure 1 illustrates how the rank multiplexing is made possible by the Multiplexing Registered Clock Driver (MRCD) and the Memory Data Buffers (MDB). The MRCD uses time-based multiplexing to redirect commands to each rank while the MDBs convert a 16-bit DRAM interface running at native speed from two ranks into an 8-bit host interface running at twice that speed.



MCRD: Simultaneously accesses 2 ranks (supports up to 4 ranks)

MDB: Includes 2 ports on the back side to enable muxing / de-muxing

Figure 1: MRDIMM (conceptual illustration)

The external data bus (DQ/DQS) is still shared on MRDIMMs, but the MDBs decouple internal DRAM from the host interface, allowing banks within both ranks to be activated to transfer data in parallel. The buffering and rank-multiplexing capability enables MRDIMMs to overlap internal operations across ranks, reducing idle cycles and improving effective bandwidth (whereas the typical RDIMMs expose all ranks directly to the channel, forcing the controller to serialize operations due to heavy electrical loading and shared timing constraints). The MDBs enhance signal integrity and timing margins at high speeds. By re-driving signals and isolating the memory channel from direct DRAM loading, they mitigate degradation caused by attenuation, reflections, and crosstalk, enabling sustained transfers at very high MT/s with cleaner, more reliable signaling. While this architecture introduces approximately two nanoseconds of latency, the benefits of rank multiplexing, such as increased effective bandwidth and improved parallelism (which brings 2x more cache lines), help offset this delay under idle and loaded conditions, as demonstrated in the benchmark results section.

MRDIMMs leverage the same DRAM devices and interface as DDR5 RDIMMs, enabling them to be a drop-in replacement in systems designed for the MRDIMM technology. This compatibility means that servers capable of MRDIMM operation can interchange RDIMMs and MRDIMMs without requiring hardware modifications, helping preserve existing infrastructure while unlocking higher bandwidth and capacity benefits. MRDIMM is a JEDEC-standardized technology.

Generation	Effective Bandwidth	DRAM Backplane Speed	Support
Gen. 1	8000 MTs & 8800 MTs	4000/4400	Intel Xeon 6 with P-Cores
Gen. 2	10400 MTs & 12800 MTs	6400	Next Generation Intel, AMD Venice

Table 3. Multiplexed Rank Registered Dual Inline Memory Device (MRDIMM) generations

Table 3 shows Micron's MRDIMM generations, speed grades, and platform support. Gen. 1 supports 8000–8800 MT/s. Micron's MRDIMM Gen. 2 is expected to reach 12800 MT/s, further increasing the bandwidth by 45% compared to MRDIMM Gen. 1. Like the first generation of MRDIMM, hosts using Gen. 2 will transparently run the memory bus 2x faster to increase bandwidth while data is transparently multiplexed across ranks. Note that the MRDIMM performance can change depending on the deployed system platform constraints (e.g., Intel® GNR-SP systems downclock MRDIMM Gen. 1 to 8000 MT/s).

Benchmarks and workloads

To evaluate the performance characteristics of RDIMM and MRDIMM memory architectures under various workloads, we designed a comprehensive benchmarking framework that captures both application and system-level metrics. On each workload evaluation, the only change is the RDIMM/MRDIMM devices, while all other hardware and software are kept the same. Table 4 identifies and summarizes the benchmarks used.

Benchmark	Characteristics
Microbenchmark	Intel® Memory Latency Checker (MLC) This benchmark, developed by Intel, measures memory latencies and bandwidths across interconnects, CPU caches, and memory hierarchies. MLC evaluates memory performance in three modes: idle latency, loaded latency, and bandwidth. **Idle latency**: Measures memory latency using dependent pointer-chase loads with a 64 B stride (128 B for random access) and a 200 MB buffer. **Loaded latency**: Measures latency while other CPUs generate bandwidth, showing latency versus bandwidth by sweeping injection delays. **Bandwidth tests:** Measures memory bandwidth by streaming through large buffers (100 MB per read/write buffer) with configurable traffic mixes.
Bandwidth-Intensive workloads	OpenFOAM (Computational Fluid Dynamics) POT3D (Solar Physics) SPECCPU (Includes several sub-benchmarks) Llama 3 Chatbot (CPU LLM inference) Workloads streaming large data volumes at high rates, with performance scaling almost linearly with memory bandwidth. HPC applications like OpenFOAM and POT3D fit this category due to their repeated read/write operations on large arrays, making bandwidth the bottleneck. Evaluating MRDIMM focuses on time-to-solution and power efficiency, as these workloads stress memory throughput.
Large Capacity workloads	Apache Spark SVM (Support Vector Machine)
	Graph Algorithm Platform Benchmark Suite (GAP BS)
	Vector DB (RAG) (Retrieval Augmented Generation) Like Apache Spark and graph analytics, need large memory to avoid disk I/O. MRDIMM's higher capacity allows in-memory processing at scale, optimizing differently than pure bandwidth. Performance is measured by training scalability (SVM), graph traversal efficiency, and query latency under large data footprints

Table 4. MRDIMM Performance Evaluation Categories and Experimental Goals

Although bandwidth influences all these workloads, the categories reflect their dominant performance constraint: throughput for HPC and capacity for big data, vector databases, and graph analytics.

System Architecture

We evaluated MRDIMM performance across multiple benchmarks using the two system configurations shown in Figure 2 and table 5. The first setup is based on a system with two Intel Xeon 6 processors with 96 P-cores, and the second setup is a single socket system with a single Xeon 6 with 128 P-cores.

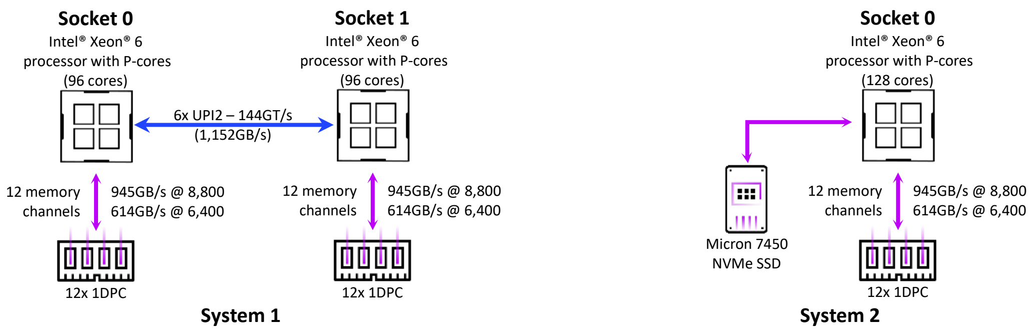


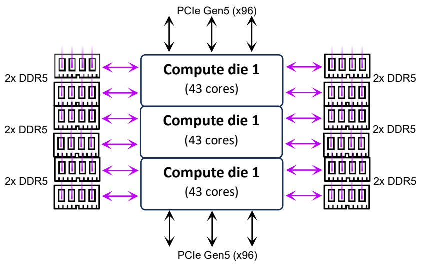
Figure 2: System configurations 1 and 2 used on MRDIMM benchmarks.

Hardware configuration		Software specifications	
CPU	1x Intel Xeon 6 with 128 P-Cores / 2x Intel Xeon 6 with 96 P-cores	**OS**	Alma 9.3, Linux Kernel 6.12
Memory	12x/24x DDR5 RDIMM/MRDIMM	**Intel MLC**	3.11b
Network	Mellanox Technologies - MT2892	**Intel Hibench**	7.1.1
Storage	Micron_7450_MTFDKBA480TFR 480GB	**OpenFoam**	11

Table 5: System Hardware and Software Configuration

All benchmarking experiments ran on an AlmaLinux 9.3 environment, configured with the latest, stable Alma Linux kernel (BIOS and firmware updated before benchmark execution). For high-performance computing (HPC) workloads, Simultaneous Multithreading (SMT) was disabled as it did not improve performance. Sub-NUMA Clustering (SNC) was enabled across all test scenarios to optimize memory locality and reduce inter-compute-die latency. Furthermore, system tuning was applied using the tuned utility, with the performance profile set to the HPC preset to ensure deterministic and optimized execution characteristics.

