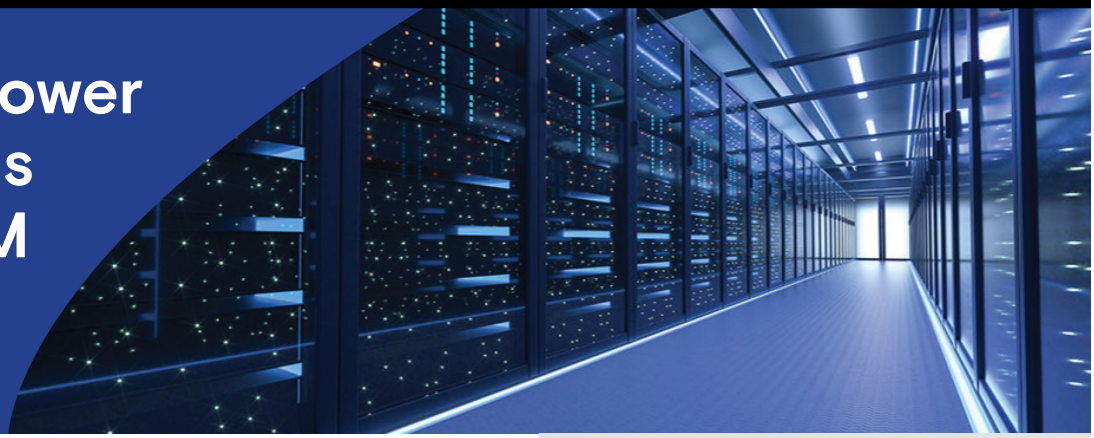
Unlock the power of more cores with MRDIMM



Unlock the performance potential of your servers

Micron® multiplexed-rank DIMM (MRDIMM) is Micron's new product line of high-performance, high-capacity DDR5[1] dual in-line memory module (DIMM) technology designed for the next generation of Intel® Xeon® processors, codenamed Granite Rapids (GNR), to expand bandwidth per core by supporting memory requirements of higher CPU core counts.

Designed to meet the rigorous demands of modern data centers, MRDIMM delivers faster data rates, lower latencies, and cost-effective high capacity for applications including high-performance computing (HPC), artificial intelligence (AI) inference, data analytics, and virtual multi-tenancy in cloud and data center environments.

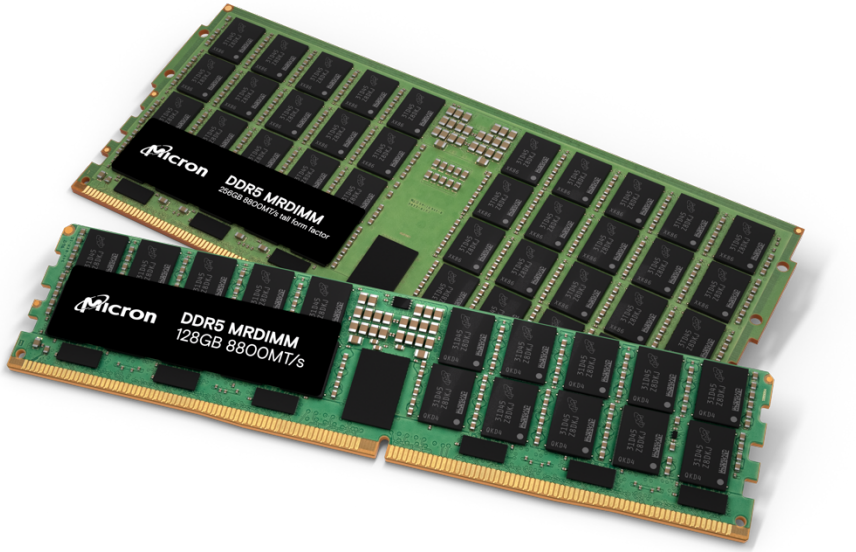


Figure 1: MRDIMM low-profile form factor (foreground) and TFF (background)

1. DDR5 is the fifth-generation double data rate (DDR) synchronous dynamic random-access memory (SDRAM).
2. As compared to 128GB RDIMM at 6400 MT/s (2Rx4), 128GB MRDIMM (4Rx4) at 8800 MT/s yields up to 39% improvement in bandwidth for 1:1 reads-writes. Data measured with Intel Memory Latency Checker (Intel MLC).
3. A 128GB TFF MRDIMM, with a height of 56.9mm, offers twice (2x) the module capacity over a standard 64GB RDIMM, which has a height of 31.25mm. 2x more placements result in 2x more capacity.
4. Up to 40% lower latency for 2R:1W sequential memory access pattern measured with Intel MLC. Run on GNR AP platform. **MRDIMM** system: 1 DIMM per channel (1DPC), 128GB TFF 4Rx4 MRDIMM (16Gb) at 8800 MT/s; **RDIMM** system: 1DPC, 128GB 2Rx4 RDIMM (32Gb) at 6400 MT/s with 1TB of system memory.

Key benefits

Designed to expand bandwidth per core by supporting memory requirements of higher CPU core counts.

39%

A 39% increase in bandwidth, providing an upgrade for platforms looking to exceed DDR5 RDIMM data rates.[2]

2x

MRDIMM tall form factor (TFF) provides 2x more capacity without the need for more complex packaging techniques such as 3DS (enabled by TSVs).[3]

40%

Delivers up to 40% lower latency for bandwidth-sensitive workloads.[4]

Standard form factor

Fits current socket (mechanical and electrical) technology for DDR5.

Enable next-gen servers

New memory architectures are needed to meet next-generation bandwidth per core requirements in server CPUs. The high bandwidth and high capacity of MRDIMM accelerates the performance of server CPUs, enabling customers to increase critical bandwidth and capacity per core.

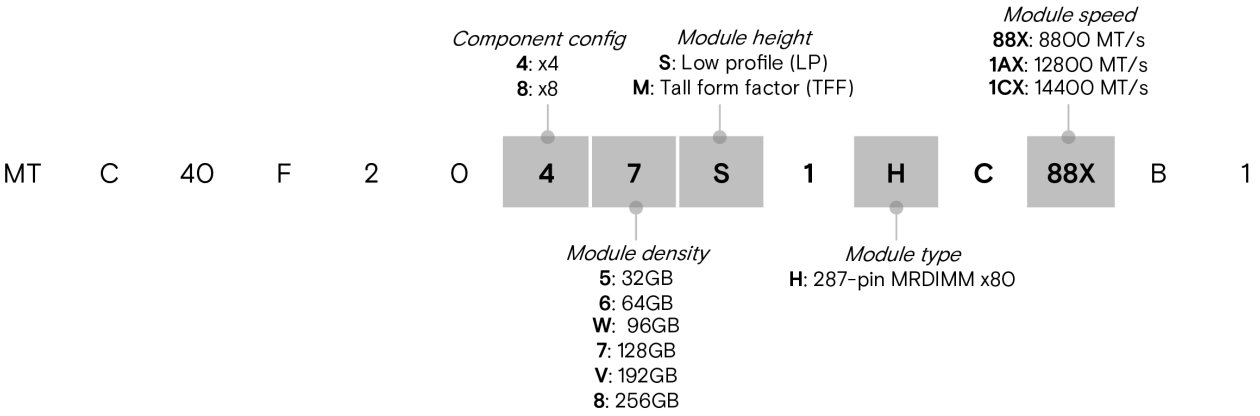


Figure 2: Micron part number (MPN) for the first generation of MRDIMM at 8800 MT/s

Form factor: Available in low-profile and tall form factor (TFF) module height options.

DDR5: Conforms to the standard DDR5 mechanical and electrical interface and is drop-in ready with systems that have enabled the protocol.

Capacity: Available in capacities up to 256GB. Delivers maximum capacity without bandwidth degradation in a low-profile module using dual-die package components (DDP) or a TFF module using single-die package (SDP) components.[5]

5. This option benefits system configurations, especially for platforms with one memory socket per memory channel restrictions, where high-capacity is enabled without complex TSV-based packaging.