CPU: Intel Xeon 6 with P-Cores



NUMA node memory	NUMA node process		
	0	**1**	**2**
0	273,923	214,245	214,113
1	213,269	274,329	214,199
2	212,641	213,342	274,379

Bandwidth (MB/s): 64GB MRDIMM @ 8,800 MT/s

NUMA node memory	Numa node process		
	0	**1**	**2**
0	109.3	124.0	149.7
1	120.4	100.6	122.3
2	140.2	120.1	102.1

Latency (ns): 64GB MRDIMM @ 8,800 MT/s

Figure 3: CPU Architecture. The left shows Intel Xeon 6 with P-cores compute-die architecture. Right shows the highest bandwidth and lowest latency in green for cross-compute die communication. Sub-Numa Cluster option is enabled.

MRDIMM was evaluated on Intel Xeon 128-core and 96-core processor variants of Intel's Xeon 6 performance lineup (P-Core). Modern Intel Xeon and AMD EPYC server CPUs employ chiplet-based architectures, where multiple smaller dies (chiplets or tiles) are interconnected to function as a single processor. Intel Xeon 6 has a multi-die architecture and, since Sapphire Rapids, introduces non-uniform memory access (NUMA) topology, where each die has local access to a subset of memory channels, resulting in varying memory access latencies depending on where the process runs and where it accesses data from. We enabled Sub-NUMA Clustering (SNC) in the system BIOS to optimize performance in such a topology as illustrated in Figure 3. SNC partitions the processor into three NUMA domains, aligning each compute die with its local memory and I/O resources. This configuration allows the operating system and applications to make topology-aware scheduling decisions, ensuring that processes are preferentially mapped to cores local to the memory they access. This is particularly critical in memory bandwidth and latency-sensitive workloads, where minimizing remote memory access can significantly improve throughput and latency.

Results and analysis

This section presents the experimental results obtained from evaluating MRDIMM and RDIMM configurations under a range of memory-intensive workloads noted earlier. The analysis focuses on key performance metrics for each workload and system configuration. Subsections provide detailed microbenchmark results, comparative performance trends, and insights into architectural implications for MRDIMM in different environments. Results are organized into subsections that detail microbenchmark outcomes, application-level performance, and architectural impact for MRDIMM deployment in high-performance computing (HPC), artificial intelligence (AI) inference, data analytics, and graph processing environments.

All experiments were conducted using rigorously controlled hardware and software environments, as described in Section 3. The only variable modified between test runs was the memory module type (RDIMM vs. MRDIMM), ensuring a fair and isolated comparison. The results presented herein are intended to elucidate the practical benefits and trade-offs of MRDIMM technology, providing quantitative insights into its impact on modern data center workloads. Please note that results may slightly vary on different system configurations, software, and even across BIOS versions in the same system.

Microbenchmark

Intel Memory Latency Checker (MLC) showed that the memory subsystem with MRDIMM could scale performance independent of capacity, offering higher sustained bandwidth and lower loaded latency in every tested access pattern compared to RDIMM. In these tests, we used the System 2 architecture.

Figures 4, 5, and 6 compare memory bandwidth across eight configurations and five memory access patterns: 1:1 Reads-Writes, 2:1 Reads-Writes, 3:1 Reads-Writes, Stream-triad (2R:1W with streaming writes), and Read-Only. The configurations include standard RDIMMs at different ranks and speeds and high-capacity MRDIMMs (128GB).

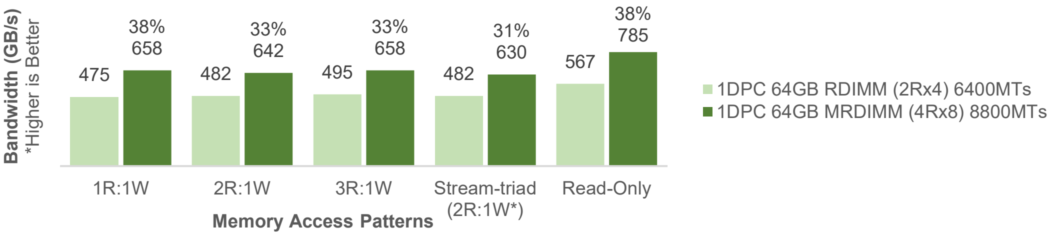


Figure 4: Microbenchmark maximum bandwidth (64GB MRDIMM)

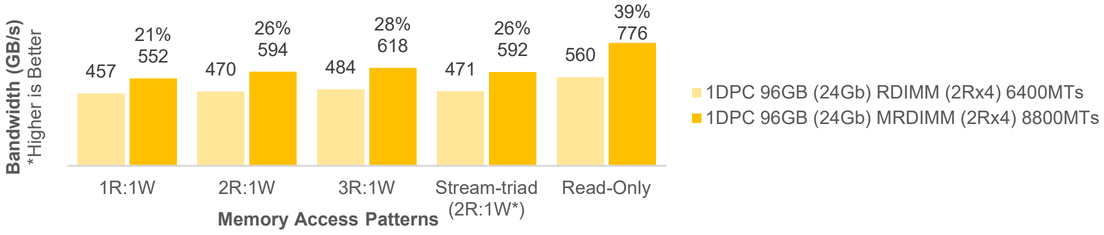


Figure 5: Microbenchmark maximum bandwidth (96GB MRDIMM)

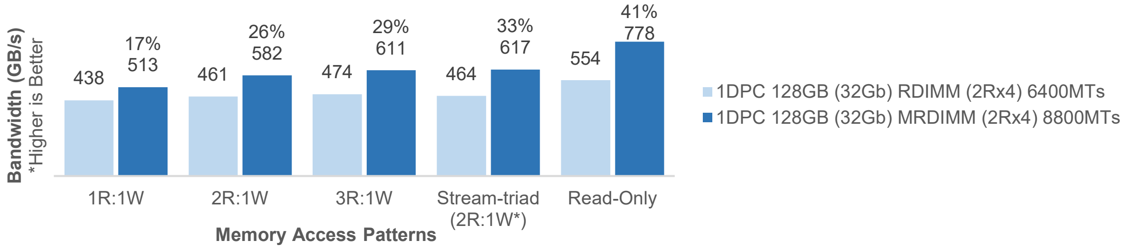


Figure 6: Microbenchmark maximum bandwidth (128GB MRDIMM)

Figures 4-6 show MRDIMM bandwidth gains with different memory access patterns. The theoretical uplift for MRDIMM operating at 8800 MT/s compared to RDIMM at 6400 MT/s is 37%. However, the actual performance can vary based on the bus efficiencies of

MRDIMM and RDIMM on different access patterns. MRDIMM showed from 38% to 41% higher bandwidth on a read-only access pattern, outpacing speed grade gains (37%) due to higher read-only bus efficiency than RDIMM devices in the evaluated platform (as defined in the System Architecture section).

Dual-rank (2R) MRDIMM can show slightly better performance in read-only since continuous reads avoid read-to-write turnaround penalties, and the MRDIMM accesses both ranks as a single entity (due to the DQ/DQS-to-MDB constraint), eliminating rank-switch overhead. Since ranks are accessed in pairs on 2R MRDIMM, each rank pair access is comparable to single rank access on a 1R RDIMM. On a quad-rank MRDIMM, the additional pair of ranks improved performance for mixed access patterns. 4R MRDIMMs allows the device to better utilize the available banks under more relaxed read-to-write turnaround compared to 2R MRDIMM, even though rank-to-rank switching introduces some latency. Effectively, going from dual rank to a four rank MRDIMM is comparable to going from one rank to a dual-rank RDIMM.

Also, note that 64GB MRDIMM devices use 16Gb DRAM components in contrast to 96GB MRDIMM and 128GB 2Rx4 MRDIMM, which uses 24Gb and 32 Gb components respectively. As standardized for DDR5 (JEDEC standard), 24Gb and 32Gb DRAM components exhibit higher Refresh Cycle Time (tRFC) relative to their 16Gb counterparts. Consequently, 16Gb based devices can exhibit a slight edge on some workloads.

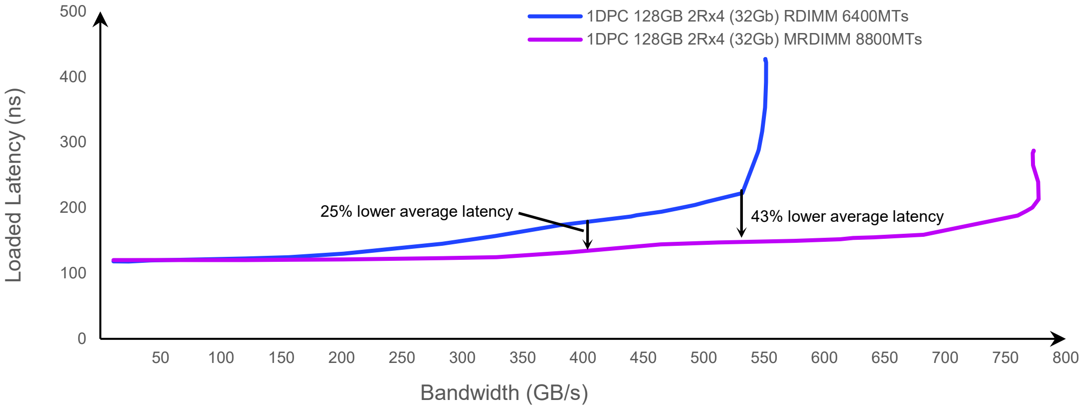


Figure 7. Intel MLC loaded Latency read-only benchmark with 64B Stride and 1GB buffer per core. Compares MRDIMM against RDIMM.

Figure 7 shows MRDIMM and RDIMM latencies under varying memory load conditions. MRDIMM demonstrates improved latency starting at approximately 150GB/s. MRDIMM lead increases as load increases, reaching ~43% at the point RDIMM latency grows exponentially (read queues saturation). Other capacity SKUs show similar improvements. Memory latencies will often exhibit three stages [2]. The initial stage is characterized by latencies that remain constant as memory requests are serviced without additional delay, executed as they arrive at the memory controller. The second stage shows linear latency growth where the command queues increase but are not full. However, we see sharp latency growth in the third stage as the memory controller queue reaches saturation.

In the third stage, full memory controller queues cause the CPU to stall, particularly when the load instructions are on the critical path and cannot be retrieved until data returns from DRAM. Modern out-of-order CPUs typically delay instruction retirement or block dispatch when internal buffers (e.g., Reorder Buffer, Load-Store Queue) are full. These stalls manifest as pipeline bubbles, idle cycles where no useful computation occurs, reducing overall instruction throughput. Note that sudden spikes in memory queue saturation can occur even in non-bandwidth-bound applications, due to DRAM locality effects. DRAM locality refers to how memory accesses are distributed across the memory system, whether concentrated in specific banks, ranks, or channels. When many accesses target the same memory region, it can lead to contention and queuing delays, even if the overall bandwidth usage remains low. This localized pressure can temporarily saturate memory queues, impacting performance unpredictably.

MRDIMM’s higher bandwidth helps mitigate memory bottlenecks by improving queuing efficiency and reducing the risk of controller queue saturation. With less congested queues, CPU stalls occur less frequently and for shorter durations than traditional RDIMMs, resulting in smoother overall system performance.

Latency Histogram. Traditional analyses of memory performance often rely on average latency metrics to characterize the trade-off between bandwidth and latency. However, such averages can obscure critical insights, particularly on highly skewed latency distributions. Figure 8 represents a latency histogram comparing the loaded latency distributions of a 128GB MRDIMM SFF 2Rx4 operating at 8800 MT/s against a 128GB RDIMM 2Rx4 at 6400 MT/s across varying levels of memory load. To illustrate, latencies are binned in 8ns intervals from low to high memory load. Figure 8 represents the distribution of MRDIMMs latency under low and heavy memory load, while Figure 9 presents the distribution of MRDIMMs latency under heavy memory load.

Note how MRDIMM shows 80% of latency measurements under ~204ns, whereas RDIMM has only 53%. Also, since the experiments are the same with only a change in memory modules, note how both have the same % of occurrences under 4ns, corresponding to CPU cache hits.

MRDIMM exhibits a significantly tighter clustering of latencies within the lower bins, accompanied by a markedly thinner tail, indicating fewer high-latency outliers. The pronounced peaks in the MRDIMM distribution reflect a higher concentration of memory transactions completed within lower latency thresholds, even as the system load increases. This behavior shows MRDIMM's ability to deliver more consistent and predictable memory access times, which can be advantageous even for latency-sensitive applications that are not strictly bandwidth-bound.

Figure 9 presents a histogram with a closer examination of the high load. It reveals that approximately 80% of MRDIMM latencies fall below 204ns (approximately 2 times the idle latency of RDIMM/MRDIMM), compared to only 50% for the RDIMM counterpart. Despite a low load, MRDIMM showed 16% fewer occurrences of extremely high latencies of 2 microseconds or more compared to RDIMM. This demonstrates MRDIMM's capability to sustain low latency on high and even low memory load scenarios, contributing to a more stable and responsive system.