Application	VIRTUAL MULTI-TENTANT							
	HIGH-PERFORMANCE COMPUTING			ARTIFICIAL INTELLIGENCE				
Capacity	32GB	64GB		96GB	128GB		256GB	
Rank	2Rx8	2Rx4	4Rx8	2Rx4	2Rx4	4Rx4	4Rx4	4Rx4
Form factor	Low profile	Low profile	Low profile	Low profile	Low profile	Tall form factor	Low profile, DDP	Tall form factor
Value	Optimized capacity	Higher RAS	Higher performance & lower power	Higher RAS	Compact capacity	Improved thermal density & performance	Compact capacity	Improved thermal density & performance

Table 1: Available capacities and configurations for tall and low-profile form factors of MRDIMM

Accelerate memory-intensive workloads

Optimal for parallel processing applications, MRDIMM accelerates workloads for HPC, AI inference and training, and virtual multi-tenant.

HPC	AI inference and training	Virtual multi-tenancy
Runtime reduction of nearly **26%** for OpenFoam[6], leading to faster execution times of computational tasks for HPC workloads.	Up to **24%** lower inference latency[7] for Llama3 to improve response times for LLMs, accelerating language processing tasks.	Enables more memory per module at faster speeds and lower latency, improving Quality of Service (QoS) for multi-tenant applications.

Table 2: Applications of MRDIMM and benchmark results

6. OpenFoam runtime for Motorbike mesh 600x300x300. Empirical data comparing 128GB MRDIMM at 8800 MT/s against 128GB RDIMM at 6400 MT/s.
7. Llama3 benchmark with 8 billion parameters, 1024 input tokens, 1 batch, and 10 streams running on GNR AP platform with 192 cores at 2.7 GHz. Empirical data comparing 128GB MRDIMM at 8800 MT/s (speed for the first generation of MRDIMM) against 128GB RDIMM at 6400 MT/s.

Workload data

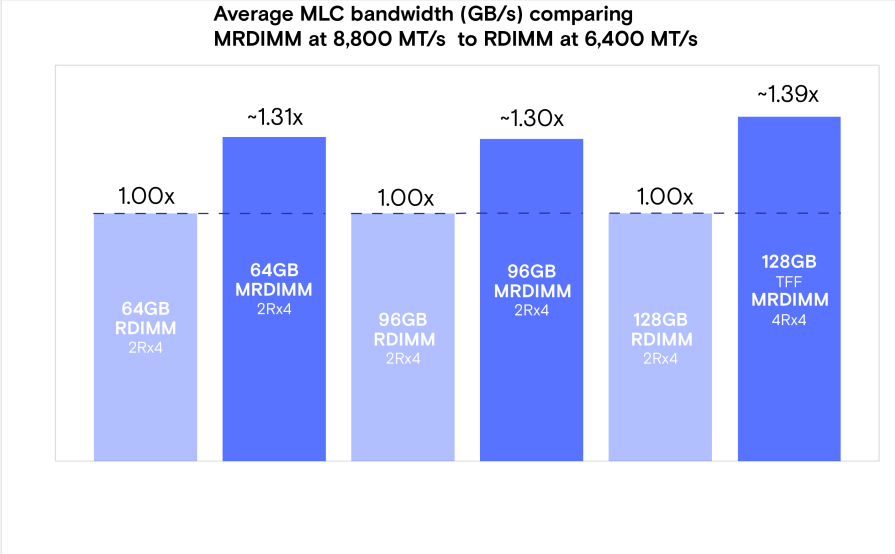


Figure 3: Average MLC bandwidth comparing MRDIMM to RDIMM for 64GB, 96GB and 128GB capacities[8]

8. Benchmark results of max bandwidth, taken using MLC, averaged across the following access patterns: ALL Reads, 3:1 Reads-to-Writes ratio, 2:1 Reads-to-Writes ratio, 1:1 Reads-to-Writes, and STREAM-triad. Dark blue bars: system configured with the first generation of MRDIMM at 8800 MT/s for 64GB/96GB (2Rx4) and 128GB (TFF, 4Rx4) capacities. Light blue bars: system configured with RDIMM at 6400 MT/s for 64GB/96GB/128GB (2Rx4) capacities. GNR AP platform with 1 DIMM per channel (1DPC) was used for both MRDIMM and RDIMM.

Sign up for the Micron Technology Enablement Program (TEP) at https://www.micron.com/secure-portal/ddr5-tep-membership-request to gain early access to Micron's technical information and support, electrical and thermal models, as well as DDR5 products to aid in the design, development and deployment of next-generation computing platforms.

micron.com/mrdimm