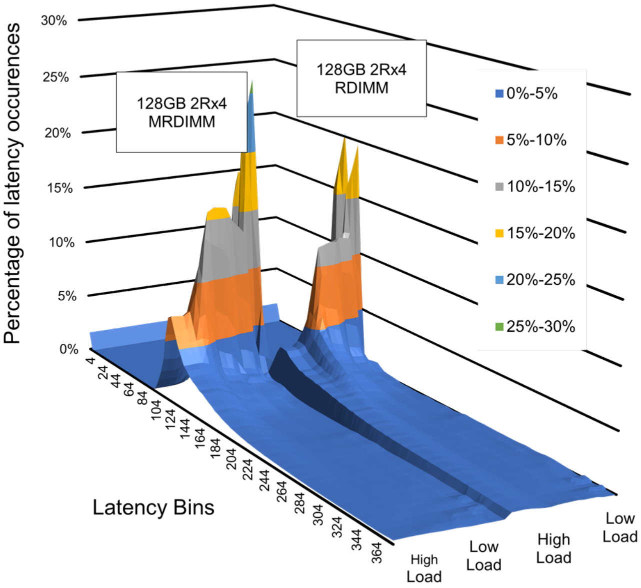


Figure 8. MRDIMM loaded latency histograms – low/heavy memory load

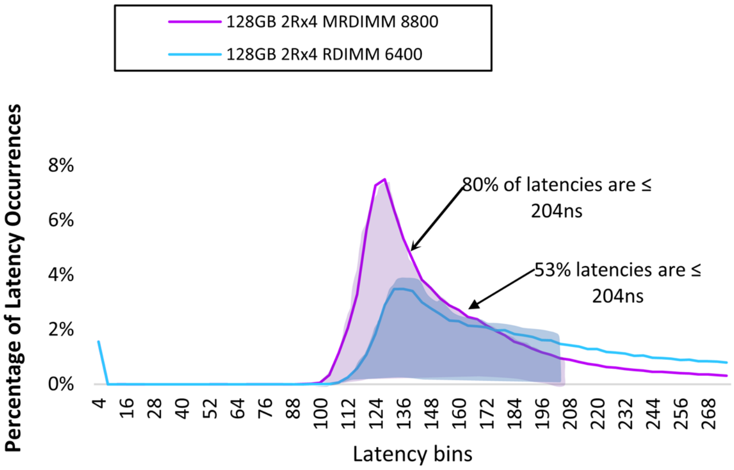


Figure 9. MRDIMM loaded latency histogram – heavy memory

<u>MRDIMM Gen. 2 Preview</u>

Micron's second-generation MRDIMM is engineered to double bandwidth compared to RDIMM at 6400MT/s. This enhanced bandwidth provides greater headroom for maintaining lower latencies under high-load conditions, even outperforming first-generation MRDIMMs at 8800 MT/s on similar memory loads. Additionally, the new Micron 128GB 4Rx8 MRDIMM Gen. 2 modules are expected to deliver even lower relative power consumption and enhanced performance, particularly for workloads with mixed access patterns, which are common in large memory databases and analytics. Together, these advancements position MRDIMM Gen. 2 to set a new benchmark for memory subsystem responsiveness and scalability in next-generation data center platforms.

Bandwidth-Intensive Workloads

This section shows the results on bandwidth-sensitive applications such as OpenFOAM.

OpenFOAM results

OpenFOAM (Open Field Operation and Manipulation) is an open-source, C++ based toolbox primarily used for Computational Fluid Dynamics (CFD) simulations and is maintained and developed under the governance of the OpenFOAM Foundation. OpenFOAM can tackle problems ranging from turbulent flows, incompressible and compressible fluids, chemical reactions, heat transfer, and electromagnetics. Also, OpenFOAM simulations often involve intricate computations on vast grids or meshes, necessitating high memory bandwidth to transfer data quickly and low latency for fast data access. Memory capacity is also paramount to hold intermediate computation results and extensive datasets, especially when simulating large or complex systems.

We executed the canonical OpenFOAM motorbike on a high-resolution 600 x 500 x 500 mesh with OpenMPI running one process per physical core. As shown in Figure 10, the 128GB TFF MRDIMM (4Rx4) showed a 1.35x speedup compared to the 128GB 2Rx4 RDIMM.

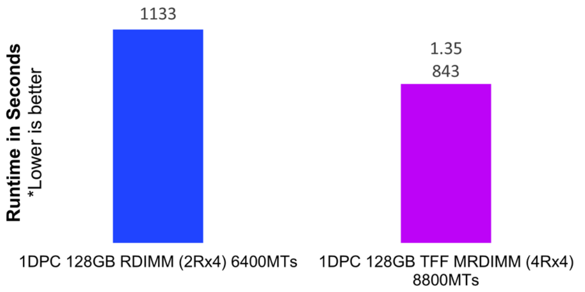


Figure 10: OpenFOAM execution time (1P) for the Motorbike mesh (600 x 500 x 500) with MRDIMM and RDIMM.

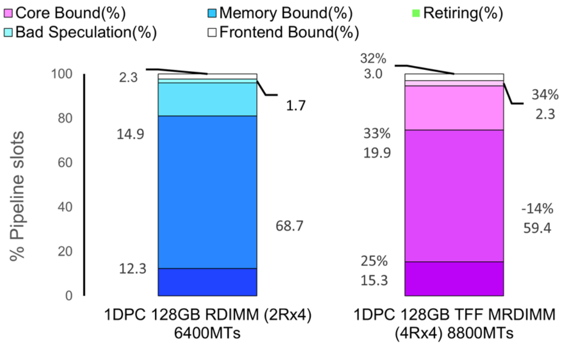


Figure 11: Top-Down Microarchitecture Analysis.

Figure 11 represents TMA (Top-Down Microarchitecture Analysis) results, which indicate that RDIMM systems have significant memory stalls, with 68.7% of pipeline slots being memory-bound. MRDIMM reduces this to 59.4%, improving memory efficiency and increasing instruction retirement from 14.9% to 19.9%. These shifts highlight MRDIMM's effectiveness in alleviating memory bottlenecks and enhancing the execution of memory-intensive workloads like OpenFOAM. Moreover, the MRDIMM 128GB TFF 4Rx4 was able to save energy since the task was completed faster.

Figure 12 shows how the reduction in runtime translates directly into lower task energy. Despite drawing more power per unit time, a server equipped with MRDIMM enables a shorter execution window, resulting in a smaller total energy footprint for the task. This efficiency is critical in high-performance computing environments, where energy consumption scales with workload duration. By accelerating task completion and reducing memory-induced stalls, MRDIMM improves performance and enhances energy efficiency at the system level.

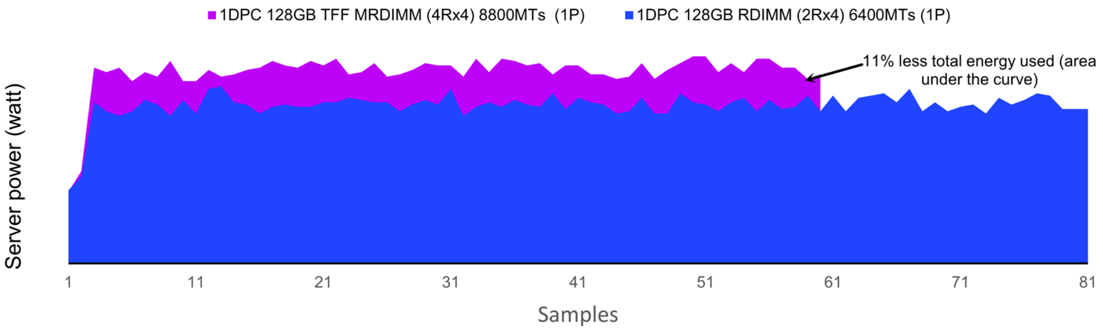


Figure 12: OpenFOAM system power consumption versus time.

POT3D results

POT3D benchmark models the solar coronal magnetic field by solving the Laplace equation in 3D spherical coordinates using a preconditioned conjugate gradient solver. It is a memory-bound workload where performance is limited more by data movement than computation. The solver's sparse matrix-vector operations generate high memory traffic with low arithmetic intensity, making

DRAM bandwidth a critical performance factor. When memory demand exceeds available bandwidth, DRAM backpressure occurs, stalling compute units and reducing throughput. As such, POT3D is also part of the SPEChpc™ 2021 suite and is a valuable benchmark for evaluating memory subsystem efficiency and bandwidth scalability in modern HPC architectures.

Like OpenFOAM, POT3D can also depend on memory bandwidth on large meshes despite using structured grids and finite difference methods, which tend to have more regular memory access patterns (which can be more cache efficient).

Figure 13 shows that we reached about 1.36x speedup with MRDIMM. TMA analysis revealed similar gains to OpenFOAM, with decreased pipeline stalls caused by DRAM bandwidth and increased retiring instructions.

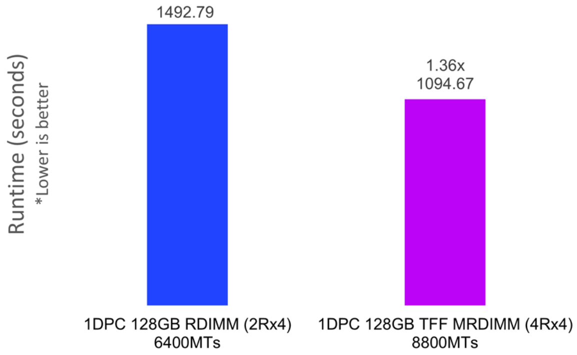


Figure 13: POT3D MRDIMM and RDIMM runtime.

SPEC CPU® 2017 results

SPEC CPU benchmarks are industry-standardized tests developed by the Standard Performance Evaluation Corporation (SPEC). The SPEC CPU 2017 benchmark package includes 43 kernels that measure compute-intensive performance, focusing on the processor, memory hierarchy (cache and main memory), and C, C++, and Fortran compilers with their optimizers. The final score of SPEC CPU is based on the geometric mean of several tests.

Table 6 represents SPEC CPU® 2017 benchmark scores across memory configurations, quantifying the effect of MRDIMM on integer and floating-point workloads (intrate: spins multiple copies of each benchmark, simulating a high-throughput environment. It is designed to measure the performance of a system by running multiple threads or processes simultaneously. FPspeed: runs one copy of each benchmark, focusing on the throughput per unit of time. It is designed to measure the performance of a single-threaded workload.

Memory config.	intrate		intspeed		fprate		fpspeed	
RDIMM 2Rx4	1520	Baseline	8.49	Baseline	1770	Baseline	297	Baseline
MRDIMM 4Rx8	1560	3%	9.52	12%	1970	11%	318	7%

Table 6: SPEC CPU® 2017 benchmark scores

Relative to RDIMM 2Rx4 as the baseline, MRDIMM 4Rx8 improves the scores, reaching up to +3% for intrate, +11% for fprate, and +7% for fpspeed. However, SPEC CPU scores are a geometric mean over multiple benchmarks; they can mask per-application variability. Notable uplifts on SPEC CPU 2017 with MRDIMM include:

- ≥30% uplift: 519.lbm_r (fluid dynamics), 620.omnetpp_s (discrete-event simulation).
- 20–30% uplift: 521.wrf_r (weather forecasting), 549.fotonik3d_r (computational electromagnetics), 554.roms_r (regional ocean modeling), 602.gcc_s (GNU C compiler), 603.bwaves_s (explosion modeling), 628.pop2_s (ocean modeling).
- 10–20% uplift: 503.bwaves_r (explosion modeling), 520.omnetpp_r (discrete-event simulation), 600.perlbench_s (Perl interpreter), 605.mcf_s (route planning), 607.cactuBSSN_s (numerical relativity), 619.lbm_s (fluid dynamics).

The 4Rx8 MRDIMM shows an additional 5-10% improvement on 519.lbm_r, 521.wrf_r, 549.fotonik3d_r, and 554.roms_r. Power measurements indicate that an average 64 GB 4Rx8 MRDIMM delivers up to 10% lower DRAM power than a 64 GB 2Rx4 MRDIMM.

MRDIMM benefits memory-bound workloads by lowering effective memory latency and increasing bandwidth. Improvements in intspeed and fpspeed underscore better single-thread performance, which is critical for many real-world applications, while throughput gains corroborate broader responsiveness across conventional CPU workloads.

Llama 3 8B results

Meta's Llama represented a significant leap forward in open-source large language model (LLM) technology. Llama 3, offered in 8B, 70B, and 400B parameter configurations, was designed to support both research and commercial applications. Its versatility extends to intelligent assistants and tasks such as content generation, translation, and question answering. The 8B model was optimized for systems with at least 16GB of RAM, while the 70B model benefits from configurations with 32GB or more.

CPU-based inference remains a practical option for smaller models like Llama 3 8B. To evaluate real-world deployment performance, we tested Llama 3 using Intel's OpenVINO toolkit, an optimization framework engineered for efficient inference on Intel platforms. OpenVINO focuses on low-latency execution and dynamic model scaling, making it well-suited for edge-to-cloud AI strategies. CPU-based inference for smaller LLMs such as Llama 8B can be sensitive to memory bandwidth and latency. During chatbot interactions, each token generation requires rapid access to model weights and attention to states. Limited bandwidth can throttle throughput and token latency, directly impacting response time and user experience. Increasing memory bandwidth accelerates token processing, while lower latency reduces response delays, resulting in faster, more fluid chatbot interactions.

Our benchmark used OpenVINO's benchmark app and Llama 3 8B model on a system equipped with dual 96-core CPUs (System Setup 1). The evaluation used 1,024 input tokens, a single batch, dynamic quantization to 8 bits, and ten concurrent streams (in OpenVINO, streams represent parallel executions of the same model, distinct from batching, which imposes different parallelization constraints). We compared two memory configurations: 128GB TFF 4Rx4 MRDIMM at 8,800 MT/s versus 64GB 2Rx4 RDIMM at 6,400 MT/s.

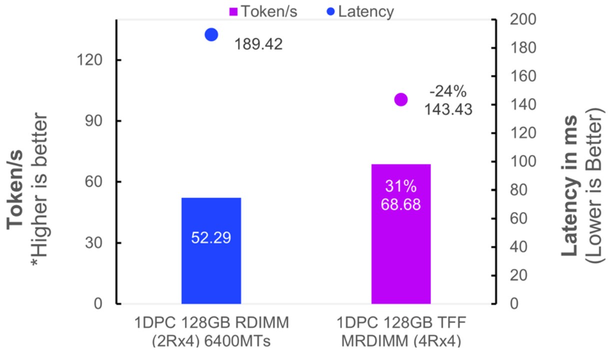


Figure 14: Average sustained token throughput and latency (Llama 3 8B)

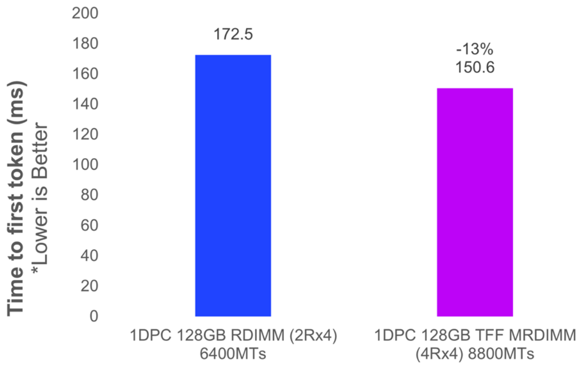


Figure 15: First token latency (Llama 3 8B)

As shown in Figure 14, using 1DPC 128GB RDIMM (2Rx4) at 6400 MT/s, the baseline configuration achieved a throughput of 52.29 tokens per second with a latency of 189.42 milliseconds. In contrast, the optimized configuration with MRDIMM, featuring 1DPC of Micron's 128GB TFF MRDIMM (4Rx4) at 8800 MT/s, delivered a 31% increase in throughput, reaching 68.68 tokens per second. Latency was also significantly reduced by 24%, dropping to 143.43 milliseconds. Also, MRDIMM showed a 13% reduction Time to First Token as shown in Figure 15.

The MRDIMM configuration enhanced token generation speed and improves responsiveness, critical for real-time applications. When combined with OpenVINO's deployment capabilities, Llama 3 can be effectively scaled across diverse environments, from low-power edge devices to high-performance cloud infrastructures.

MRDIMM Gen. 2 Preview

Performance uplift for memory bandwidth-bound workloads typically scales nearly linearly with increases in available memory bandwidth. Micron's MRDIMM Gen. 2 is expected to show faster task completion times and improved energy for demanding applications such as OpenFOAM and Pot3D. MRDIMM Gen. 2 is also expected to deliver higher scores on SPEC CPU benchmarks than Gen. 1, as many SPEC workloads are highly memory bandwidth bound. For CPU-based inference workloads, including models like Llama 3, MRDIMM Gen. 2 is also projected to deliver performance gains proportional to its bandwidth uplift. This enables larger batch sizes, faster time-to-first-token, and higher token throughput, outperforming both RDIMM at 6400 MT/s and MRDIMM Gen. 1 at 8800 MT/s configurations. These improvements position MRDIMM Gen. 2 as a critical enabler of next-generation data center efficiency and scalability.

Large Capacity Workloads

As CPUs and workloads expand in complexity and scale, the need for bandwidth and reduced latency grows commensurately with capacity. While most HPC applications have long emphasized memory bandwidth performance, workloads such as large-scale data analytics, graph processing, and vector databases demand architectures capable of accommodating expansive, memory-resident datasets while supporting high bandwidth to realize optimal computational efficiency.

Vector database results

A vector database is a specialized database designed to store and search high-dimensional vector embeddings, representing data like text, images, audio, or video. Unlike traditional databases that rely on exact matches, vector databases enable similarity search using distance metrics (e.g., cosine similarity or Euclidean distance), making them ideal for semantic search applications, recommendation systems, image, and speech recognition. Also, Vector databases play a central role in Retrieval-Augmented Generation (RAG) systems by enabling fast and scalable similarity search across embedded representations of documents. The performance of these systems is heavily influenced by the underlying hardware, particularly memory bandwidth, capacity, and access latency, which affect both retrieval speed, search, and overall throughput.

To evaluate MRDIMM performance, we utilized the Milvus vector database and a curated subset of the Falcon-RefinedWeb dataset using System Setup 1. This dataset contains 968 million deduplicated and stringently filtered rows sourced from CommonCrawl, totaling 2.8TB (1.68TB compressed) and encompassing an estimated 500 to 650 billion tokens. We selected a subset specifically sized to fit within the memory footprint of RDIMM/MRDIMM modules. We employed the Hierarchical Navigable Small World (HNSW) algorithm for indexing, an in-memory strategy that enables high-speed access times. Embeddings were generated using the all-MiniLM-L6-v2 model, which produces 384-dimensional vectors with 23 million parameters. The resulting index size was approximately 1.5TB. Lastly, we executed 15,000 parallel queries against the vector database to stress-test the memory subsystem.

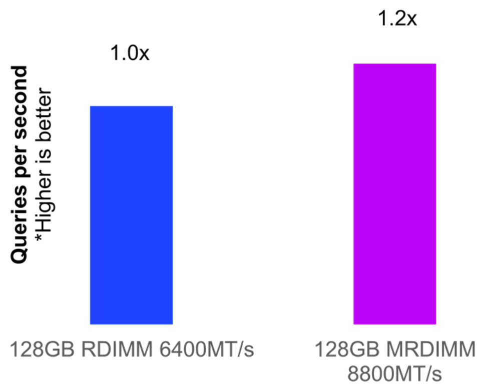


Figure 16: Vector database queries per second (QPS) improvement with MRDIMM (15,000 parallel queries).

Figure 16 shows the evaluation comparing 128GB RDIMM at 6400 MT/s (baseline, or 1.0x in Figure 16) to 128GB MRDIMM at 8800 MT/s. MRDIMM showed 20% throughput uplift in query per second.

Apache Spark results

Apache Spark is a distributed computing framework designed for large-scale data processing. It supports many workloads, including batch processing, real-time streaming, machine learning, graph analytics, and SQL-based querying. Unlike traditional Hadoop MapReduce, which relies heavily on disk I/O, Spark leverages in-memory computation to significantly accelerate performance by minimizing disk access. We used the HiBench benchmark suite to evaluate the Apache Spark MLlib implementation of the Support Vector Machine (SVM) algorithm and provide a standardized methodology for assessing Spark workloads across various domains such as machine learning, graph processing, and SQL queries. This study focuses on Spark's Support Vector Machine (SVM) implementation running on a 3TB dataset. This workload highlights the benefits of MRDIMM's increased capacity, showing how MRDIMM enables the system to handle larger in-memory datasets while improving effective bandwidth.

To assess the effect of MRDIMM on large-scale data analytics workloads, we conducted a comparative evaluation of Apache Spark workloads using the System 1 configuration with three memory configurations: 128GB RDIMM, 128GB MRDIMM, and 256GB MRDIMM. Apache Spark exhibits pronounced sensitivity to memory capacity, bandwidth, and latency, particularly during shuffle-intensive and iterative operations. We extended SVM's default Big Data input size to stress-test scalability from 500GB to 2.3TB.

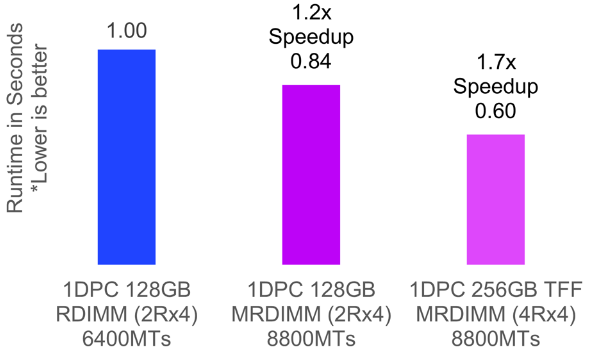


Figure 17: Apache SPARK AI SVM relative runtime performance.

Figure 17 shows Apache Spark runtime comparisons across the three memory configurations. MRDIMM at 8800 MT/s with 2Rx4 reduces runtime by 16% over RDIMM, while the 4Rx4 256GB MRDIMM achieves a 40% reduction. These results highlight MRDIMM's ability to accelerate data processing by improving memory bandwidth and parallelism, making it highly effective for memory-bound analytics workloads.

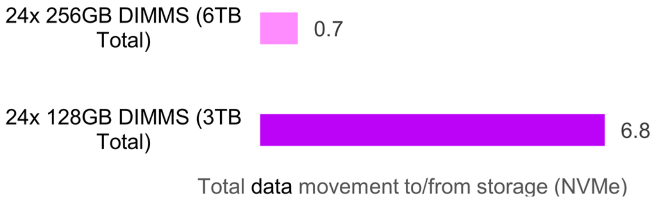


Figure 18: Total amount of data transferred between main memory and storage

In Figure 18, the relative I/O to storage is normalized to 1.0 for the 3TB configuration, while the 6 TB configuration shows a tenfold reduction (0.1×). This indicates that systems with higher memory capacity can buffer and cache significantly more data, reducing the frequency and volume of I/O operations to slower storage tiers. This reduction in I/O improves performance, reduces wear on storage devices, and lowers energy consumption associated with disk access.

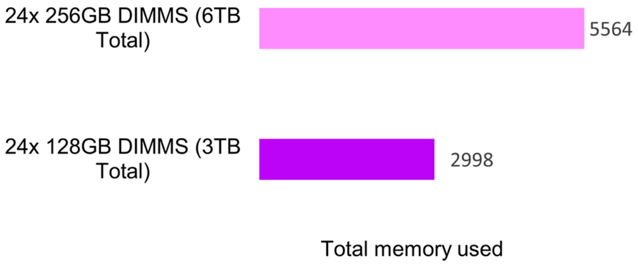


Figure 19: Average DRAM memory usage

Figure 19 complements this by showing that the 6TB configuration enables nearly double the memory usage (5564GB vs. 2998GB), demonstrating that applications can scale their in-memory working sets when more capacity is available. This is particularly beneficial for data-intensive workloads such as Apache Spark, where larger memory footprints reduce the need for disk-based shuffling and intermediate storage.

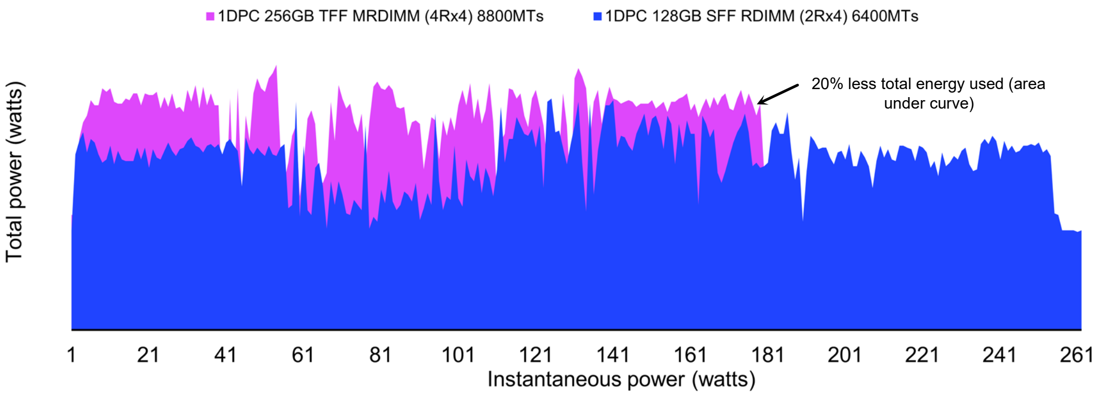


Figure 20: Comparing 1DPC 256GB (32Gb) TFF MRDIMM (4Rx4) 8800 MT/s and 1DPC 128GB (32Gb) RDIMM (2Rx4) 6400 MT/s energy to complete Apache Spark SVM AI workload (2.9TB dataset).

Figure 20 represents MRDIMM power energy task savings. Since we see a 1.7x speedup, the system completes the task fast enough to reduce energy consumed by 20%. Together, these results show the value of high-capacity MRDIMM configurations in enabling more efficient memory-resident computation, reducing the I/O overhead, and improving overall system throughput and energy efficiency.

Graph Algorithm Platform Benchmark Suite (GAP BS)

The following results demonstrate Micron's MRDIMM improves execution efficiency and scalability across graph-centric analytics tasks compared to traditional RDIMM configurations using system configuration 2.

We compare 128GB 32Gb 2Rx4 MRDIMM 8800 with 128GB 32Gb 2Rx4 RDIMM 6400 MT/s, analyzing graph workloads at the GAP BS Scale Factor (SF) of 31 (which corresponds to 2.1 billion vertices and 34.1 billion edges). The system exhibits 1TB memory occupancy in this configuration, utilizing 66% of a 1.5TB overall memory capacity. These figures reflect unweighted graph structures; memory usage is expected to increase significantly for weighted graphs, where each edge carries additional metadata.

Graphs of this size highlight the importance of memory bandwidth and latency optimizations, areas where MRDIMM demonstrates clear advantages over RDIMM, enabling more efficient processing of large-scale graph analytics.

GAP BS Kernel	**Average execution time** (seconds, lower is better)		**Speed increase** (higher is better)	**Task-energy efficiency** (higher is better)
	RDIMM	MRDIMM		
Betweenness Centrality (BC)	305	264	1.15x	7%
Breadth First Search (BFS)	1.10	1.04	1.05x	5%
Connected Components (CC)	1.18	1.09	1.08x	20%
Page Rank (PR)	22.78	22.02	1.03x	5%
Single Source Shortest Path (SSSP)	13.58	11.92	1.14x	28%
Triangle Counting (TC)	1,961	1,763	1.11x	3%

Table 7: GAP BS performance on six kernels of the suite, comparing average runtime (in seconds) between RDIMM and MRDIMM configurations.

Table 7 shows the performance gains on 6 GAP-BS kernels: Betweenness Centrality (BC), Breadth-First Search (BFS), Connected Components (CC), PageRank (PR), Single Source Shortest Path (SSSP) and Triangle Counting (TC), MRDIMM consistently outperforms RDIMM, with a runtime speedup ranging from 3% to 15%. Micron's MRDIMM enabled up to 28% energy savings on task energy efficiency compared to RDIMM.

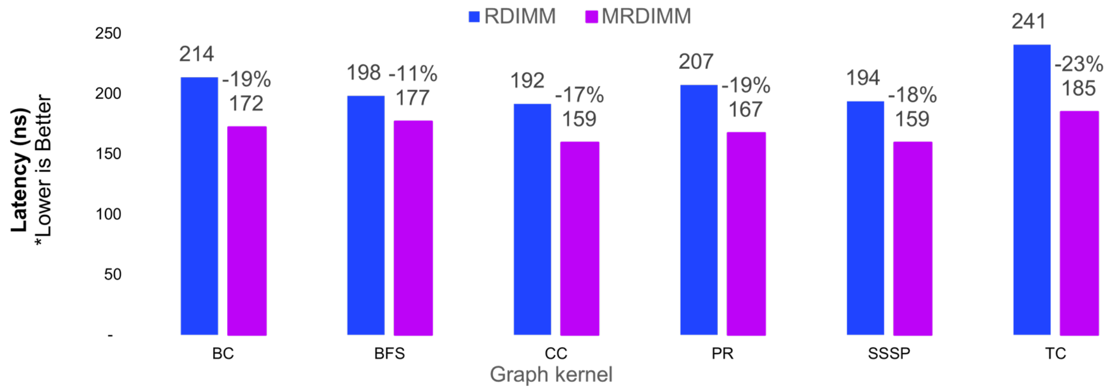


Figure 21. RDIMM and MRDIMM last level cache read miss latency (ns, lower is better)

Figure 21 represents the runtime speedup of six core kernels from the GAP Benchmark Suite, correlating the runtime speedup (in seconds) and CPI between RDIMM and MRDIMM configurations. This figure highlights the *Last Level Cache* (LLC) miss latency metric collected from the CPU performance counters with the Linux perf utility. The LLC miss latency represents the time it takes for the processor to retrieve data from main memory (DRAM) after a cache miss occurs in the last level of the CPU cache hierarchy (typically L3 cache). Micron's MRDIMM improves the latency of read requests to DRAM across all graph kernels, with improvements ranging from 11% to 23% lower latencies than RDIMM (these values are negative percentages).

These results highlight how Micron's MRDIMM can accelerate graph analytics workloads by reducing memory latency and improving bandwidth efficiency, which are critical for the irregular access patterns typical of graph processing.

MRDIMM Gen. 2 Preview

Next-generation platforms will expand memory channel support to 16, enabling significantly higher aggregate memory capacity for data-intensive workloads. MRDIMM Gen. 2, featuring higher bandwidth and lower latency across 16 memory channels, will allow CPUs with more cores to work more efficiently on big data applications such as Apache Spark, Vector, and Graph databases. Also, MRDIMM Gen. 2 will expand the bandwidth headroom to support PCIe Gen6 devices that need to perform direct memory access to main memory. All this will place MRDIMM technology as a key enabler for future data center deployments, offering greater throughput, energy-task efficiency, and scalability for memory-intensive and large capacity scenarios.

Discussion

The evaluation of MRDIMM across diverse workloads, ranging from HPC and AI inference to data analytics and graph processing, demonstrates its potential for consistent performance and power efficiency uplift over traditional RDIMM configurations. Micron's MRDIMM architectural bandwidth, latency, and capacity innovations translate into measurable improvements in real-world applications.

On bandwidth-bound applications, MRDIMM showed improved performance across all tested domains and consistently reduced execution time. In OpenFOAM, a memory bandwidth-bound CFD HPC application, MRDIMM achieved a 25.6% reduction in runtime and significantly improved CPU pipeline efficiency, reducing memory-bound stalls by over 23%. These findings validate Micron's MRDIMM bandwidth advantage while highlighting its ability to alleviate DRAM backpressure, a common bottleneck in HPC workloads.

In Apache Spark, MRDIMM's impact was twofold. The high-capacity 256GB TFF MRDIMM provided extra bandwidth and capacity compared with the 128GB 2Rx4 RDIMM configuration, reducing runtime by 40%. These gains were accompanied by a 10x decrease in I/O to storage and nearly double the memory utilization, underscoring the importance of bandwidth and capacity in data analytics. Caching larger working sets in memory directly reduced disk access, improving performance and task-energy efficiency.

Moreover, Micron's MRDIMM combination of high bandwidth, low latency, and large capacity directly addresses the core challenges of vector and graph databases. For vector databases, Micron's MRDIMM enables faster, more scalable similarity search by supporting larger in-memory indexes and delivering higher query throughput, which is critical as AI workloads and datasets expand. In graph analytics, MRDIMM's reduced latency and improved efficiency on graph traversal and computation on large, irregular datasets make real-time analytics feasible at scale. These advances position Micron's MRDIMM as a key enabler for next-generation, memory-resident data platforms, supporting performance and energy efficiency in demanding enterprise and AI applications.

One of the most compelling findings is MRDIMM's task-energy efficiency. Despite the higher instantaneous power draw, MRDIMM's faster task completion results in lower total energy consumption. This was clearly illustrated on OpenFOAM, Apache Spark, and Graph Analytics power profiles, where MRDIMM completed the tasks faster, resulting in a smaller energy footprint. The lack of sufficient memory bandwidth on RDIMM setups forces the CPU to idle, wasting precious cycles and time waiting for memory requests to complete. As such, the entire server will remain consuming power but not actually doing any work. Micron's MRDIMM low latency and higher bandwidth enable faster CPU access to data, effectively increasing its power efficiency. This has significant implications for sustainable computing, especially in data centers where energy costs and thermal constraints are critical.

Finally, the microbenchmark latency experiment demonstrated that even in compute-bound applications, where the primary bottleneck is the CPU's ability to process and retire instructions, MRDIMM can help reduce worst-case latencies. This can happen because even when the overall memory load is low, localized contention can still occur, such as when multiple threads simultaneously access the same memory banks within a module, causing occasional latency spikes. For mission-critical applications with strict latency requirements, a faster and more responsive memory subsystem with Micron's MRDIMM can mitigate these outliers, reducing extreme tail latencies (e.g., 99.9999th percentile) and delivering more consistent and predictable performance.

The Future Demand for DRAM Bandwidth

In addition to the escalating bandwidth demands driven by increasing core counts, PCIe-connected devices impose greater pressure on main memory bandwidth and capacity concurrently.

With the advent of PCIe 6.0, per-lane throughput will double to 8 GB/s, meaning a typical ×16 slot now offers up to 128 GB/s of bandwidth (bi-directional). High-end server CPUs feature up to 128 PCIe lanes, potentially on the order of 1 TB/s of aggregate I/O bandwidth flowing into or out of the processor. Meanwhile, next-generation DDR5 memory is reaching data rates of DDR5-8000 or DDR5-8800 (8.0–8.8 GT/s), which are expected on upcoming platforms. Coupled with wider memory interfaces (e.g., 16 memory channels in advanced server configurations), the main memory subsystem can likewise supply roughly 1 TB/s of throughput (for context, DDR5-8800 delivers ≈70 GB/s per channel, so 16 channels would provide on the order of 1.1 TB/s). This parity means that PCIe 6.0 devices on a 128-lane server can push data as fast as the memory can absorb it. Large NVMe arrays, NICs, or GPUs could collectively saturate the system’s memory bandwidth, where previously either the I/O or PCIe was the limiting factor. In short, next-gen servers are moving terabytes per second on both the I/O and memory fronts, so a balanced design is crucial to avoid bottlenecks.

Micron’s MRDIMM can help close the memory gap. To prevent main memory from becoming a performance bottleneck amid the exponential growth in I/O and CPU throughput, MRDIMM technology will be a critical enabler. Future iterations are projected to achieve data rates of up to 12,800 MT/s, delivering aggregate bandwidths approaching ≈1.6 TB/s in next-generation server platforms.

PCIe Gen6 servers may coincide with MRDIMM Gen. 2 (12,800 MT/s) support in next-generation servers planned for 2026/2027, optimizing memory subsystems for mixed CPU and DMA workloads. The extra bandwidth from Micron’s MRDIMM will improve bandwidth consistency during bursty accelerator phases such as checkpointing, RDMA, or RAID parity operations. Also, MRDIMM's higher per-socket memory bandwidth reduces reliance on PCIe oversubscription or DMA throttling, enhancing quality of service isolation via controller-level scheduling between accelerator streams and CPU cores.

Conclusion

We showed that Micron’s MRDIMM consistently outperformed RDIMM configurations across various workloads, including high-performance computing, AI inference, data analytics, and graph processing. These findings confirm MRDIMM’s capability to deliver higher bandwidth, lower latency, and increased capacity, translating into measurable improvements in application performance and overall system efficiency.

Key findings

- **Performance uplift:** MRDIMM modules operating at 8800 MT/s consistently outperformed RDIMMs at 6400 MT/s, with up to 41% higher bandwidth and 40% lower latency observed in microbenchmark testing. These improvements correlated with real-world applications, with notable reductions in runtime for memory-bound workloads such as OpenFOAM, Vector Databases, and Apache Spark AI workloads. MRDIMM Gen. 2 at 12800 MT/s devices are expected to show bandwidth gains commensurate with the theoretical increase of 2x compared to RDIMM at 6400 MT/s.
- **Improved energy efficiency**: Despite higher instantaneous power draw, MRDIMM's ability to accelerate task completion resulted in lower total energy consumption for a wide range of workloads. This is particularly significant for data centers seeking to optimize performance and sustainability. MRDIMM Gen. 2 will feature improved design which will increase power efficiency compared to MRDIMM Gen. 1.
- **Scalability and flexibility:** The availability of MRDIMM modules in capacities up to 256GB, including Tall Form Factor variants, allows for significant scalability in memory-intensive applications. This flexibility supports larger memory datasets, reduces dependence on slower storage tiers, and enhances overall system performance. On next-generation platforms supporting PCIe Gen6, MRDIMM Gen. 2 will play a critical role in expanding the bandwidth headroom, enabling both CPU and devices that depend on direct memory access (e.g., SSDs, GPUs, and NICs) to function efficiently.
- **Predictable latency**: MRDIMM's tighter latency distributions and reduced tail latencies contribute to more deterministic system behavior, essential for latency-sensitive and mission-critical workloads.

Micron's MRDIMM raises delivered GB/s/W and GB/s/pin, smoothing the performance on both cores and PCIe attached devices, reducing tail latency during contention, and extending the useful life of existing memory controllers and topologies as I/O speeds outpace conventional RDIMM scaling. By aligning memory subsystem capabilities with the escalating demands of PCIe 6.0 and increasingly core-dense processors, Micron's MRDIMM ensures sustained data delivery at scale. In doing so, it preserves architectural balance, mitigating the risk of memory starvation and enabling CPUs and PCIe devices to operate at full computational potential. MRDIMM is a foundational enabler for the next generation of intelligent, scalable, and energy-efficient AI, analytics, and database infrastructure. Its adoption accelerates innovation across a broad spectrum of AI and data-intensive applications, positioning memory as a strategic asset in the evolving landscape of data-centric computing.

References

Scott Beamer, Krste Asanovi´c, David Patterson. "The GAP Benchmark Suite". (2017) 1508.03619

Shang Li, Dhiraj Reddy, and Bruce Jacob. 2018. "A performance & power comparison of modern high-speed DRAM architectures." In Proceedings of the International Symposium on Memory Systems (MEMSYS '18). Association for Computing Machinery, New York, NY, USA, 341–353.

micron.com